Computer Networks 57 (2013) 404–421

Contents lists available at SciVerse ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

The SIC botnet lifecycle model: A step beyond traditional epidemiological models

Masood Khosroshahy *, Mustafa K. Mehmet Ali, Dongyu Qiu

Electrical and Computer Engineering Dept., Concordia University, Montreal, Canada

ARTICLE INFO

Article history:
Available online 20 December 2012

Keywords:
Analytical models
Markov processes
Peer-to-peer
Epidemic models
Malware propagation
Botnets

ABSTRACT

Botnets, overlay networks built by cyber criminals from numerous compromised network-accessible devices, have become a pressing security concern in the Internet world. Availability of accurate mathematical models of population size evolution enables security experts to plan ahead and deploy adequate resources when responding to a growing threat of an emerging botnet. In this paper, we introduce the *Susceptible-Infected-Connected* (SIC) botnet model. Prior botnet models are largely the same as the models for the spread of malware among computers and disease among humans. The SIC model possesses some key improvements over earlier models: (1) keeping track of only key node stages (*Infected* and *Connected*), hence being applicable to a larger set of botnets; and (2) being a Continuous-Time Markov Chain-based model, it takes into account the stochastic nature of population size evolution. The SIC model helps the security experts with the following two key analyses: (1) estimation of the global botnet size during its initial appearance based on local measurements; and (2) comparison of botnet mitigation strategies such as disinfection of nodes and attacks on botnet's Command and Control (C&C) structure. The analysis of the mitigation strategies has been strengthened by the development of an analytical link between the SIC model and the P2P botnet mitigation strategies. Specifically, one can analyze how a random sybil attack on a botnet can be fine-tuned based on the insight drawn from the use of the SIC model. We also show that derived results may be used to model the sudden growth and size fluctuations of real-world botnets.

1. Introduction

Botmasters, the cyber criminals behind botnets, leverage a wide range of malware vectors to infect network-accessible devices, with the majority of the devices being personal computers in homes, businesses, schools, and governments. Once infected, these devices (or *nodes*) form botnets and are remotely controlled by the botmasters for illicit activities such as sending e-mail spam and extortion by threats of launching Distributed Denial-of-Service (DDoS) attacks.

In recent years, the number of infected and remotely controlled nodes in each of the major botnets has reached the order of millions, e.g., the Mariposa botnet has been estimated to have 13 million computers across 190 countries [1]. Indicating how much botnets are responsible for e-mail spams on the Internet, a single takedown of a rogue ISP which hosted the main infrastructure of few botnets in November 2008 led to an instant drop of 80% in the level of e-mail spams [2]. In another incident, the country of Estonia came under a politically-motivated DDoS attack in April 2007 which knocked off critical infrastructure and the media [3]. The cumulative processing and bandwidth resources at the disposal of cyber criminals are therefore enough to severely attack any entity or temporarily knock

* Corresponding author. Tel.: +1 514 465 8769.

E-mail addresses: m.kh@ieee.org (M. Khosroshahy), mustafa@ece.concordia.ca (M.K. Mehmet Ali), dongyu@ece.concordia.ca (D. Qiu).

1389-1286/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.comnet.2012.07.020

entire countries off the Internet; this has resulted in the designation of botnets as a major security threat.

Analytical models may provide significant benefits in the fight against botnets. When either a new botnet threat emerges or an existing botnet goes into a rapid growth period due to a new infection, then there are two main questions that we would like to have answers to. One of them will be the determination of seriousness of the threat, which requires prediction of the size of the botnet as a function of time. This will let us know the number of nodes that eventually may be compromised. The other will be to determine the appropriate mix of mitigation strategies that need to be deployed to stop the growth of the botnet and possibly reverse it. In both cases, a good analytical model will be helpful if the estimates of its parameters are available. As a result of the growing botnet threat, new organizations are emerging that continuously keep track of botnets and measure their sizes. Thus, it is expected that the estimates of the model parameters will become available so that analytical models may be used to give answers to the above questions.

To this end, we develop an analytical model tailored to botnet, its expansion and evolution behaviors. Each Internet node/host goes through several stages during the lifetime of the botnet. The stages, and the back-and-forth transition between them, associated with an Internet node that can join a botnet are more complex compared to those of an infected computer (node) which remains isolated. These complex node stage characteristics lead to a botnet expansion behavior that cannot be explained or predicted using the available analytical models for computer malware propagation. Further, as shown in the Related Work section, recent analytical botnet models have not addressed this issue adequately. In this work, we intend to fill this gap.

The contribution of this paper is twofold: (1) the SIC botnet model which captures the key node stages relevant to botnets; we derive important results such as mean and variance of the number of nodes in different stages based on this model; and (2) development of a link between a botnet lifecycle/propagation/population model (the SIC model) and mitigation strategies aimed at Distributed Hash Table (DHT)-based Peer-to-Peer (P2P) botnets; with this analytical link, a security expert would be able to evaluate different mitigation strategies (disinfection, Sybil attack, Index Poisoning, etc.) prior to their implementation.

The paper is organized as follows: in Section 2, we examine the prior modeling efforts in this domain by introducing and analyzing both stochastic and deterministic botnet models. In Section 3, the SIC model is introduced by describing the Continuous-Time Markov Chain (CTMC) model as well as justifying the modeling assumptions. We then present an extensive performance modeling of the SIC Model in Section 4. First, the fundamental probability flow equations resulting from the CTMC model are presented. We then proceed to derive the means, variances, and Basic Reproduction Number of the SIC model. Afterwards, we introduce the developed link between the SIC model and mitigation strategies aimed at DHT-based P2P botnets. As a case study, we analyze a random sybil attack on a P2P botnet and examine how the attack can be fine-tuned based on the information provided by the SIC model. Next, we study in Section 5 how the results estimated by the SIC model would relate to some of the reported botnet size measurements. In Section 6, numerical results are provided showing the kinds of insight that can be drawn from the SIC model based on the aforementioned derived analytical results. Finally, we conclude the paper in Section 7 by providing some final thoughts as well as mentioning our future work.

2. Related work

Abstracting away the name of *actors* in the *system* to be modeled, developing analytical models for spread of computer virus, expansion of botnets and disease spread (biology) are similar problems. In the past two decades, researchers have adapted the analytical results from epidemiology to malware propagation and, recently, to botnet lifecycle modeling. We limit, however, the overview in this section to studies regarding botnet population/lifecycle modeling to ensure that the models can be reasonably compared to one another.

In computer science, the term *virus* was first used in the late 1980s to refer to a "self-replicating" code intended to do damage. Facing this new phenomenon, [4] was the first study that suggested the application of epidemiology for studying the propagation of computer virus. In the course of the two decades that followed, numerous other analytical models based on the same premises were proposed such as [5–8]. Before examining the related work, a few definitions are due:

> *Node Stage* A node (an arbitrary network-accessible device in the Internet) can be in either of the *stages* defined in the analytical model (e.g., *Susceptible* and *Infected* stages). With time, depending on the model, nodes can usually transition from one stage to another. In this paper, we use the term *stage* in the context of a node and the term *state* in the context of the whole system to avoid confusion; the terminology of the cited works has been adapted to be compatible with ours. *State* of the system, therefore, is used to indicate the number of nodes that are in each *stage* at any given time.
>
> *Lifecycle* indicates the fact that nodes change stage in the lifetime of the botnet. Botnet refers to the nodes that are in a certain stage, e.g., in the *Connected* stage in the SIC model. *Botnet lifecycle*, on the other hand, indicates the fact that the botnet itself appears, expands, shrinks, and disappears, as a collection of nodes that are in a certain stage within the overall system which is the Internet.

2.1. Stochastic vs. deterministic modeling

When considering the analytical models, it is important to consider that every analytical model for botnet expansion/lifecycle falls into either of the following two broad categories: deterministic and stochastic. While a deterministic model is easy to develop and analyze, it does not allow some critically important analysis permitted by a stochastic model which is relatively more difficult to construct and analyze. Specifically, the botnet population size is a stochastic process since dynamics of botnet expansion is probabilistic. In the deterministic models, the botnet population size is assumed to be a deterministic variable and the arrivals/departures to/from the population are also assumed to have deterministic values. As a result, the population size as a function of time is governed by an ordinary differential equation which is written in an ad hoc manner. The deterministic models may capture the mean population size accurately, however, this approach neither gives the distribution of the population size nor its higher moments. On the other hand, increasing the number of node stages causes a stochastic model to become intractable far more quickly in comparison to a deterministic model; therefore, when developing a stochastic model, it becomes imperative to limit the number of node stages considered. In what follows, stochastic models are introduced first, followed by deterministic ones.

2.1.1. Stochastic models

In [9], the population size of the Storm botnet has been studied through simulation of a Stochastic Activity Network (SAN) model (a variant of stochastic Petri nets). The SAN model and its parameters have been loosely based on the information gathered on the Storm Worm botnet. The SAN models the lifecycle of a node with four stages: *Susceptible*, *InitialBotInfection*, *ConnectedBot*, and *FullyConnectedBot*. It is assumed that the number of nodes in the *Susceptible* stage is *infinite* and the time intervals for a node to move from one stage to the next one in the last three stages are exponentially distributed with different parameters. It has been also assumed that the move of a node between stages succeeds with certain probability and unsuccessful nodes are removed from the experiment. Success probabilities may be used to account for the impact of mitigation strategies on the growth of the botnet. The paper presents simulation results for the mean population size of nodes in *FullyConnectedBot* stage as a function of time for different success probabilities between stages. It may be seen that when success probability is one, the botnet grows exponentially.

Ref. [10] has introduced “genetic mechanism” as the topology construction mechanism of botnets. Through this modeling method, they study “in-degree distribution”, shortest distance, and clustering coefficient of the constructed topology. The study, however, lacks results regarding botnet size and various parameters thereof.

Ref. [11] investigated P2P botnet topologies using the stochastic Monte Carlo *simulation*. Under worm infection and user countermeasures, the metrics of “number of peers” and “botnet size” have been studied which leads to the determination of robustness and effectiveness of the formed P2P botnets. Like [9], the usefulness of this study is limited due to the used simulation environment and the study lacks formulas to examine the botnet size, which in general limits any botnet analysis by a third party.

A probability model to estimate the number of machines infected per hour with the Conficker-C worm has been presented in [12]; the work includes derivation of the distribution of the number of hourly UDP connection attempts made by an infected host and the conditional distribution of the number of observed hits in the monitored IP space. While being a solid analytical study, this is a *one-stage* model, i.e., it cannot keep track of more than one node stage. An analytical botnet model, however, needs to simultaneously keep track of at least two sets of nodes: (1) nodes which are infected by the initial malware; and (2) the infected nodes which subsequently managed to join the botnet.

A schematic diagram representing the movement of nodes between several stages was presented in [13] in order to model “botnet propagation”. The study, however, lacks an analytical, simulation, or measurement component, nor does it have an accompanying quantitative or qualitative analysis. It therefore does not seem possible to evaluate this study in the current form.

Finally, [14] presented a model of worm’s propagation probability in a P2P overlay network using a fully-connected graph. This model is limited to small networks, however, as having a square matrix of dimension n, with n being the number of nodes in the network, to define and examine the network topology and botnet size leads to the model being unusable for Internet-scale scenarios.

2.1.2. Deterministic models

Inspired by epidemic models, there have been several deterministic models proposed in recent years [15–18,8] based on ordinary differential equations describing the flow of nodes from one stage to another; these are briefly described as follows: [15] includes a model for the growth of the presented P2P botnet which is dependent on the number of target hosts that can be infected at any one time. [16] extended the classic Susceptible-Infectious-Removed (SIR) model by taking into account the diurnal pattern, i.e., the effect of time zones in malware propagation. It is important to note that the SIC model proposed in this paper is a model to estimate the botnet footprint/total size (i.e., not just live/awake nodes) at any given time; diurnal patterns do not affect botnet footprint/total size. Using the domain name redirection technique to gather data on the Conficker botnet, [17] customized the SIR epidemic model. [18], on the other hand, analyzes the relationship between the number of infected hosts and propagation ratio based on the SIR model, drawing an insight regarding the effects of different propagation ratios on botnet scale and stability.

We conclude the Related Work section by introducing and examining [8], as an example of prior botnet models, as follows: this model assumes finite node population of size N and the lifecycle of nodes consists of four stages: (1) S stage: susceptible nodes that can become infected; (2) I stage: infectious nodes that can infect the susceptible nodes; (3) V stage: infectious nodes that can infect the susceptible nodes on top of being active in botnet’s illicit activities (nodes autonomously and probabilistically change stage between V and I); and (4) R stage: removed/disinfected nodes that remain immune to all future infection vectors utilized by the botmasters. The nodes in stage V can either transition to stage R with the rate γ or transition back to stage *S* with the rate ρ. After the derivation of a system of equations for the rates of change of (normalized) number of nodes in various stages, the authors then proceed to present some figures regarding the evolution of variable values, focusing in each case on changing a specific parameter; one such equation is as follows: $\frac{ds(t)}{dt} = -\beta[i(t) + v(t)]s(t) + \rho v(t)$.

The above work has several limitations: (1) this is a deterministic model and does not account for the stochastic nature of botnet node population changes; (2) the analysis only leads to the mean number of nodes in different stages of node lifecycle and higher moments cannot be obtained. Further, the results may only be calculated numerically and no closed form results are obtained for the mean values; and (3) in the model, new infections depend on the number of nodes in *Infected* stages (*I* and *V*), which is not usually the case in botnets (*Infected* nodes not yet part of the botnet are generally not able to cause new infections, e.g., see [19]).

3. The SIC model

In this section, we present our botnet lifecycle model and then develop its mathematical representation. We first introduce the model basics and later elaborate on the main assumptions of the model.

