International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 02
February 2017

The Spread of Malicious Software in Large Scale Networks

CHINTAKUNTA VIJAY ANAND 1*,S G NAWAZ 2* ,R RAMACHANDRA 3*

1. M.tech,dept of cse sri krishna devaraya engineering college.

2.Assoc.prof,head-dept of cse, sri krishna devaraya engineering college.

3.Principal, sri krishna devaraya engineering college

Abstract: Malware is pervasive in networks, and poses a critical threat to network security. However, we have very limited understanding of malware behavior in networks to date. In this paper, we investigate how malware propagate in networks from a global perspective. We formulate the problem, and establish a rigorous two layer epidemic model for malware propagation from network to network. Based on the proposed model, our analysis indicates that the distribution of a given malware follows exponential distribution, power law distribution with a short exponential tail, and power law distribution at its early, late and final stages, respectively. Extensive experiments have been performed through two real-world global scale malware data sets, and the results confirm our theoretical findings

1. INTRODUCTION

MALWARE are malicious software programs deployed by cyber attackers to compromise computer systems by exploiting their security vulnerabilities. Motivated by extraordinary financial or political rewards, malware owners are exhausting their energy to compromise as many networked computers as they can in order to achieve their malicious goals. A compromised computer is called a bot, and all bots compromised by a malware form a botnet. Botnets have become the attack engine of cyber attackers, and they pose critical challenges to cyber defenders. In order to fight against cyber criminals, it is important for defenders to understand malware behavior, such as propagation or membership recruitment patterns, the size of botnets, and distribution of bots. To date, we do not have a solid understanding about the size and distribution of malware or botnets.

Researchers have employed various methods to measure the size of botnets, such as botnet infiltration [1], DNS redirection [3], external information [2]. These efforts indicate that the size of botnets varies from millions to a few thousand. There are no dominant principles to explain these variations. As a result, researchers desperately desire effective models and explanations for the chaos. Dagon et al. [3] revealed that time zone has an obvious impact on the number of available bots. Mieghem et al. [4] indicated that network topology has an important impact on malware spreading through their rigorous mathematical analysis. Recently, the emergence of mobile malware, such as Cabir [5], Ikee [6], and Brador [7], further increases the difficulty level of our understanding on how they propagate. More details about mobile malware can be found at a recent survey paper [8]. To the best of our knowledge, the best finding about malware distribution in large-scale networks comes from Chen and Ji [9]: the distribution is non-uniform. All this indicates that the research in this field is in its early stage. The epidemic theory plays a leading role in malware propagation modelling. The current models for malware spread fall in two categories: the epidemiology model and the control theoretic model. The control system theory based models try to detect and contain the spread of malware [10], [11]. The epidemiology models are more focused on the number of compromised hosts and their distributions, and they have been explored extensively in the computer science community [12], [13], [14]. Zou et al. [15] used a susceptible-infected (SI) model to predict the growth of Internet worms at the early stage. Gao and Liu [16] recently employed a susceptible-infected-recovered (SIR) model to describe mobile virus propagation. One critical condition for the epidemic models is a large vulnerable population because their principle is based on differential equations.

More details of epidemic modelling can be find in [17]. As pointed by Willinger et al. [18], the findings, which we extract from a set of observed data, usually reflect parts of the studied objects. It is more reliable to extract theoretical results from appropriate models with confirmation from sufficient real world data set experiments. We practice this principle in this study.

In this paper, we study the distribution of malware in terms of networks (e.g., autonomous systems (AS), ISP domains, abstract networks of smartphones who share the same vulnerabilities) at large scales. In this kind of setting, we have a sufficient volume of data at a large enough scale to meet the requirements of the SI model. we break our model into two layers. First of all, for a given time since the breakout of a malware, we calculate how many networks have been compromised based on the SI model. Second, for a compromised network, we calculate how many hosts have been compromised since the time that the network was compromised. With this two layer model in place, we can determine the total number of compromised hosts and their distribution in terms of networks. Through our rigorous analysis, we find that the distribution of a given malware follows an exponential distribution at its early stage, and obeys a power law distribution with a short exponential tail at its late stage, and finally converges to a power law distribution. We examine our theoretical findings through two large-scale real-world data sets: the Android based malware [19] and the Conficker. The experimental results strongly support our theoretical claims. To the best of our knowledge, the proposed two layer epidemic model and the findings are the first work in the field.

