A Time-dependent SIS-model for Long-term Computer Worm Evolution

Marcus Märtens*, Hadi Asghari†, Michel van Eeten† and Piet Van Mieghem*
*Network Architectures and Services Group, Faculty of Electrical Engineering, Mathematics and Computer Science
†Economics of Cybersecurity group, Faculty of Technology, Policy and Management
Delft University of Technology, Delft, The Netherlands
Email: {m.maertens, h.asghari, m.j.g.vaneeten, p.f.a.vanmieghem}@tudelft.nl

***Abstract*—Epidemic models like the SIS or SIR model enable us to describe simple spreading processes over networks but are often not sufficient to accurately capture more complex network dynamics as exhibited by sophisticated and malicious computer worms. Many of the common assumptions behind epidemic models do not necessary hold if the process under investigation spans big networks or large scales of time. We extend the standard SIS network model by dropping the assumption of a constant curing rate in favour of a time-dependent curing rate function, which enables us to reflect changes in the effectiveness of the active worm removal process over time. The resulting time-dependent mean-field SIS model allows us to study the evolution of the size of computer worm bot-nets. We exemplify the complete procedure, including data-processing, needed to obtain a reliable model on data from Conficker, an extremely resilient computer worm. Using empirical data obtained from the Conficker sinkhole, we fit long time periods of up to 6 years on multiple scales and different levels of noise. We end by reflecting on the limits of epidemic models in empirical analysis of malware threats.**

I. Introduction

Computer worms have become a serious threat for individuals and organizations in today's Internet. While their local mechanisms of propagation can be reverse engineered and are well understood, their global impact remains hard to estimate. Armed with the potential to spread indefinitely, not even the authors of the worms might be able to predict how many machines will end up compromised as part of a worm's botnet.

While mathematical models for epidemics can be applied to estimate the size of botnets over time [1], [2], many of them rely on strong assumptions which might not be fulfilled by the networks or the worms. This makes it especially hard to fit empirically obtained data from measurements with those models. Firstly, most worms start spreading undiscovered and apply camouflaging techniques, so that data from the early infectious periods are often lacking. Since infected machines are subject to different up and down-times found in various environments, a worm might appear or disappear any time. Worms can also hibernate undetected on media like USB-sticks, possibly allowing them to reinfect cleaned up machines. To complicate matters even more, the spread of computer worms is additionally influenced by humans that apply various sorts of counter-measures, like patches, blocking of certain IP address ranges or re-routing and filtering network traffic, sometimes in very disruptive manners. Irregardless of counter-measures, reinfection is always possible: reimaging or roll-backs of work environments are common business procedures who can easily return machines into susceptible states by reintroducing vulnerabilities.

All these complex behaviours were observable for one of the prime examples of a long-lasting continuous battle against a maliciously growing computer worm: *Conficker* [3]. At its highest peak, Conficker was estimated to had infected over 9 million of Windows machines worldwide[1], creating one of the largest botnets in the history of the Internet. The command and control structure of this botnet was disrupted by using a sinkhole [4], a server that intercepted all calls from infected machine originally addressed to reach the bot-masters. The log-files of the sinkhole allow us to view Conficker's spread on very different levels of granularity by filtering the infected IP addresses by their autonomous systems (AS) and, for example, aggregating them again to the level of individual countries.

Especially on AS-level, the data can be noisy and sometimes exhibit unexpected patterns, as the worm was removed with varying effectiveness over time. To properly address this data, an extension of the traditional epidemic models is needed, which is able to describe the evolution of a worm over long periods of time. Our main contribution is to propose a new time-dependent mean-field SIS-model and apply it to the case of Conficker. In particular, our work is structured as follows:

- In Section II, we describe some traditional epidemic models used for computer worms and propose our new and general time-dependent mean-field SIS-model, which takes the aspects of reinfection and applied countermeasures into account.
- Section III explains how the general model needs to be further adjusted to the specific case of Conficker, how the sinkhole data was processed and critically reviews the legitimacy of our underlying model assumptions when put into practice.
- Afterwards, we fit our model to the actual Conficker data in Section IV. We show that our model deals better with the inherent noise of the data by providing high quality fits comparable to previously introduced models. For special cases in which the decline of the computer worm does not follow a strictly monotonically decreasing pattern, our model is still able to give

[1] http://goo.gl/9oaHEc (www.theregister.co.uk, Nov.2015)

a reasonable explanation of the data as it allows for changes in the effective worm removal process, which is not possible for monotonous models.

- Finally, Section V relates our model to previous work on computer worm research, Section VI discusses limitations of epidemic models for network security in practice and Section VII concludes with possible applications of the model and ideas on future research.

II. Epidemic models

This section describes classical epidemic models, which we will later use as a reference, and introduces our main contribution: the time-dependent mean-field SIS-model.

A. *The population-based SIR-model*

The population-based Susceptible-Infected-Removed (SIR) model, originally described by Kermak and Mckendrick [5], describes a spreading process in a fully mixed population, for which the corresponding underlying graph is the complete graph, in which each of the N individuals can be in one out of three possible compartments: I for *infected*, S for *susceptible* (to infection) or R for *removed*. The dynamics are described by the following set of differential equations

$$\frac{dS}{dt} = \frac{-\beta SI}{N}, \quad \frac{dI}{dt} = \frac{\beta SI}{N} - \delta I, \quad \frac{dR}{dt} = \delta I \tag{1}$$

where β is the infection rate and δ is the rate at which infected individuals are removed from the population. Both β and δ are assumed to be constant in classical SIR theory, in which case set (1) was already solved analytically [5] in 1927.

Individuals in SIR either stay in compartment S or make the transition $S \to I \to R$. Consequently, the number of susceptible hosts over time is always monotonically decreasing within this model. Similar to the SIS-model, which we introduce next, also the SIR model can be generalized to contact networks [6]. In this work however, we will use the simple original SIR model defined over fully mixed populations as a base-line for comparison with our more sophisticated time-dependent model.

