Detecting P2P Botnet by Analyzing Macroscopic Characteristics with Fractal and Information Fusion

SONG Yuanzhang

Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, Jilin Province, P. R. China

Abstract: Towards the problems of existing detection methods, a novel real-time detection method (DMFIF) based on fractal and information fusion is proposed. It focuses on the intrinsic macroscopic characteristics of network, which reflect not the "unique" abnormalities of P2P botnets but the "common" abnormalities of them. It regards network traffic as the signal, and synthetically considers the macroscopic characteristics of network under different time scales with the fractal theory, including the self-similarity and the local singularity, which don't vary with the topology structures, the protocols and the attack types of P2P botnet. At first detect traffic abnormalities of the above characteristics with the nonparametric CUSUM algorithm, and achieve the final result by fusing the above detection results with the Dempster-Shafer evidence theory. Moreover, the side effect on detecting P2P botnet which web applications generated is considered. The experiments show that DMFIF can detect P2P botnet with a higher degree of precision.

Keywords: P2P botnet; fractal; information fusion; CUSUM algorithm

I. INTRODUCTION

The botnet, mainly formed by compromised machines which are infected by various ways such as worms and Trojans, is a type of malicious host group. The botmaster can easily change the load of the bot nodes with the secondary injection, so as to easily launch various malicious activities, such as DDoS, spamming and so on. The new P2P botnet achieves its Command and Control (C&C) mechanism by adopting the decentralized structure of P2P network, and there is no control center in the decentralized structure, and every peer serves as both client and server, which made it influence less by single-point (server) failure. The remaining bot nodes are still able to launch attacks effectively with the help of P2P network, while a certain number of bot nodes are excluded. Storm is a representative type of P2P botnet, which uses the Overnet/eDonkey network to build and maintain the C&C mechanism [1].The decentralized structure of botnet is the evolution trend of botnet, and the detection method of it has become the burning focus of the network security research.

Based on the analysis of the life cycle process and the characteristics of Storm, a novel real-time detection method based on the fractal theory and the information fusion theory is proposed, which is named as DMFIF. It regards network flow as the signal, and focus on the intrinsic characteristics of network, and synthetically takes into account the macroscopic characteristics of network traffic under the different time scales with the fractal theory, including the self-similarity and the

local singularity, which don't vary with the topology structures, the protocols and the attack types of P2P botnet. Firstly, the self-similarity and the local singularity are used to describe accurately the characteristics of network traffic, and two detection results are acquired by using the nonparametric CUSUM algorithm to detect the traffic abnormalities of the above characteristics, and the final detection result is acquired by fusing the above two detection results with the Dempster-Shafer evidence theory. Moreover, DMFIF uses the characteristics of the TCP flow to weaken the side effect on detecting P2P botnet which web application programs generated in some degree, especially the application programs based on P2P protocols. The experiments show that the method was able to detect the new P2P botnet with a higher degree of precision.

II. Related Work

Currently, the studies related to the detection and analysis of P2P botnet are as follows.

Sarat *et al.*[2] analyzed the life cycle process of Storm and the abnormal characteristics, such as that the distribution of peer IDs of Storm is irregular and there are some unreachable IPs. And it serves as a solid foundation for further studies.

Steggink *et al.*[3] analyzed the mechanism that how the botnets evade the detection methods, and found some unique characteristics of Storm, and proposed a method based on the characteristics of the network flow, such as the specific length of the IP packet.

Porras *et al.*[4] proposed a penetrating analysis on the logic of Storm, and proposed a dialog-based detecting method, which deals with the dialog in the pattern matching theory to detect P2P botnet.

Holz *et al.*[5] found a method to mitigate the botnet, which is to infiltrate into the Overnet network as a peer and publish some fake keys to reroute or disturb the communication traffic between bot nodes.

Chen Wei *et al.*[6] proposed a method to detect the encrypted botnet traffic, which uses the spatial-temporal correlation in suspicious botnet traffic after a large amount of non-malicious traffic is filtered with a payload feature exaction method.

Qiao Yong *et al.*[7] proposed a P2P botnet detection model based on mining and evaluating the periodic patterns to identify the P2P bots traffic, considering that P2P bots periodically send requests to update the peer lists or receive commands from the botmaster in the process of C&C.

Zang Tianning *et al.*[8] presented an approach for analyzing the relationship among botnets. It extracts some botnet communication characteristics, including the amount of data flows, the number of packets per data flow, the payload of communication and data packets, and it deals with them with the cloud model and the statistical similarity functions. The experiments show that the presented method is valid and efficient, even in the case of encrypted botnet communication messages.

In the area of the collaborative botnet detection model, Wang Hailong *et al.*[9] proposed a hierarchical collaborative model, which shares information and cooperates in the three levels of information, characteristic, and decision-making. Zang Tianning *et al.*[10] presented a Coordinative Running Model based on Universal Turing Machine, which is able to analyze the potential relationships existing among network security incidents which occurrs at different positions and time. On the basis of the model, a Collaborative Running System is implemented, and has been used in botnet tracking, correlation analysis for alerts of DDoS and relationship analysis between DDoS source and botnet.