Our contributions are summarized as follows. We propose a two layer malware propagation model to describe the development of a given malware at the Internet level. Compared with the existing singlelayer epidemic models, the proposed model represents malware propagation better in large-scale networks. We find the malware distribution in terms of networks varies from exponential to power law with a short exponential tail, and to power law distribution at its early, late, and final stage, respectively. These findings are first theoretically proved based on the proposed model, and then confirmed by the experiments through the two large-scale real-world data sets.

2 RELATED WORK

The basic story of malware is as follows. A malware programmer writes a program, called bot or agent, and then installs the bots at compromised computers on the Internet using various network virus-like techniques. All of his bots form a botnet, which is controlled by its owners to commit illegal

tasks, such as launching DDoS attacks, sending spam emails, performing phishing activities, and collecting sensitive information. There is a command and control (C&C) server(s) to communicate with the bots and collect data from bots. In order to disguise himself from legal forces, the botmaster changes the url of his C&C frequently, e.g., weekly. An excellent explanation about this can be found in [1]. With the significant growing of smartphones, we have witnessed an increasing number of mobile malware. Malware writers have develop many mobile malware in recent years. Cabir [5] was developed in 2004, and was the first malware targeting on the Symbian operating system for mobile devices. Moreover, it was also the first malware propagating via Bluetooth. Ikee [6] was the first mobile malware against Apple iPhones, while Brador [7] was developed against Windows CE operating systems. The attack victors for mobile malware are diverse, such as SMS, MMS, Bluetooth, WiFi, and Web browsing. Peng et al. [8] presented the short history of mobile malware since 2004, and surveyed their propagation models. A direct method to count the number of bots is to use botnet infiltration to count the bot IDs or IP addresses. Stone- Gross et al. [1] registered the URL of the Torpig botnet before the botmaster, and therefore were able to hijack the C&C server for ten days, and collect about 70G data from the bots of the Torpig botnet. They reported that the footprint of the Torpig botnet was 182,800, and the median and average size of the Torpig's live population was 49,272 and 48,532, respectively. They found 49,294 new infections during the ten days takeover. Their research also indicated that the live population fluctuates periodically as users switch between being online and offline. This issue was also tacked by Dagon et al. in [3]. Another method is to use DNS redirection. Dagon et al. [3] analyzed captured bots by honypot, and then identified the C&C server using source code reverse engineering tools. They then manipulated the DNS entry which is related to a botnet's IRC server, and redirected the DNS requests to a local sinkhole. They therefore could count the number of bots in the botnet. As discussed previously, their method counts the footprint of the botnet, which was 350,000 in their report.

In this paper, we use two large scale malware data sets for our experiments. Conficker is a well-known and one of the most recently widespread malware. Shin et al.collected a data set about 25 million Conficker victims from all over the world at different levels. At the same time, malware targeting on Android based mobile systems are developing quickly in recent years. Zhou and Jiang [19] collected a large data set of Android based malware. In [2], Rajab et al. pointed out that it is inaccurate to count the unique IP addresses of bots because DHCP and NAT techniques are employed extensively on the Internet ([1] confirms this by their observation that 78.9 percent of the infected machines were behind a NAT, VPN, proxy, or firewall). They therefore proposed to examine the hits of DNS caches to find the lower bound of the size of a given botnet. Rajab et al. [21] reported that botnets can be categorized into two major genres in terms of membership recruitment: worm-like botnets and variable scanning botnets. The latter weights about 82 percent in the 192 IRC bots that they investigated, and is the more prevalent class seen currently. Such botnets usually perform localized and non-uniform scanning, and are difficult to track due to their intermittent and continuously changing behavior. The statistics on the lifetime of bots are also reported as 25 minutes on average with 90 percent of them staying for less than 50 minutes. Malware propagation modelling has been extensively explored. Based on epidemiology research, Zou et al. [15] proposed a number of models for malware monitoring at the early stage. They pointed out that these kinds of model are appropriate for a system that consists of a large number of vulnerable hosts; in other words, the model is effective at the early stage of the outbreak of malware, and the accuracy of the model drops when the malware develops further. As a variant of the epidemic category, Sellke et al. [12] proposed a stochastic branching process model for characterizing the propagation of Internet worms, which especially focuses on the number of compromised computers against the number of worm scans, and presented a closed form expression for the relationship. Dagon et al. [3] extended the model of [15] by introducing time zone information aðtÞ, and built a model to describe the impact on the number of live members of botnets with diurnal effect.