B. *The network-based SIS-model*

The network-based Susceptible-Infected-Susceptible (SIS) model [7]–[9] is a Markovian model that describes a spreading process with possible reinfection for an underlying contact network. Each node of the network can be in two possible compartments: I for *infected* or S for *susceptible* (to infection). A network of N nodes can thus be in 2^N different states.

Usually, two independent Poisson processes, each with constant rate, determine the transitions between these states. The *infection process* determines for each susceptible node its transition to the infected compartment dependent on the number of infected neighbors. A node can only become infected if it shares a link to an already infected node. Each infected neighbor contributes with an infection rate of β to the infection. The *curing process* determines for each infected node its transition from the infected to the susceptible compartment with a corresponding curing rate δ.

As the state-space of the SIS Markov model grows exponentially in N, computing the probabilities of infection per node becomes quickly intractable for large networks. Mean-field approximations [7], [8], [10] are a common tool to reduce the size of the governing equations and make them amenable for analytic solutions. The N-Intertwined mean-field approximation model [11] (NIMFA) is currently the best continuous-time, first-order mean-field approximation. Given a fixed network, NIMFA approximates the probability $v_i(t)$ that a node i is infected at a certain time t by

$$\frac{dv_i(t)}{dt} = -\delta v_i(t) + \beta(1 - v_i(t)) \sum_{j=1}^{N} a_{ij} v_j(t). \tag{2}$$

These equations can be solved to determine the steady-state infection probability $v_{i\infty}$ of each node i, where $\frac{dv_i(t)}{dt} = 0$ and from which the average steady-state fraction of infected nodes

$$y_\infty = \frac{1}{N} \sum_{i=1}^{N} v_{i\infty}$$

can be computed.

If an r-regular graph[2] is considered as the underlying contact network, the infection probability is the same for each node due to symmetry: $v_i(t) = v(t) = y(t)$. Thus, the equation (2) simplifies to

$$\frac{dy(t)}{dt} = \beta r y(t)(1 - y(t)) - \delta y(t) \tag{3}$$

The particular equation (3) was studied by Kephard and White [12], who gave the solution

$$y(t) = \frac{y_0 y_\infty}{y_0 + (y_\infty - y_0)e^{-(\beta r - \delta)t}} \tag{4}$$

where the evolution of the fraction $y(t)$ of infected nodes is described by the initial fraction y_0 of infected nodes and the steady-state fraction of infected nodes $y_\infty = \lim_{t\to\infty} y(t)$. While equation (3) only holds for regular networks, it gives an excellent starting point for the development of a time-dependent model, as we will see in the next subsection.

C. *The time-dependent mean-field SIS-model*

If the source for an infection or for curing is not constant, the fixed rates β and δ have to be transformed into functions $\beta(t)$ and $\delta(t)$ to describe the rates for the infection and curing processes at any time t. The time-dependent extension of (3) is

$$\frac{dy(t)}{dt} = \beta(t) r y(t)(1 - y(t)) - \delta(t) y(t). \tag{5}$$

While the exact Markovian SIS dynamics seem to be impossible to solve for time-dependent rates, even for highly symmetric cases as the complete graph, Van Mieghem [13] shows that the differential equation (5) can be solved exactly

[2] A graph G is r-regular if each node in G has degree r.

to determine the evolution of the fraction of infected nodes $y(t)$ over time by

$$y(t) = \frac{\exp\left(\int_0^t (r\beta(u) - \delta(u))du\right)}{\frac{1}{y_0} + r\int_0^t \beta(s)\exp\left(\int_0^s (r\beta(u) - \delta(u))du\right)ds}. \tag{6}$$

A convenient short-hand is to define the net dose as

$$\rho(t) = \int\limits_0^t (r\beta(u) - \delta(u))du \tag{7}$$

which equals the net average number of infections reduced by all curings in a time interval $[0, t]$ for a particular node in the r-regular graph. Using the net dose (7), equation (6) becomes

$$y(t) = \frac{e^{\rho(t)}}{\frac{1}{y_0} + r\int_0^t \beta(s)e^{\rho(s)}ds}. \tag{8}$$

The main quantities in (8) are the degree of the regular network r, the initial fraction of infected nodes y_0, the time-dependent infection rate function $\beta(t)$ and the curing rate function $\delta(t)$. We describe in subsection IV-A how the parameters $\{r, y_0\}$ and the functions $\{\beta(t), \delta(t)\}$ can be determined to match the infection curve of the Conficker worm. For the remainder, we will refer to this model in short as the time-dependent SIS-model.

III. Methodology

This section outlines the necessary steps before applying any epidemic model (like the ones introduced before) to measured data from a sinkhole. We will first describe the Conficker data, how it was processed and why we think that the time-dependent SIS-model is a reasonable choice to describe the propagation of the worm.

A. Datasets

All data of the Conficker worm is based on logfiles from the sinkhole. The sinkhole was used to disrupt the update mechanism of Conficker, which connected to 250 pseudorandomly generated URLs in order to get payload (i.e. instructions, malware or new functionality) from its original authors. By registering the domain names before the botmasters, and redirecting every access to a central server (the *sinkhole*), the worm was effectively cut off from its authors.

There were some partially successful attempts of the botmasters to regain control over the Conficker botnet, made possible due to mistakes during the sinkholing process. This resulted in new variants of the worm, which employed a more sophisticated update mechanism. However, after April 2009 the botnet remained under control. From this perspective, Conficker provides an interesting case study of the propagation of an unaltered computer worm over a reasonably long period of time.

In total, the sinkhole logfiles provide us data from February 2009 to September 2014 and contain over 178 million unique IP addresses. With the help of GeoIP-databases[3] and IP-to-ASN-lookup[4], these IP addresses were associated with the corresponding ISO country code and autonomous system (AS). Thus, the data can be viewed at different levels of granularity:

- **global:** all unique IPs for the complete sinkhole worldwide
- **country:** all unique IPs belonging to a specific ISO country code
- **autonomous system:** all unique IPs belonging to a specific AS

In total, the IPs belong to 241 different ISO country codes and to over 34.000 different autonomous systems.