3.1. Introduction

As reported extensively in the literature [20,21,9,22], a node, when infected by a botnet-related malware, goes through multiple stages in the lifetime of the botnet, with the main stages being *Susceptible*, *Infected*, and *Connected*. Here are the definitions of these terms, as used in this paper:

Susceptible (S) A node is considered to be in the Susceptible stage, if it is healthy, whether or not vulnerable. A vulnerable node can be infected through at least one of the possibly many infection vectors (worm scans, e-mail attachments, etc.) deployed simultaneously or sequentially by the botmasters of a single botnet. On the other hand, a node is invulnerable if either it cannot be infected by any infection vector or the address is either unused or unroutable/unreachable. As defined, the Susceptible node population corresponds to the entire population of the Internet. The term Susceptible refers to the fact that until probed, one usually cannot determine whether or not the node is vulnerable. A Susceptible node may either get infected with the small probability *p* and possibly later become part of the botnet or remain healthy throughout the whole period with the large probability of $1 - p$. All nodes are initially considered to be in the Susceptible stage.

Infected (I) The *Infected* stage denotes a stage in which a node has been infected by any of the infection vectors that have been utilized by the botmasters. In this stage, the node usually does not have the full malware code to engage in illicit activities; this is primarily for keeping the payload small. The minimal malware code serves only to connect the node to the botnet and pass the node to the *Connected* stage.

The *Connected (C)* stage refers to the stage when the node is connected to the botnet, can download the full malware code and receive the botmasters' Command & Control (C&C) traffic, and therefore, it is part of the army of bots controlled by the botmasters.

As we model the lifecycle of a node with the aforementioned three stages, the model is referred to as the Susceptible-Infected-Connected (SIC) model. In Fig. 1, we show the stages of the model and the transitions between the stages. As shown in the figure, we let n_1 and n_2 denote the number of nodes in *Infected* and *Connected* stages, respectively, and the state of the system is represented by the vector (n_1, n_2). In Fig. 2, we show all the transitions from and to state (n_1, n_2).

In this model, we consider that each node in the botnet (nodes in *Connected* stage) infects one *Susceptible* node (increases n_1 by one) with probability $\lambda_1 \Delta t + o(\Delta t)$ in any Δt interval (cf. Fig. 1). Thus the time interval for a Connected node to infect a Susceptible node is exponentially distributed with parameter λ_1 and the transition rate between Susceptible and Infected stages is given by $\lambda_1 n_2$. Further, each Infected node can transition to Connected stage (which increases n_2 and decreases n_1) with probability $\lambda_2 \Delta t + o(\Delta t)$ in any Δt interval. Finally, there is a transition rate ($\lambda_a n_2$) from Connected stage to Infected stage. This transition rate represents an attack on the botnet, attacks such as index poisoning and sybil attacks in the case of P2P botnets. Under such attacks, nodes do not transition back to Susceptible stage; they just lose the ability to communicate and might be able to reconnect again (hence the rate from Connected stage back to Infected stage). We further assume the rate of disinfection of nodes which are in Infected stage and Connected stage to be $\lambda_{r1} n_1$ and $\lambda_{r2} n_2$, respectively.

3.2. Model assumptions

In this sub-section, we put forward the reasoning behind the assumptions made in the development of the SIC model. To the best of our knowledge, these assumptions are reasonable mathematically as well as consistent with precedence and evidence from closely-related phe-

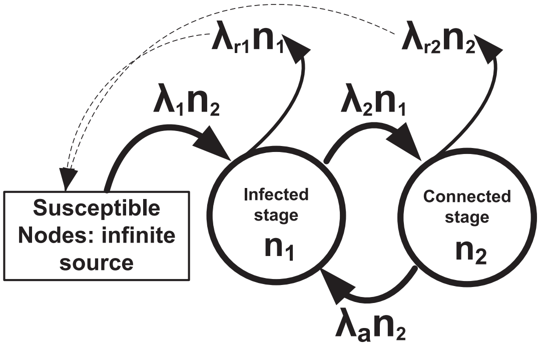


Fig. 1. SIC model: 2-dimensional birth–death CTMC.

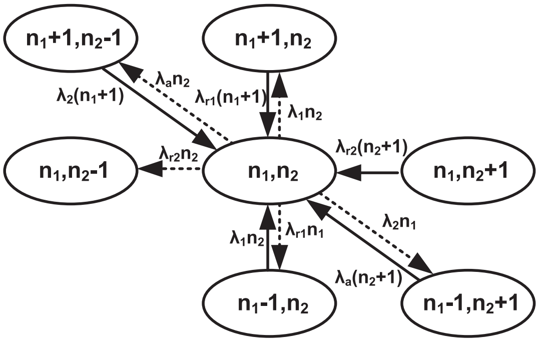


Fig. 2. SIC model: State-transition-rate diagram.

nomena such as malware propagation and spread of human disease.

3.2.1. CTMC (Exponential probability distributions) modeling

Continuous-time Markov Chain (CTMC) models are based on the assumption that the time intervals for the transitions of nodes from one stage to the next one are exponentially distributed with different parameters. In this part, we first provide the mathematical basis for the use of exponential distributions and then describe how this assumption is in agreement with precedence.

3.2.1.1. Mathematical basis. Each attempt to make a node transition from any stage to another stage is a Bernoulli trial with success probability of p. We explain how this concept of Bernoulli trial corresponds to physical reality first for the transition from Susceptible to Infected (S → I) which is the most important transition in the model leading to the exponential growth of the number of nodes in the Infected and Connected stages. At the end, we briefly explain how the same concept of Bernoulli trial also corresponds to physical reality for the rest of the transitions in the SIC model.

Most botnets apply worm-scanning methods to recruit new bots [23]. Further, it has been reported that 66.5% of scan patterns are *uniform* random scanning [24]. To explain the process with a concrete example, we therefore consider *uniform random scanning* as the infection vector used by the botnet node. Using the terminology presented in [25], for a uniform scan worm, η is the average scan rate, i.e., the average number of scans a botnet node sends out per unit of time. Each scan corresponds to an attempt to infect a susceptible node. If the susceptible node is vulnerable to this specific worm scan, then, it will be infected, otherwise the attack will fail and the node will remain healthy. η is therefore equal to m, which is the number of aforementioned Bernoulli trials. The campaign of a single botnet node to infect can then be viewed as a series of Bernoulli trials with few successes/infections among many failures.

The above series of Bernoulli trials has therefore a Binomial distribution with parameters p (success probability) and m (number of Bernoulli trials). A Binomial distribution can be approximated by a Poisson distribution with parameter $\lambda = mp$, when p is small and m is large [26, pp. 111–113]. Note that m is different from n_1 and n_2 which denote the numbers of Infected and Connected nodes, respectively, however, the λ parameter refers to λ_1 indicated in Fig. 1. The conditions on the values of m and p are consistent with the S → I transition, as the success probability is low and the number of trials is large. Therefore, the probability distribution of the number of nodes making transitions in a unit time period can be approximated by this Poisson distribution. Further, as sum of processes each having a Poisson distribution with parameter λ_1 also has a Poisson distribution, the whole arrivals into the Infected stage due to all botnet nodes then have a Poisson distribution with parameter $\lambda_1 n_2$. From the Poisson distribution, it follows that the time intervals between node arrivals to the Infected stage are exponentially distributed.

As noted at the beginning, we provide a brief explanation regarding how the same concept of Bernoulli trial also corresponds to physical reality for the rest of the transitions in the SIC model as follows:

I → C Each Infected node, which has the minimal malware code to help it to connect itself to the botnet, makes, on average, several attempts to either connect to the central C&C server or find peers in a P2P botnet. As such, we can designate a success probability of p for the successful connection to the botnet for these attempts each of which can be considered a Bernoulli trial.

C → I When the botnet is under attack, the effort to disconnect each botnet node can also be considered a Bernoulli trial with a success probability of p which is the probability of disconnection. As botnet mitigation strategies are generally complicated and hard to implement with often limited impact on the botnet, on average, this per-node success probability is small.

I → S & C → S Similar to the attack on the botnet, each attempt to disinfect a node that is in either stages of Infected or Connected can be considered a Bernoulli trial with a success probability of p, i.e., the probability of disinfection. As the identification of most nodes as well as the physical access to them are hard, on average, this success probability is small.

With the aforementioned descriptions for the characteristics of all the inter-stage transitions, the CTMC model can be considered a reasonable approximation.

3.2.1.2. Accordance with precedence. CTMC as a modeling tool in epidemiology has a proven track record [27] that deals with the phenomenon of spread of an element within a susceptible population which has a close resemblance to the spread of malware and the expansion of a botnet. Further, successful use of CTMC models in the study of spread of malware has also been documented [28]. Expansion, and size evolution, of botnets happen under the influence of the same physical processes as the ones affecting the spread of malware; therefore, the use of the same CTMC theory for botnets is a natural extension. To our knowledge, the only case of application of CTMC-like models to botnets is the work of [9] which is a simulation model that has been developed based on the measurement data of the

Storm botnet. Finally, in terms of the choice of Poisson distribution for the arrival of nodes into a stage (i.e., the exponentially-distributed inter-arrival times), similar to the SIC model, [12] has also determined this assumption to be reasonable in the study of Conficker-C botnet/worm for the distribution of the number of UDP connection attempts made by an infected host.

3.2.2. Node stages and transitions

3.2.2.1. Main node stages considered. As described in Section 3.1, the main dynamics of botnets can be captured by keeping track of the two main node stages, i.e., *Infected* and *Connected*. On the other hand, as mentioned in Section 2.1, the number of stages considered in a stochastic model, and in our CTMC model in particular, must be limited if we are to avoid an intractable model caused by consideration of several node stages. Based on our extensive investigations and considering the prior work done in this field, the optimal tradeoff has been determined to be the consideration of the aforementioned two node stages (i.e., *Infected* and *Connected*), each being a dimension in the CTMC (hence the number of nodes in each of these two stages is tracked) with *Susceptible* stage having infinite number of nodes (hence the number of Susceptible nodes need not be tracked). An infinite susceptible population is a reasonable assumption, since this population corresponds to the population of the entire Internet which is an assumption made also in [9].

3.2.2.2. No Immune/Removed stage considered. As botmasters use a plethora of methods to infect (and re-infect) the nodes, it is reasonable to assume that a node is never in Immune (or Removed) stage; therefore, we do not consider this stage in our model. It is important to remember that existence and maintenance of a botnet is independent of any infection vector (e-mail attachments, file sharing sites, worm scans, etc.) used by the botmaster and obtaining immunity against one infection vector still leaves the node susceptible to be re-infected through other infection vectors.

3.2.2.3. Botnet's footprint vs. live population. Using the terminology presented in [29], we emphasize that the SIC model tracks the botnet's "footprint" and not its "live population". As such, effects such as day/night differences and time zones which impact the number of live botnet nodes at any given time, are not taken into account. In the SIC model, *Connected* nodes represent the total number of botnet nodes, i.e., botnet's footprint. On the other hand, it is possible to use the SIC model and take into account the effects of time zones and day/night differences on λ parameters' values as follows depending on the length of the analysis period: (1) if the analysis period is around or less than 24 h, then, piecewise time-invariant parameters can be used, i.e., we use different sets of values for the λ parameters in each 12-h analysis period to account for the day/night differences and/or the time zones; and (2) if the analysis period is significantly more than 24 h, e.g., weekly size variations are important as is the case in the analysis of FourLakeRiders botnet in Section 5, then, the variations due to time zones and day/night differences are insignificant and average parameter values will yield accurate results.

3.2.2.4. Accommodating time-variant parameters. As presented, the λ parameters are considered constant throughout the analysis period. It is however possible to use the SIC model if these parameters change over time using piecewise time-invariant parameters, i.e., in each piece of the analysis period, we consider the parameters to be constant. The duration of each piece can be decided upon on a case-by-case basis; an example of this kind of analysis, with each piece duration to be a week, is presented in Section 5. Another example of this kind of analysis, as suggested in the above point, is to accommodate the effects of time zones and the day/night differences when the analysis period is less than 24 h. In this case, we can choose a 12-h analysis period during which we consider the λ parameters to be constant and can set low values for the λ parameters during night time.

As described above, the SIC model and its main assumptions are similar to the model in [9] which has been based on the gathered information about the Storm botnet. These assumptions were further justified mathematically and through comparison to other similar works. As a result, we believe that we have a realistic model, which leads us to two-dimensional Markovian birth–death processes. Using the model, we can study the size evolution of a botnet as well as effectiveness of mitigation strategies by monitoring the number of nodes that are in Infected and Connected stages at any given time.

4. Performance modeling of the SIC model

In this section, we provide an extensive performance modeling of the SIC model. First, botnet size evolution phases and initial state values for the SIC model are explained. We then proceed to derive the probability flow equations based on the two-dimensional CTMC of the SIC model. These probability flow equations are further reduced to a partial differential equation (PDE) of the probability generating function (PGF). Directly from this PDE, we then derive the mean and variance of the SIC model. Next, the derivation of the Basic Reproduction Number, which is a widely used parameter in epidemiology and the study of malware propagation, is documented. We conclude this section by deriving a novel analytical result which is a link between the SIC model and the mitigation strategies against Distributed Hash Table (DHT)-based P2P botnets.

4.1. Botnet size evolution phases and initial state values

A botnet may go through many phases during its lifecycle, where a phase will refer to a period that system parameters ($\lambda_1, \lambda_2, \lambda_{r1}, \lambda_{r2}$, and λ_a) remain constant. For example, when a botnet appears for the first time, it will probably experience unhindered expansion as there will not be any active mitigation strategies to counter its growth, thus $\lambda_{r1}, \lambda_{r2}$, and λ_a will be zero. Typically, the botnet's population will alternate between sawtooth growth period followed by a period of relatively stable population size [30]. The sawtooth growth begins with the release of a new infection; after sometime, it will be reversed with the deployment of new counter measures until an equilibrium is reached. Probably, the new equilibrium population will have a size greater than previous equilibrium size. In any phase, the SIC model will apply with the end results of the preceding phase providing the initial conditions (state values) to the next phase.