Zhuge Jianwei *et al.*[11,12] and Wang Tianzuo *et al.*[13] presented the evolution process, concept, functional structure and execution mechanism of botnet, analyzed the evolving of botnet's propagation, attack and the C&C mechanism, and summarized the recent advances of botnet research, including botnet monitoring, infiltration, analysis of botnet characteristics, detection, disruption and so on. At last the paper discussed the limitation

This paper proposed a novel real-time detection method (DMFIF) based on fractal and information fusion. It focuses on the intrinsic macroscopic characteristics of network, which reflect not the "unique" abnormalities of P2P botnets but the "common" abnormalities of them.

of current botnet studies, the evolving trend of botnet, and some possible directions for future research about botnet.

In conclusion, there are some problems in the analysis and detection research of P2P botnet:

(1) Current studies mainly focus on certain or several "unique" abnormalities of P2P botnet. When a new type of P2P botnet emerges, these detection methods will no longer work.

DMFIF proposed in the paper focus on the intrinsic macroscopic characteristics of network traffic, which reflect the "common" abnormalities of P2P botnet. It regards the network traffic as the signal, and synthetically takes into account the characteristics of network traffic under different time scales with the fractal theory, including the self-similarity and the local singularity, which don't vary with the topology structures, the protocols and the attack types of P2P botnet. Therefore, DMFIF will be able to detect the new emerging P2P botnet, and shows higher extensibility.

(2) The side effect on detecting P2P botnet is not considered in most studies, which is generated by the web applications, especially the P2P applications. In essence, P2P botnet is a type of P2P network used to launch malicious attacks by the botmaster, so the characteristics of them are similar, which leads to a larger false positive rate of P2P botnet detection.

DMFIF uses the characteristics of TCP flow to overcome the problem.

(3) Some detection methods require large amount of historical data, and are not suitable for real-time detection.

DMFIF mainly focuses on the macroscopic characteristics of network traffic under the different time scales, and does not require too much prior knowledge, so that it meets the requirements of the real-time detection of P2P botnet.

III. Analysis of P2P Botnet

Storm is a typical representative of P2P botnets, and many unique network characteristics of its life cycle process are worth noticing:

(1) The number of UDP packets is detected sharply increasing, because the UDP flow is used to build and maintain the Command and Control mechanism of P2P botnet, such as publishing itself, keeping alive, peer discovery and so on.

(2) The bot randomly sends requests to connect to other peers to bootstrap, which leads to a rarely seen amount of ICMP "destination unreachable" packets in regular circumstances.

(3) The number of SMTP packets is in a rising trend when botnets are spamming [5].

In conclusion, the botnet will make the number of IP packets increase, which leads to the change of the macroscopic characteristics of network traffic. DMFIF of the paper used the self-similarity and the local singularity to accurately describe the macroscopic characteristics of network under the large time scale and the small time scale, and detect P2P botnet with the change of the above characteristics.

IV. The Detection Method of P2P Botnet—DMFIF

4.1 Overview of DMFIF

The process of the detection method proposed in the paper is shown in Fig. 1. The detection method proposed in the paper focuses on the intrinsic characteristics of network, which reflects not the "unique" abnormalities of P2P botnet but the "common" abnormalities of them. It regards network traffic as the signal, and synthetically considers the characteristics of network under different time scales with the fractal theory, including the self-similarity and the local singularity, which don't vary with the topology structures, the protocols and the attack types of P2P botnet. Firstly, the self-similarity and the local singularity are used to describe accurately the characteristics of network, and two detection results are acquired by using the nonparametric CUSUM algorithm to detect the traffic abnormalities of the above characteristics, and the final detection result is acquired by fusing the above

detection results with the Dempster-Shafer evidence theory.

Moreover, DMFIF uses the characteristics of the TCP flow to identify that whether the abnormalities of network traffic are caused by P2P botnet or P2P application programs, so as to weaken the side effect on detecting P2P botnet which P2P application programs generate.

4.2 Fractal theory

Fractal refers to a rough or fragmented geometric shape which is able to split into parts, each of which is (at least approximately) a reduced-size copy of the whole shape[14]. Network traffic exhibits the characteristics of the signal, and it can be regarded as the signal. The studies showed that network traffic exhibits an inherent characteristic[15-19]—fractal, including the self-similarity (Single fractal) under the large time scale and the local singularity (Multi-fractal) under the small time scale. As mentioned in Section III, the botnet will make the number of IP packets increase, which leads to the change of the self-similarity and the local singularity of network traffic, so P2P botnet can be detected with the help of the above characteristics.

4.2.1 Self-similarity

Studies have demonstrated that the self-similarity process is able to better describe the characteristics of network traffic than the traditional short-term correlation model[20,21]. Self-similarity refers to that there is a certain degree of consistency between the local structure and the overall structure. Let $X(t)$ denote a continuous-time random process, if equation (1) is established for all $a > 0$, then $X(t)$ exhibits the self-similarity.

$$X(at) = a^H X(t) \quad \forall a > 0 \tag{1}$$

The "=" of (1) indicates that the variables on both sides of it are equal statistically, and the parameter $H(0.5 \le H < 1)$ is referred to as the Hurst exponent. It is referred to as the index of long-range dependence, and indicates a time series with long-term positive autocorrelation.