B. Preprocessing

Botnet size estimation: Accurate estimations of the amount of infected machines is a difficult problem (see Abu Rajab *et al.* [14]), as long as our only way for identification of a machine is via its IP address. On the one hand, it is possible to *undercount* because multiple infected machines might share a common IP address due to Network address translation (NAT). On the other hand, a single infected machine might be represented by multiple IP addresses due to different ISP policies. To avoid this *overcounting*, the number of IP-addresses needs to be corrected by a DHCP-churn rate, which varies over countries and ISP. Determining accurate DCHP-churn rates is a challenge in itself (see Moreira *et al.* [15]), which we will not undertake here.

Instead, we aggregate the unique IP-addresses over short time slots of one hour. We consider the DHCP-churn effect on this time scale to be minimal. Using short time slots introduces another source of undercouting because not every infected machine might be contacting the sinkhole every hour (for example they might not be powered). The hourly values are then averaged out over a time slot of a week, eliminating biases introduced by diurnal patterns. Accordingly, together with the NAT-effects, our estimate of the infected machines should be considered as a lower bound.

Data cleansing: While analyzing the sinkhole data, missing measurements become apparent as there are several periods ranging from a few hours to a few days in which the number of IPs drops down to zero. We account technical malfunctions of the infrastructure (i.e. downtime of the sinkhole) for these artifacts. Consequently, we remove these outliers before applying any model fitting by the following procedure: for each datapoint z of week w, we compute the difference between z and the median of all datapoints in a time window spanning ± 10 weeks from w. From all datapoints, we exclude the 10% with the highest differences. This procedure does not remove all outliers for all cases, but reduces their impact on the fitting procedure considerably.

Normalization: In order to apply the time-dependent epidemic model (5), the data need to be normalized, because $y(t)$ describes the average fraction of infected nodes and not the number of infected nodes in the networks. An accurate

[3] http://maxmind.com/app/geoip_country

[4] https://github.com/hadiasghari/pyasn

normalization would use the amount of vulnerable machines, which is not known to us. In fact, the Conficker worm spreads only in unpatched versions of all major Microsoft Windows versions up to Windows Vista and Server 2008, for which we have not found reliable estimates. Instead, we use the *peak point* of infection to generate a relative scaling. After the aggregation and data cleansing, we determine the maximum number p_{max} of infections over the whole infectious period and compute the scaling factor s_y by

$$s_y = \frac{k}{p_{max}} \tag{9}$$

where k is a real number between 0 and 1. For our fitting procedure, a value of $k = 0.9$ proved to be sufficient and was used if not stated otherwise. The scaling factor s_y can be used to fit the original data as we will discuss in subsection IV-A.

We use a bin size of one week to count the unique IPs, resulting in 280 bins for the complete infectious period. This period is linearly transformed on the horizontal axis so that the starting point of the infectious period maps to 0 and the end point maps to 1.

C. Model Assumptions

The dynamics of the Conficker-spread are heterogeneous, because different infection vectors are invoked to infect new machines and networks. For example, a person might obtain the worm by plugging out his USB-stick from an infected computer. Much like an actual biological disease, this person could traverse large amounts of space and time before he triggers a new infection with his USB-stick on a different machine. This and similar effects make the construction of an actual (dynamic) contact network impossible. Considering the extremely large scale of the Internet and the long period of time (6 years), it seems completely unreasonable to assume that any (simple) model would be able to reflect this degree of complexity.

However, the time-dependent mean-field SIS-model (5) can be refined to capture the basic observable infection patterns. In order to justify the refined model, we review the basic assumptions and argue to which extent they are adequate in the case of Conficker.

Constant spreading rate: As Conficker was disconnected from its authors by the sinkhole, its code remained largely constant for the whole infectious period. There exist some updated versions of Conficker (named Conficker C, D and E) at the very beginning of the logged infectious period (up to April 2009), but these updates were used to improve the command and control structure of the worm and to add a scareware payload to it. The main infection vectors (NetBIOS vulnerability, USB-sticks and Shared Folders) remained largely unaffected by these changes. It is thus safe to assume, that the spreading rate of Conficker remained constant. For this reason, we will set $\beta(t) = \beta$ in equation (5).

Time-dependent curing rate function: Contrary to the infection rate $\beta(t)$, we assume that the curing and removal forces were not constant. In general, the clean-up of Conficker was regarded to be rather involved as the worm possessed several counter-measures. The curing rate function $\delta(t)$ of our model (5) reflects the combined effort that was spent to fight Conficker, i.e. by patching the vulnerability, use of removal tools and also the replacement of infected machines. As some of the countermeasures did not provide complete immunity, reinfection with Conficker was possible and is well documented [3]. This effect is reflected by the basic SIS dynamics. Next to SIS, our model is able to simulate removal as in the SIR-model [8] by increasing the curing rate of a node. Once a node's curing rate is very high, its infectious periods become very small, which can be regarded as a removal or immunization effect. Although our model (5) never *explicitly* removes nodes from the network, the time-dependent curing rate blends both SIS and SIR-dynamics and thus captures effects like permanent removal of machines or an acquired immunity, for example by system upgrades without excluding reinfection dynamics like SIR.

Network topology: The equation (5) of the time-dependent mean-field SIS-model demands an underlying and constant contact network of degree r. This is a necessity from a model point of view, as computation would quickly become intractable otherwise. While the Internet is clearly not constant in its size, we justify the regularity assumption by one of the infection vectors of Conficker. The technical reports (see Porras *et al.* [16]) suggest that Conficker used a scan-and-infect subroutine that occasionally scanned random IP-addresses[5] for new victims. The worm did not flood the complete IP-space but concealed itself by connecting only to a limited amount of possible new victims. Thus, for the fixed allocated time-slots that we investigate, we assume that there is an upper bound on the possible scan-attempts for an infected machine, which is independent of the configuration or network properties (e.g. bandwidth). This upper bound translates into an estimate on the degree of the underlying contact network, which is our parameter r.