4.2. Probability flow equations and PDE of PGF

In this section, we determine the probability flow equations and then, the partial differential equation (PDE) of the probability generating function (PGF) describing the system. We write probability flow equations through inspection from the state-transition-rate diagram given in Fig. 2 by equating the rate of change of probabilities at any state to the difference between the total input/output flows to/from that state. Let $P_{n_1,n_2}(t)$ denote the probability that the system is in state (n_1, n_2) at time t, then the probability flow equations are given by:

$$\begin{cases} \frac{dP_{n_1,n_2}(t)}{dt} = \lambda_1 n_2 P_{n_1-1,n_2}(t) + \lambda_{r1}(n_1+1)P_{n_1+1,n_2}(t) \\ \quad +\lambda_{r2}(n_2+1)P_{n_1,n_2+1}(t) \\ \quad +\lambda_2(n_1+1)P_{n_1+1,n_2-1}(t) + \lambda_a(n_2+1)P_{n_1-1,n_2+1}(t) \\ \quad -(\lambda_1 n_2 + \lambda_{r1} n_1 + \lambda_{r2} n_2 + \lambda_2 n_1 + \lambda_a n_2)P_{n_1,n_2}(t) \\ \quad \langle n_1 > 0, n_2 > 0 \rangle\,(a) \\ \frac{dP_{0,n_2}(t)}{dt} = \lambda_{r1} P_{1,n_2}(t) + \lambda_{r2}(n_2+1)P_{0,n_2+1}(t) + \lambda_2 P_{1,n_2-1}(t) \\ \quad -(\lambda_1 n_2 + \lambda_{r2} n_2 + \lambda_a n_2)P_{0,n_2}(t) \\ \quad \langle n_1 = 0, n_2 > 0 \rangle\,(b) \\ \frac{dP_{n_1,0}(t)}{dt} = \lambda_{r1}(n_1+1)P_{n_1+1,0}(t) + \lambda_{r2} P_{n_1,1}(t) + \lambda_a P_{n_1-1,1}(t) \\ \quad -(\lambda_{r1} n_1 + \lambda_2 n_1)P_{n_1,0}(t) \\ \quad \langle n_1 > 0, n_2 = 0 \rangle\,(c) \\ \frac{dP_{0,0}(t)}{dt} = \lambda_{r1} P_{1,0}(t) + \lambda_{r2} P_{0,1}(t) \\ \quad \langle n_1 = 0, n_2 = 0 \rangle\,(d) \end{cases} \tag{1}$$

In order to solve (1) and derive the probability distribution $P_{n_1,n_2}(t)$, a known method is to transform the equations of probability flows to a partial differential equation (PDE) of the probability generating function (PGF) which can be tackled using known methods to solve PDEs. The relationship between the PGF $P(z_1, z_2, t)$ and the probability distribution $P_{n_1,n_2}(t)$ is as follows: $P(z_1, z_2, t) = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} P_{n_1,n_2}(t) z_1^{n_1} z_2^{n_2}$.

The initial probability distribution is denoted by $P_{k_1,k_2}(0)$. Here, we assume that the initial number of nodes in each stage is constant (k_1, k_2). Though the initial derivations are conditional, we will suppress the conditions for simplicity in expressing the PDE. This aspect, however, has been fully taken care of in the derivation of means (e.g., see (C.5) and (C.6)) and later in the derivation of variances.

We multiply each of the equations in (1) by $z_1^{n_1} z_2^{n_2}$, sum over the respective ranges of n_1 and n_2, and then add them together. After some simplifications and manipulations (detailed derivation provided in Appendix A), we arrive at the following PDE of the PGF:

$$\begin{aligned} &(\lambda_{r1} + \lambda_2 z_2 - \lambda_{r1} z_1 - \lambda_2 z_1)\frac{\partial P(z_1, z_2, t)}{\partial z_1} + (\lambda_1 z_1 z_2 + \lambda_{r2} \\ &\quad + \lambda_a z_1 - \lambda_1 z_2 - \lambda_{r2} z_2 - \lambda_a z_2)\frac{\partial P(z_1, z_2, t)}{\partial z_2} \\ &\quad - \frac{\partial P(z_1, z_2, t)}{\partial t} = 0 \end{aligned} \tag{2}$$

Our efforts to solve the preceding PDE, however, have not been successful, as detailed in Appendix B. Nonetheless, there are publications reporting new solved cases of Abel/Lienard equations (differential equations encountered in the process of solving the PDE). Thus, it is possible that we may have the solution of the PDE in the near future. We can still obtain from the PDE the moments of botnet population size, as presented next.

4.3. Derivation of the time-dependent mean and variance of botnet population size

In this section, we derive the means and variances of the number of nodes in *Infected* stage and *Connected* stage (botnet population size) as a function of time. Let $E_t[n_1]$ and $E_t[n_2]$ denote the mean number of nodes that are in *Infected* and *Connected* stages at time t, respectively, then:

$$\begin{aligned} E_t[n_1] &= \frac{\partial P(z_1, z_2, t)}{\partial z_1}\Big|_{z_1=z_2=1}, \\ E_t[n_2] &= \frac{\partial P(z_1, z_2, t)}{\partial z_2}\Big|_{z_1=z_2=1} \end{aligned} \tag{3}$$

We take the derivatives of the PDE given in (2) with respect to z_1 and z_2, separately. By setting $z_1 = z_2 = 1$ in each resulting equation, we arrive at a set of ODEs of $E_t[n_1]$ and $E_t[n_2]$. To emphasize the time dependency of the means from here on, we will denote $E_t[n_1]$ and $E_t[n_2]$ by $E_1(t)$ and $E_2(t)$, respectively. Note that only the important steps of derivation are provided here; the rest of the steps is in Appendix C. After the initial steps outlined above and detailed in the appendix, we arrive at the following set of ODEs:

$$\begin{cases} \frac{dE_1(t)}{dt} = (\lambda_1 + \lambda_a)E_2(t) - (\lambda_2 + \lambda_{r1})E_1(t) \\ \frac{dE_2(t)}{dt} = \lambda_2 E_1(t) - (\lambda_{r2} + \lambda_a)E_2(t) \end{cases} \tag{4}$$

We then proceed to derive $E_1(t)$ and $E_2(t)$ from the previous set of ODEs as detailed in Appendix C; the final results are as follows:

$$\begin{aligned} E_1(t) = \Big[&\exp\left(-\frac{1}{2}t(\lambda_{T3} + \lambda_{T1})\right)\Big(\overline{k_1}\lambda_2(-\exp(t\lambda_{T3})) \\ &+ \left(\overline{k_1}\lambda_a - \overline{k_1}\lambda_{r1} + \overline{k_1}\lambda_{r2} + \overline{k_1}\lambda_{T3} + 2\lambda_1\overline{k_2}\right)\exp(t\lambda_{T3}) \\ &+ 2\overline{k_2}\lambda_a\exp(t\lambda_{T3}) + \overline{k_1}\lambda_{T3} + \overline{k_1}\lambda_2 - \overline{k_1}\lambda_a + \overline{k_1}\lambda_{r1} \\ &- \overline{k_1}\lambda_{r2} - 2\lambda_1\overline{k_2} - 2\overline{k_2}\lambda_a\Big)\Big]/(2\lambda_{T3}) \end{aligned} \tag{5}$$

$$\begin{aligned} E_2(t) = \Big[&\exp\left(-\frac{1}{2}t(\lambda_{T3} + \lambda_{T1})\right)\Big(2\overline{k_1}\lambda_2\exp(t\lambda_{T3}) \\ &+ \left(\lambda_2\overline{k_2} - \overline{k_2}\lambda_a + \overline{k_2}\lambda_{r1} - \overline{k_2}\lambda_{r2} + \overline{k_2}\lambda_{T3}\right)\exp(t\lambda_{T3}) \\ &- 2\overline{k_1}\lambda_2 + \overline{k_2}\lambda_{T3} - \lambda_2\overline{k_2} + \overline{k_2}\lambda_a - \overline{k_2}\lambda_{r1} + \overline{k_2}\lambda_{r2}\Big)\Big]/(2\lambda_{T3}) \end{aligned} \tag{6}$$

where $\lambda_{T1} = \lambda_2 + \lambda_a + \lambda_{r1} + \lambda_{r2}$, $\lambda_{T2} = -\lambda_1\lambda_2 + \lambda_{r2}(\lambda_2 + \lambda_{r1}) + \lambda_a\lambda_{r1}$, and $\lambda_{T3} = \sqrt{\lambda_{T1}^2 - 4\lambda_{T2}}$.

Next, we describe the derivation of variances, which are given by:

$$\sigma_1^2(t) = E_t[n_1^2] - (E_1(t))^2, \quad \sigma_2^2(t) = E_t[n_2^2] - (E_2(t))^2 \tag{7}$$

where

$$\begin{aligned} E_t[n_1^2] &= \frac{\partial^2 P(z_1,z_2,t)}{\partial z_1^2}\Big|_{z_1=z_2=1} + \frac{\partial P(z_1,z_2,t)}{\partial z_1}\Big|_{z_1=z_2=1} \\ E_t[n_2^2] &= \frac{\partial^2 P(z_1,z_2,t)}{\partial z_2^2}\Big|_{z_1=z_2=1} + \frac{\partial P(z_1,z_2,t)}{\partial z_2}\Big|_{z_1=z_2=1} \end{aligned} \tag{8}$$

Let us define

$$\begin{aligned} \psi_1(t) &\triangleq \frac{\partial^2 P(z_1,z_2,t)}{\partial z_1^2}\Big|_{z_1=z_2=1} \\ \psi_2(t) &\triangleq \frac{\partial^2 P(z_1,z_2,t)}{\partial z_2^2}\Big|_{z_1=z_2=1} \\ \psi_{12}(t) &\triangleq \frac{\partial^2 P(z_1,z_2,t)}{\partial z_1 \partial z_2}\Big|_{z_1=z_2=1} \end{aligned} \tag{9}$$

Considering that $E_1(t=0) = \overline{k_1}$ and $E_2(t=0) = \overline{k_2}$, the preceding functions have the following initial values:

$$\begin{aligned} &\psi_1(t=0) = \overline{k_1^2} - \overline{k_1}, \quad \psi_2(t=0) = \overline{k_2^2} - \overline{k_2}, \\ &\psi_{12}(t=0) = \overline{k_1 k_2} \end{aligned} \tag{10}$$

The variances are then given by:

$$\begin{aligned} \sigma_1^2(t) &= \psi_1(t) + E_1(t) - (E_1(t))^2, \\ \sigma_2^2(t) &= \psi_2(t) + E_2(t) - (E_2(t))^2 \end{aligned} \tag{11}$$

Next, we take the 2nd derivatives of the PDE in (2) with respect to z_1 and z_2, separately. Further, we take the derivative of the PDE with respect to z_1 and then with respect to z_2 (see Appendix D). By setting $z_1 = z_2 = 1$ in each resulting equation, we arrive at a set of ordinary differential equations, which if written in terms of $\psi_1(t), \psi_2(t)$, and $\psi_{12}(t)$ is, as follows:

$$\begin{cases} \frac{d\psi_1(t)}{dt} = 2(\lambda_1 + \lambda_a)\psi_{12}(t) - 2(\lambda_{r1} + \lambda_2)\psi_1(t) \\ \frac{d\psi_2(t)}{dt} = 2\lambda_2\psi_{12}(t) - 2(\lambda_{r2} + \lambda_a)\psi_2(t) \\ \frac{d\psi_{12}(t)}{dt} = -(\lambda_{r1} + \lambda_2 + \lambda_{r2} + \lambda_a)\psi_{12}(t) + \lambda_2\psi_1(t) + \lambda_1 E_2(t) + (\lambda_1 + \lambda_a)\psi_2(t) \end{cases} \tag{12}$$

Finally, from the preceding set of ODEs, we obtain the variances, as explained in Appendix D.

4.4. Epidemiological threshold: basic reproduction number

Basic Reproduction Number[1] (R_0) is a widely used parameter in epidemiology as well as in the study of malware propagation. In the context of botnets, this number is the mean number of infections that any single botnet node can cause among the population of susceptible nodes. The measurement of the mean number is assumed to happen with the presence of mitigation strategies that bring down the number of botnet nodes while the remaining botnet nodes cause new infections. R_0 is calculated based on the rates used in the model. If $R_0 < 1$, the botnet will eventually disappear with probability one. If $R_0 > 1$, however, there is a probability that the botnet size will continue to increase exponentially.

Based on (4), the Basic Reproduction Number (R_0) can be derived in terms of various SIC model's parameters using the "Next Generation Matrix" method as follows (detailed derivation in Appendix E):

$$R_0 = \sqrt{\frac{\lambda_2(\lambda_1 + \lambda_a)}{(\lambda_{r2} + \lambda_a)(\lambda_2 + \lambda_{r1})}} \tag{13}$$

4.5. P2P botnet mitigation strategies and the SIC model

As our last analytical result, we present a link between lifecycle (or propagation/population) models and the P2P botnet mitigation strategies. Mitigation strategies aimed at Distributed Hash Table (DHT)-based P2P botnets include sybil, index poisoning, and eclipse attacks. We base the discussion on random sybil attack; however, the process is similar for other attack types.

Sybil attack, first presented in [31], is an attack method under which numerous *clean* nodes (sybils) are injected into the P2P botnet, posing themselves as "legitimate" botnet nodes. They then try to re-route, block, and corrupt the Command & Control (C&C) traffic, thereby lowering the efficiency of the C&C mechanism of the botnet. In a DHT-based P2P botnet, nodes find each other, construct their routing tables, and relay the traffic to, or closer to, its intended destination based on normal DHT methods. The botmaster also relies on the aforementioned methods for the C&C of the botnet; therefore, the decreased efficiency of the C&C mechanism as a result of the sybil attack translates into an inefficient botnet.

