ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Two Layer Epidemic Model for Malware Detection in Large Scale Network

K.Varsha
PG Scholar,
Sree Nidhi Institute of Science and Technology,
Ghatkesar,TS, India.

R.Soujanya
Assistant Professor,
Sree Nidhi Institute of Science and Technology,
Ghatkesar,TS, India.

Abstract:

Malware is a malicious program of software deployed by cyber attackers to compromise computer. The solution to this problem is mainly desired by cyber attackers as the network security community does not yet have solid answers. The main subject of the project is to investigate how malware propagates in networks from a global perspective. It proposes a two layer propagation model of malware and describes the development of a given malware model at the Internet level. Compared with the other existing single layer epidemic models, the model proposed represents malware propagation better in large-scale networks. Malware exploits security defects in the design of operating system, in applications such as browsers. Various factors make a system more vulnerable to malware. Hence, homogeneity can be a vulnerability.

Keywords:

Malware, Malware Propagation, Two Layers, Power Law, Supervised Classification.

1. INTRODUCTION:

A network is a classification of two or more computer systems linked together. There are many different types of con mputer networks. A computer network is a telecommunication network which allows computers to exchange data. In computer networks, networked computing devices send data to each other along data connections using a data link(network links). Data is transferred in the form of tiny small packets. The connections between nodes are established using either cable media or wireless media. The best-known common and most essential computer network is the Internet.

Network computer devices which mainly originate route and terminate the data are called network nodes or devices. Nodes can also include peer to peer connections such as personal computers, mobile phones, servers or hosts as well as networking hardware. Mainly, two such devices are said to be connected together when one device is able to exchange or share information with the other device, whether they have a direct connection or not to each other. Computer networks vary in the transmission media to carry their signals, the communications protocols arrange systematically the network traffic, the network's size, topology and organizational intent.

In most cases, communications protocols are organized on (i.e. work using) various other more commonly specific or more generally used communication protocols, except for the transmission media that directly deals with the physical layer. Computer networks support applications such as access to the World Wide Web, which has shared use of application and storage servers, printers, and fax machines, and use of email and instant messaging applications**.** It suggests a two layer malware propagation model mainly to narrate the development of a given malware at the Internet level. Compared with the single layer epidemic models, the proposed model represents malware propagation better in large-scale networks. The spread of malware may be in terms of networks (e.g., autonomous systems, ISP domains, which share the same vulnerabilities) at large scale. In this paper, SI model is preferred, which is the simplest for epidemic analysis. This paper proposes a two layer epidemic model technique over the existing single layer epidemic model technique.

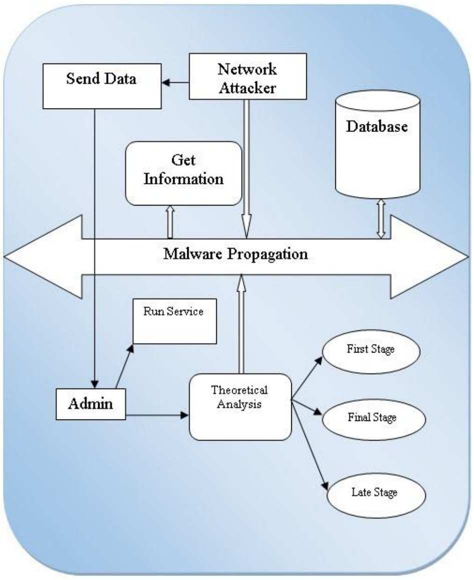
ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Two layer epidemic model: the upper layer focuses on networks of a large scale networks, for example, domains of the Internet; the lower layer particularly focuses on the peer to peer connections/ hosts of a given network. This also finds the distribution of given malware in terms of networks varies from various mathematical expressions such as exponential to power law with a short exponential tail, and to power law distribution at its early, late, and final stage, respectively.

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 various compromised computers/hosts on the Internet using different network virus-like techniques. All of his bots form a botnet, which is mainly controlled by its owners to commit illegal tasks, such as launching DDoS attacks, quick delivery of spam emails too easily, performing phishing activities, and collecting sensitive information by lawful means. There is a command and control (C&C) server(s) to interact with the bots and collect data from bots. In order to impersonate himself from legal forces, the botmaster mostly changes the url of C&C frequently, e.g., weekly. An excellent explanation about malware data can be found in [1]. Technologically, with the significant growing of most smartphones, it also has witnessed an increasing number of mobile malware.

Malware writers has developed many mobile malwares in recent years. Cabir [5] also has developed in 2004, and was the first malware targeting on the various Symbian operating system for various mobile devices. Moreover, it was also the first malware which propagates or spreads via Bluetooth. Ikee [6] was the first malware spread against popular Apple iPhones, while Brador [7] has developed against Windows CE operating systems. The attack victors for mobile malware are very different and shows a great deal of variety, such as SMS, MMS, Bluetooth, WiFi, and Web browsing. Peng et al. [8] gave the short history of mobile malware in a clear approach since 2004, and surveyed propagation models more effectively. A most direct method to count about the number of bots is to use botnet infiltration to count number of the bot IDs or IP addresses. Stone- Gross et al.[1] also registered the URL of the Torpig botnet before the botmaster was introduced, and therefore were able to hijack the C&C server for ten days, and also collected about 70G data from the bots of the Torpig botnet. This was reported that the footprint of the Torpig botnet was 182,800, and the average size of the Torpig's live population was 49,272 and 48,532, respectively. They found 49,294 new different infections from the mobile malware during the ten days takeover. Even the research also indicated that the live population fluctuates periodically as the users switch between online and offline being. This issue was also tracked by Dagon et al. in [3].



