SComF and SComI Botnet Models: The Cases of Initial Unhindered Botnet Expansion

Preprint (accepted version)

Presented in 25th Annual Canadian Conference on Electrical and Computer Engineering (CCECE12), April 29-May 2, 2012, Montréal, Canada, DOI: 10.1109/CCECE.2012.6334871

SCOMF AND SCOMI BOTNET MODELS: THE CASES OF INITIAL UNHINDERED BOTNET EXPANSION

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

Electrical and Computer Engineering Department
Concordia University, Montreal, Canada
m.kh@ieee.org, {mustafa, dongyu}@ece.concordia.ca

ABSTRACT

Botnets have become platforms to launch distributed denial-of-service attacks and coordinate massive e-mail spam campaigns, to name just a few of botnet-related nefarious activities. Apart from the wired networks, the increasingly Internet-enabled cellular wireless networks are also vulnerable to botnet attacks; a situation which motivates a thorough study of botnet expansion and the mathematical models thereof. In this paper, we propose the following two Continuous-Time Markov Chain-based models for prediction of the botnet size in the initial phase of botnet lifecycle: SComF for the case of finite number of susceptible nodes (suitable for a botnet expanding in a closed environment such as an administrative domain, or a LAN) and SComI for the case of infinite number of susceptible nodes (suitable for a botnet expanding in the larger Internet). Having access to such models would enable security experts to have reliable size estimates and therefore be able to defend against an emerging botnet with adequate resources. We derive the probability distributions for both models and provide some numerical results as well as a simulation study accompanying the numerical analysis of the SComF model using the GTNetS network simulator.

***Index Terms*—** Analytical models, Computer security, Epidemic models, Malware propagation, Botnets.

1. INTRODUCTION

Botnets, which are overlay networks of compromised computers built by cybercriminals known as botmasters, are pressing security concerns in the Internet world. As the size of some of the recent botnets has reached millions [1, 2], their firepower has become a major security threat. An important aspect of botnets that needs to be understood and predicted is their size; the bigger the size, the higher the threat level.

We present two Continuous-Time Markov Chain (CTMC) models of botnet expansion. CTMC models take into account *stochastic* population size changes and the appropriateness of their use has been confirmed [3] for malware propagation which happens under the influence of the same physical processes affecting botnets. Each dimension in the CTMC models represents a node *stage*, with the considered stages being *Susceptible* (i.e., susceptible to be compromised) and *Compromised* (i.e, *Infected* and *Connected* to the botnet). As botmasters use a plethora of methods to infect the nodes, it is reasonable to assume that a node is never in Immune/Removed stage. Further, we do not track the number of *Infected-only* nodes; these nodes are not important threats as long as they are not *Connected* to the botnet. Limiting the number of stages allows the development of tractable CTMC models.

We first model the unhindered growth of botnet when the population size is *infinite*. An infinite population size is a realistic assumption considering the total number of devices that are connected to the Internet today. As we consider **S**usceptible and **Com**promised stages and the population size to be **I**nfinite, we name the model **SComI**. Next, we model the unhindered growth of botnet when the population size is *finite*. The assumption of finite population size makes the model more suitable in case a segment of Internet or a local/wide area network is the environment in which the botnet can expand. As we consider **S**usceptible and **Com**promised stages and the population size to be **F**inite, we name the model **SComF**.

1.1. Related Work

Due to having similar problem structures, researchers have approached botnet population size modeling by adapting analytical results from the domain of malware propagation [4, 5] which in turn has borrowed from mathematical epidemiology [6]. There have been several deterministic models for botnet population size modeling [7–9], however, these models are inherently unable to capture the stochastic nature of population size changes. On the other hand, there have been few efforts recently to tackle the problem using stochastic mathematical methods: [10] modeled a botnet using a Stochastic Activity Network in an *analytical simulator* and [11] used stochastic Monte Carlo *simulation* to model a Peer-to-Peer botnet. Both of these studies are limited due to the used simulation environment which in general limits easy replication of the results and further analysis by others. Finally, [12] derived a probability model for a botnet, however, it equated *botnet expansion* with *worm spread*.

