Deterministic epidemic modeling of future botnet malware with a contact process

Hideyoshi Miura¹, Tomotaka Kimura², and Kouji Hirata¹a)

¹ Faculty of Engineering Science, Kansai University
Yamate-cho, Suita, Osaka 564-8680, Japan
² Faculty of Science and Engineering, Doshisha University
Tataramiyakodani, Kyotanabe, Kyoto 610-0934, Japan

a) hirata@kansai-u.ac.jp

Abstract: This paper proposes deterministic epidemic modeling of future botnet malware named self-evolving botnets, assuming a contact process. The self-evolving botnets discover vulnerabilities of hosts by exploiting the computing resources of infected hosts, which could become serious threats to the future Internet society. The proposed epidemic model represents their infection dynamics on an overlay network consisting of hosts. It predicts the infection spreading of the self-evolving botnets with the use of mean field approximation using the degree distribution of the overlay network. Through numerical calculations, we show that the proposed epidemic model well approximates the infection spreading in short time.

Keywords: Infection dynamics, deterministic epidemic model, malware, contact process, degree distribution

Classification: Internet

References

[1] MATLAB, https://jp.mathworks.com/products/matlab.html, accessed Nov. 26, 2020.
[2] A. Barabasi and R. Albert, “Emergence of scaling in random networks,” Science, vol. 286, pp. 509–512, 1999.
[3] J. Borello and L. Me, “Code obfuscation techniques for metamorphic viruses,” Journal in Computer Virology, vol. 4, no. 3, pp. 211–220, 2008.
[4] A. Cani, M. Gaudesi, E. Sanchez, G. Squillero, and A. Tonda, “Towards automated malware creation: code generation and code integration,” in Proc. Symposium on Applied Computing, Gyeongju, Korea, Mar. 2014.
[5] A. Dadlani, M. Kumar, S. Murugan, and K. Kim, “System dynamics of a refined epidemic model for infection propagation over complex networks,” IEEE Systems Journal, vol. 10, no. 4, pp. 1316–1325, 2016.
[6] T. Kudo, T. Kimura, Y. Inoue, H. Aman, and K. Hirata, “Stochastic modeling of self-evolving botnets with vulnerability discovery,” Computer Communications, vol. 124, pp. 101–110, 2018.
[7] R. Scandariato, J. Walden, A. Hovsepian, and W. Joosen, “Predicting vulnerable software components via text mining,” IEEE Transactions on Software Engineering, vol. 40, no. 10, pp. 993–1006, 2014.
1 Introduction

There exist some researches that discover vulnerabilities in software with the use of machine learning to protect software against malware [7]. However, they can be exploited by malicious attackers. Furthermore, malware has rapidly evolved. For example, [3] has introduced metamorphic malware that rewrites and obfuscates its own source codes and [4] has discussed generation of new malware by combining known malware source codes. Accordingly, Kudo et al. [6] have introduced a new concept of future malware named self-evolving botnets. They discover vulnerabilities by means of distributed machine learning with the use of computing resources of infected hosts. By using discovered vulnerabilities, susceptible hosts get infected and then are embedded into the self-evolving botnets. The authors in [6] have provided a stochastic epidemic model of self-evolving botnets, which represents the infection dynamics as a continuous-time Markov chain. Through Monte Carlo simulation experiments on the Markov chain, the authors have shown that the infectability of the self-evolving botnets is too strong.

Epidemic models are categorized into the stochastic model [6] and the deterministic model [5]. In simulation experiments based on the stochastic model, we can see detailed events randomly occurring with time elapsed such as the state transition of each host. However, to calculate performance metrics such as the change in the average number of infected hosts, we need to repeat simulation experiments many times, so that the computation time becomes relatively long. To overcome this problem, this paper proposes a deterministic epidemic model of self-evolving botnets. It provides infection dynamics, assuming a contact process [8] on an overlay network. By calculating them with the use of mean field approximation using the degree distribution of the overlay network, the proposed deterministic model predicts the infection spreading of the self-evolving botnets. Through numerical calculations, we show that the proposed model can provide good approximation solution in relatively short time.