Random sybil attack on P2P botnets has been studied in [21]. The derived formula therein can be used to construct a relationship between the number of sybils inserted in the network and $\lambda_a n_2$, the transition rate from *Connected* stage to *Infected* stage (cf. Fig. 1). The obtained formula for the random sybil attack is the following [21]:

$$P_s(n_s) = \left(1 - \frac{n_s}{n_s + n}\right)^{\frac{\log_2(n_s+n)}{b}} \tag{14}$$

where $P_s(n_s)$ is the probability that a botnet node successfully obtains the commands of the botmaster. n_s is the number of sybils inserted randomly in the network. b is the number of bits improved per step for a lookup (set to a mid-range value of 5 in our study [21]). n is the botnet size which is the value of n_2 in our model.

We therefore note that $1 - P_s(n_s)$ is the probability that a botnet node is no longer able to receive the commands of the botmaster as a result of the attack on the botnet (insertion of sybils). This probability is therefore equal to

[1] In the theoretical epidemiology literature [27], Basic Reproduction Number (R_0) generally refers to the onset of disease spread. Once the epidemic is underway, and especially when control measures (mitigation strategies) are put into effect, other terminologies such as "Control Reproduction Number (R_c)" and "Effective Reproduction Number (R_e)" are used instead to refer to essentially the same threshold parameter. In this paper, we use the phrase "Basic Reproduction Number (R_0)" in all instances.

$\lambda_a \Delta t$, as the latter is the approximate probability that a botnet node transitions from *Connected* stage to *Infected* stage (i.e., the node gets disconnected).

The aforementioned link between lifecycle models and the P2P botnet mitigation strategies is therefore demonstrated using the following formula:

$$\lambda_a \Delta t = 1 - P_s(n_s) \tag{15}$$

As seen in (14), $P_s(n_s)$ is a function of n_s. At any instant of time, a change in λ_a (i.e., $\Delta\lambda_a$) is a result of a change in the number of sybils (i.e., Δn_s). Based on (15), we can then analyze the relationship between the amount of change of λ_a with respect to a change in the number of sybils inserted in the network as follows:

$$\frac{\lambda_a + \Delta\lambda_a}{\lambda_a} = \frac{1 - P_s(n_s + \Delta n_s)}{1 - P_s(n_s)} \tag{16}$$

5. SIC model vs. reported botnet measurements

In this section, we show that our results can be used to model the botnets in the real-world. Measurements of the size of some botnets have been reported on a weekly basis by Damballa [30]. Assuming that the employed measurement techniques capture correctly the global size of the botnets, in this section we examine how such measurement results would compare to the results predicted by the SIC model. First, we examine a case of initial unhindered botnet expansion, based on available data from a *Zeus*-based botnet called *GreenAlienRiders*. Next, we will examine a case of deployment of mitigation strategies, based on available data from another *Zeus*-based botnet called *FourLakeRiders*.

GreenAlienRiders is a botnet for which the initial unhindered expansion phase has been captured and reported by Damballa [30]. From the Damballa report, it appears that the botnet has reached the size of about 6000 nodes at hour 12 of its appearance. To reach this size, using the SIC model, we can set $\lambda_1 = 6.85$ and $\lambda_2 = 0.1$ (both nodes/h). The result is shown in Fig. 3. Further, Fig. 3 also shows the SIC model's estimate of the existing *Infected* nodes during this period.

FourLakeRiders, on the other hand, is a botnet for which deployment of mitigation strategies can be analyzed based on a portion of data of the botnet size evolution over time, a 5-week period from week 36 to week 40, as captured and reported by Damballa [30]. The data reported for this 5-week period lends itself to an analysis with clear separation of effects of each of the mitigation strategies. The scenario that follows, however, represents one of potentially many possibilities. The reported data on botnet size during this 5-week period is depicted in Fig. 4a. A scenario that fits this pattern of rise-and-fall is as follows: during week 36, the botnet size has reached an equilibrium; on one side, the number of *Infected* and *Connected* nodes grow, and on the other side, some mitigation strategies are reducing the number of *Infected* and *Connected* nodes ($\lambda_{r1}, \lambda_{r2}$). During week 37, the aforementioned mitigation strategies weaken and, during week 38, they completely disappear, which results in a steep growth of the size of the botnet. During weeks 39 and 40, all mitigation strategies are employed ($\lambda_{r1}, \lambda_{r2}$, and λ_a), which results in a dramatic reduction in the size of the botnet. The described scenario, and the chosen parameter values to make it happen, are depicted and mentioned in Fig. 4b. The potential number of *Infected* nodes are estimated using the SIC model as well, as depicted in Fig. 5. As may be seen, during both expansion and shrinkage, our results follow quite well the reported data.

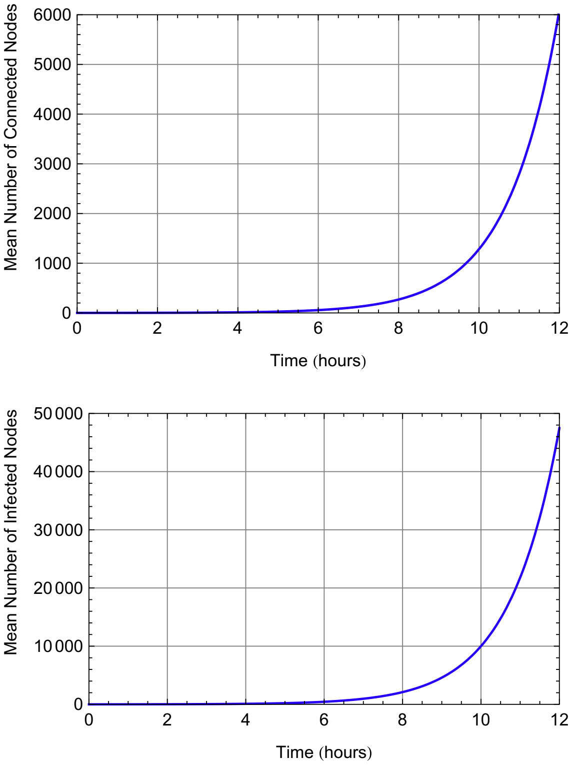
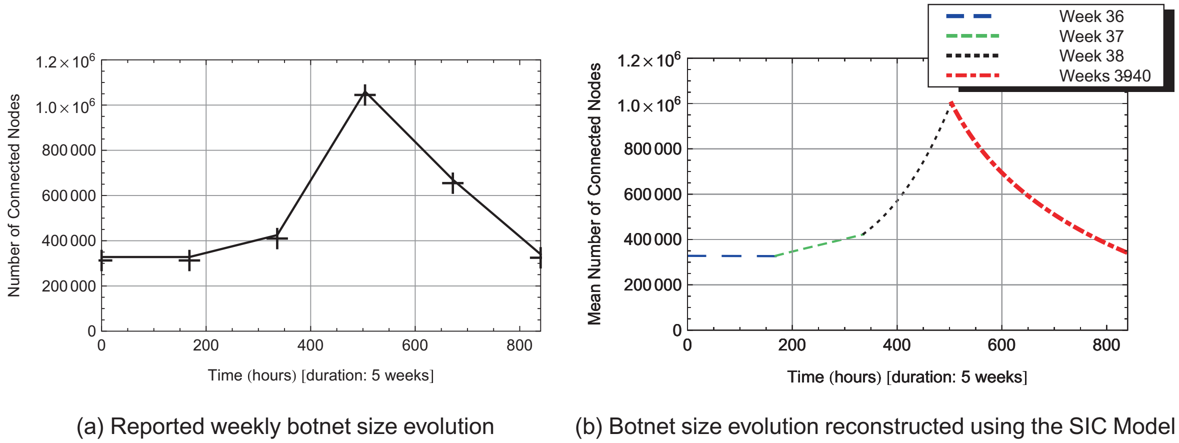


Fig. 3. GreenAlienRiders (a Zeus-based botnet): initial unhindered botnet expansion estimated using the SIC model.

6. Numerical analysis

In this section, we present some numerical results to further illustrate the usefulness of the SIC model. First, we briefly introduce some parameter estimation techniques which help with the use of the SIC model. The first set of numerical results are with regard to the analysis of the initial unhindered expansion of a botnet. We then show how the SIC model could help with the evaluation and comparison of mitigation strategies. Botnet size standard deviation and utilization of Basic Reproduction Number are then depicted and examined next. We conclude this section by examining the developed analytical link between the SIC model and the P2P botnet mitigation strategies through an analysis of a random sybil attack on a P2P botnet. Throughout this section, we plot the previously-derived analytical results by assigning values to various parameters ($\lambda_1, \lambda_2, \lambda_{r1}, \lambda_{r2}$, and λ_a), all with the unit of nodes/time unit (time unit can be hour, day, week, or any



(a) Reported weekly botnet size evolution

(b) Botnet size evolution reconstructed using the SIC Model

Fig. 4. FourLakeRiders (a Zeus-based botnet): botnet mitigation strategies analyzed using the SIC model. To produce (b) and Fig. 5, parameter values have been chosen as follows: During the whole 5-week period, λ_1 and λ_2 are constant and set as follows: $\lambda_1 = 0.042$ and $\lambda_2 = 0.001$. λ_{r1}, λ_{r2}, and λ_a are chosen as follows for each week: Part 1 (week 36): $\lambda_{r1} = 0.0082, \lambda_{r2} = 0.0046$, and $\lambda_a = 0$; Part 2 (week 37): $\lambda_{r1} = 0.0082, \lambda_{r2} = 0.0027$, and $\lambda_a = 0$; Part 3 (week 38): $\lambda_{r1} = 0, \lambda_{r2} = 0$, and $\lambda_a = 0$; Part 4 (weeks 39–40): $\lambda_{r1} = 0.0082, \lambda_{r2} = 0.0046$, and $\lambda_a = 0.0057$. All λ parameters are nodes/h.

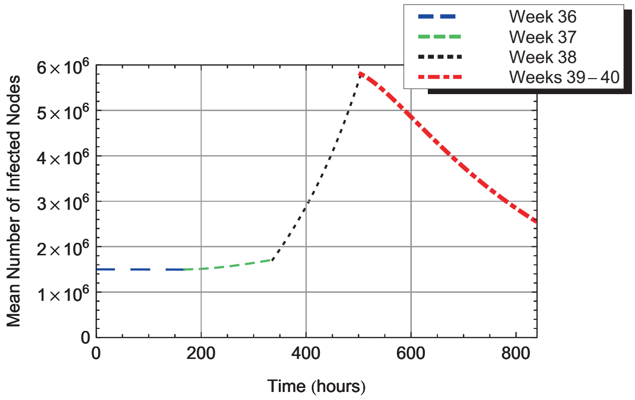


Fig. 5. FourLakeRiders botnet: size evolution of the number of Infected nodes estimated using the SIC Model. Parameter values are mentioned in the caption of Fig. 4.

other period). The plotted results are therefore general as parameter values may be assumed to be nodes per any time unit and then the plotted time-dependent performance measures will be interpreted as functions of that time unit.

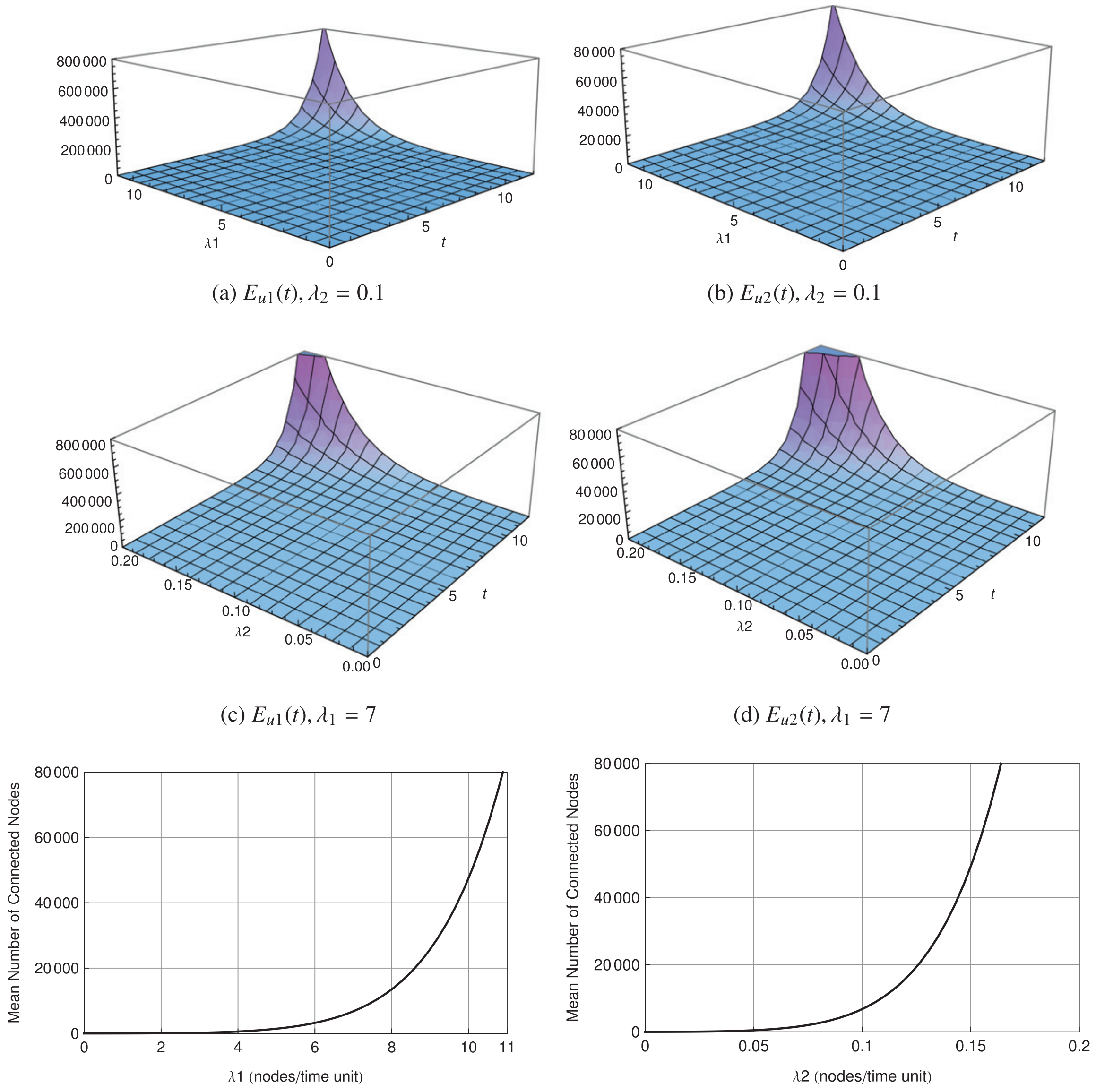
6.1. Model's parameter estimation techniques

Using the SIC model, the botnet size estimation problem has been reduced from having to estimate the global size of the botnet to the estimation of the model's parameters (λ_1 and λ_2) which requires only local knowledge. On the other hand, values for $\lambda_{r1}, \lambda_{r2}$, and λ_a depend on the type of disinfection and attack on the botnet; as the mitigation strategies are being conducted by the security experts, they will be able to reliably choose values for these latter parameters.