IV. Model application

A. Modeling the spread of Conficker

The key to a good epidemic model of Conficker is to determine the time-varying parts of equation (8), namely the spreading rate function $\beta(t)$ and the curing rate function $\delta(t)$. As argued before, we assume $\beta(t) = \beta$ as constant, so that the net dose $\rho(t)$ in (7) simplifies to

$$\rho(t) = r\beta t - D(t)$$

where $D(t) = \int_0^t \delta(u)du$ is the *accumulated curing dose*. If we assume $D(t)$ to be an analytic function, there exists a Taylor series that allows us to express $D(t)$ precisely. The Taylor expansion is truncated after d terms to retrieve a polynomial approximation

$$D(t) = \int_0^t \delta(u)du \approx \sum_{i=0}^{d} a_i t^i \tag{10}$$

[5] Due to a bug in the pseudo-random number generator of Conficker, only one fourth of the complete IP-address space was susceptible for this attack vector.

with $a_d \neq 0$. We use the last two equations to transform equation (8) as

$$
\begin{aligned}
y(t) &\approx \frac{\exp\left(r\beta t - \sum\limits_{i=0}^{d} a_i t^i\right)}{\frac{1}{y_0} + r\int_0^t \beta \exp\left(r\beta s - \sum\limits_{i=0}^{d} a_i s^i\right) ds} \\
&= \frac{\exp(-a_0)\cdot \exp\left(r\beta t - \sum\limits_{i=1}^{d} a_i t^i\right)}{\frac{1}{y_0} + \exp(-a_0)\cdot r\beta\int_0^t \exp\left(r\beta s - \sum\limits_{i=1}^{d} a_i s^i\right) ds} \\
&= \frac{\exp\left(r\beta t - \sum\limits_{i=1}^{d} a_i t^i\right)}{\frac{\exp(a_0)}{y_0} + r\beta\int_0^t \exp\left(r\beta s - \sum\limits_{i=1}^{d} a_i s^i\right) ds}.
\end{aligned}
$$

We define the products $y_0^a = y_0 \cdot e^{-a_0}$ and $\beta_r = r\beta$ to simplify the model further. Thus, the free parameters of our model are y_0^a, β_r and $a_d, \ldots, a_1$. In order to apply the model to absolute values rather than fractions, we apply the scaling-factor from equation (9):

$$
y(t) = \frac{e^{\widetilde{\rho(t)}}}{s_y\left(\frac{1}{y_0^a} + \beta_r \int_0^t e^{\widetilde{\rho(s)}} ds\right)} \quad \text{with} \quad \widetilde{\rho(t)} = \beta_r t - \sum_{i=1}^{d} a_i t^i. \tag{11}
$$

Since we are interested in the dynamics and not in absolute values, we will set $s_y = 1$ for simplicity. Consequently, in all the Figures showing model fittings, the maximum value will always be found at $y = 0.9$ (recall Section III-B, Normalization). In summary, the model is given by

$$
y(t) = \frac{e^{\widetilde{\rho(t)}}}{\frac{1}{y_0^a} + \beta_r \int_0^t e^{\widetilde{\rho(s)}} ds} \quad \text{with} \quad \widetilde{\rho(t)} = \beta_r t - \sum_{i=1}^{d} a_i t^i. \tag{12}
$$

Given the preprocessed and cleansed data as input, the trust-region non-linear least squares method provided by MATLAB is used to fit our model. The initial values for all parameters were picked randomly between zero and one. Then, the polynomial of degree 3 was fitted first. The parameters found for this fit were used as initial guesses to fit the polynomial of degree 4, which significantly improved the convergence speed. The same procedure was applied to create the fit for degree 5, based on the fit of degree 4.

B. Quality of fits at global scale

The degree of freedom of the proposed model for Conficker is $2 + d$, where d is the order of the polynomial curing rate function (10). Thus, we are able to trade the complexity of the model with its accuracy. In a first experiment, we examine which value of d is useful for modelling the underlying data. We use the global aggregation of all worm infections as input, because it has less noise than data from country- or ASN-level granularity. Figure 1 shows how the quality of the fit and of the prediction bounds improves by using higher degree polynomials. The adjusted R^2-value is 0.96 for $d = 2$ and 0.99 for $d \geq 3$, indicating a good fit overall. However, a visual inspection reveals that the fit for $d = 2$ is not good enough to accurately describe the decline of the worm. On the other hand, $d = 4$ does not provide much more quality, but requires an additional fitting parameter. Thus, guided by parsimony as a modelling rule, we believe that $d = 3$ is a good choice for this particular case. Table I gives the actual values of the fitted parameters.

TABLE I. FITTING PARAMETERS FOR GLOBAL MODEL ON CONFICKER, 95% CONFIDENCE BOUNDS IN PARENTHESIS.