2 Proposed epidemic model

2.1 Behavior of self-evolving botnets

The detailed behavior of self-evolving botnets has been discussed in [6]. Therefore, we here briefly explain the behavior. We represent the state of each host with a Susceptible-Infected-Recovered-Susceptible (SIRS) model shown in Fig. 1(a), where “S” indicates that the host has vulnerabilities, “I” indicates that the host is infected with botnet malware, and “R” indicates that the host has no known vulnerabilities. Each host belongs to one of the

[8] K. Sharkey, “Deterministic epidemic models on contact networks: Correlations and unbiological terms, Theoretical Population Biology, vol. 79, no. 4, pp. 115–129, 2011.
states. Susceptible hosts could get infected with the botnet malware, and then transition to state I as shown in Fig. 1(b). Susceptible hosts and infected hosts transition to state R when they repair their vulnerabilities and eliminate the botnet malware from themselves, respectively. All hosts in state R transition to state S when the self-evolving botnet discovers a new vulnerability. In this case, the botnet malware could infect the hosts with the use of the newly discovered vulnerability and take in them (see Fig. 1(b)), which can be exploited to discover further new vulnerabilities.

2.2 Deterministic modeling

We assume a contact process as an infection spreading model, where infected hosts can attack only susceptible hosts adjacent to each other on a given network as shown in Fig. 1(b). Based on the SIRS model and the contact process, we provide the deterministic epidemic model of a self-evolving botnet on an overlay network consisting of hosts, which considers the mean field approximation using the degree distribution. Let $N$ and $K$ denote the number of hosts and the maximum degree of hosts, respectively, on the overlay network. Let $S_k(t)$, $I_k(t)$, and $R_k(t)$ denote the numbers of hosts whose degree is equal to $k$ ($k = 1, 2, \ldots, K$) and belong to states S, I, and R, respectively, at time $t$, where $N = \sum_{k=1}^{K} S_k(t)$, $I(t) = \sum_{k=1}^{K} I_k(t)$, and $R(t) = \sum_{k=1}^{K} R_k(t)$. Our deterministic model represents the infection dynamics with the following differential equations, where $k = 1, 2, \ldots, K$:

$$\frac{dS_k(t)}{dt} = -\alpha k S_k(t) \Phi(f(t)) - \delta S_k(t) + \eta \sum_{k'=1}^{K} I_{k'}(t) + 1) R_k(t), \quad (1)$$
\[
\frac{dI_k(t)}{dt} = \alpha k S_k(t) \Phi(f(t)) - \beta I_k(t), \tag{2}
\]
\[
\frac{dR_k(t)}{dt} = \beta I_k(t) + \delta S_k(t) - \eta \left( \sum_{k'=1}^{K} I_{k'}(t) + 1 \right) R_k(t), \tag{3}
\]

where \( \alpha \) denotes the malware infection rate per infected host, \( \delta \) denotes the repair rate per susceptible host, \( \beta \) denotes the malware elimination rate per infected host, and \( \eta \) denotes the new vulnerability discovery rate per infected host. \( \Phi(f(t)) \) denotes the probability that an adjacent host of a given host is infected, which is given by \( \Phi(f(t)) = \sum_{k'} k' p(k') f_{k'}(t) \bar{k} \), where \( \bar{k} \) denotes the average degree of the hosts, \( p(k') \) denotes the probability that the degree of a host is equal to \( k' \) (i.e., degree distribution), and \( f_{k'}(t) \) denotes the ratio of the number of infected hosts with degree of \( k' \) at time \( t \) to the total number of hosts with degree of \( k' \) at time \( t \).