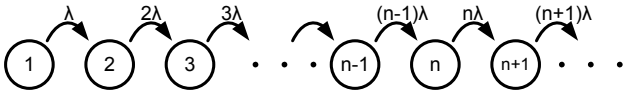

Fig. 1: SComI botnet model: 1-dimensional CTMC

1.2. Paper Contribution and Organization

The contribution of this paper is twofold: (1) two *analytical stochastic* botnet models, SComI and SComF, that cover both cases of infinite and finite node population sizes ; and (2) the method of examination of the interaction between the botnet expansion and the worm spread using Georgia Tech Network Simulator (GTNetS).

The paper is organized as follows: Sections 2 and 3 present the model, the probability distribution derivation, and some numerical results regarding the SComI model and the SComF model, respectively. We present the simulation study in Section 4 and conclude the paper in Section 5. Some of the intermediary steps in the derivations of this paper are provided in [13] due to space constraints.

2. SCOMI: UNHINDERED BOTNET EXPANSION MODEL - INFINITE POPULATION SIZE

In this section, we model unhindered growth of the botnet and define the state of the system to be the number of nodes that are in the botnet (nodes in *Compromised* stage). Our development leads to a solution for the time-dependent probability distribution of the number of nodes in the botnet.

2.1. State-transition-rate Diagram

The state-transition-rate diagram for the SComI model is depicted in Fig. 1. As initial condition, we assume that the size of the botnet is one. In this model, we consider that each node in the botnet recruits one node (grows the size of the botnet by one) with probability $\lambda\Delta t + o(\Delta t)$ in any Δt interval.

2.2. Probability Distribution Derivation

2.2.1. Differential-Difference Equations

For this pure-birth process, the rate of change of probability at any state is determined by setting it equal to the difference of probability flows into and out of that state as follows:

$$\frac{dP_n(t)}{dt} = (n-1)\lambda P_{n-1}(t) - n\lambda P_n(t) \qquad n \geq 1 \quad (1)$$

The initial condition is $P_1(0) = 1$.

2.2.2. Probability Generating Function

To determine $P_n(t)$, the probability distribution, we first derive the Probability Generating Function (PGF). For that, we need to start from the aforementioned differential-difference

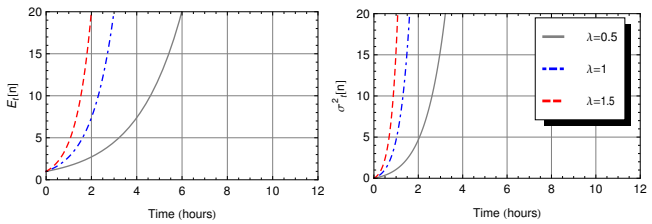

Fig. 2: SComI model: Mean and variance of botnet size

equation. The relationship between $P(z,t)$, the PGF, and $P_n(t)$, the probability distribution, is as follows: $P(z,t) = \sum_{n=0}^{\infty} P_n(t)z^n$. We therefore have $\frac{\partial P(z,t)}{\partial t} = \sum_{n=0}^{\infty} \frac{dP_n(t)}{dt} z^n$ and can write the initial condition in terms of PGF as $P(z,0) = z$. We multiply (1) by z^n and make a summation over the range of values of n to yield (after simplification):

$$\frac{\partial P(z,t)}{\partial t} + \lambda z(1-z)\frac{\partial P(z,t)}{\partial z} = 0 \qquad (2)$$

We need to solve this first-order Partial Differential Equation (PDE) in order to derive $P(z,t)$. We use the Method of Characteristics [14] to solve this PDE as follows: we define the auxiliary variable s which represents the scaled distance along a characteristic curve. Based on Method of Characteristics, z, t, and therefore $P(z,t)$ are effectively all functions of s. We therefore can write the following equations based on (2):

$$\begin{cases} \frac{\partial t}{\partial s} = 1 \\ \frac{\partial z}{\partial s} = \lambda z(1-z) \\ \frac{dP(z,t)}{ds} = 0 \end{cases} \qquad (3)$$