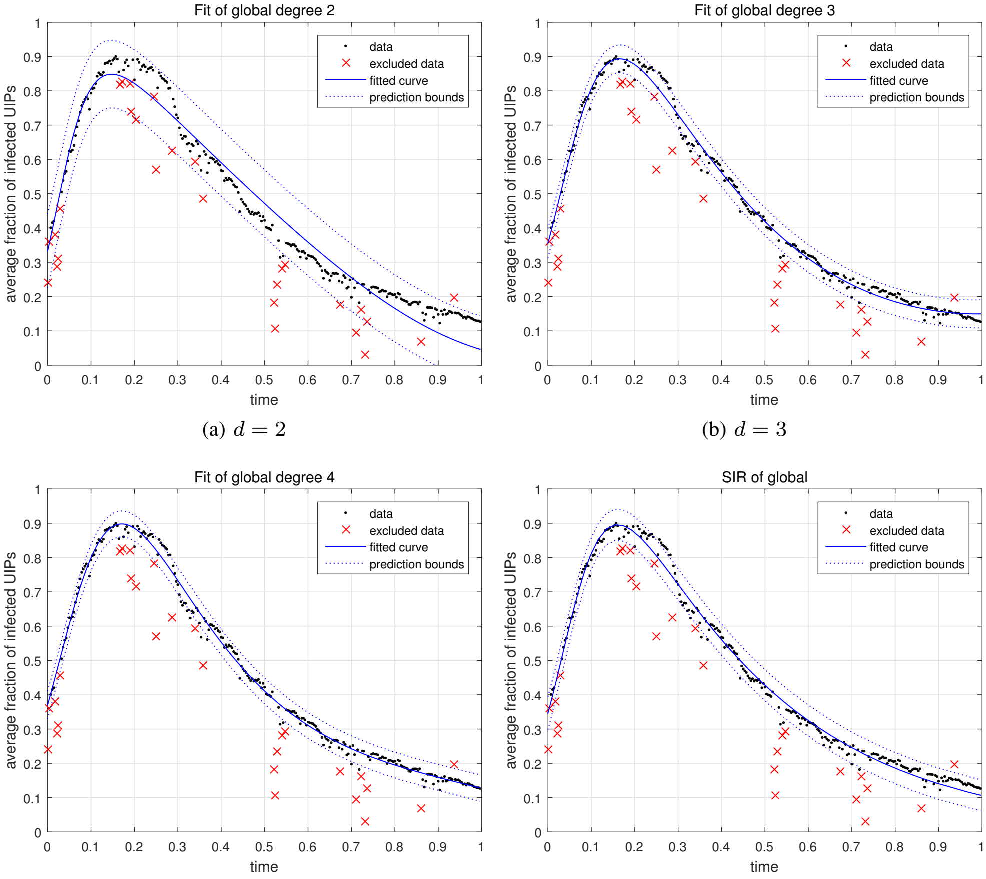
d	2	3	4
β_r	23.5 (19.55, 27.45)	12.03 (10.98, 13.08)	7.304 (5.656, 8.952)
v_0^a	0.333 (0.287, 0.379)	0.3554 (0.3396, 0.3712)	0.3703 (0.3556, 0.385)
a_1	-1.17 (-1.64, -0.692)	-5.572 (-5.813, -5.331)	-7.597 (-8.23, -6.964)
a_2	16.06 (14.1, 18.02)	23.29 (22.48, 24.11)	31.05 (28.61, 33.49)
a_3	-	-10.33 (-10.79, -9.863)	-28.07 (-33.43, -22.7)
a_4	-	-	9.646 (6.713, 12.58)

Fixing $d = 3$ results in 5 parameters that need to be determined by the fitting procedure. This is one parameter more than we would use to fit the population-based SIR-model, which is also shown in Figure 1d for comparison. More specifically, the parameters used for the SIR-model are the size of the population N, the constant infection rate β, the constant curing rate δ and the initial number of infected individuals v_0. As the adjusted R^2-value for SIR is with 0.99 qualitatively very high and on the same level as for $d = 3$, SIR is an even more parsimonious model that works very well on global scale.

C. Quality of fits at subglobal scales

Moving from global-level to country-level, we have more noise and variation in the data since not all countries were affected by Conficker in the same way. Out of the 241 different ISO countries, we picked a subset of 40 countries (belonging to the OECD and the European Union) to analyse. We used again SIR as a baseline and compared it with fits of order $d = 3, 4$ and 5. We chose the adjusted R^2-value and the sum of squared errors (SSE) as indicators for the quality of fit. Figure 2 shows the distribution of those indicators for the 40 chosen countries sorted by their corresponding quality in the SIR-model. The fits of the Conficker-model are of high quality and only in 3 cases worse than the corresponding SIR-model. We believe that the fitting procedure converged sub-optimally in those cases, as 7 parameters needed to be determined, while lower order fits were still better than SIR or equally good. A visual inspection of the fits showed that the time-dependent SIS-model is able to fit the tail better than SIR. The latter is forced to monotonically decline in this area while the time-dependent SIS-model can better adapt to nearly constant viral levels which are observable in the tails of the data for some of the countries.

Moving down to the ASN-level imposes a bigger challenge, since the number of infected IPs becomes so low that the



(a) $d = 2$

(b) $d = 3$

(c) $d = 4$

(d) SIR

Fig. 1. Global-level fitting of Conficker in time-dependent SIS and SIR model. Dotted curves are the 95% prediction bound

influence of noise grows more significant. To circumvent this problem, we selected the 30 ASNs with the highest number of infected unique IPs out of 34000 available for our analysis. Still, the data for some of these ASNs is considerably more distorted than any country-level data. While gaps of missing data on country-level usually span a couple of weeks, they can spent months or even years for some of the ASN datasets. We expect that some sort of ISP-wide countermeasures were used to prevent infected machines during those times to connect to the sinkhole, though we did not find evidence for this claim.

The data-cleansing procedure is not sufficient to remove all outliers completely, so they inevitably impact the quality of the models. Figure 3 shows the results in the same way as we did for the country-level. Some of the fits are not visible as they are of so low quality ($R^2 < 0.9$ or SSE > 2) that they are outside our scale. We ordered all data after the quality of the SIR-fit nevertheless. The variation in quality is much higher than it was for the country-level: 4 out of 30 ASNs were so degenerated that every model produced only poor fits. Similar to the country-level, we observe the time-dependent SIS-model to be better than SIR, though with a higher relative qualitative difference than on country-level. This is not unexpected as models with more degrees of freedom adapt more easily to noise in general. A visual inspection of the fits on the ASN-level showed that in not degenerated cases (i.e. high *jitter* or very large gaps) the models still give a fairly good description of the spreading pattern.

D. Determining the effectiveness of worm removal

While fitting the absolute or relative number of infected IPs gives a comprehensive overview about the prevalence $y(t)$ of Conficker, the time-dependent SIS-model allows for an additional perspective. To gain more insight in the worm removal, we look closer to the curing rate function $\delta(t)$ and understand, what it actually means. While changes in the curing rate can be easily interpreted as changes in the applied counter-measures against the worm, we have to keep in mind that the most effective counter-measures were already in place *before* the sinkhole recorded its data. More precisely, the NetBIOS vulnerability was patched very fast in November 2009, 4 months before our data collection starts. However, Conficker was remarkably resilient, raising the question why the worm could survive despite the patch for years?