Let t denote the current time, as mentioned in Section III, the botnet will make the number of IP packets increase, which leads to weaken the self-similarity of network traffic and make $Hurst_t$ decrease, so the abnormalities can be discovered by detecting $Hurst_t$. In order to improve the sensitivity of the detection, $Hurst_t$ is input to the nonparametric CUSUM algorithm in order to rapidly detection the abnormalities of the self-similarity of network traffic.

There are many Hurst exponent estimation methods, and researchers found that R/S method is less affected by noise and other factors[22, 23]. In order to ensure real-time and further improve the estimation accuracy, R/S method was improved in the paper by bringing in the sliding window, shown in Fig. 2. Let W denote the length of the sliding window, the sliding window is moved forward by a certain step to calculate the new Hurst exponent after the current Hurst exponent is calculated by R/S method. Let $\{X_1,\ldots,X_W\}$ denote the data of

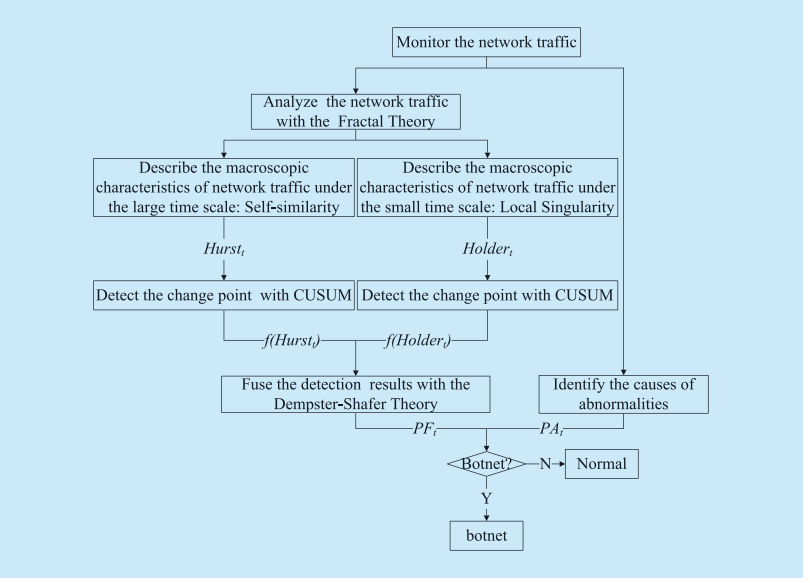

Fig.1 *The process of DMFIF*

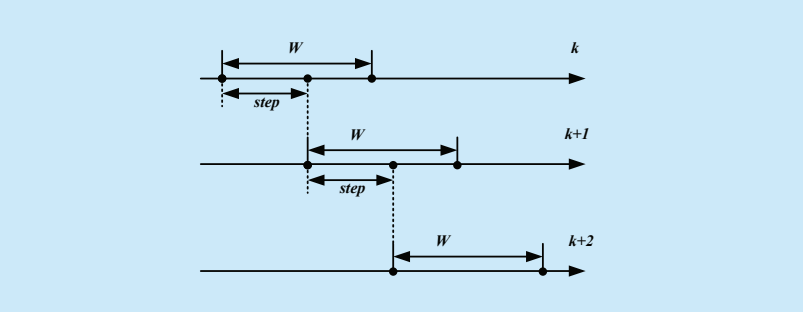

Fig.2 *The sliding window*

the sliding window, how to estimate the value of the Hurst exponent with the R/S method is described as follows.

The data of the sliding window are divided into many sub-windows, let m denote the length of the sub-window, then the number of the sub-window is $d=W/m$. For each sub-window, $n=1,\ldots, d$:

(1) Calculate the mean E_n.

$$E_n = \frac{1}{m} \times \sum_{i=(n-1)\times m+1}^{n\times m} X_i \quad (2)$$

(2) Calculate the standard deviation S_n.

$$S_n = \sqrt{\frac{1}{m} \times \sum_{i=(n-1)\times m+1}^{n\times m} (X_i - E_n)^2} \quad (3)$$

(3) Calculate the range R_n.

$$Rn = \max_{1\le i\le m}\{\sum_{j=1}^{i} (Y_{j,n} - E_n)\} - \min_{1\le i\le m}\{\sum_{j=1}^{i} (Y_{j,n} - E_n)\} \quad (4)$$

$Y_{j,n}$ represents the value of the j-th element in the n-th sub-window.

