Investigation of State Division in Botnet Detection Model

Wei WAN*,**, Jun LI*,**

*University of Chinese Academy of Sciences, Beijing, China

**Computer Network Information Center of Chinese Academy of Sciences, Beijing, China

wanwei@cstnet.cn, jlee@cstnet.cn

Abstract— **Botnet as a new technology of attacks is a serious threat to Internet security. With the rapid development of the botnet, botnet based several protocols came into being. In accordance with the feature of botnet, the Hidden Markov Model has application in botnet detection. Firstly, according to the situation and problems of the botnet recently, the life cycle and behaviour characteristics of the botnet have been analysed. After that a mathematical model based on state division has been built to describe the botnet. Meanwhile, a method of botnet detection based on this model has been proposed. Finally, we analyzed and summarized the experimental results, and verified the reliability and rationality of the detection method.**

Keywords— **Botnet, Hidden Markov Model, State Division**

I. Introduction

A new kind of attack technologies, botnet, is a rise trend posted on Internet security. Statistics of sampling in year 2012 from CNCERT/CC shows that, there are more than 1,460,000 victims controlled by botnets in Chinese mainland and more than 3,800,000 victims out of Chinese mainland. Meanwhile, in all of the botnets that have been detected, the scale of 100 to 1000 accounted for 79.2% or more in year 2012. In addition, there are 98 botnets contained more than 100,000 victims as in [1]. These botnets whose control servers are almost abroad threatens China's public network security seriously.

Botnets use one or more methods to spread bots in order to infect numerous hosts. So that, a one-to-many control network is generated between attacker and victims. With the rapid spreading of bots, a lot of new victims are joining the botnets continuously because of the distributive botnets. Attackers distribute all kinds of commands to victims via botnets to launch attacks. The command and control protocols are so diverse that the botnets are commonly separated into three broad categories: botnet based on IRC protocol (IRC-based botnet), botnet based on HTTP protocol (HTTP-based botnet) and botnet based on P2P protocol (P2P-based botnet) as in [2].

II. Analysis of Botnet Characteristics

Both botnet based on IRC protocol and botnet based on HTTP protocol belong to botnet based on concentrated C/S framework. Their command and control mechanisms are efficient in execution. However, as a concentrated communication centre, once the control server is controlled or destroyed by others, the whole botnet will be collapsed. With the development of P2P technology, the P2P-based botnet comes into being. This kind of botnet is different from that based on concentrated C/S framework. For instance, P2P-based botnet has no centre control server. On the other hand, every victim in P2P-based botnet can spread and receive commands and controls independently. So that, there will be little effect to P2P-based botnet if some of the victims are detected or controlled.

A. Lifecycle of a Botnet

The lifecycle of a typical botnet can be separated into 5 phases: phase of spread, phase of infection, phase of control, phase of attack and phase of destruction as in [3].

*1) **Phase of Spread:*** An attacker of botnet must possess a load of victims to attain some goals to attack. The larger the scale of the botnet is, the more obvious attack effect performs. Therefore, attackers should make numerous bots and spread them to victims who once infected will spread bots automatically, so that they can achieve the goal of expanding the botnets.

*2) **Phase of Infection:*** Once a victim has downloaded a bot, it will execute the bot automatically to infect the system. Meanwhile, the infected victim will hide itself and join the control and command channel.

*3) **Phase of Control:*** A victim will acquire the commands of attacker from control and command channel. In an IRC-based botnet, the victims will translate and execute the messages as commands which have been received as normal users via the IRC chat channels from the attackers of a botnet. In an HTTP-based botnet, the victims will execute the scripts in a website which contains some malicious commands set up by attackers. Regarding the P2P-based botnet, there are two methods to acquire commands, active acquirement and passive acquirement. The former means that the victims acquire commands at a particular place regularly. And the latter means that the victims spread the commands with P2P protocols to make many more victims to get the commands, after they received commands from attackers.

4) Phase of Attack: In this phase, attackers use botnets to initiate a DDoS attack to a target website, or to send numerous spams, or even to make phishing attacks by sending fake emails.

5) ***Phase of Destruction:*** For the purpose of better protection to botnets, sometimes attackers will destruct part of victims according to their states actively. In some cases, botnets will be destructed after they have been detected by some technological means.

B. Characteristics of Botnet

The IRC-based botnet have some special actions. For example, victims are always in idle state after logging in the IRC channel, and they keep connected by "Ping-Pong" command. Meanwhile, there are some certain statistical properties about the IRC data packet length in IRC-based botnet. Moreover, most of the clients in a IRC-based botnet will execute a large number of similar attack commands, and the nicknames of clients also have some similarity.