With the initial condition $P(z,0) = z$. From the previous equations, we can derive the PGF as follows:

$$P(z,t) = \frac{ze^{-\lambda t}}{1 - z + ze^{-\lambda t}} \qquad (4)$$

2.2.3. Probability Distribution

To get the probability distribution $P_n(t)$, we need to derive the inverse PGF of (4). Using the transform properties $A\alpha^n \Leftrightarrow \frac{A}{1-\alpha z}$ and $f_{n-k} \Leftrightarrow z^k F(z)$ (for $k > 0$), we derive $P_n(t)$, the probability distribution of the number of nodes in the botnet at time t, as follows:

$$P_n(t) = e^{-\lambda t}(1 - e^{-\lambda t})^{n-1} \qquad n \geq 1 \qquad (5)$$

2.3. Numerical Analysis

We now present some numerical results, depicted in Figs. 2 and 3, to illustrate how the derived probability distribution could be used in the study of any particular parameter of interest in the process of botnet expansion. Time-dependent mean and variance have been calculated and drawn, depicted in Fig. 2, which show how quickly botnet expansion can happen if the botnet is able to expand throughout the Internet. In each

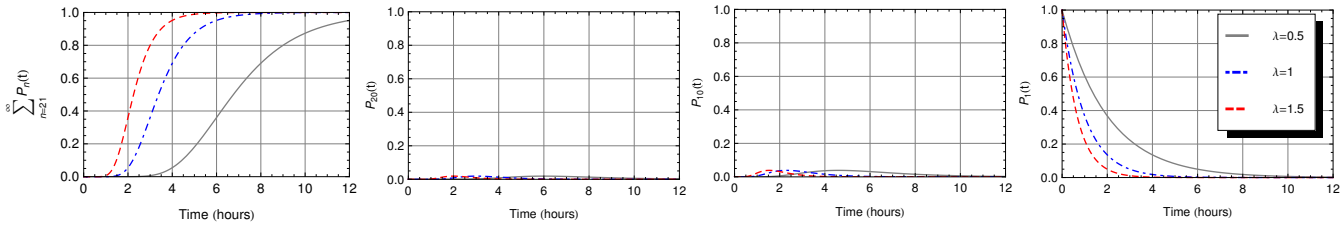

Fig. 3: SComI model: Time-dependent probability of the number of *Compromised* nodes (botnet size) being one (right), ten (center-right), twenty (center-left), or greater than twenty (left).

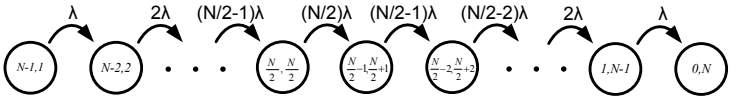

Fig. 4: SComF botnet expansion model: 2-dimensional CTMC. In the middle, the expansion rate starts to decrease.

figure, we demonstrate the effect of varying values of λ (with the unit of nodes/hour) which provides insight on how this parameter affects the speed of botnet expansion. In the numerical analysis of the SComF model (introduced in the next section), we set the total number of nodes (N) to 20 (N can be set to any arbitrarily large value of interest as well). To have a comparison between the two models, we draw the time-dependent probabilities for several values of N, including the time-dependent probability of number of *Compromised* nodes being greater than 20 (left sub-figure of Fig. 3).

3. SCOMF: UNHINDERED BOTNET EXPANSION MODEL - FINITE POPULATION SIZE

In this section, we model unhindered growth of the botnet with the finite population size assumption. In this model, there is a fixed number of nodes in *Susceptible* stage and these nodes move to *Compromised* stage as time goes by. State of the system is therefore defined to be the number of nodes in the aforementioned stages. Our development leads to a solution for the time-dependent probability distribution of the number of nodes in the botnet (nodes in *Compromised* stage).