As a starting point, we would suggest a consideration of the following methods when trying to estimate values for λ_1 and λ_2: (1) real botnet size measurements, if available, can be used to estimate the parameter values (as done in Section 5); (2) local measurements through Honeynet log analysis [32], for example; and (3) a statistical approach to botnet virulence estimation (vulnerability and infection rates estimation) [33].

6.2. Initial unhindered botnet expansion

We first examine the unhindered botnet expansion that happens when the botnet first appears. In Fig. 6, we consider a 12-time-unit period during which the botnet expands. In this initial phase, there is neither any attack on the botnet, nor any removal (disinfection) from *Infected/Connected* stages; hence we set $\lambda_{r1} = \lambda_{r2} = \lambda_a = 0$. We choose $\lambda_1 = 7$ and $\lambda_2 = 0.1$ as the center values for these parameters; these values are based on the values derived from the analysis of GreenAlienRiders botnet (cf. Fig. 3). We then examine how the mean values of the number of nodes in *Infected* stage and *Connected* stage (botnet size) would change over this initial expansion period by varying



(a) $E_{u1}(t), \lambda_2 = 0.1$

(b) $E_{u2}(t), \lambda_2 = 0.1$

(c) $E_{u1}(t), \lambda_1 = 7$

(d) $E_{u2}(t), \lambda_1 = 7$

(e) $E_{u2}(t = 12), \lambda_2 = 0.1$

(f) $E_{u2}(t = 12), \lambda_1 = 7$

Fig. 6. Initial unhindered botnet expansion. Mean number of nodes in *Infected* stage ($E_{u1}(t)$) and *Connected* stage ($E_{u2}(t)$). Initial state values: $E_{u1}(0) = \overline{k_1} = 0, E_{u2}(0) = \overline{k_2} = 1$. The subscript **u** refers to the **U**nhindered expansion.

the parameter values in the following ranges: $0 \leqslant \lambda_1 \leqslant 11$ and $0 \leqslant \lambda_2 \leqslant 0.2$. In Fig. 6a and b, we set $\lambda_2 = 0.1$ and examine the change of mean values over time by varying λ_1 over $[0, 11]$. In Fig. 6c and d, on the other hand, we set $\lambda_1 = 7$ and examine the change of mean values over time by varying λ_2 over $[0, 0.2]$. Slicing Fig. 6b and d at $t = 12$, Fig. 6e and f closely show how mean numbers would change over the respective ranges of values for λ_1 and λ_2. Finally, Fig. 7 shows the means along with the standard deviations.

6.3. Comparison of mitigation strategies

One of the main advantages of the SIC model is that it enables security experts to compare and analyze mitigation strategies *before* deployment. In this sub-section, we study the case where botnet faces attack and/or removal (disinfection) and observe how severe these interventions must be in order to contain or dismantle the botnet. In all scenarios, we assign $\lambda_1 = 7$ and $\lambda_2 = 0.1$; their choice has no bearing on the following analysis regarding $\lambda_{r1}, \lambda_{r2}$, and λ_a. Further, we assume the mean number of *Infected* nodes and *Connected* nodes to be as follows: $E_1(0) = \overline{k_1} = 53,484$ and $E_2(0) = \overline{k_2} = 6786$; these values are derived from Fig. 7 at $t = 12$ when $\lambda_1 = 7$ and $\lambda_2 = 0.1$. We can then proceed to analyze how this particular botnet would react to different mitigation strategies.

In Fig. 8, the solid line depicts the scenario where there are no mitigation strategies and the number of Infected nodes and the botnet size continue to increase. Dotted/

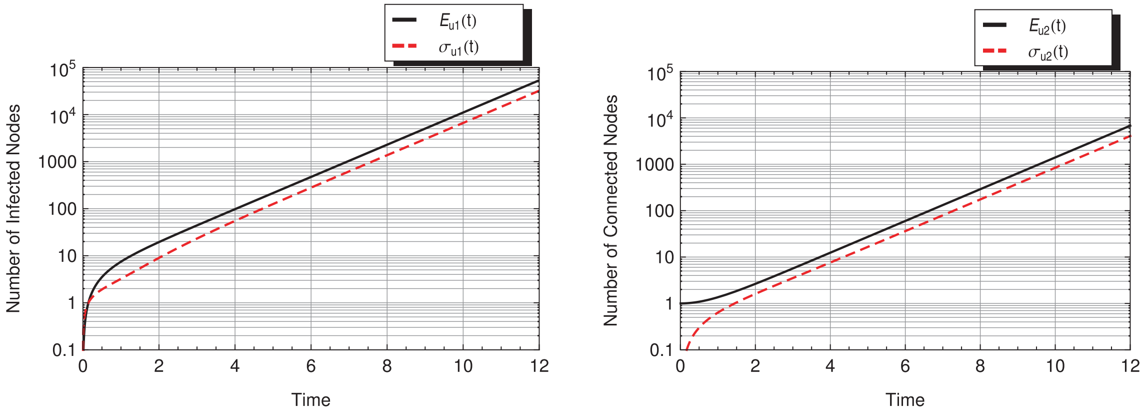
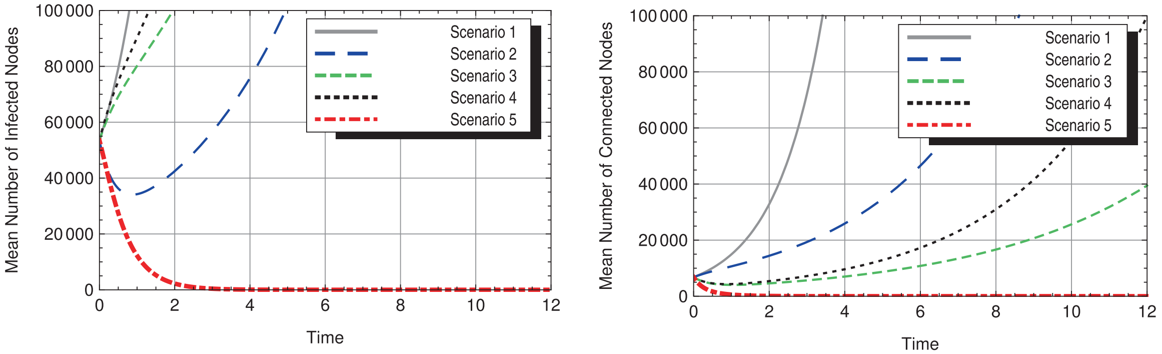


Fig. 7. Initial unhindered botnet expansion. Mean and standard deviation of the number of nodes in *Infected* and *Connected* stages. Initial state values: $E_{u1}(0) = \overline{k_1} = 0, E_{u2}(0) = \overline{k_2} = 1$. Parameter values: $\lambda_1 = 7, \lambda_2 = 0.1$ (both nodes/time unit).



(a) $E_1(t)$

(b) $E_2(t)$

Fig. 8. Comparison of mitigation strategies. Mean number of nodes in *Infected* stage ($E_1(t)$) and *Connected* stage ($E_2(t)$). Initial state values: $E_1(0) = \overline{k_1} = 53,484, E_2(0) = \overline{k_2} = 6786$; Parameter values: $\lambda_1 = 7, \lambda_2 = 0.1$; Scenario 1: unhindered expansion ($\lambda_{r1} = 0, \lambda_{r2} = 0, \lambda_a = 0$); Scenario 2: only removal of Infected nodes ($\lambda_{r1} = 2, \lambda_{r2} = 0, \lambda_a = 0$); Scenario 3: only removal of Connected nodes ($\lambda_{r1} = 0, \lambda_{r2} = 2, \lambda_a = 0$); Scenario 4: only attack on botnet ($\lambda_{r1} = 0, \lambda_{r2} = 0, \lambda_a = 2$); Scenario 5: three strategies simultaneously ($\lambda_{r1} = 2, \lambda_{r2} = 2, \lambda_a = 2$). All λ parameters are nodes/time unit.

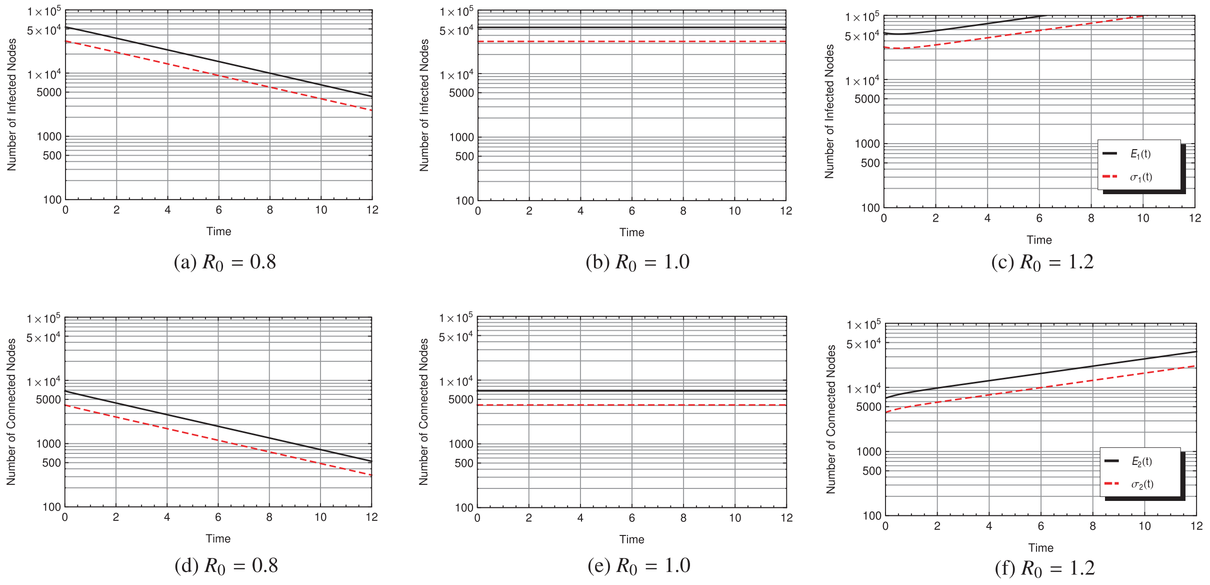
dashed lines denote scenarios under which different values chosen for $\lambda_{r1}, \lambda_{r2}$, and λ_a result in different trajectories for the mean. In Fig. 8a, we observe that the mean eventually goes to zero in only one scenario, i.e., when all three strategies are employed at the same time. Note that a large enough value chosen for λ_{r1} would make the mean number of nodes in *Infected* stage go to zero as well. Fig. 8b depicts the same scenarios as in Fig. 8a, but this time, the mean is for the nodes in *Connected* stage (botnet size). In this particular case, we observe that the mean number of nodes in *Connected* stage also eventually goes to zero in only one scenario, i.e., when all three strategies are employed at the same time.

We can therefore state that, all things being equal, removal/disinfection from *Connected* stage (λ_{r2}) has the most effect on containing the size of the botnet (nodes in *Connected* stage). Further, we intuitively deduce that it would be less costly to combat a botnet if we implement all three strategies at the same time, as we can choose moderate disinfection/attack rates. Concentrating on a single strategy (disinfection or attack) would mean that we need to choose a very high rate to achieve a comparable effect. Having to choose a high rate is usually associated with high cost in the real world (e.g., the plan of malware removal from near 100% of computers is either infeasible or extremely costly to implement).

6.4. Standard deviation and basic reproduction number

In Fig. 9, we draw the mean along with the standard deviation in each sub-figure. Inclusion of standard deviation helps put the mean in its proper context; the higher the standard deviation gets, the less should be the importance of the precise value of the mean in our interpretations. Since we consider that all mitigation strategies are being implemented, the sub-figures of Fig. 9 would be comparable to Fig. 8, as the chosen initial state values (values for $\overline{k_1}$ and $\overline{k_2}$) are the same.