The HTTP-based botnet publish a web page as a centralized control command platform. It is covert in communication mechanism, because the bots in HTTP-based botnet visit the centralized control websites by normal HTTP protocol. However, there are some certain characteristics in the websites which publish centralized control commands. For example, the centralized control web pages contain some abnormal strings or suspicious scripts and so on. In addition, there are some similarities in communication messages and communication frequencies between bots in HTTP-based botnet and the centralized control website.

The P2P-based botnet applies P2P protocols which manifested in spreading bots with P2P protocols to expand the botnet scale and in building control and command channel with P2P protocols as in [4]. Therefore, some characteristic of P2P networks besides those of some typical botnets are considered as the characteristic of P2P-based botnet. For example, the nodes of the P2P-based botnet may join or quit frequently, and there would be little effect to the whole P2P-based botnet if some nodes were close. In a P2P-based botnet, attackers send attack commands to victims at a certain hour, which makes many victims have very similar network behaviours at the certain time, such as probing, spamming, downloading some executive files, DDoS attacks and so on. Moreover, a P2P-based botnet may have some other characteristic as a P2P network such as sending SYN packets and ARP request frequently and attempting to connect to some IP addresses periodically as in [5].

III. Modelling of Botnet

By analysing the botnets in depth, a modelling based on state division with hidden Markov model will be plan to investigate the detection of botnets in this paper.

A. Hidden Markov Model

A hidden Markov model is one of the Markov chains. In a hidden Markov model, the state is not directly visible, but it is visible by observing vector sequence. Each observing vector expresses as various states by several probability distributions. Therefore each observing vector is generated by a state sequence which has a probability distribution as in [6]. So the hidden Markov model is a double stochastic process. One of it is a finite state Markov chain which can describe the transfer of the states. And another one is used to describe the statistics relationship between states and observed values. A hidden Markov model can be defined as a quintuple:

$$\lambda = (X, O, A, B, \Pi)$$

$$X = \{S_1, S_2, \cdots, S_N\}$$

$$O = \{V_1, V_2, \cdots, V_M\}$$

$$A = \left\{a_{ij} = P\left(q_{t+1} = S_j \middle| q_t = S_i\right)\right\}, i, j \in [1, N]$$

$$B = \left\{b_j(k) = P\left(O_t = V_k \middle| q_t = S_j\right)\right\}, j \in [1, N], k \in [1, M]$$

$$\Pi = \left\{\pi_i = P(q_1 = S_i)\right\}, i \in [1, N]$$

In this quintuple, X means a state set which have N states. All of these states which meet Markov property are implied the performance of states in hidden Markov model. Here q_t means the certain state at t time, $q_t \in X$. O means the set of observed values. M means the number of the different observed values which output from each state. A is a state transition probability matrix, which means the probability of S_i at t time transferring to S_j at $t+1$ time. B is a probability distribution matrix, which means the conditional probability of the observed value of S_j as V_k at t time. Π is a initialization state probability distribution matrix, which means the conditional probability of the state as S_i at the initialization time. Generally speaking, a hidden Markov model can be simplified as a triad as follows:

$$\lambda = (A, B, \Pi)$$

B. Model Building Based on State Division

According to the description of the lifecycle of the botnet, it can be divided into 4 states with the finite state machine as shown in Figure 1.

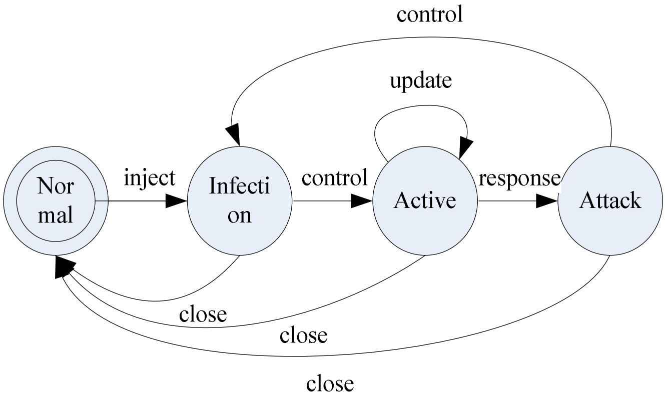

Figure 1. A finite state machine of a botnet

The finite state machine of a botnet can be described as follows:

$$\mathrm{M} = \left(Q, \Sigma, \delta, q_0, F\right)$$

$$Q = \left\{Normal, Infection, Active, Attack\right\}$$

$$\Sigma = \left\{inject, control, update, response, close\right\}$$

$$\delta\left(Normal, inject\right) = Infection$$

$$\delta\left(Infection, control\right) = Active$$

$$\delta\left(Active, update\right) = Active$$

$$\delta\left(Active, response\right) = Attack$$

$$\delta\left(Active, control\right) = Infection$$

$$\delta\left(Infection\middle| Active\middle| Attack, close\right) = Normal$$

$$q_0 = Normal$$

$$F = \left\{Attack\right\}$$

In this finite state machine, Q is the finite state set. Σ is the finite input set, and at any certain moment, the finite state machine can only be accepted one certain input. δ is the state transition function. q_0 is the initial state, $q_0 \in Q$. F is the accepted state set, $F \subseteq Q$.

Normal State: It means the state when the host has never been infected by bot program or the bot program on the host has been closed or cleared already. The state is the normal state of the host.

Infection State: It means the state when the host has been embedded bot program. In this state, the host may perform several abnormal behaviors, such as modifying the registry, opening certain ports, accessing certain IP addresses or URLs, probing certain ports, establishing lots of connections, sending lots of ARP requests, playing some abnormal P2P feature and so on.

Active State: It means the state when the host infected by bot program is communicating with other victims. In this state, the host may perform several abnormal behaviors, such as sending update packets periodically, accessing certain IP addresses or URLs periodically, listening to the certain ports, exploiting other's vulnerabilities, playing some abnormal P2P feature and so on.

Attack State: It means the state when the victims attack after they have received commands, such as spamming, DDoS attacking and so on.

According to the analysis above, the state transition of the botnet is also a stochastic process. The behavior feature or the host attribute of the botnet is visible for observers, but the state of the botnet is not directly visible. On the other hand, the state transition of the botnet fays in with the Markov rules. Consequently, the characteristics of the botnet can be modeled by the hidden Markov model.

With a view to the mathematics description of the hidden Markov model and the definition of the states of the botnet, the four states of the botnet can be considered as the state set X of the hidden Markov chain, and the feature of behaviours, characteristics of network traffics and attributes of the botnet can be considered as the observing value set O. Each victim in a same botnet should perform a similar hidden Markov chain process, so the parameters of the hidden Markov model which make the likelihood function of the training model to maximize can be found by Baum-Welch algorithm iteration on the training set of the botnet as in [7]. Several hidden Markov models can be built to a certain botnet by repeating this process with clustering algorithm. Then these hidden Markov models can be merged into a final hybrid model. After that we find the most probable state sequence and build an optimizing model to a certain observing value set O and hidden Markov model λ. It can be described as:

$$class\left(O\right) = \arg\max_c P_{viterbi}\left(O, \lambda_c\right)$$

C. Detection Method

According to the modelling and state division of the botnets, the observing value set O of the detection objects needs to be acquired, namely, abnormal characteristics of basic attribute of hosts, network traffics, network behaviours and so on. Then a host who has been detected will be decided if it belongs to a botnet by analysing the similarity between host state transition models and botnet models which have been trained. And a system construction drawing is shown in Figure 2.

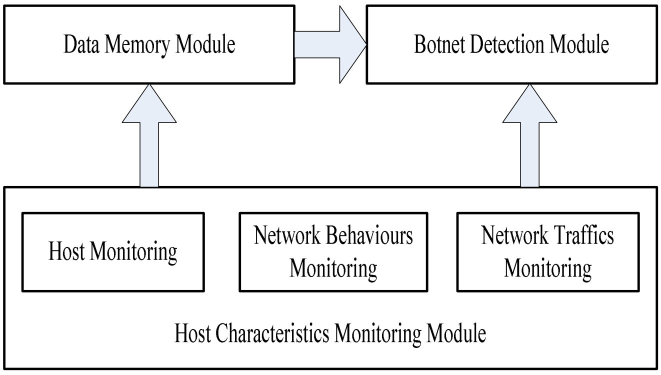
Data Memory Module
Botnet Detection Module
Host Monitoring
Network Behaviours Monitoring
Network Traffics Monitoring
Host Characteristics Monitoring Module

Figure 2. A system construction drawing

In principle, the system is divided into 3 parts, namely, host characteristics monitoring module, data memory module and botnet detection module.