3.1. State-transition-rate Diagram

The state-transition-rate diagram for the SComF model is depicted in Fig. 4. As initial condition, we assume that the size of the botnet is one and there are $N-1$ nodes in *Susceptible* stage. A state in the 2-dimensional CTMC is denoted by the duplet (n_0, n_1) (n_0 is the number of nodes in *Susceptible* stage and n_1 is the number of nodes in *Compromised* stage, as indicated in the diagram). One variable between n_0 and n_1, however, is a dependent variable since $n_0+n_1 = N$. For simplicity of notation, let us denote $P_{n0,n1}(t)$ as $P_{n1}(t)$ by dropping n_0 (i.e., considering n_0 to be the dependent variable). Finally, let us use n instead of n_1, thus $P_{n1}(t)$ is replaced by $P_n(t)$. $P_n(t)$ is therefore the time-dependent probability distribution of the number of nodes in the botnet.

In this model, we consider that each node in the botnet recruits one node (grows the size of the botnet by one) with probability $\lambda\Delta t + o(\Delta t)$ in any Δt interval. The expansion rate continues to increase up to the point where half of the susceptible population has left this stage. At this point, there are less nodes in *Susceptible* stage in the neighborhood of each node of the botnet; this would lead to a decrease in the expansion rate from that point on. The rate will continue to decrease until all nodes are in *Compromised* stage.

3.2. Probability Distribution Derivation

3.2.1. Differential-Difference Equations

For this birth process, the equations for the rate of change of probabilities are as follows:

$$\begin{cases} \frac{dP_1(t)}{dt} = -\lambda P_1(t) & n=1 \\ \frac{dP_n(t)}{dt} = (n-1)\lambda P_{n-1}(t) - n\lambda P_n(t) & 2 \le n \le \frac{N}{2} \\ \frac{dP_n(t)}{dt} = (N-n+1)\lambda P_{n-1}(t) - (N-n)\lambda P_n(t) & \frac{N}{2}+1 \le n \le N \end{cases} \tag{6}$$

Initial condition is $P_1(0) = 1$. Without loss of generality, we assume N to be even.

3.2.2. Laplace Transform of the Probability Distribution

From (6), letting $P_n^*(s)$ denote the Laplace transform of $P_n(t)$, we can derive the following expressions for $P_n^*(s)$:

$$P_n^*(s) = \begin{cases} \frac{1}{s+\lambda} & n=1 \\ \frac{(n-1)\lambda}{s+n\lambda} P_{n-1}^*(s) & 2 \le n \le \frac{N}{2} \\ \frac{(N-n+1)\lambda}{s+(N-n)\lambda} P_{n-1}^*(s) & \frac{N}{2}+1 \le n \le N \end{cases} \tag{7}$$

Using the Induction method, we can recursively determine $P_n^*(s)$ from (7) as follows:

$$P_n^*(s) = \begin{cases} \frac{1}{s+\lambda} & n=1 \\ \frac{(n-1)!\lambda^{n-1}}{\prod_{k=1}^{n}(s+k\lambda)} & 2 \le n \le \frac{N}{2} \\ \frac{\frac{N}{2}!}{(N-n)!}\lambda^{(n-\frac{N}{2})} \Big/ \prod_{k=1}^{n-\frac{N}{2}}\left(s+(\frac{N}{2}-k)\lambda\right) P_{\frac{N}{2}}^*(s) & \frac{N}{2}+1 \le n \le N \end{cases} \tag{8}$$

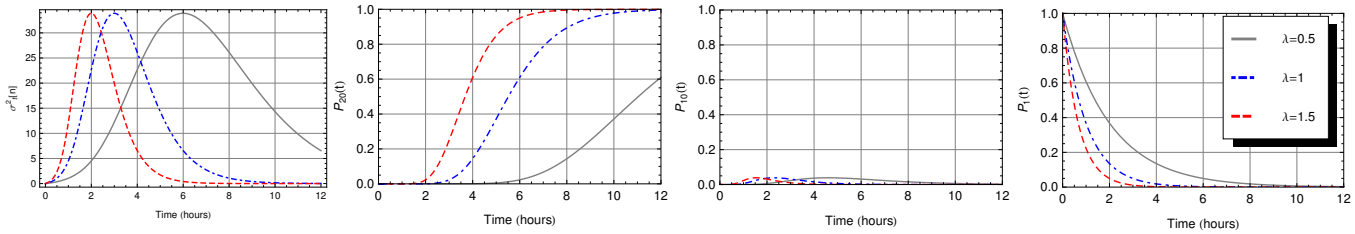

Fig. 5: SComF model: Time-dependent probabilities (right sub-figs.) and variance (left sub-fig.) of the botnet size