(4) Calculate the mean of the R_n/S_n, for $n=1,\ldots,d$

$$(R/S)_m = \frac{1}{d} \times \sum_{n=1}^{d} (R_n/S_n) \quad (5)$$

The relationship between $(R/S)_m$ and the length of sub-window m can be expressed as equation (6).

$$(R/S)_m = C \times m^H \quad (6)$$

C is a constant, and the value of the parameter H is referred to as the Hurst exponent. Equation (6) is transformed as follows. If the $log\ m$ is taken as the x-axis, the $log(R/S)_m$ is taken as the y-axis, and points are fitted to a straight line, the slope of the line gives the estimated value of the Hurst exponent.

$$log(R/S)_m = log\ C + H \times log\ m \quad (7)$$

4.2.2 Local singularity

For many non-uniform fractal processes, one dimension is not able to describe all the characteristics of it. After in-depth studies on TCP flow, Riedi et al.[24] found that the self-similarity is only one aspect of the fractal of network traffic under the large time scale. The network traffic exhibits self-similarity under the large time scale, but it behaves differently under the small time scale. Compared with the constant scale parameter under the large time scale, signals under the small time scale possess an irregular changing exponent, so these signals have been called multi-fractal. Generally speaking, self-similarity focus on the global characteristics of a process, and the self-similarity describes how the whole process changes from one scale to another. In other word, it characterizes the characteristic of a process under the large time scale. On the contrary, multi-fractal cares more about the local singularity of a process, which means the characteristic of a process under the small time scale. Multi-fractal is the extension and refinement of self-similarity, and is able to flexibly describe the irregular phenomenon under the local time scale that has little connection to the self-similarity of the network traffic under the large time scale.

Let $X(t)$ denote a continuous-time random process, if equation (8) is established for all $a> 0$, then $X(t)$ exhibits multi-fractal.

$$X(at) = a^{H(t)}X(t) \forall a > 0 \quad (8)$$

The "=" of (8) indicates that the variables on both sides of it are equal statistically, and the parameter $H(t)$ is referred to as the Holder exponent. The degree of local singularity of X at a given point t can be characterized by its local Holder exponent $H(t)$.

Let t denote the current time, as mentioned in Section III, the botnet will make the number of IP packets increase, which leads to strengthen the local singularity of network traffic and make $Holder_t$ decrease, so the abnormalities can be discovered by detecting $Holder_t$. In order to improve the sensitivity of the detection, $Holder_t$ is input to the nonparametric CUSUM algorithm in order to rapidly detection the abnormalities of the local singularity of network traffic.

Let $X(t)$ denote the total number of packets captured up to time t, $X(0),\ldots,X(t)$ are divided into many sub-interval, and the length of the sub-interval is d, so $Holder_t$ can be calculated as follows[25,26].

$$Holder_t = \lim_{d\to 0} \frac{\log\left(\left|X\left(t+\frac{d}{2}\right) - X\left(t-\frac{d}{2}\right)\right|\right)}{\log(d)} \quad (9)$$

Without loss of generality, if the number of the sub-interval is 2^n, $Holder_t$ can be calculated as equation (10).

$$Holder_t = \lim_{n\to\infty} \frac{log\left(\left|X\left(\frac{i+1}{2^n}\right) - X\left(\frac{i}{2^n}\right)\right|\right)}{log\left(\frac{1}{2^n}\right)} = \lim_{n\to\infty}\left[-\frac{log\left(\left|X\left(\frac{i+1}{2^n}\right) - X\left(\frac{i}{2^n}\right)\right|\right)}{n}\right] \quad i = 0, 1, ..., 2^n - 1 \tag{10}$$

4.3 Nonparametric CUSUM algorithm

Non-parametric CUSUM (Cumulative Sum) algorithm is able to detect the changes of the mean of a statistical process[27,28]. It cumulates the small offset so as to achieve the effect of amplifying the offset. And it is typically used for monitoring change detection, and it is a sequential analysis technique, and it satisfies the requirements of detecting P2P botnets.

The Hurst exponent $Hurst_t$ of the self-similarity and the Holder exponent $Holder_t$ of the local singularity are separately input into CUSUM to detect the abnormalities. Let $X=\{x_1,...,x_n\}$ denote the input of CUSU, the output of CUSUM is defined as follows.

$$f(X) = \begin{cases} 1 & X \text{ is abnormal} \\ 0 & X \text{ is not abnormal} \end{cases} \tag{11}$$

4.4 Dempster-shafer evidence theory

Since the characteristics of P2P botnet are complex and changeable, using a single network characteristic to describe the details of the network changes to detect botnet can lead to high false negative rate and false positive rate, so the information fusion method of decision level is adopted to solve the problem.

There are many information fusion methods of decision level, include the Bayesian theory and the Dempster-Shafer evidence theory[29, 30]. The Bayesian theory requires the priori probability and the conditional probability for each question of interest, and requires that all the hypotheses are mutually independent and that all the hypotheses construct a complete set. The Dempster-Shafer evidence theory is a generalization of the Bayesian theory of subjective probability and it doesn't require the priori probability and conditional probability, and it is able to reduce the hypothesis set by combining the evidences gradually.