Another different method is to use DNS redirection. Dagon et al. [3] analyzed and captured bots by honeypot, and then identified the C&C (command and control) server which uses source code reverse engineering tools. And then, manipulated the DNS entry which is more related to botnet's IRC server, and has redirected the DNS requests to a local sink hole. So that they could count the number of bots in the botnet. As discussed below, this method counts the footprint of the botnet, which was 350,000 given in the report.

In this paper, it mostly uses two large scale malware data sets for experiments. Conficker is a most well-known and one of the most recently used widespread malware. Shin et al. [20] also collected a data set about 25 million Conficker also victims malware from all over the world at different levels. Simultaneously, malware targeting on Android based mobile systems are developing quickly in recent years. Zhou and Jiang [19] has collected a large amount of data set for Android based malware. In [2], Rajab et al. pointed out that it is not accurate to count the number of unique IP addresses of bots because DHCP and NAT techniques are employed more extensively on the Internet ([1] confirms it by the observation that 78.9 percent of the infected machines were behind a NAT, VPN, proxy, or firewall). Therefore, it also proposed and examined the hits of DNS caches to find the lower bound of the size of a given botnet.

3. DETERMINISTIC EPIDEMIC MODELLING:

After nearly 100 years of growth in this development, the epidemic models [17] have proved effectively and appropriately for a system that possesses a wide range of vulnerable hosts. In other different words, they are suitable at a macro level. Zou et al. [15] demonstrates by giving evidence that they were suitable for the studies of Internet based virus propagation at the early stage. It also records that there are many different factors that impact the given malware propagation of large scale networks or botnet membership recruitment, such as network topology, recruitment frequency, and connection status of vulnerable end to end hosts. All these type of factors mainly contribute to the speed of malware propagation. Unexpectedly, it also includes all these factors into one parameter as infection rate b in epidemic theory. Therefore, in this study, let N be the total number of vulnerable hosts in a large-scale network (e.g., the Internet) for a given malware propagation and detection.

$$\frac{dI(t)}{dt} = \beta(t)[N - R(t) - I(t) - Q(t)]I(t) - \frac{dR(t)}{dt} \rightarrow (1)$$

where R(t), and Q(t) represent the number of removed hosts from the infected population, and also various number of removed hosts at time t which are from the susceptible population. The variable β(t) is the infection rate at time t. In this study, model (1) is much 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.

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

where the infection rate β 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 I(t) at discrete time points. Therefore, t = 0,1,2,.... It can interpret at each 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.

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

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

$$I(t) = (1 + \alpha)I(t-1) - \beta\,(I(t-1))^2 \rightarrow (4)$$

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

$$\Delta I(t) \cong I(t) - I(t-1), t = 1,2,... \rightarrow (5)$$

To date, many researches are confined to the “early stage” of an epidemic, such as [16].

Under the early stage condition, I(t) <<N, therefore, N −I(t) ≈ N. As a result, a closed form solution is obtained as follows.

$$I(t) = I(0)e^{\beta Nt} \rightarrow (6)$$

When we take the ln operation on both sides of equation (6), we have

$$\ln I(t) = \beta Nt + \ln I(0) \rightarrow (7)$$

For a given vulnerable network, β, N and I(0) are constants, therefore, the graphical representation of equation (7) is a clear straight line. Based on the definition of equation (5) given, we obtain the increase of new members of a malware at the early stage as

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

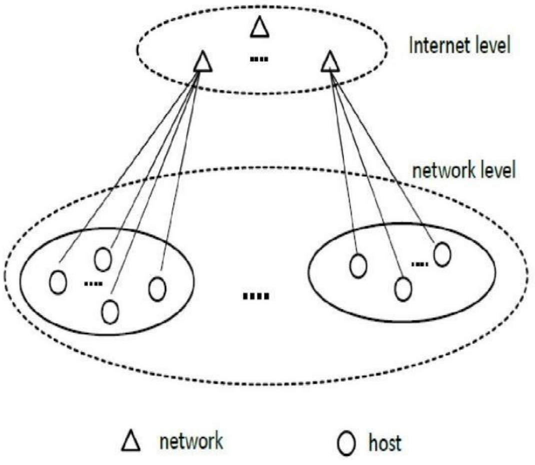
Taking the ln operation on both side of (8), we have

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

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.

3.1 Complex Networks:

Researches on various complex networks have been demonstrated that the number of hosts of networks present will follow the power law method. Mostly, people found that the size distribution usually follows the power law method, such as high population in the cities in a country or personal income from a nation [24]. In terms of the Internet, researchers have also implemented many power law phenomenon techniques, such as the size distribution of web files [25]. Recent progresses reported in [26] have further demonstrated that the size of networks follows the power law phenomenon.