3 PRELIMINARIES

Preliminaries of epidemic modelling and complex networks are presented in this section as this work is mainly based on the two fields. For the sake of convenience, we summarize the symbols that we use in this paper .

3.1 Deterministic Epidemic Models

After nearly 100 years development, the epidemic models [17] have proved effective and appropriate for a system that possesses a large number of vulnerable hosts. In other words, they are suitable at a macro level. Zou et al. [15] demonstrated that they were suitable for the studies of Internet based virus propagation at the early stage. We note that there are many factors that impact the malware propagation or botnet membership recruitment, such as network topology, recruitment frequency, and connection status of vulnerable hosts. All these factors contribute to the speed of malware propagation. Fortunately, we can include all these factors into one parameter as infection rate b in epidemic theory. Therefore, in our study, let N be the total number of vulnerable hosts of a large-scale network (e.g., the Internet) for a given malware. There are two statuses for any one of the N hosts, either infected or susceptible. Let IðtÞ be the number of infected hosts at time t, then we where RðtÞ, and QðtÞ represent the number of removed hosts from the infected population, and the number of

removed hosts from the susceptible population at time t. The variable bðtÞ is the infection rate at time t. For our study, model (1) is too detailed and not necessary

as we expect to know the propagation and distribution of a given malware. As a result, we employ the following susceptible-infected model:

$$\frac{dI(t)}{dt} = \beta I(t)[N - I(t)] \qquad (2)$$

where the infection rate b is a constant for a given malware for any network.

We note that the variable t is continuous in model (2) and (1). In practice, we measure IðtÞ at discrete time points. Therefore, t ¼ 0; 1; 2; . . . . We can interpret each time point as a new round of malware membership recruitment, such as vulnerable host scanning. As a result, we can transform model (2) into the discrete form as follows:

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 02
February 2017

$$I(t) = (1 + \alpha\Delta)I(t-1) - \beta\Delta I(t-1)^2, \quad (3)$$

where t ¼ 0; 1; 2; . . . ; D is the unit of time, Ið0Þ is the initial number of infected hosts (we also call them seeds in this paper), and a ¼ bN, which represents the average number of vulnerable hosts that can be infected by one infected host per time unit. In order to simplify our analysis, let D ¼ 1, it could be one second, one minute, one day, or one month, even one year, depending on the time scale in a given context. Hence, we have a simpler discrete form given by

$$I(t) = (1 + \alpha)I(t-1) - \beta(I(t-1)) \quad (4)$$

Based on Equation (4), we define the increase of infected hosts for each time unit as follows.

$$\Delta I(t) , I(t) - I(t-1), t = 1, 2, \ldots \quad (5)$$

To date, many researches are confined to the "early stage" of an epidemic, such as [15]. Under the early stage condition, IðtÞ << N, therefore, N _ IðtÞ _ N. As a result,

a closed form solution is obtained as follows: $I(t) = I(0)e^{\beta N t}$. (6)

Whenwe take the ln operation on both sides of Equation (6),

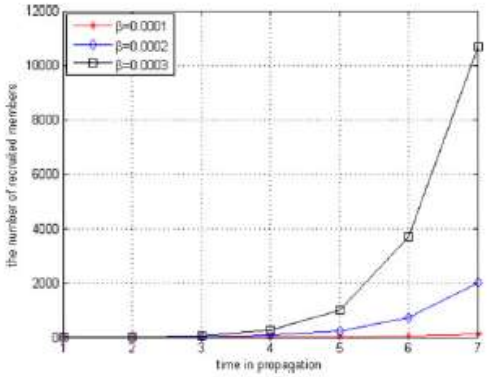

Fig. 1. The impact from infection rate β on the recruitment progress for a given vulnerable network with $N = 10,000$.

we have $I(t) = \beta N t + \ln I(0)$. (7) For a given vulnerable network, b, N and Ið0Þ are constants, therefore, the graphical representation of Equation (7) is a straight line. Based on the definition of Equation (5), we obtain the increase of new members of a malware at the early stage as

$$\Delta I(t) = (e^{\beta N} - 1)I(t-1)$$

$$= (e^{\beta N} - 1)I(0)e^{\beta N (t-1)} .(8)$$

Taking the ln operation on both side of (8), we have

$$\ln \Delta I(t) = \beta N(t-1) + \ln (e^{\beta N} - 1)I(0) . \quad (9)$$

Similar to Equation (7), the graphical representation of equation (9) is also a straight line. In other words, the number of recruited members for each round follows an exponential distribution at the early stage.