If we assume that the slow decline of the worm is, to a large amount, not caused by security patches but by long-term effects like failure or substitution of infected machines with newer ones, it seems reasonable that the population-based SIR-model gives a fairly good description of the process. By looking at the curing rate function $\delta(t)$ and the accumulated

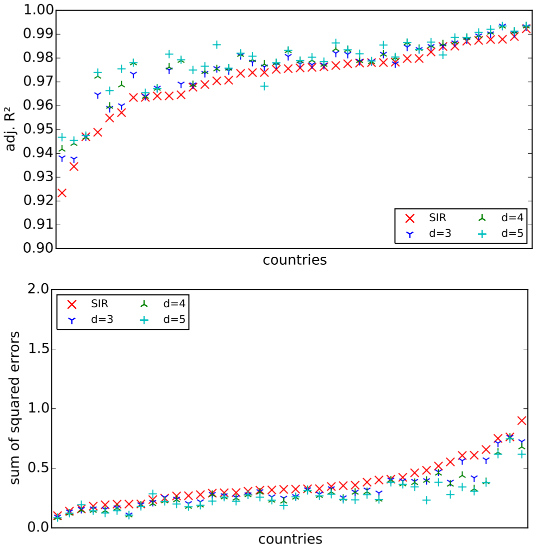


Fig. 2. Overview of Country-level fitting quality, ordered by SIR.

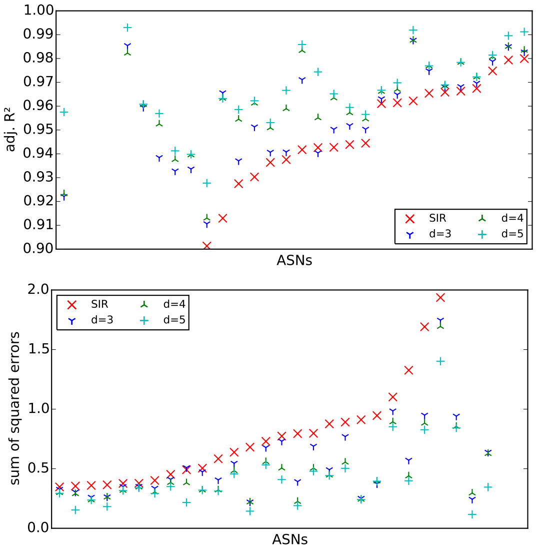


Fig. 3. Overview of ASN-level fitting quality, ordered by SIR.

curing dose $D(t)$ obtained by the fits of the time-dependent SIS-model, we can analyse at which periods in time the removal of the worm changed. We call the time pattern given by $\delta(t)$ the *effective worm removal*.

Figure 4 shows $\delta(t)$ and $D(t)$ obtained by the fits on the global aggregation level. It is interesting to see that $\delta(t)$ seems to approach a sigmoidal shape in the interval $[0, 1]$ once we increase the degree of the polynomial. A sigmoidal effective worm removal starts very low, describing a time-period in which the worm is persistent and spreads unhampered. However, this is followed by a sharp increase in curing rate, which leads to a rapid decline in the worm prevalence. Finally, the sigmoidal shape reaches a saturation of high curing rates, which explain the low levels of persistence in the later phases of the worm evolution and the long period of final decline.

E. *Sensitivity of parameters*

For $d = 3$, the time-dependent SIS-model is determined by 5 parameters: a_3, a_2, a_1, β_r and y_0^a. Since $y(0) = y_0^a$, the meaning of y_0^a is clear: it defines the initial fraction of infected IPs and thus the starting point for the spreading process. However, it is not obvious how the course of the infection is influenced for $t > 0$ by a_3, a_2, a_1 and β_r? To investigate this further, we collected all values that occurred for those parameters while fitting each country-level dataset. To avoid outliers, we computed for each parameter the 10% and 90%-percentile. The range between both values was divided linearly so that we end up with 5 evenly spread out values for each parameter, which are representative for the fits. We used the median of those 5 values to define a reference curve and adjusted each parameter separately to see how sensitive the model is to changes. Results are shown in Figure 5.

The coefficients a_3, a_2, a_1 have dominant influence on different phases, with a_3 being dominant at the later stages, a_2 in the middle and a_1 at the beginning. The curvature of the infection itself is strongly influenced by β_r, which regulates the height of the peak and the decline of infection. During the fitting procedure, these 4 parameters are balanced out against each other. For example: increasing β_r makes for a much flatter decline in worm prevalence, but also moves the peak higher. If the data suggests a flat decline but a low peak, β_r should be high but also with a high a_1 to correct the peak.

An intriguing property of the time-dependent SIS-model is the fact that it is not monotonous unlike the SIR model: by decreasing a_3 or increasing a_2, it is possible to have an increase in worm prevalence *after* the maximum peak. While this behaviour is - for the case of Conficker - not observable on country-level, there are some rare occurrences on AS-level that suggest such a behaviour may occur in practice. Figure 6 shows with AS8452 such an example and shows the corresponding fits of the non-monotonous time-dependent SIS model in comparison to the monotonous SIR model.

V. Related Work

In a previous work by Asghari et al. [17], the same dataset for Conficker was analyzed to show the effectiveness of anti-bot net campaigns in different countries. The focus of our current work is different: we develop a very general epidemic model to describe time-dependent propagation and use the Conficker data as an example to show the applicability of such an approach. In contrast, the previous work started with the data of Conficker and developed a non-epidemic descriptive model to extract features of the worm prevalence on country-level. Those features were then correlated with different institutional factors (broadband access, operating system market shares, software piracy, etc.) to explain regional difference between the countries.