Host characteristics monitoring module contains 3 sub-modules as host monitoring, network behaviours monitoring and network traffics monitoring. Host monitoring sub-module needs to be deployed on the host, responsible for recording system critical files changes, registry changes, critical processes changes and port status. And abnormal information should be picked up as abnormal characteristics of host basic attributes in this sub-module. Network behaviours monitoring sub-module detects the abnormal network behaviours based on snort, such as spamming and DDoS attacks. Network traffics monitoring sub-module analyses the abnormal traffic characteristics by comparing the historical data with the traffic samples based on netflow technology.

Data memory module stores the abnormal characteristic information generated by host characteristics monitoring module.

Botnet detection module decides if a host belongs to a certain botnet by analysing the real-time abnormal characteristic information generated by host characteristics monitoring module and the historical data stored in data memory module.

IV. EXPERIMENTS RESULTS

In the experiments, each 50 kinds of IRC-based botnet programs, HTTP-based botnet programs, P2P-based botnet programs, other malicious programs, normal programs based on IRC protocol, normal programs based on HTTP protocol, normal programs based on P2P protocol and other normal programs have been chosen to run on the testing hosts in order to analyse the recognition rate. The results of the experiments are shown in Table 1 and Table 2.

TABLE 1. RESULTS OF BOTNETS DETECTION

Type of Botnet Programs	Sample Size	Hit Result
IRC-based Botnet Programs	50	48
HTTP-based Botnet Programs	50	38
P2P-based Botnet Programs	50	47

TABLE 2. RESULTS OF OTHER PROGRAMS DETECTION

Type of Programs	Sample Size	Hit Result
Other Malicious Programs	50	4
Normal Programs Based on IRC Protocol	50	2
Normal Programs Based on HTTP Protocol	50	1
Normal Programs Based on P2P Protocol	50	4
Other Normal Programs	50	2

According to the results of the experiments, with this botnet detection method, the false negative rate is 4% of IRC-based botnet detection, 24% of HTTP-based botnet and 6% of P2P-based botnet, and the false alarm rate is 5.2% of non-botnets.

Analytically, it is ineffective in HTTP-based botnet detection, because it is very similar of the actions between HTTP-based botnet and normal web page visit, and the communication data are usually encrypted of most HTTP-based botnet. On the other hand, the false negative rate of this system can be reduced effectively by increasing in the number of training set. Meanwhile, the false alarm rate caused by non-botnet programs which contain with observing characteristic values can be reduced by adjusting the parameters of the model.

V. CONCLUSIONS

A botnet detection method is proposed in this article. In this method, the suspicious hosts are detected by modelling to botnets based on state division with the hidden Markov model. According to the results of the experiments, it is indicated that the IRC-based botnets and P2P-based botnets can be detected effectively by this method, but it is ineffective on HTTP-based botnets. In the future, the model will be optimized by researching the botnets persistently in order to increase the efficiency of the detection.

REFERENCES

[1] CNCERT/CC, *China Internet Security Report 2012,* POST & TELECOM PRESS. Beijing, China: 2013.

[2] ZHUGE Jianwei, HAN Xinhui, Zhou Yonglin, et al, “Research and Development of Botnets”, *Journal of Software,* Vol.19, No.3, pp.702-715, 2008.

[3] Lee WK, Wang C, Dagon D, et al, *Botnet Detection: Countering the Largest Security Threat*, New York: Springer-Verlag, 2007.

[4] ZHANG Chen, WANG Liang, XIONG Wenzhu, “Technologies of P2P Botnet detection”, *Journal of Computer Applications,* Vol.30, pp.117-120, 2010.

[5] ZHOU Jipeng, ZHU Liangyuan, “P2P System Model with Topology of Physical Network”, *Microelectronics & Computer,* Vol.23, No.10, pp.65-67, 2006.

[6] SUN Yongqiang, XU Xin, HUANG Zunguo, “A DDoS Attack Detection Method Based on Hidden Markov Model”, *Microelectronics & Computer,* Vol.10, No.23, pp.176-186, 2006.

[7] LIN Guoyuan, GUO Shanqing, HUANG Hao, CAO Tianjie, “An Anomaly Detection Model Based on Dynamic Behavior and Character Patterns”, *Chinese Journal of Computers,* Vol.29, No.9, pp.1553-1560, 2006.

Wei WAN received the B.S. degrees from Beijing Information Technology Institute in 2004, and M.S. degrees from Graduate University of Chinese Academy of Sciences. Now he is currently working toward Ph.D. degree at University of Chinese Academy of Sciences. He is working at Computer Network Information Center of Chinese Academy of Sciences. His research interests include network security.

Jun LI is Ph.D, research fellow-professor, Ph.D tutor, Vice Chief Engineer of Computer Network Information Center of Chinese Academy of Sciences. His research interests include network security, network architecture.