In (1) and (2), \( \alpha k S_k(t) \Phi(f(t)) \) is the average number of susceptible hosts with degree of \( k \) per unit time that get infected by adjacent infected hosts on the overlay network. In (1) and (3), \( \delta S_k(t) \) indicates the average number of susceptible hosts with degree of \( k \) per unit time that repair their vulnerabilities. Also, in (2) and (3), \( \beta I_k(t) \) means the average number of infected hosts with degree of \( k \) per unit time that eliminate the botnet malware from themselves. When the self-evolving botnet discovers unknown vulnerabilities, recovered hosts transition to state S. In (1) and (3), \( \eta \left( \sum_{k'=1}^{K} I_{k'}(t) + 1 \right) R_k(t) \) indicates the average number of recovered hosts with degree of \( k \) per unit time that the self-evolving botnet discovers their new vulnerabilities. Note that as discussed above, all recovered hosts transition to state S when the self-evolving botnet discovers a new vulnerability [6]. We here approximate this event by \( \eta \left( \sum_{k'=1}^{K} I_{k'}(t) + 1 \right) R_k(t) \), which indicates that the occurrence frequency of vulnerability discovery events is proportional to the number of recovered hosts. By calculating these differential equations, we can observe the behavior of the self-evolving botnet on the overlay network.

### 3 Numerical calculations

We here conduct numerical calculations based on (1)-(3). We use MATLAB [1] to calculate these differential equations. For the sake of comparison, we obtain results from simulation experiments based on the stochastic model [6]. In these calculations and experiments, we use DELL Precision 5820 with an Intel Xeon Silver 4110 2.10 GHz CPU and 16 GB memory, running Ubuntu 20.04. We assume that there exist \( N = 5,000 \) hosts in an overlay network. The overlay network is constructed based on the Barabasi-Albert (BA) model [2] with the average degree \( \bar{k} = 4 \) and maximum degree \( K \) is 142. One host is infected and all the other hosts are susceptible at time \( t = 0 \), i.e., \( (S(0), I(0), R(0)) = (4999, 1, 0) \). As the initial infected host, we select a host with degree of \( k = 107 \), which is an initial placement host of the BA model. The parameters \( \beta \) and \( \delta \) are set to 0.005 and 0.05 in (1)-(3), respectively.
Fig. 2. Numbers of hosts in respective states.

Fig. 2(a) shows the numbers of hosts in respective states against time elapsed, which are obtained from simulation experiments, where the malware infection rate $\alpha = 0.01$ and the vulnerability discovery rate $\eta = 0.01$. These values are the average of 100 samples. Also, Figs. 2(b) shows the result obtained from the proposed deterministic model. From these figures, we observe that the number of infected hosts rapidly increases with time elapsed and does not decrease, which indicates the high infectability of the botnet malware. We also observe that these two models have similar behavior. The number of recovered hosts in Fig. 2(a) is slightly larger than that in Fig. 2(b). This is because we approximate vulnerability discovery events by the botnet malware in the proposed deterministic model, as mentioned in Section 2. However, the numbers of hosts in the respective states of each model become almost the same with time elapsed.

Figs. 2(c) and (d) show the numbers of hosts in respective states against time elapsed, where $\alpha = 0.01$ and $\eta = 0.02$. Compared with the results in Figs. 2(a) and (b), the vulnerability discovery rate of the botnet malware increases. Therefore, the number of recovered hosts decreases in early stage of the spreading. Also, Figs. 2(e) and (f) show the numbers of hosts in respective states against time elapsed, where $\alpha = 0.02$ and $\eta = 0.01$. We here set the malware infection rate $\alpha$ to a larger value, and thus the infection
spreading speed is high and the number of infected hosts becomes very large with time elapsed. From these figures, we also observe that the proposed deterministic model well approximates the behavior in different situations.

Figs. 3(a) and (b) show the computation time against the number of hosts and \( \eta \), respectively, where \( \alpha = 0.01 \). As we can see from these figures, the computation time of the proposed deterministic model is much smaller than that of the simulation experiments. Note that the computation time increases with \( \eta \) because vulnerability discovery events frequently occur when \( \eta \) is large. As shown in Fig. 2, although the respective results of the simulation experiments are the average of 100 samples, they are not stable especially in early stage of the spreading. We thus conclude that the proposed deterministic model well approximates the behavior of the botnet malware in short time.

4 Conclusion

This paper proposed the deterministic epidemic model of self-evolving botnets to predict the infection spreading, assuming a contact process on an overlay network. Through numerical calculations, we showed that the proposed epidemic model well approximates the infection spreading in short time.

Acknowledgments

This research was partially supported by JSPS KAKENHI Grant No. 20H04184.