Epidemic models have already been investigated for describing spread of computer worms [2], [18]–[20], mostly by extending SIS or SIR-models. Some of those models are not directly comparable to this work, as they are based on very different assumptions. For instance, the WPM model by Peng et al. [21] emphasizes strongly on network topology and

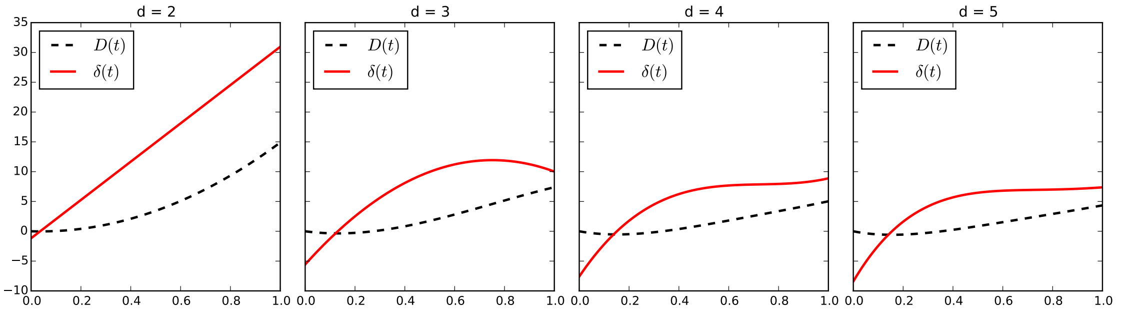


Fig. 4. $\delta(t)$ and $D(t)$ for global-level fitting with polynomial degrees 2,3,4 and 5.

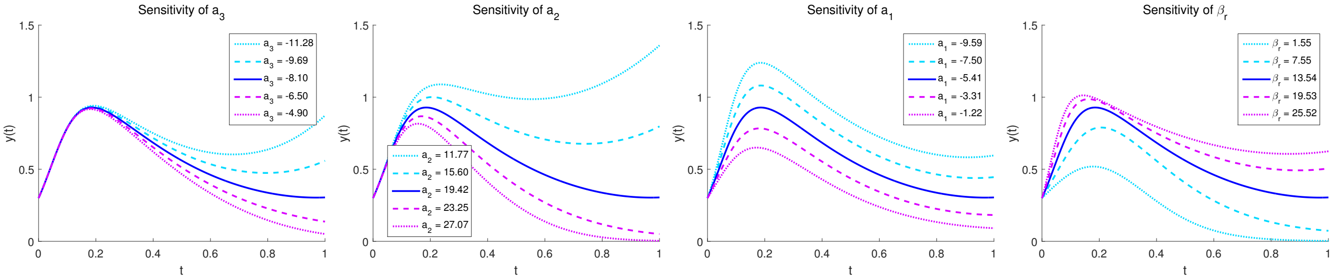


Fig. 5. Influence of the polynomial coefficients a_3, a_2, a_1 and β_r on the worm prevalence. The solid curve corresponds to $a_1 = -5.41$, $a_2 = 19.42$, $a_3 = -8.10$ and $\beta_r = 13.54$ in each subfigure and is used as a reference to study the change of its shape by adjusting each parameter separately.

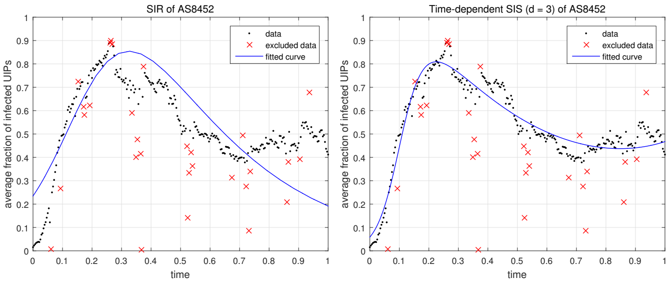


Fig. 6. Left: The SIR model has difficulties to fit the noisy and non-monotonous data from AS 8452. Right: The time-dependent SIS model with $d = 3$ already provides more flexibility to address those issues.

uses $2d$-cellular automatons to describe propagation. Another model, the two-factor worm model by Zou et al. [22] assumes dynamic changes in the spreading rate, as it was used to describe the propagation of the Code Red computer worm, who literally *exploded* over the Internet within a day compared to the more subtle Conficker that was designed to be much more hideous. Moreover, the two-factor worm model needs 4 different compartments and 6 free parameters, which makes the actual fitting for this model much more involved than for the time-dependent SIS.

Zhang et al. [23] extend the basic SIR model to a hybrid epidemic model, which is evaluated on data from Conficker as well. Their model considers multiple ways of infection (globally and locally) and thus does not assume - like our and many other works - that Conficker only spread via its scan-and-infect subroutine. However, the counter-measures and the curing was not in their focus, as no change in curing rate was incorporated. Our work is complementary in the sense that it could be part of new hybrid epidemic models to give a stronger focus on multiple sources of countermeasures.

VI. Discussion

Epidemic models have many proponents in the security community. The success of these models in studying human epidemics might have served as the inspiration. They have certainly enabled researchers to think and theorize about countermeasures, and simulate their effects on malware outbreaks. This paper also contributes to this literature. It presented a time-dependent SIS-model that approximates the Conficker worm over long periods quite well, even at the AS-level.

Can epidemic models also help in analysing security practices from past malware outbreaks? For this ambition to be realized, you need to be able to interpret the parameters of any model, after applying it to empirical data. Herein lie two fundamental challenges: model assumptions and parameter sensitivity.

Our time-dependant SIS-model has fewer assumptions than traditional models. Yet it still assumes constant spreading rate and topology. These assumptions are unlikely to hold over long timeframes and many networks. What do the parameters mean in such a case? Furthermore, multiple combinations of parameter values often generate similar trends. This makes it nearly impossible to draw security insights with confidence. Increasing model parameters to reduce assumptions will make matters worse.

The underlying problem are the latent model parameters that cannot be measured from existing infection data. It should not be a surprise that papers using epidemic models are often theoretical, or use datasets that are spatially or temporally limited (e.g. weeks, or one network). To bridge the gap between these models and empirical data, research on how to measure or estimate the latent parameters is needed.

VII. Conclusion and Future Work

In this work, we showed how epidemic models can be designed to describe the complex propagation and decline patterns of long-lasting computer worms with Conficker as a showcase. The time-dependent SIS model has shown to be useful on scales where the SIR model could no longer provide the best fit due to its inherent limitations. SIR might still be a good choice if the worm data is smooth and shows a clean monotonous decline, but the time-dependent SIS model should be preferred if any of these conditions is violated. Changing the degree of the polynomial curing rate function allows for a more flexible approach, when it is needed.