3.2.3. Probability Distribution

We use the Partial Fraction Expansion method on expressions obtained for $P_n^*(s)$ in (8) in order to perform the Laplace Inversion to derive $P_n(t)$ (see [13] for a detailed derivation). The probability distribution of the number of nodes in the botnet (nodes that are in *Compromised* stage) at time t (i.e., $P_n(t)$) is therefore obtained as follows:

$$\begin{cases} n{=}1: & e^{-\lambda t} \\ 2{\le}n{\le}\frac{N}{2}: & \sum_{k=0}^{n-1}\left((-1)^k\binom{n-1}{k}e^{-(k+1)\lambda t}\right) \\ \frac{N}{2}{+}1{\le}n{<}N: & \sum_{\substack{k=1 \\ k\notin[N-n,\frac{N}{2}-1]}}^{\frac{N}{2}}\left(\frac{T_1}{T_2}e^{-k\lambda t}\right) \\ & + \sum_{\substack{k=1 \\ k\in[N-n,\frac{N}{2}-1]}}^{\frac{N}{2}-1}\left(\frac{T_1}{T_2}te^{-k\lambda t}+\frac{d(\frac{T_1}{T_3})}{ds}\Big|_{s=-k\lambda}e^{-k\lambda t}\right) \\ n{=}N: & \frac{T_1}{T_2|_{k=0}}+\sum_{k=1}^{\frac{N}{2}-1}\left(\frac{T_1}{T_2}te^{-k\lambda t}+\frac{d(\frac{T_1}{T_3})}{ds}\Big|_{s=-k\lambda}e^{-k\lambda t}\right) \\ & + \frac{T_1}{T_2|_{k=\frac{N}{2}}}e^{-\frac{N}{2}\lambda t} \end{cases} \quad (9)$$

where T_1, T_2, and T_3 are given as follows:

$$\begin{aligned} T_1 &= \frac{\frac{N}{2}!}{(N-n)!}\lambda^{(n-\frac{N}{2})}(\tfrac{N}{2}-1)!\lambda^{\frac{N}{2}-1} \\ T_2 &= \prod_{\substack{i=1 \\ i\neq k}}^{\frac{N}{2}}(i-k)\lambda\prod_{\substack{j=N-n \\ j\neq k}}^{\frac{N}{2}-1}(j-k)\lambda \\ T_3 &= \prod_{\substack{i=1 \\ i\neq k}}^{\frac{N}{2}}(s+i\lambda)\prod_{\substack{j=N-n \\ j\neq k}}^{\frac{N}{2}-1}(s+j\lambda) \end{aligned}$$

3.3. Numerical Analysis

Comparable to the numerical analysis for the SComI model, Fig. 5 depicts the three respective time-dependent probabilities for the SComF model as well as the time-dependent variance (left sub-fig.). Fig. 6 depicts the time-dependent mean number of nodes in *Compromised* stage (botnet size) for the SComF model. Like before, the effect of varying values of λ on the speed of botnet expansion can be observed in the figures. It is also interesting to observe the "saturation effect" as the botnet expands to all nodes, depicted in Fig. 6.

4. SIMULATION STUDY

To our knowledge, there are neither publicly available measurement data nor simulation models of initial botnet expansion. As botnets are inherently designed to remain invisible, it is extremely difficult, if not impossible, to be able to observe and measure in real-time the expansion of a real-world botnet. In this section, we provide a brief simulation study which is complementary to the presented numerical analysis. The simulation study concerns the SComF model, as the SComI model does not lend itself to comparison to a simulated network with limited number of nodes.