In the paper the Dempster-Shafer evidence theory is used to fuse the above two detection results, which allows one to combine evidences from different sources and arrive at a degree of belief that takes into account all the available evidences. Often used to fuse the information of sensor, the Dempster-Shafer evidence theory is based on two ideas: obtaining degrees of belief for one question from subjective probabilities for a related question, and Dempster's rule for combining such degrees of belief when they are based on independent items of evidence.

Let U denote the universal set, which represents all possible values of a random variable X, and the elements of U are inconsistent, then U is called the discernment frame. The power set 2^U is the set of all subsets of U. In the paper, U is defined as $U=\{normal, abnormal\}$, and "*normal*" represents the captured network traffic is normal, and "*abnormal*" represents the captured network traffic exhibits abnormalities.

Let U denote the discernment frame, if a function m: $2^U \to [0,1]$ exhibits the two follow properties:

$$(1) \quad m(\phi) = 0; \tag{12}$$

$$(2) \quad \sum_{A\subset U} m(A) = 1; \tag{13}$$

then $m(A)$ is called the basic belief assignment of the set A.

Let U denote the discernment frame, and the function m: $2^U \to [0,1]$ is the basic belief assignment of U, if the function BEL: $2^U \to [0,1]$ is defined as the sum of all the masses of subsets of the set of interest,

$$BEL(A) = \sum_{B\subset A} m(B) \quad (\forall A \subset U) \tag{14}$$

then BEL is called the belief function. If $BEL(A)>0$, A is called the focus element of BEL.

The following problem is how to combine two independent sets of probability mass assignments. Different sources may express their beliefs over the frame in terms of belief

constraints, Dempster's rule of combination is the best appropriate fusion operator, which derives common shared belief between multiple sources and ignores all the conflicting belief through a normalization factor. Let BEL_1 and BEL_2 denote the two belief functions of U, the corresponding focus elements of them are $A_1,\ldots,A_k$ and $B_1,\ldots,B_k$, the corresponding basic belief assignments are m_1 and m_2, and the combination of m_1 and m_2 is calculated as follows.

$$m(C)=\begin{cases}\dfrac{\sum\limits_{\substack{i,j\\A_i\cap B_j=C}} m_1(A_i)m_2(B_j)}{1-K_1} & \forall C\subset U \quad C\neq\phi \\ 0 & C=\phi\end{cases} \tag{15}$$

where

$$K_1=\sum_{\substack{i,j\\A_i\cap B_j=\phi}} m_1(A_i)m_2(B_j)<1 \tag{16}$$

K_1 is a measure of the amount of conflict between the two mass sets. If K_1=1, m_1 and m_2 are conflict, and they are not able to be combined. If $K_1\neq1$, m is the combination of m_1 and m_2.

The Dempster's rule of combination is used to fuse the above two detection results. The former is the parameter $f(Hurst_t)$, which is acquired by the self-similarity of network and the nonparametric CUSUM algorithm. The latter is the parameter $f(Holder_t)$, which is acquired by the local singularity of network and the nonparametric CUSUM algorithm.

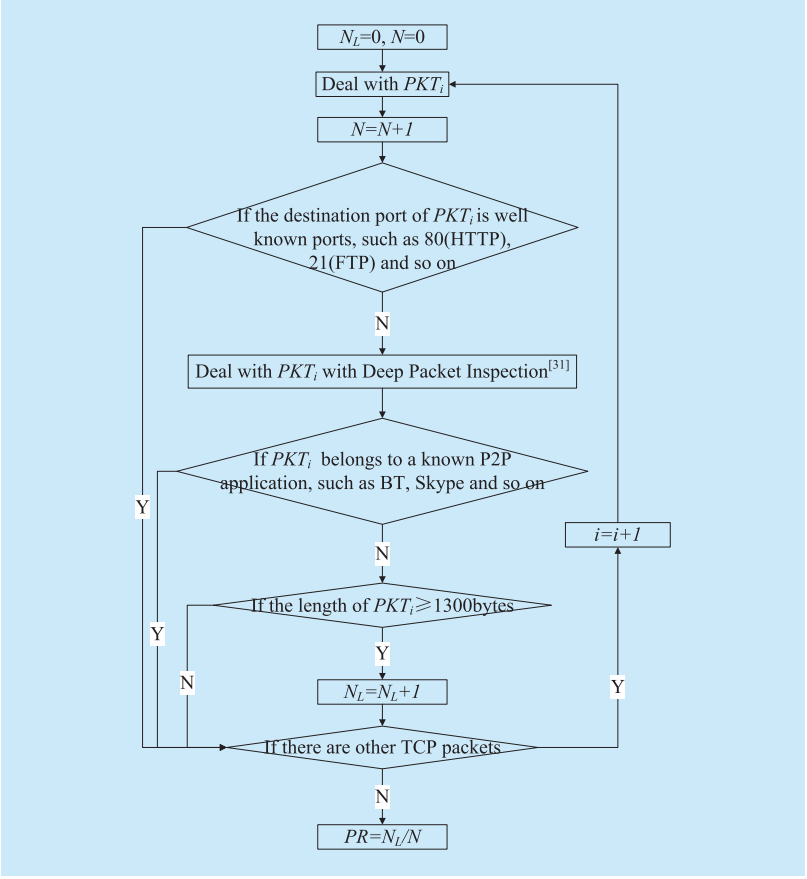

Fig.3 *The process of dealing with the TCP flow*

4.5 Identify the causes of abnormalities

In essence, P2P botnet is a type of P2P network which is used to launch malicious attacks by the botmaster, so the flow characteristics of them are very similar, which leads to a higher false positive rate of the P2P botnet detection. So the characteristics of the TCP flow are used to identify that whether the abnormalities of network traffic mentioned above are caused by the P2P botnet or the P2P application programs, so as to weaken the side effect on detecting the P2P botnet which the P2P application programs generate.