Furthermore, in Fig. 9, we use the derived formula for Basic Reproduction Number (R_0) to choose values for different parameters in a way that leads to the size of the botnet shrinking (left sub-figs.), remaining constant (center



(a) $R_0 = 0.8$ (b) $R_0 = 1.0$ (c) $R_0 = 1.2$

(d) $R_0 = 0.8$ (e) $R_0 = 1.0$ (f) $R_0 = 1.2$

Fig. 9. Number of nodes in *Infected*stage (upper row figs.) and *Connected* stage (lower row figs.). Parameter values are as follows: $\overline{k_1} = 53,484$, $\overline{k_2} = 6786$, $\lambda_1 = 7, \lambda_2 = 0.1$, and $\lambda_a = 0.2$ for all sub-figures; for left sub-figures: $R_0 = 0.8, \lambda_{r1} = 1$,and $\lambda_{r2}(\text{determined}) = 0.8227$; for center sub-figures: $R_0 = 1, \lambda_{r1} = 0.8135$,and $\lambda_{r2}(\text{determined}) = 0.5880$; and for right sub-figures: $R_0 = 1.2, \lambda_{r1} = 1$, and $\lambda_{r2}(\text{determined}) = 0.2545$. All λ parameters are nodes/time unit.

sub-figs.) or growing (right sub-figs.). To achieve this, we choose sample values for various parameters (except for λ_{r2}) and for R_0; therefore, the value of λ_{r2} would be determined in order to satisfy (13).

6.5. Random sybil attack on DHT-based P2P botnets

Finally, we provide a numerical analysis of the developed relationship between the SIC model's attack rate (λ_a) and the number of sybils inserted in the P2P botnet. The analysis will be the case of adding the sybils at $t = 0$ in Fig. 8, assuming an instantaneous effect on the P2P botnet, and examining the situation in the next Δt. The numerical result is derived from (16) and depicted in Fig. 10.[2] The figure demonstrates the relationship between the percentage increase in the number of inserted sybils and the resulting percentage increase in the value of λ_a. The demonstrated relationship leads to the following insight: once the sybil attack is underway, the value of n_s is known and the resulting λ_a can be measured. The security expert can then determine, for example, how many sybils should be added in order to arrive at a desired λ_a to have the intended mitigation effect.

7. Concluding remarks and future work

There is a lack of appropriate analytical models on botnets in the literature. The prior work on botnets mostly consists of either deterministic analytical or simulation-based models. The deterministic models have the drawback of treating the botnet size as a deterministic variable, which neglects the stochastic nature of the evolution of botnets. These models only lead to determination of the mean botnet population size and not to the probability distribution of size or its higher moments. Further, the existing models determine the mean botnet size numerically and they have not obtained closed-form expressions. On the other hand, simulation-based models can be designed to capture the details of botnet lifecycle, but their results cannot be easily replicated or used by others. Finally, we have shown that our results may be used to model the size evolution of botnets, including their sudden growth, in the real world.

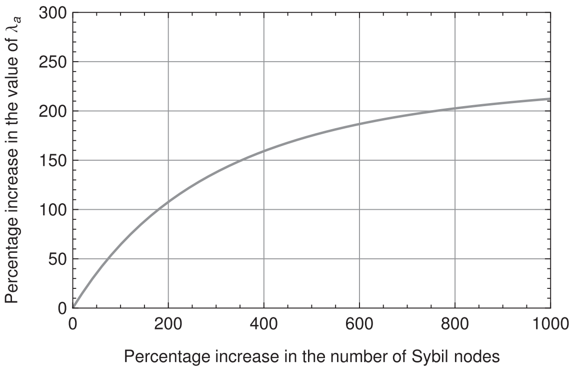


Fig. 10. Relationship between the attack rate (λ_a) and the number of sybils (Initial $n_s = 1000, n = \overline{k_2} = 6786$, and $b = 5$).

In this paper, we have developed a stochastic analytical model that captures the dynamics of a botnet's lifecycle. We have modeled the lifecycle of a node in the system with three stages referred to as, *Susceptible*, *Infected*, and *Connected*. Further, we have assumed that the nodes in the Infected and Connected stages may go back to Susceptible

[2] As the size of botnet changes with time, it is necessary to update the respective calculated values at regular intervals to keep a close approximation.

stage. We have modeled the system using a two-dimensional Markov process and derived a partial differential equation for the joint distribution of the number of nodes in each stage. Though this equation could not be solved, we were able to obtain closed-form expressions for the time dependent mean and variance of the population size in each stage. It is possible to obtain even higher moments of the botnet population size, but the results get too complicated.

To our knowledge, the demonstrated relationship between a lifecycle/population model and the P2P botnet mitigation strategies is the first of its kind presented in the open literature. The developed relationship leads to a two-step, or recursive, analysis process: (1) examining the effect of the chosen λ_a on the botnet size based on Eq. 4 for the means; and (2) examining the relationship between a change of λ_a and the associated change in the number of sybils based on Eq. 16. We are currently working on the integration of these two steps which entails changes to the SIC model itself and leads to an analytical model specific to DHT-based P2P botnets.

Appendix A. Deriving a PDE from the differential-difference equations

We can write (1.a) as follows:

$$\begin{aligned}\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\frac{dP_{n_1,n_2}(t)}{dt}z_1^{n_1}z_2^{n_2} &= \sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\lambda_1 n_2 P_{n_1-1,n_2}(t)z_1^{n_1}z_2^{n_2}\\ &+\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\lambda_{r1}(n_1\\ &+1)P_{n_1+1,n_2}(t)z_1^{n_1}z_2^{n_2}\\ &+\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\lambda_{r2}(n_2\\ &+1)P_{n_1,n_2+1}(t)z_1^{n_1}z_2^{n_2}\\ &+\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\lambda_2(n_1\\ &+1)P_{n_1+1,n_2-1}(t)z_1^{n_1}z_2^{n_2}\\ &+\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\lambda_a(n_2\\ &+1)P_{n_1-1,n_2+1}(t)z_1^{n_1}z_2^{n_2}\\ &-\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}(\lambda_1 n_2+\lambda_{r1}n_1\\ &+\lambda_{r2}n_2+\lambda_2 n_1\\ &+\lambda_a n_2)P_{n_1,n_2}(t)z_1^{n_1}z_2^{n_2} \qquad (A.1)\end{aligned}$$

and write (1.b) as follows:

$$\begin{aligned}\sum_{n_2=1}^{\infty}\frac{dP_{0,n_2}(t)}{dt}z_2^{n_2} &= \sum_{n_2=1}^{\infty}\lambda_{r1}P_{1,n_2}(t)z_2^{n_2}+\sum_{n_2=1}^{\infty}\lambda_{r2}(n_2\\ &+1)P_{0,n_2+1}(t)z_2^{n_2}\\ &+\sum_{n_2=1}^{\infty}\lambda_2 P_{1,n_2-1}(t)z_2^{n_2}-\sum_{n_2=1}^{\infty}(\lambda_1 n_2\\ &+\lambda_{r2}n_2+\lambda_a n_2)P_{0,n_2}(t)z_2^{n_2} \qquad (A.2)\end{aligned}$$

Finally, we write (1.c) as follows:

$$\begin{aligned}\sum_{n_1=1}^{\infty}\frac{dP_{n_1,0}(t)}{dt}z_1^{n_1} &= \sum_{n_1=1}^{\infty}\lambda_{r1}(n_1+1)P_{n_1+1,0}(t)z_1^{n_1}\\ &+\sum_{n_1=1}^{\infty}\lambda_{r2}P_{n_1,1}(t)z_1^{n_1}\\ &+\sum_{n_1=1}^{\infty}\lambda_a P_{n_1-1,1}(t)z_1^{n_1}-\sum_{n_1=1}^{\infty}(\lambda_{r1}n_1\\ &+\lambda_2 n_1)P_{n_1,0}(t)z_1^{n_1} \qquad (A.3)\end{aligned}$$

We now add together (A.1)–(A.3), and (1.d). Here is the result:

$$\begin{aligned}\frac{\partial P(z_1,z_2,t)}{\partial t} &= \sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\lambda_1 n_2 P_{n_1-1,n_2}(t)z_1^{n_1}z_2^{n_2} \qquad (A.4)\\ &+\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\lambda_{r1}(n_1+1)P_{n_1+1,n_2}(t)z_1^{n_1}z_2^{n_2} \qquad (A.5)\\ &+\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\lambda_{r2}(n_2+1)P_{n_1,n_2+1}(t)z_1^{n_1}z_2^{n_2} \qquad (A.6)\\ &+\sum_{n_1=0}^{\infty}\sum_{n_2=1}^{\infty}\lambda_2(n_1+1)P_{n_1+1,n_2-1}(t)z_1^{n_1}z_2^{n_2} \qquad (A.7)\\ &+\sum_{n_1=1}^{\infty}\sum_{n_2=0}^{\infty}\lambda_a(n_2+1)P_{n_1-1,n_2+1}(t)z_1^{n_1}z_2^{n_2} \qquad (A.8)\\ &-\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}(\lambda_1 n_2+\lambda_{r1}n_1+\lambda_{r2}n_2+\lambda_2 n_1\\ &+\lambda_a n_2)P_{n_1,n_2}(t)z_1^{n_1}z_2^{n_2} \qquad (A.9)\end{aligned}$$

We write (A.4) as follows:

$$\begin{aligned}&\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}\lambda_1 n_2 P_{n_1-1,n_2}(t)z_1^{n_1}z_2^{n_2}\\ &=\lambda_1 z_1\sum_{n_1=1}^{\infty}\sum_{n_2=1}^{\infty}n_2 P_{n_1-1,n_2}(t)z_1^{n_1-1}z_2^{n_2}\\ &=\lambda_1 z_1 z_2\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}n_2 P_{n_1,n_2}(t)z_1^{n_1}z_2^{n_2-1}\\ &=\lambda_1 z_1 z_2\frac{\partial P(z_1,z_2,t)}{\partial z_2} \qquad (A.10)\end{aligned}$$

and (A.5) as follows:

$$\begin{aligned}&\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\lambda_{r1}(n_1+1)P_{n_1+1,n_2}(t)z_1^{n_1}z_2^{n_2}\\ &=\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\lambda_{r1}\frac{n_1+1}{z_1}P_{n_1+1,n_2}(t)z_1^{n_1+1}z_2^{n_2}\\ &=\sum_{n_1=1}^{\infty}\sum_{n_2=0}^{\infty}\lambda_{r1}n_1 P_{n_1,n_2}(t)z_1^{n_1-1}z_2^{n_2}=\lambda_{r1}\frac{\partial P(z_1,z_2,t)}{\partial z_1} \qquad (A.11)\end{aligned}$$

and (A.6) as follows:

$$\begin{aligned}&\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\lambda_{r2}(n_2+1)P_{n_1,n_2+1}(t)z_1^{n_1}z_2^{n_2}\\ &=\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\lambda_{r2}\frac{n_2+1}{z_2}P_{n_1,n_2+1}(t)z_1^{n_1}z_2^{n_2+1}\\ &=\sum_{n_1=0}^{\infty}\sum_{n_2=1}^{\infty}\lambda_{r2}n_2 P_{n_1,n_2}(t)z_1^{n_1}z_2^{n_2-1}=\lambda_{r2}\frac{\partial P(z_1,z_2,t)}{\partial z_2} \qquad (A.12)\end{aligned}$$

and (A.7) as follows:

$$\begin{aligned}
&\sum_{n_1=0}^{\infty}\sum_{n_2=1}^{\infty}\lambda_2(n_1+1)P_{n_1+1,n_2-1}(t)z_1^{n_1}z_2^{n_2}\\
&=\sum_{n_1=0}^{\infty}\sum_{n_2=1}^{\infty}\lambda_2(n_1+1)\frac{z_2}{z_1}P_{n_1+1,n_2-1}(t)z_1^{n_1+1}z_2^{n_2-1}\\
&=\sum_{n_1=1}^{\infty}\sum_{n_2=0}^{\infty}\lambda_2 n_1\frac{z_2}{z_1}P_{n_1,n_2}(t)z_1^{n_1}z_2^{n_2}\\
&=\lambda_2 z_2\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}n_1P_{n_1,n_2}(t)z_1^{n_1-1}z_2^{n_2}\\
&=\lambda_2 z_2\frac{\partial P(z_1,z_2,t)}{\partial z_1}
\end{aligned}\tag{A.13}$$

and (A.8) as follows:

$$\begin{aligned}
&\sum_{n_1=1}^{\infty}\sum_{n_2=0}^{\infty}\lambda_a(n_2+1)P_{n_1-1,n_2+1}(t)z_1^{n_1}z_2^{n_2}\\
&=\sum_{n_1=1}^{\infty}\sum_{n_2=0}^{\infty}\lambda_a(n_2+1)\frac{z_1}{z_2}P_{n_1-1,n_2+1}(t)z_1^{n_1-1}z_2^{n_2+1}\\
&=\sum_{n_1=0}^{\infty}\sum_{n_2=1}^{\infty}\lambda_a n_2\frac{z_1}{z_2}P_{n_1,n_2}(t)z_1^{n_1}z_2^{n_2}\\
&=\lambda_a z_1\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}n_2P_{n_1,n_2}(t)z_1^{n_1}z_2^{n_2-1}\\
&=\lambda_a z_1\frac{\partial P(z_1,z_2,t)}{\partial z_2}
\end{aligned}\tag{A.14}$$

Finally, (A.9) as follows:

$$\begin{aligned}
&\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}(\lambda_1 n_2+\lambda_{r1}n_1+\lambda_{r2}n_2+\lambda_2 n_1\\
&\quad+\lambda_a n_2)P_{n_1,n_2}(t)z_1^{n_1}z_2^{n_2}\\
&=\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}(\lambda_{r1}+\lambda_2)n_1P_{n_1,n_2}(t)z_1^{n_1}z_2^{n_2}+\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}(\lambda_1\\
&\quad+\lambda_{r2}+\lambda_a)n_2P_{n_1,n_2}(t)z_1^{n_1}z_2^{n_2}\\
&=(\lambda_{r1}+\lambda_2)z_1\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}n_1P_{n_1,n_2}(t)z_1^{n_1-1}z_2^{n_2}+(\lambda_1\\
&\quad+\lambda_{r2}+\lambda_a)z_2\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}n_2P_{n_1,n_2}(t)z_1^{n_1}z_2^{n_2-1}\\
&=(\lambda_{r1}+\lambda_2)z_1\frac{\partial P(z_1,z_2,t)}{\partial z_1}+(\lambda_1+\lambda_{r2}+\lambda_a)z_2\\
&\quad\times\frac{\partial P(z_1,z_2,t)}{\partial z_2}
\end{aligned}\tag{A.15}$$

Replacing (A.4)–(A.9) with the ones derived in (A.10) through (A.15), after simplification, we arrive at (2).