We borrow the term "spearhead" and the notion of "two-stage worm" used in [7] to describe our simulation study. *Spearhead* is the initial infection (e.g., a worm) that spreads in the network after which the infected nodes start to connect to the botnet which is the process of botnet expansion. The botnet expansion therefore happens with a delay compared to the spread of the initial infection. In this simulation study, we examine the spread of the initial infection (the worm spread) and compare it to the determined analytical result for the botnet expansion.

The infection of nodes (e.g., worm spread) is independent from the connection of infected nodes to the botnet. Specifically, there is no relationship between the infection rate and the connection rate (λ), i.e., the latter could be lower or higher than the former. It is, however, interesting to observe how the botnet could expand in relation to the spread of the worm, in case these rates are equal.

The worm models developed in the GTNetS simulator [15] seem to be the most well-developed publicly-available simulation models for worm spread. With these models incorporated, the simulator models the activities of the worms in packet-level detail. In this simulator, a UDP-based worm can be configured using parameters such as worm "scan rate", whereas a TCP-based worm can be configured by setting the number of simultaneous TCP connections that each infectious node can create. We use a UDP-based worm in this simulation in order to be able to have a comparison between λ of the SComF model and the "scan rate" of the worm model[1].

Comparable to the numerical results for the SComF model, we simulated a 20-host topology and tracked the number of infected hosts over time. The worm "scan rate" is the number of infectious packets each infected host sends out each hour (here, scan rate is set to one). The number of infected hosts over time is depicted in Fig. 7, which is one typical result

[1]The simulation scenario is a slightly modified version of *wormsim.cc* in GTNetS; see [13] for the code and for a screen shot of the simulated topology.

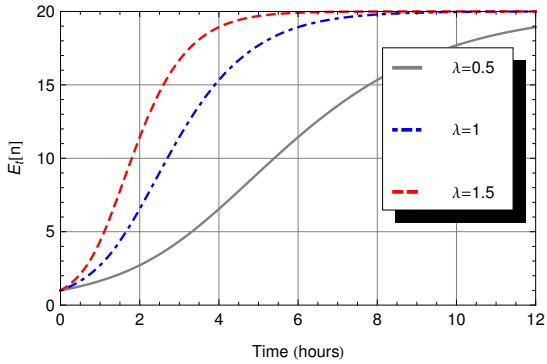

Fig. 6: SComF model: Mean botnet size

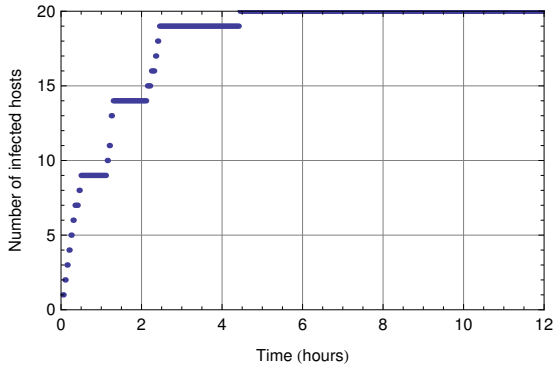

Fig. 7: Count of infected hosts over time in the simulation

from several runs of the simulation. The curve indicating the number of infected hosts over time can be superimposed over Fig. 6 which shows the botnet growth. We therefore can observe that the botnet growth (for $\lambda = 1$) correctly lags the spread of infection (for scan rate $= 1$).

5. CONCLUDING REMARKS

Using the developed analytical models, the botnet size estimation problem has reduced from having to estimate the global size of the botnet to the estimation of the model's parameter (λ) which requires only local knowledge. In order to use the models in the real world, one could consider the following methods when trying to estimate a value for λ: (1) local measurements through Honeynet log analysis [16], for example; and (2) a statistical approach to botnet virulence estimation which has recently been proposed [17]; this latter method improves the reliability of the process of estimating λ.