P2P application programs generally use the long TCP packets to transmit data, the length of which is longer than 1300 bytes. But the TCP packets of P2P botnet is generated by the "secondary injection" process, which is implemented in the way of transmitting or updating the load of bot nodes by the HTTP protocol and the data volume is small. So the proportions of the long TCP packets—*PR* is able to identify the causes of abnormalities of network traffic, the length of which is above 1300 bytes, and the TCP packets belonging to the known application programs must be filtered. The smaller the parameter *PR* is, the more possibly the abnormalities of network traffic mentioned above are caused by the P2P botnet. Let PKT_i denote the TCP packet to be dealt with, the total number of the TCP packets is N, the number of the long TCP packets is N_L, and the process is shown in Fig. 3.

The decision function is defined as follows.

$$PA=\begin{cases}1 & PR<T_{TCP}\\ 0 & PR\geq T_{TCP}\end{cases} \tag{17}$$

If $PR<T_{TCP}$, the abnormalities of network traffic mentioned above are more possibly

caused by P2P botnet. The threshold T_{TCP} is dynamically adjusted with the Kaufman algorithm [32].

4.6 Process of DMFIF

Let t denote the current time, the process of DMFIF is:

(1) Capture network data from monitoring device, and get the number of IP packets. Meantime, get PA_t of TCP flow to identify that whether the abnormalities of network traffic are caused by P2P botnet or P2P application programs, so as to weaken the side effect on detecting the P2P botnet which the P2P application programs generate.

(2) Detect the abnormities of traffic flow under different time scale with the fractal theory.

① Detect the abnormities of traffic flow under the large time scale

Calculate the Hurst exponent with the method based on the sliding window, and then get the parameter $Hust_t$ to describe the self-similarity of network traffic under the large time scale, and finally get the output $f(Hust_t)$ after input $Hust_t$ into the nonparametric CUSUM algorithm.

② Detect the abnormities of traffic flow under the small time scale

Calculate the $Holder_t$ exponent to describe the local singularity of network traffic under the small time scale, and get the output $f(Holder_t)$ after input $Holder_t$ into the nonparametric CUSUM algorithm.

(3) Get the fused detection result PF_t with the Dempster-Shafer evidence theory.

(4) Synthesize the outputs and make the decision. The decision method is

$$R_t = \alpha_t \times PF_t + \beta_t \times PA_t$$
$$\alpha_t + \beta_t = 1 \tag{18}$$

Where α_t and β_t are the weight values generated by the Exponential Weighted Moving Average algorithm. If $R \geq T$, it is judged abnormal, and considers that botnet exists, otherwise not. The threshold T is dynamically adjusted with the Kaufman algorithm[32] to overcome the different network situations.

V. Performance Evaluation

5.1 Experiment environment

The experiment environment is built by the virtual machine technology and referred to Ref.[33]. Set up a number of virtual machines, and select a virtual machine to act as the router, which is called the monitor VM. And set up the packet analyzer Wireshark in the monitor VM.

The sample packets are captured by the monitor VM every 10 seconds. After running normally for a while, Storm bots are injected into some virtual machines.

5.2 Self-similarity experiment

The change of the Hurst exponent is observed in the experiment, which reflects the self-similarity of network traffic. In normal circumstances, the network traffic shows significant self-similarity, and the Hurst exponent will keep a relatively stable value interval, which ranges from 0.65 to 0.85. It may fluctuate, but usually in a small range.

From Fig. 4, after the storm bots are injected to the network, the self-similarity of network traffic became weaker and weaker, and the Hurst exponent started to decrease when t=400s, and even to 0.41 at the lowest point when t=460s. With bot nodes of P2P botnet are becoming more and more, a new "sick" self-similarity exists in the network traffic, which is actually abnormal in normal circum-

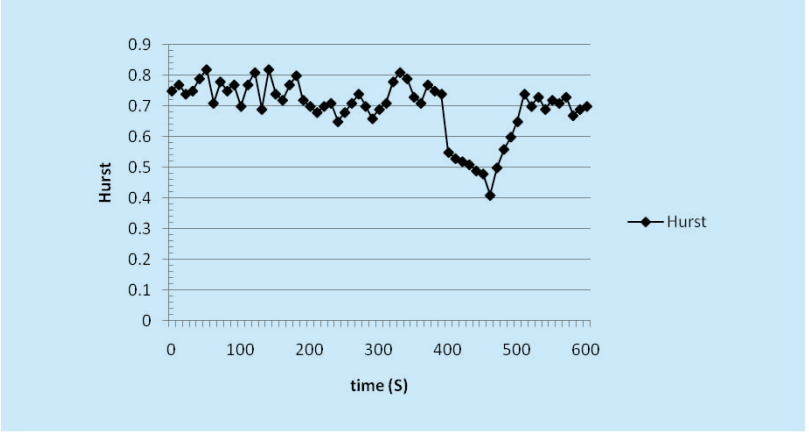

Fig.4 *The Hurst exponent*

stances.