Appendix B. Attempt to solve the PDE using method of characteristics

We describe our efforts to solve the partial differential Eq. (2) describing the system. Following the Method of Characteristics [34, p. 432] to solve PDEs, based on (2), we can write:

$$\begin{cases}
\frac{\partial t}{\partial s}=-1 & (a)\\
\frac{dP}{ds}=0 & (b)\\
\frac{\partial z_1}{\partial s}=\lambda_{r1}+\lambda_2 z_2-\lambda_{r1}z_1-\lambda_2 z_1 & (c)\\
\frac{\partial z_2}{\partial s}=\lambda_1 z_1 z_2+\lambda_{r2}+\lambda_a z_1-\lambda_1 z_2-\lambda_{r2}z_2-\lambda_a z_2 & (d)
\end{cases}\tag{B.1}$$

where s is a parametric variable and $P=P(z_1,z_2,t)$ is the PGF. With the initial condition $P(z_1,z_2,0)=z_1^{k_1}z_2^{k_2}$, we therefore have:

$$\begin{cases}
t(s=0)=0 & (a)\\
z_1(s=0)=i_1 & (b)\\
z_2(s=0)=i_2 & (c)\\
P(s=0)=i_1^{k_1}i_2^{k_2} & (d)
\end{cases}\tag{B.2}$$

From (B.1.a) and (B.2.a), we have:

$$t=-s\tag{B.3}$$

Likewise, from (B.1.b) and (B.2.d), we have:

$$P=(i_1(z_1,z_2,t))^{k_1}(i_2(z_1,z_2,t))^{k_2}\tag{B.4}$$

Eqs. (B.1.c) and (B.1.d) are "non-separable", i.e., we cannot derive z_1 and z_2 from 1st order ordinary differential equations (ODEs). We therefore proceed as follows: from (B.1.c), we derive z_2:

$$z_2=\frac{1}{\lambda_2}\left[\frac{dz_1}{ds}+(\lambda_{r1}+\lambda_2)z_1-\lambda_{r1}\right]\tag{B.5}$$

Replacing z_2 in (B.1.d) with the expression given in (B.5), after some simplifications, we can write (B.1.d) as follows:

$$\begin{aligned}
&\frac{d^2z_1}{ds^2}+(\lambda_{r1}+\lambda_2+\lambda_1+\lambda_{r2}+\lambda_a)\frac{dz_1}{ds}-\lambda_1 z_1\frac{dz_1}{ds}\\
&\quad-\lambda_1(\lambda_{r1}+\lambda_2)z_1^2\\
&\quad+((\lambda_1+\lambda_{r2}+\lambda_a)(\lambda_{r1}+\lambda_2)+\lambda_1\lambda_{r1}-\lambda_a\lambda_2)z_1\\
&\quad-(\lambda_1+\lambda_{r2}+\lambda_a)\lambda_{r1}-\lambda_{r2}\lambda_2=0
\end{aligned}\tag{B.6}$$

Eq. (B.6) has the form of a second order Lienard equation [35] given below:

$$\frac{d^2z_1}{ds^2}+(A+Bz_1)\frac{dz_1}{ds}+Cz_1^2+Dz_1+E=0\tag{B.7}$$

Eq. (B.7) is not in the form of solvable cases presented in [35, Section 2.2.3–2], [36, pp. 204–5], and [37]. As a further attempt to solve (B.7), we have used the following substitution suggested in [35, Section 2.2.3–1]:

$$w=\frac{dz_1}{ds},\quad \frac{d^2z_1}{ds^2}=w'_s=w'_{z_1}\frac{dz_1}{ds}=w'_{z_1}w\tag{B.8}$$

The above substitution transformed (B.7) into an Abel equation of the 2nd kind given below:

$$ww'_{z_1}+(A+Bz_1)w+Cz_1^2+Dz_1+E=0\tag{B.9}$$

Eq. (B.9) is also not among the solvable cases presented in [35, Section 1.3.3–2].

Appendix C. Derivation of means from the PDE of the PGF

We take the derivative of (2) with respect to z_1 as follows:

$$\begin{aligned}
&(-\lambda_{r1}-\lambda_2)\frac{\partial P(z_1,z_2,t)}{\partial z_1}+(\lambda_{r1}+\lambda_2 z_2-\lambda_{r1}z_1-\lambda_2 z_1)\\
&\quad\times\frac{\partial^2 P(z_1,z_2,t)}{\partial z_1^2}+(\lambda_1 z_2+\lambda_a)\frac{\partial P(z_1,z_2,t)}{\partial z_2}\\
&\quad+(\lambda_1 z_1 z_2+\lambda_{r2}+\lambda_a z_1-\lambda_1 z_2-\lambda_{r2}z_2-\lambda_a z_2)\frac{\partial^2 P(z_1,z_2,t)}{\partial z_2\partial z_1}\\
&\quad-\frac{\partial^2 P(z_1,z_2,t)}{\partial t\partial z_1}=0
\end{aligned}\tag{C.1}$$

Setting $z_1=z_2=1$ in (C.1) gives us the following equation:

$$\frac{dE_1(t)}{dt}+(\lambda_2+\lambda_{r1})E_1(t)-(\lambda_1+\lambda_a)E_2(t)=0\tag{C.2}$$

We then take the derivative of (2) with respect to z_2 as follows:

$$\begin{aligned}
&\lambda_2\frac{\partial P(z_1,z_2,t)}{\partial z_1}+(\lambda_{r1}+\lambda_2 z_2-\lambda_{r1}z_1-\lambda_2 z_1)\\
&\quad\times\frac{\partial^2 P(z_1,z_2,t)}{\partial z_2\partial z_1}+(\lambda_1 z_1-\lambda_1-\lambda_{r2}-\lambda_a)\frac{\partial P(z_1,z_2,t)}{\partial z_2}\\
&\quad+(\lambda_1 z_1 z_2+\lambda_{r2}+\lambda_a z_1-\lambda_1 z_2-\lambda_{r2}z_2-\lambda_a z_2)\\
&\quad\times\frac{\partial^2 P(z_1,z_2,t)}{\partial z_2^2}-\frac{\partial^2 P(z_1,z_2,t)}{\partial t\partial z_2}\\
&\quad=0
\end{aligned}\tag{C.3}$$

Setting $z_1=z_2=1$ in (C.3) gives us the following equation:

$$\frac{dE_2(t)}{dt}-\lambda_2 E_1(t)+(\lambda_{r2}+\lambda_a)E_2(t)=0\tag{C.4}$$

Re-arranging (C.2) and (C.4) gives us (4).

Taking (4) to Laplace domain, we can write:

$$\begin{cases}sE_1(s|k_1,k_2)-k_1+(\lambda_{r1}+\lambda_2)E_1(s|k_1,k_2)-(\lambda_1+\lambda_a)E_2(s|k_1,k_2)=0\\ sE_2(s|k_1,k_2)-k_2-\lambda_2 E_1(s|k_1,k_2)+(\lambda_{r2}+\lambda_a)E_2(s|k_1,k_2)=0\end{cases}\tag{C.5}$$

where k_1 and k_2 are values of n_1 and n_2 at $t=0$, respectively. Note that k_1 and k_2 are variables themselves and their means are obtained as follows:

$$\begin{cases}\overline{k_1}=\sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}k_1 P_{k_1,k_2}(t=0)\\ \overline{k_2}=\sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}k_2 P_{k_1,k_2}(t=0)\end{cases}\tag{C.6}$$

$\overline{k_1}$ and $\overline{k_2}$ are therefore the values of the means at $t=0$. We then proceed to uncondition (C.5), i.e., we take $\sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\{X\}P_{k_1,k_2}(t=0)$, with X being each element of the equation set. After simplification, we have:

$$\begin{cases}sE_1(s)-\overline{k_1}+(\lambda_{r1}+\lambda_2)E_1(s)-(\lambda_1+\lambda_a)E_2(s)=0\\ sE_2(s)-\overline{k_2}-\lambda_2 E_1(s)+(\lambda_{r2}+\lambda_a)E_2(s)=0\end{cases}\tag{C.7}$$

$E_1(s)$ and $E_2(s)$ are then obtained as follows:

$$E_1(s)=\frac{\overline{k_1}s+\overline{k_2}(\lambda_1+\lambda_a)+\overline{k_1}(\lambda_{r2}+\lambda_a)}{s^2+(\lambda_{r2}+\lambda_a+\lambda_{r1}+\lambda_2)s+(\lambda_{r2}+\lambda_a)(\lambda_{r1}+\lambda_2)-\lambda_2(\lambda_1+\lambda_a)}\tag{C.8}$$

$$E_2(s)=\frac{s+\lambda_2+\lambda_{r1}}{\lambda_1+\lambda_a}E_1(s)-\frac{\overline{k_1}}{\lambda_1+\lambda_a}\tag{C.9}$$

Finally, the inverse Laplace of $E_1(s)$ and $E_2(s)$ are obtained as shown in (5) and (6).

Appendix D. Derivation of variances from the PDE of the PGF

Taking the derivative of (C.1) with respect to z_1 (i.e., taking the 2nd derivative of (2) with respect to z_1), we have:

$$\begin{aligned}
&(-\lambda_{r1}-\lambda_2)\frac{\partial^2 P(z_1,z_2,t)}{\partial z_1^2}+(-\lambda_{r1}-\lambda_2)\frac{\partial^2 P(z_1,z_2,t)}{\partial z_1^2}\\
&\quad+(\lambda_{r1}+\lambda_2 z_2-\lambda_{r1}z_1-\lambda_2 z_1)\frac{\partial^3 P(z_1,z_2,t)}{\partial z_1^3}\\
&\quad+(\lambda_1 z_2+\lambda_a)\frac{\partial^2 P(z_1,z_2,t)}{\partial z_2\partial z_1}+(\lambda_1 z_2+\lambda_a)\frac{\partial^2 P(z_1,z_2,t)}{\partial z_2\partial z_1}\\
&\quad+(\lambda_1 z_1 z_2+\lambda_{r2}+\lambda_a z_1-\lambda_1 z_2-\lambda_{r2}z_2-\lambda_a z_2)\frac{\partial^3 P(z_1,z_2,t)}{\partial z_2\partial z_1^2}\\
&\quad-\frac{\partial^3 P(z_1,z_2,t)}{\partial t\partial z_1^2}=0
\end{aligned}\tag{D.1}$$

Likewise, taking the derivative of (C.3) with respect to z_2 (i.e., taking the 2nd derivative of (2) with respect to z_2), we have:

$$\begin{aligned}
&\lambda_2\frac{\partial^2 P(z_1,z_2,t)}{\partial z_1\partial z_2}+\lambda_2\frac{\partial^2 P(z_1,z_2,t)}{\partial z_1\partial z_2}+(\lambda_{r1}+\lambda_2 z_2-\lambda_{r1}z_1-\lambda_2 z_1)\\
&\quad\times\frac{\partial^3 P(z_1,z_2,t)}{\partial z_2^2\partial z_1}+(\lambda_1 z_1-\lambda_1-\lambda_{r2}-\lambda_a)\frac{\partial^2 P(z_1,z_2,t)}{\partial z_2^2}\\
&\quad+(\lambda_1 z_1-\lambda_1-\lambda_{r2}-\lambda_a)\frac{\partial^2 P(z_1,z_2,t)}{\partial z_2^2}\\
&\quad+(\lambda_1 z_1 z_2+\lambda_{r2}+\lambda_a z_1-\lambda_1 z_2-\lambda_{r2}z_2-\lambda_a z_2)\frac{\partial^3 P(z_1,z_2,t)}{\partial z_2^3}\\
&\quad-\frac{\partial^3 P(z_1,z_2,t)}{\partial t\partial z_2^2}=0
\end{aligned}\tag{D.2}$$

Finally, taking the derivative of (C.1) with respect to z_2, we have:

$$\begin{aligned}
&(-\lambda_{r1}-\lambda_2)\frac{\partial^2 P(z_1,z_2,t)}{\partial z_1\partial z_2}+\lambda_2\frac{\partial^2 P(z_1,z_2,t)}{\partial z_1^2}\\
&\quad+(\lambda_{r1}+\lambda_2 z_2-\lambda_{r1}z_1-\lambda_2 z_1)\frac{\partial^3 P(z_1,z_2,t)}{\partial z_1^2\partial z_2}\\
&\quad+\lambda_1\frac{\partial P(z_1,z_2,t)}{\partial z_2}+(\lambda_1 z_2+\lambda_a)\frac{\partial^2 P(z_1,z_2,t)}{\partial z_2^2}\\
&\quad+(\lambda_1 z_1-\lambda_1-\lambda_{r2}-\lambda_a)\frac{\partial^2 P(z_1,z_2,t)}{\partial z_2\partial z_1}\\
&\quad+(\lambda_1 z_1 z_2+\lambda_{r2}+\lambda_a z_1-\lambda_1 z_2-\lambda_{r2}z_2-\lambda_a z_2)\frac{\partial^3 P(z_1,z_2,t)}{\partial z_2^2\partial z_1}\\
&\quad-\frac{\partial^3 P(z_1,z_2,t)}{\partial t\partial z_1\partial z_2}=0
\end{aligned}\tag{D.3}$$

Setting $z_1 = z_2 = 1$ in (D.1), (D.2), and (D.3) gives us (12).