For example, if we assume that future worms will have the ability to adapt (for example by some evolutionary process), they might also develop an immunity against certain countermeasures in which case a non-monotonous curing rate function could be used to model this behaviour. It is possible, to apply the time-dependent SIS model also for worms with changing spreading rates: one only needs to define a time-dependent infection rate function $\beta(t)$ analogously to $\delta(t)$.

From a theoretical perspective, the fact that both the SIR and the time-dependent SIS model can fit the same data is surprising as well and might suggest SIR being a special case of SIS with time-dependent curing or infection rate functions. Possible future work could investigate the relationship between both models. Because of the monotonicity of SIR it is clear that not every propagation pattern of the time-dependent SIS model can be reproduced by SIR. However, it might be possible that there exists a mapping from the parameters of the SIR model to the parameters of the time-dependent SIS model in a way that the SIR propagation patterns can be (approximately) reproduced.

On a more practical level, our work could provide some building blocks for the development of novel worm tracking systems that would monitor the current effective removal over time. There might also be suitable applications outside the domain of computer worms for which a time-dependent epidemic model like ours could be applicable. Examples could include diffusion of technologies, spread of memes, or fighting darknet websites – where various measures and counter-measures over time influence the diffusion pattern. While very different scenarios, it would be interesting to see whether network epidemic models are still applicable or not.

References

[1] S. Qing and W. Wen, "A survey and trends on Internet worms," *Computers & Security*, vol. 24, no. 4, pp. 334–346, 2005.

[2] Y. Wang, S. Wen, Y. Xiang, and W. Zhou, "Modeling the propagation of worms in networks: A survey," *Communications Surveys & Tutorials, IEEE*, vol. 16, no. 2, pp. 942–960, 2014.

[3] J. Shearer. (2008, November) Symantec. [Online]. Available: http://www.symantec.com/security_response/writeup.jsp?docid=2008-112203-2408-99

[4] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna, "Your botnet is my botnet: analysis of a botnet takeover," in *Proceedings of the 16th ACM conference on Computer and communications security*. ACM, 2009, pp. 635–647.

[5] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 115, no. 772. The Royal Society, 1927, pp. 700–721.

[6] M. E. Newman, "Spread of epidemic disease on networks," *Physical review E*, vol. 66, no. 1, p. 016128, 2002.

[7] R. M. Anderson and R. M. May, *Infectious diseases of humans*. Oxford University Press, Oxford, 1991, vol. 1.

[8] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," *Reviews of Modern Physics*, vol. 87, no. 3, pp. 925–979, 2015.

[9] P. Van Mieghem, *Performance Analysis of Complex Networks and Systems*. Cambridge University Press, 2014.

[10] ——, "Epidemic phase transition of the SIS type in networks," *EPL (Europhysics Letters)*, vol. 97, no. 4, p. 48004, 2012.

[11] P. Van Mieghem, J. Omic, and R. Kooij, "Virus spread in networks," *IEEE/ACM Transactions on Networking*, vol. 17, no. 1, pp. 1–14, 2009.

[12] J. O. Kephart and S. R. White, "Directed-graph epidemiological models of computer viruses," in *Research in Security and Privacy, 1991. Proceedings., 1991 IEEE Computer Society Symposium on*. IEEE, 1991, pp. 343–359.

[13] P. Van Mieghem, "SIS epidemics with time-dependent rates describing ageing of information spread and mutation of pathogens," Delft University of Technology, Report20140615 (www.nas.ewi.tudelft.nl/people/Piet/TUDelftReports), 2014.

[14] M. Abu Rajab, J. Zarfoss, F. Monrose, and A. Terzis, "My botnet is bigger than yours (maybe, better than yours): why size estimates remain challenging," in *Proceedings of the 1st USENIX Workshop on Hot Topics in Understanding Botnets, Cambridge, USA*, 2007.

[15] G. Moreira Moura, C. Ganan, Q. Lone, P. Poursaied, H. Asghari, and M. Van Eeten, "How dynamic is the ISPs address space? towards internet-wide DHCP churn estimation," in *Networking 2015, 14th International Conference on Networking, 20–22 May 2015, Toulouse, France*. IFIP, 2015.

[16] P. Porras, H. Saidi, and V. Yegneswaran, "An analysis of Conficker's logic and rendezvous points," *Computer Science Laboratory, SRI International, Tech. Rep*, 2009.

[17] H. Asghari, M. Ciere, and M. J. van Eeten, "Post-mortem of a zombie: Conficker cleanup after six years," in *24th USENIX Security Symposium (USENIX Security 15), Washington, DC*, 2015.

[18] G. Serazzi and S. Zanero, "Computer virus propagation models," in *Performance Tools and Applications to Networked Systems*. Springer, 2004, pp. 26–50.

[19] T. M. Chen and J.-M. Robert, "Worm epidemics in high-speed networks," *Computer*, vol. 37, no. 6, pp. 48–53, 2004.

[20] S. Fei, L. Zhaowen, and M. Yan, "A survey of internet worm propagation models," in *Broadband Network & Multimedia Technology, 2009. IC-BNMT'09. 2nd IEEE International Conference on*. IEEE, 2009, pp. 453–457.

[21] S. Peng, S. Yu, and A. Yang, "Smartphone malware and its propagation modeling: A survey," *Communications Surveys & Tutorials, IEEE*, vol. 16, no. 2, pp. 925–941, 2014.

[22] C. C. Zou, W. Gong, and D. Towsley, "Code red worm propagation modeling and analysis," in *Proceedings of the 9th ACM conference on Computer and communications security*. ACM, 2002, pp. 138–147.

[23] C. Zhang, S. Zhou, and B. M. Chain, "Hybrid Epidemics A Case Study on Computer Worm Conficker," *PLoS ONE*, vol. 10, no. 5, 2015.