5.3 Local singularity experiment

The change of the Holder exponent is observed in the experiment, which reflects the local singularity of network traffic. In normal circumstances, the network traffic shows the significant local singularity, and the Holder exponent will keep a relatively stable value interval.

From Fig. 5, after the storm bots are injected to the network, the local singularity of network traffic became stronger and stronger, and the Holder exponent started to decrease when t=350s.

The Holder exponent is more sensitive than the Hurst exponent, but will lead to the higher false positive rate. The self-similarity reflects the long-range dependence of the network traffic, and it will change after data accumulated for a period of time since the outbreak of the botnet. The local singularity reflects the irregular over the small time scale, and it will change shortly after the outbreak of the botnet.

Table I *False-positive and False-negative*

Detection Results	Real	Detection Methods		
		Only use the self-similarity under the large time scale	Only use the local singularity under the small time scale	DMFIF
1	0	15	19	11
2	0	41	49	23
3	1000	729	893	882
4	1000	1543(796)	1696(815)	1021(783)

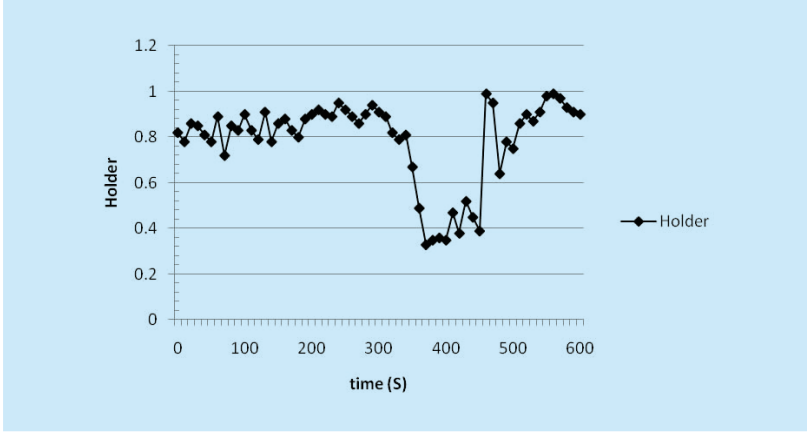

Fig.5 *The Holder exponent*

5.4 False-positive and false-negative comparison experiment

Without loss of generality, 4 groups of data are selected to compare the false negative and false positive rate of DMFIF with other detection methods, which are using different combination of protocols and net flow rates, especially the variation of flow intensity in P2P applications. The first 2 samples are in the environment without bots, the last 2 samples are collected when bots are injected, and the 2nd and 4th samples contain a number of packets from P2P application programs and other web application programs.

From Table I, the detecting precision of our DMFIF is desirable, as its false-positive and false-negative rate in Sample 1 and Sample 3 (network traffic without P2P applications) are approximating to the real situation. In Sample 2 and Sample 4, the P2P applications are added such that the network background flow consists of a large amount of P2P protocols packets. In this case, all detection methods show error reports. However, DMFIF, equipped with TCP deep packet inflection functions, can identify that whether the abnormalities of network traffic mentioned above are caused by the P2P botnet or the P2P application programs, so as to weaken the side effect on detecting P2P botnet which P2P application programs generate. Even under the extreme condition of Sample 4, DMFIF performs well in a relatively lowest false-positive and false-negative rate. Only use the self-similarity under the large time scale and only use the local singularity under the small time scale lead to the higher false negative and false positive rate. Because the characteristics of P2P botnet are complex and changeable, using a single network characteristic to describe the details of the network changes to detect botnet is not enough. So in the detection method of DMFIF, the information fusion theory of decision level is adopted to solve the problem. The "1543(796)" of Table I represents that the detection method of DMFIF detects 1543 times of attack in total and 796 times of them

are correct.

VI. Conclusions

How to detect P2P botnet quickly and efficiently is a challenging problem. In this paper, we propose a novel real-time detection method (DMFIF) based on the fractal and information fusion theory. It synthetically considered the characteristics of network under different time scales with the fractal theory, including the self-similarity and the local singularity, which reflect the intrinsic characteristics of network and don't vary with the topology structures, the protocols and the attack types of P2P botnet. Therefore, DMFIF will be able to detect the new emerging P2P botnet. Firstly, the self-similarity and the local singularity are utilized to describe the characteristics of network. Secondly, two detection results are acquired by using the nonparametric CUSUM algorithm to detect the traffic abnormalities of the above characteristics. Finally, the final detection result is acquired by fusing the above two detection results with the Dempster-Shafer evidence theory. Furthermore, the characteristics of the TCP flow are utilized to weaken the side effect on detecting the P2P botnet which the web application programs generated, especially the P2P application programs. The experiments show that the method is able to detect the new P2P botnet with a relatively lower false-positive and false-negative rate.