In (12), we have three ODEs and three variables ($\psi_1(t), \psi_{12}(t)$, and $\psi_2(t)$); therefore, we can find a unique solution by solving this system of linear ODEs. Taking (12) to Laplace domain, we have:

$$\begin{cases} s\psi_1(s|k_1,k_2) - k_1^2 + k_1 = 2(\lambda_1+\lambda_a)\psi_{12}(s|k_1,k_2) \\ \quad -2(\lambda_{r1}+\lambda_2)\psi_1(s|k_1,k_2) \\ s\psi_2(s|k_1,k_2) - k_2^2 + k_2 = 2\lambda_2\psi_{12}(s|k_1,k_2) \\ \quad -2(\lambda_{r2}+\lambda_a)\psi_2(s|k_1,k_2) \\ s\psi_{12}(s|k_1,k_2) - k_1k_2 = -(\lambda_{r1}+\lambda_2+\lambda_{r2}+\lambda_a)\psi_{12}(s|k_1,k_2) \\ \quad +\lambda_2\psi_1(s|k_1,k_2) + \lambda_1E_2(s|k_1,k_2) + (\lambda_1+\lambda_a)\psi_2(s|k_1,k_2) \end{cases} \tag{D.4}$$

Like before, we then proceed to uncondition (D.4). After simplification, we have:

$$\begin{cases} s\psi_1(s) - \overline{k_1^2} + \overline{k_1} = 2(\lambda_1+\lambda_a)\psi_{12}(s) - 2(\lambda_{r1}+\lambda_2)\psi_1(s) \\ s\psi_2(s) - \overline{k_2^2} + \overline{k_2} = 2\lambda_2\psi_{12}(s) - 2(\lambda_{r2}+\lambda_a)\psi_2(s) \\ s\psi_{12}(s) - \overline{k_1k_2} = -(\lambda_{r1}+\lambda_2+\lambda_{r2}+\lambda_a)\psi_{12}(s) + \lambda_2\psi_1(s) + \lambda_1E_2(s) + (\lambda_1+\lambda_a)\psi_2(s) \end{cases} \tag{D.5}$$

The solution of (D.5) (i.e., the expressions for $\psi_1(s)$ and $\psi_2(s)$) as well as the expressions for $\sigma_1^2(t)$ and $\sigma_2^2(t)$ are extremely lengthy; hence, they are provided in [38] instead due to space constraints.

Appendix E. Basic reproduction number calculation through the "next generation matrix" method

Based on the steps of the "Next Generation Matrix" method [27, pp. 160–5], we proceed as follows: From SIC model's differential equations for means (i.e., (4)), we extract the f and v matrices:

$$f = \begin{bmatrix} (\lambda_1+\lambda_a)E_2(t) \\ \lambda_2E_1(t) \end{bmatrix} \quad v = \begin{bmatrix} (\lambda_2+\lambda_{r1})E_1(t) \\ (\lambda_{r2}+\lambda_a)E_2(t) \end{bmatrix} \tag{E.1}$$

F and V matrices would be therefore as follows:

$$F = \begin{bmatrix} 0 & \lambda_1+\lambda_a \\ \lambda_2 & 0 \end{bmatrix} \quad V = \begin{bmatrix} \lambda_2+\lambda_{r1} & 0 \\ 0 & \lambda_{r2}+\lambda_a \end{bmatrix} \tag{E.2}$$

The next generation matrix (K) would be as follows:

$$\begin{aligned} K &= F \times V^{-1} \\ &= \begin{bmatrix} 0 & \lambda_1+\lambda_a \\ \lambda_2 & 0 \end{bmatrix} \times \frac{1}{(\lambda_2+\lambda_{r1})(\lambda_{r2}+\lambda_a)} \begin{bmatrix} \lambda_{r2}+\lambda_a & 0 \\ 0 & \lambda_2+\lambda_{r1} \end{bmatrix} \\ &= \frac{1}{(\lambda_2+\lambda_{r1})(\lambda_{r2}+\lambda_a)} \begin{bmatrix} 0 & (\lambda_1+\lambda_a)(\lambda_2+\lambda_{r1}) \\ \lambda_2(\lambda_{r2}+\lambda_a) & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & \frac{\lambda_1+\lambda_a}{\lambda_{r2}+\lambda_a} \\ \frac{\lambda_2}{\lambda_2+\lambda_{r1}} & 0 \end{bmatrix} \end{aligned} \tag{E.3}$$

To derive R_0, we proceed as follows:

$$\det(K - R_0 \times I) = 0 \tag{E.4}$$

where I is an identity matrix. We therefore have:

$$\det \begin{bmatrix} -R_0 & \frac{\lambda_1+\lambda_a}{\lambda_{r2}+\lambda_a} \\ \frac{\lambda_2}{\lambda_2+\lambda_{r1}} & -R_0 \end{bmatrix} = R_0^2 - \frac{\lambda_1+\lambda_a}{\lambda_{r2}+\lambda_a} \times \frac{\lambda_2}{\lambda_2+\lambda_{r1}} = 0$$

Basic Reproduction Number (R_0) is therefore derived as noted in (13).

References

[1] S. Mansfield-Devine, Battle of the botnets, Network Security (2010) 4–6.

[2] D. Bleaken, Botwars: the fight against criminal cyber networks, Computer Fraud & Security 201 (2010) 17–19.

[3] C.J. Mielke, H. Chen, Botnets, and the cybercriminal underground, in: Proceedings of the IEEE International Conference on Intelligence & Secured Informatics ISI, 2008, pp. 206–211.

[4] W.H. Murray, The application of epidemiology to computer viruses, Computers & Security 7 (1988) 139–145.

[5] J.O. Kephart, S.R. White, Directed-graph epidemiological models of computer viruses, in: Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy, IEEE Comput. Soc. Press, 1991, pp. 343–359.

[6] G. Serazzi, S. Zanero, Computer virus propagation models, in: M.C. Calzarossa, E. Gelenbe (Eds.), Performance Tools and Applications to Networked Systems, Lecture Notes in Computer Science, vol. 2965, Springer, Berlin Heidelberg, 2004, pp. 26–50.

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

[8] M. Ajelli, R.L. Cigno, A. Montresor, Modeling botnets and epidemic malware, in: Proceedings of the IEEE International Communication Conference (ICC), IEEE, 2010, pp. 1–5.

[9] E.V. Ruitenbeek, W.H. Sanders, Modeling peer-to-peer botnets, in: QEST '08: Proceedings of the 2008 Fifth International Conference on Quantitative Evaluation of Systems, IEEE Comp. Soc., 2008, pp. 307–316.

[10] X. Li, H. Duan, W. Liu, J. Wu, The growing model of botnets, in: Proceedings of the Int Green Circuits and Systems (ICGCS) Conference, IEEE, 2010, pp. 414–419.

[11] Q. Wang, Z. Chen, C. Chen, N. Pissinou, On the robustness of the botnet topology formed by worm infection, in: Proceedings of the IEEE Global Telecommunications Conference, GLOBECOM 2010, pp. 1–6.

[12] R. Weaver, A probabilistic population study of the conficker-c botnet, in: Passive and Active Measurement, Lecture Notes in Computer Science, vol. 6032, Springer, Berlin/Heidelberg, 2010, pp. 181–190.

[13] S.B. Banks, M.R. Stytz, Advancing botnet modeling techniques for military and security simulations, in: Proceedings of the SPIE International Society of Optical Engineering, SPIE – The International Society for Optical Engineering, vol. 8060, Orlando, FL, United States, 2011.

[14] Y. Wang, S. Wen, W. Zhou, W. Zhou, Y. Xiang, The probability model of peer-to-peer botnet propagation, in: Y. Xiang, A. Cuzzocrea, M. Hobbs, W. Zhou (Eds.), Algorithms and Architectures for Parallel Processing, Lecture Notes in Computer Science, vol. 7016, Springer, Berlin/Heidelberg, 2011, pp. 470–480.

[15] C.C. Zou, R. Cunningham, Honeypot-aware advanced botnet construction and maintenance, in: Proceedings of the International Conference on Dependable Systems and Networks DSN 2006, pp. 199–208.

[16] D. Dagon, C. Zou, W. Lee, Modeling botnet propagation using time zones, in: Proceedings of the 13th Network and Distributed System Security Symposium, NDSS, Internet Society, 2006.

[17] R. Li, L. Gan, Y. Jia, Propagation model for botnet based on conficker monitoring, in: Proceedings of the Second International Information Science and Engineering (ISISE) Symposium, IEEE, 2009, pp. 185–190.

[18] W. Xin-liang, C. Lu-Ying, L. Fang, L. Zhen-ming, Analysis and modeling of the botnet propagation characteristics, in: Proceedings of the 6th Interantional Wireless Communication Network & Mobile Computation (WiCOM) Conference, IEEE, 2010, pp. 1–4.

[19] P. Porras, H. Saidi, V. Yegneswaran, A Multi-perspective Analysis of the Storm (Peacomm) Worm, CSL Technical Note, Computer Science Laboratory, SRI International, 2007 <http://www.cyber-ta.org/pubs/StormWorm/>.

[20] M. Bailey, E. Cooke, F. Jahanian, Y. Xu, M. Karir, A survey of botnet technology and defenses, in: CATCH '09: Proceedings of the 2009 Cybersecurity Applications & Technology Conference for Homeland Security, IEEE Comp. Soc., 2009, pp. 299–304.

[21] P. Wang, L. Wu, B. Aslam, C. Zou, A systematic study on peer-to-peer botnets, in: Proceedings of 18th Interantional Conference on Computer Communications and Networks, 2009, ICCCN 2009, pp. 1 –8.

[22] J.B. Grizzard, V. Sharma, C. Nunnery, B.B. Kang, D. Dagon, Peer-to-peer botnets: overview and case study, in: HotBots'07: 1st Workshop on Hot Topics in Understanding Botnets, USENIX Ass., 2007.

[23] Q. Wang, Z. Chen, C. Chen, Characterizing internet worm infection structure, in: Proceedings of the 4th USENIX Conference on Large-Scale Exploits and Emergent Threats, LEET'11, USENIX Association, Berkeley, CA, USA, 2011.

[24] Z. Li, A. Goyal, Y. Chen, V. Paxson, Automating analysis of large-scale botnet probing events, in: Proceedings of the 4th International Symposium on Information, Computer, and Communications Security, ASIACCS '09, ACM, New York, NY, USA, 2009, pp. 11–22.

[25] C.C. Zou, D. Towsley, W. Gong, On the performance of internet worm scanning strategies, Performance Evaluation 63 (2006) 700–723.

[26] A. Papoulis, S.U. Pillai, Probability, Random Variables and Stochastic Processes, forth ed., McGraw-Hill, 2002.

[27] F. Brauer, P. van den Driessche, J. Wu (Eds.), Mathematical Epidemiology, Springer-Verlag, Berlin Heidelberg, 2008.

[28] H. Okamura, H. Kobayashi, T. Dohi, Markovian modeling and analysis of internet worm propagation, in: Proceedings of the 16th IEEE International Symposium on Software Reliability Engineering ISSRE, 2005.

[29] M.A. Rajab, J. Zarfoss, F. Monrose, A. Terzis, My botnet is bigger than yours (maybe, better than yours): why size estimates remain challenging, in: Proceedings of the first conference on First Workshop on Hot Topics in Understanding Botnets, USENIX Association, 2007.

[30] Top 10 Botnet Threat Report - 2010, Technical Report, Damballa Inc., 2011 <http://www.damballa.com/downloads/r_pubs/Damballa_2010_Top_10_Botnets_Report.pdf>.

[31] J.R. Douceur, The sybil attack, in: Revised Papers from the First International Workshop on Peer-to-Peer Systems, IPTPS'01, Springer-Verlag, London, UK, 2002. 521-260.

[32] The honeynet project, 2011 <http://www.honeynet.org/>.

[33] J. Rrushi, E. Mokhtari, A.A. Ghorbani, A statistical approach to botnet virulence estimation, in: Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security, ASIACCS '11, ACM, 2011, pp. 508–512.

[34] D. Zwillinger, Handbook of Differential Eqautions, third ed., Academic Press, 1997.

[35] A.D. Polyanin, V.F. Zaitsev, Handbook of Exact Solutions for Ordinary Differential Equations, second ed., Chapman & Hall/CRC, 2003.

[36] P.L. Sachdev, A Compendium on Nonlinear Ordinary Differential Equations, John Wiley & Sons Inc., 1997.

[37] S. Kondratenya, E. Prolisko, The existence and the form of solutions of Lienard equations with a moving algebraic singularity, Differential Equations 9 (1973) 198–201.

[38] M. Khosroshahy, M.K. Mehmet-Ali, D. Qiu, The sic botnet lifecycle model: a step beyond traditional epidemiological models (accompanying tech report: Mathematica derivations), 2012 <http://www.masoodkh.com/files/papers/SIC/SIC-TechReport.pdf>.

Masood Khosroshahy is currently a Ph.D. candidate in the Electrical and Computer Engineering Department of Concordia University, Montreal, Canada. He received his B.Sc. in "Electrical Engineering-Telecommunications" from Iran University of Science and Technology, Tehran, Iran (2004) and his M.Sc. in "Networked Computer Systems" from Télécom ParisTech (École nationale supérieure des télécommunications), Paris, France (2007). His research interests include computer and telecom networks security, cellular wireless networks, peer-to-peer traffic optimization, and congestion control.

Mustafa K. Mehmet-Ali received the B.Sc. and M.Sc. degrees in electrical engineering from Bogazichi University, Istanbul, Turkey, in 1977 and 1979, respectively, and the Ph.D. degree in electrical engineering from Carleton University, Ottawa, ON, Canada, in 1983. Until the end of 1984, he was a Research Engineer with Telesat Canada. Since 1985, he has been with the Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada, where he is currently a Professor. His current research interest is the performance modeling of wireless networks.

Dongyu Qiu received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, and the Ph.D. degree from Purdue University, West Lafayette, Indiana, USA, in 2003. He is currently an Associate Professor in the Department of Electrical and Computer Engineering, Concordia University, Montreal, Quebec, Canada. His research interests are in the areas of peer-to-peer networks, TCP/IP networks, network congestion control, queueing analysis, network security, and wireless networks.