3.2 Problem Description:

This studies malware propagation and detection issue at two levels, the Internet level and the Network level. Note that at the Network level, a network could be defined in many different ways, it could be an ISP domain, a country network, the group of specific mobile device applications, and so on. At the Internet level, it also treats every network of the network level as one element. At the Internet level, for suppose, there are M networks, each network is denoted as $L_i(1 \leq i \leq M)$. For any network Li, for suppose it physically possesses Ni hosts. Moreover, if suppose the possibility of vulnerable hosts of Li is denoted as pi(0 _ pi _ 1).

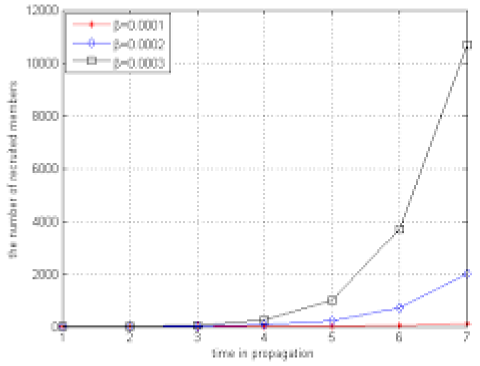
3.3 Malware Propagation Modelling:

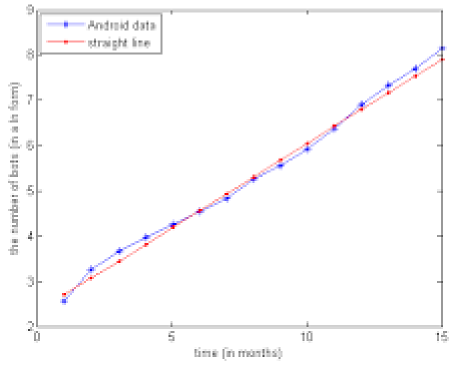
M networks of the Internet into M basic elements in our model. As a result, any two large networks, Li and Lj (i 6= j), are similar to each other at this level. Therefore, we can model the studied problem as a homogeneous system. Namely, all the M networks share the same vulnerability probability (denoted as p), and the same infection rate (denoted as _). A simple way to obtain these two parameters is to use the means.

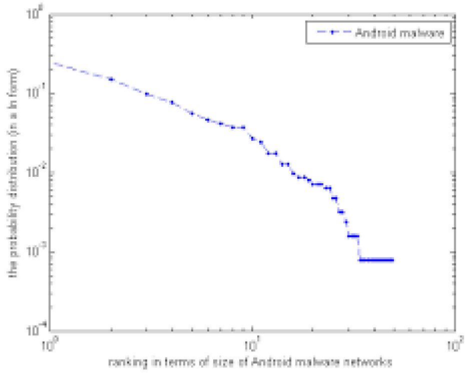
4. PERFORMANCE EVALUATION:

In this context, it examines all theoretical analysis through two well-known large-scale malware: Android malware and Conficker.

Android malware is a most popular and recent fast developing and dominant smart phone based malware [19]. Its significantly different from Android malware, Conficker worm is a well known Internet based state-of-the-art botnet [20]. Both the data sets have been widely used by the community. From the most popular Android malware data set, it has a brief overview of the malware development from August 2010 to October 2011. There are mostly 1,260 samples in total from 49 different Android malware data sets in common. For a given Android malware program, mainly it focuses on one or a various number of specific vulnerabilities. Therefore, all smartphones share these vulnerabilities and form a specified network for that Android malware. In other words, mostly there are 49 networks in the complete data set, and it is reasonable that the population of each network is too huge.







5.CONCLUSION:

The solution for problem is desperately desired by cyber defenders as the network security community does not yet have solid answers. It is different from previous modeling methods, as such as two layer epidemic model has deviced. It is upper layer focuses on networks of a large scale networks, the lower layer focuses on the hosts of a given network. This two layer epidemic model improves the accuracy compared with the available single layer epidemic models in malware modeling. Moreover, the proposed two layer model offers us the distribution of malware in terms of the low layer networks.

Finally, distribution for a given malware detection evaluates in terms of network follows exponential distribution, power law distribution method with a short mathematical exponential tail, and power law distribution, at its early, late, and final stage, respectively. In order to examine the theoretical findings, they have conducted many mathematical extensive experiments based on two real-world malware detection in large scale networks, and the results confirm our theoretical claims. In regards to future work, firstly further investigate the dynamics of the late stage. More details of the findings are expected to be further studied, such as the length of the exponential tail of a power law distribution at the late stage. Secondly, defenders may care more about their own network, the distribution of a given malware at ISP domains, where the conditions for the two layer epidemic model may not hold.

Need to seek appropriate models to address this problem. Finally, to studying the distribution of multiple malware on large-scale networks such as only focus on one malware in this paper. It is not a simple linear relationship in the multiple malware case compared to the single malware.

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 network aware 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,” IEEE Trans. Dependable Secure Comput., vol. 5, no. 2, pp. 71–86, Apr.– Jun. 2008.