The analytical models as well as the accompanying numerical results provide some indication as to how a botnet would expand in various scenarios. On the other hand, simulation results concerning the SComF model shed light on the interaction between the spread of the initial infection and the subsequent botnet expansion. As both SComF and SComI models account for the most important node stages (i.e., *Susceptible* and *Compromised*), they are useful and sufficient models in this particular application area, i.e., prediction and analysis of initial unhindered botnet expansion. The insight derived from the use of these two models leads to security practitioners being able to estimate the size of the botnet in order to adequately deploy mitigation strategies.

6. REFERENCES

[1] Phillip Porras, "Inside risks: Reflections on conficker," *Commun. ACM*, vol. 52, pp. 23–24, October 2009.

[2] Steve Mansfield-Devine, "Battle of the botnets," *Network Security*, vol. 2010, no. 5, pp. 4 – 6, 2010.

[3] H. Okamura, H. Kobayashi, and T. Dohi, "Markovian modeling and analysis of internet worm propagation," in *Proc. 16th IEEE Int. Symp. Software Reliability Engineering ISSRE 2005*, 2005.

[4] Giuseppe Serazzi and Stefano Zanero, "Computer virus propagation models," in *Performance Tools and Applications to Networked Systems*, M. C. Calzarossa and E. Gelenbe, Eds., vol. 2965 of *Lecture Notes in Comp. Science*, pp. 26–50. Springer Berlin Heidelberg, 2004.

[5] Su Fei, Lin Zhaowen, and Ma Yan, "A survey of internet worm propagation models," in *Proc. 2nd IEEE Int. Conf. Broadband Network & Multimedia Technology IC-BNMT '09*, 2009, pp. 453–457.

[6] Fred Brauer, Pauline van den Driessche, and Jianhong Wu, Eds., *Mathematical Epidemiology*, Springer Berlin Heidelberg, 2008.

[7] C. C. Zou and R. Cunningham, "Honeypot-aware advanced botnet construction and maintenance," in *Proc. Int. Conf. Dependable Systems and Networks DSN 2006*, 2006, pp. 199–208.

[8] David Dagon, Cliff Zou, and Wenke Lee, "Modeling botnet propagation using time zones," in *Proceedings of the 13th Network and Distributed System Security Symposium, NDSS*, 2006.

[9] Marco Ajelli, Renato Lo Cigno, and Alberto Montresor, "Modeling botnets and epidemic malware," in *Proc IEEE Int Comm Conf ICC*, 2010, pp. 1–5.

[10] Elizabeth Van Ruitenbeek and William H. Sanders, "Modeling peer-to-peer botnets," in *QEST '08: Proc. of the Fifth Int'l Conf. on Quantitative Evaluation of Systems*. 2008, pp. 307–316, IEEE Comp. Soc.

[11] Qian Wang, Zesheng Chen, Chao Chen, and N. Pissinou, "On the robustness of the botnet topology formed by worm infection," in *Proc. IEEE Global Telecom. Conf. (GLOBECOM)*, 2010, pp. 1–6.

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

[13] Masood Khosroshahy, Mustafa K. Mehmet-Ali, and Dongyu Qiu, "Scomf and scomi botnet models: The cases of initial unhindered botnet expansion (accompanying tech report)," [Online]. Available: http://www.masoodkh.com/files/papers/SCom/SCom-TechReport.pdf.

[14] Daniel Zwillinger, *Handbook of Diff. Eqs.*, Academic Press, 1997.

[15] G. E. Riley, M. L. Sharif, and Wenke Lee, "Simulating internet worms," in *Proc. IEEE Computer Society's 12th Annual Int. Symp. Modeling, Analysis, and Simulation of Computer and Telecommunications Systems (MASCOTS 2004)*, 2004, pp. 268–274.

[16] "Honeynet project," [Online]. Available: http://www.honeynet.org/.

[17] Julian Rrushi, Ehsan Mokhtari, and Ali A. Ghorbani, "A statistical approach to botnet virulence estimation," in *Proc. of the 6th ACM Symposium on Info., Comp. and Comm. Security*, 2011, pp. 508–512.