We have to note that it is hard for us to know whether an epidemic is at its early stage or not in practice. Moreover, there is no mathematical definition about the term early stage. In epidemic models, the infection rate b has a critical impact on the membership recruitment progress, and b is usually a small positive number, such as 0.00084 for worm Code Red [12]. For example, for a network with N ¼ 10;000 vulnerable hosts, we show the recruited members under different infection rates in Fig. 1. From this diagram, we can see that the recruitment goes slowly when b ¼ 0:0001, however, all vulnerable hosts have been compromised in less than 7 time units when b ¼ 0:0003, and the recruitment progresses in an exponential fashion. This reflects the malware propagation styles in practice. For malware based on "contact", such as blue tooth contacts, or viruses depending on emails to propagate, the infection rate is usually small, and it takes a long time to compromise a large number of vulnerable hosts in a given network. On the other hand, for some malware, which take active actions for recruitment, such as vulnerable host scanning, it may take one or a few rounds of scanning to recruit all or a majority of the vulnerable hosts in a given network. We will apply this in the following analysis and performance evaluation.

3.2 Complex Networks

Research on complex networks have demonstrated that the number of hosts of networks follows the power law. People found that the size distribution usually follows the power law, such as population in cities in a country or personal income in a nation . In terms of the Internet, researchers have also discovered many power law phenomenon, such as the size distribution of web files . Recent progresses reported in further demonstrated that the size of networks follows the power law. The power law has two expression forms: the Pareto distribution and the Zipf distribution. For the same objects of the power law, we can use any one of them to represent it. However, the Zipf distributions are tidier than the expression of the Pareto distributions. In this paper, we will use Zipf distributions to represent the power law.

4 PERFORMANCE EVALUATION

In this section, we examine our theoretical analysis through two well-known large scale malware: Android malware and Conficker.

Android malware is a recent fast developing and dominant smartphone based malware [19]. Different from Android malware, the Conficker worm is an Internet based state-of-the-art botnet . Both the data sets have been widely used by the community. From the Android malware data set, we have an overview of the malware development from August 2010 to October 2011. There are 1,260 samples in total from 49 different Android malware in the data set. For a given Android malware program, it only focuses on one or a number of specific vulnerabilities. Therefore, all smartphones share these vulnerabilities form a specific network for that Android malware. In other words, there are 49 networks in the data set, and it is reasonable that the population of each network is huge. We sort the malware subclasses according to their size(number of samples in the data set), and present them in a loglog format in we can say that the Android malware distribution in terms of networks follows the power law. We now examine the growth pattern of total number of compromised hosts of Android malware against time, namely, the pattern of IðtÞ. We extract the data from the data set and present We have to note that our experiments also indicate that this data does not fit the power law (we do not show them here due to space limitation). we match a straight line to the real data through the least squares method. Based on the data, we can estimate that the number of seeds (Ið0Þ) is 10, and a ¼ 0:2349. Following our previous discussion, we infer that the propagation of Android malware was in its early stage. It is reasonable as the size of each Android vulnerable network is huge and the infection rate is quite low (the infection is basically based on contacts). We also collected a large data set of Conficker from various aspects. Due to the space limitation, we can only present a few of them here to examine our theoretical analysis. First of all, we treat AS as networks in the Internet. In general, ASs are large scale elements of the Internet.

5.Concusion

In this paper, we thoroughly explore the problem of malware distribution at large-scale networks. The solution to this problem is desperately desired by cyber defenders as the network security community does not yet have solid answers. Different from previous modelling methods, we propose a two layer epidemic model: the upper layer focuses on networks of a large scale

networks, for example, domains of the Internet; the lower layer focuses on the hosts of a given network. This two layer model improves the accuracy compared with the available single layer epidemic models in malware modelling. Moreover, the proposed two layer model offers us the distribution of malware in terms of the low layer networks. We perform a restricted analysis based on the proposed model, and obtain three conclusions: The distribution for a given malware in terms of networks follows exponential distribution, power law distribution with a short exponential tail, and power law distribution, at its early, late, and final stage, respectively. In order to examine our theoretical findings, we have conducted extensive experiments based on two real-world large-scale malware, and the results confirm our theoretical claims.

6.References

[1]B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna, "Your botnet is my botnet: Analysis of a botnet takeover," in Proc. ACM Conf. Comput. Commun. Security, 2009, pp. 635–647.

[2] M. A. Rajab, J. Zarfoss, F. Monrose, and A. Terzis, "My botnet is bigger than yours (maybe, better than yours): Why size estimates remain challenging," in Proc. 1st Conf. 1st Workshop Hot Topics Understanding Botnets, 2007, p. 5.

[3] D. Dagon, C. Zou, and W. Lee, "Modeling botnet propagation using time zones," in Proc. 13th Netw. Distrib. Syst. Security Symp., 2006.

[4] P. V. Mieghem, J. Omic, and R. Kooij, "Virus spread in networks," IEEE/ACM Trans. Netw., vol. 17, no. 1, pp. 1–14, Feb. 2009.

[5] Cabir. (2014). [Online]. Available: http://www.f-secure.com/en/web/labs_global/2004-threat-summary

[6] Ikee. (2014). [Online]. Available: http://www.f-secure.com/vdescs/worm_iphoneos_ikee_b.shtml

[7] Brador. (2014). [Online]. Available: http://www.f-secure.com/vdescs/brador.shtml

[8] S. Peng, S. Yu, and A. Yang, "Smartphone malware and its propagation modeling: A survey," IEEE Commun.

Surveys Tuts., vol. 16, no. 2, pp. 925–941, 2014.

[9] Z. Chen and C. Ji, "An information-theoretic view of networkaware malware attacks," IEEE Trans. Inf. Forensics Security, vol. 4,

no. 3, pp. 530–541, Sep. 2009.

[10] A. M. Jeffrey, X. Xia, and I. K. Craig, "When to initiate HIV therapy: A control theoretic approach," IEEE Trans. Biomed. Eng., vol. 50, no. 11, pp. 1213–1220, Nov. 2003.

[11] R. Dantu, J. W. Cangussu, and S. Patwardhan, "Fast worm containment using feedback control," IEEE Trans. Dependable Secure Comput., vol. 4, no. 2, pp. 119–136, Apr.–Jun. 2007.

[12] S. H. Sellke, N.B. Shroff, and S. Bagchi, "Modeling and automated containment of worms,"

[13] P. De, Y. Liu, and S. K. Das, "An epidemic theoretic framework for vulnerability analysis of broadcast protocols

in wireless sensor networks," IEEE Trans. Mobile Comput., vol. 8, no. 3, pp. 413–425, Mar. 2009.

[14] G. Yan and S. Eidenbenz, "Modeling propagation dynamics of bluetooth worms (extended version)," IEEE Trans. Mobile Comput., vol. 8, no. 3, pp. 353–368, Mar. 2009.

[15] C. C. Zou, W. Gong, D. Towsley, and L. Gao, "The monitoring and early detection of internet worms," IEEE/ACM Trans. Netw., vol. 13, no. 5, pp. 961–974, Oct. 2005.

[16] C. Gao and J. Liu, "Modeling and restraining mobile virus propagation," IEEE Trans. Mobile Comput., vol. 12, no. 3, pp. 529–541, Mar. 2013.

[17] D. J. Daley and J. Gani, Epidemic Modelling: An Introduction. Cambridge, U.K. Cambridge Univ. Press, 1999.

[18] W. Willinger, D. Alderson, and J. C. Doyle, "Mathematics and the internet: A source of enormous confusion and great potential," Notices Amer. Math. Soc., vol. 56, no. 5, pp. 586–599, 2009.

[19] Y. Zhou and X. Jiang, "Dissecting android malware: Characterization and evolution," in Proc. IEEE Symp. Security Privacy, 2012, pp. 95–109.