The future work is about how to describe the characteristics more accurately and how to further weaken the side effect on detecting the P2P botnet which P2P application programs generate.

Acknowledgements

This work was supported by National High Technical Research and Development Program of China (863 Program) under Grant No. 2011AA7031024G; National Natural Science Foundation of China under Grant No. 90204014.

References

[1] STEWART J. Storm Worm DDOS Attack[R]. SecureWorks, Inc, Atlanta GA, 2007.

[2] SARAT S, TERZIS A. Measuring the Storm Worm Network[R]. Technical Report 01-10-2007, HiNRG Johns Hopkins University, 2007.

[3] STEGGINK M, IDZIEJCZAK I. Detection of peer-to-peer botnets[D]. University of Amsterdam, Netherlands, 2007.

[4] PORRAS P, SAIDI H, YEGNESWARAN V. A Multi-perspective Analysis of the Storm (Peacomm) Worm[R]. Computer Science Laboratory, SRI International, CA, 2007.

[5] Holz T, Steiner M, Dahl F. Measurements and Mitigation of Peer-to-Peer-based Botnets: A Case Study on Storm Worm[C]//1st USENIX Workshop on Large-Scale Exploits and Emergent Threats. San Francisco, 2008.

[6] CHEN Wei, YU Le, YANG Geng. Detecting Encrypted Botnet Traffic Using Spatial-Temporal Correlation[J]. China Communications, 2012, 9(10): 49-59.

[7] QIAO Yong, YANG Yuexiang, HE Jie, *et al*. Detecting P2P bots by mining the regional periodicity[J]. Journal of Zhejiang University SCIENCE C, 2013, 14(9): 682-700.

[8] ZANG Tianning, YUN Xiaochun, ZHANG Yongzheng, *et al*. A Botnet Relationship Analyzer Based on Cloud Model. Geomatics and Information Science of Wuhan University, 2012(37): 247-251.

[9] WANG Hailong, HU Ning, GONG Zhenghu. Bot_CODA: botnet collaborative detection architecture[J]. Journal on Comm-unications, 2009, 30: 15-22.

[10] ZANG Tianning, YUN Xiaochun, ZHANG Yongzheng, *et al*. A Model of Network Device Coordinative Run[J]. Journal of Computers, 2011, 34: 216-228.

[11] ZHUGE Jianwei, HAN Xinhui, ZHOU Yonglin, *et al*. Research and Development of Botnets[J]. Journal of Software, 2008, 19: 702-715.

[12] JIANG Jian, ZHUGE Jianwei, DUAN Haixin, *et al*. Research on Botnet Mechanisms and Defenses[J]. Journal of Software, 2012, 23: 82-96.

[13] WANG Tianzuo, WANG Huaimin, LIU Bo, *et al*. Some Critical Problems of Botnets[J]. Chinese Journal of Computers, 2012, 35: 1192-1208.

[14] MANDELBROT. Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier[J]. J Fluid Mech, 1974, 62: 331-358.

[15] PAUL B, JEFFERY K, DAVID P, *et al*. A signal analysis of network traffic anomalies[J]. ACM SIGCOMM Internet Measurement Workshop, 2002: 71-82.

[16] KIM J S, KAHNG B, KIM D, *et al*. Self-similarity in fractal and non-fractal networks[J]. Journal of the Korean Physical Society, 2008, 52: 350-356.

[17] GIORGI G, NARDUZZI C. A study of measurement-based traffic models for network diagnostics[J]. IEEE Instrumentation & Measurement Technology Conference, 2007: 1-3.
[18] XU Y. A network traffic model based on fractal. International Conference on Wireless Communications[J]. Networking and Mobile Computing, 2007: 1921-1924.
[19] LI M, LIM S C, HU B J, *et al.* Towards describing multi-fractality of traffic using local hurst function[C]//Computational Science-ICCS 2007-7th International Conference, 2007: 1012-1020.
[20] LELAND W E, TAQQU M S, WILLINGER W. On the self–similar nature of Ethernet traffic (extended version)[J]. IEEE/ACM Trans on Networking, 1994, 2: 1-15.
[21] BERAN J, SHERMAN R, TRAQQU S. Long range dependence in variable bit rate video traffic[J]. IEEE Trans on Communication, 1995, 43: 1566-1579.
[22] KARAGIANNIS T, MOLLE M, FALOUTSOS M. Understanding the limitations of estimation methods for long-range dependence[R]. University of California, Tech ReP: TRU-CR-CS-2006-10245, 2006.
[23] KARAGIANNIS T, MOLLE M, FALOUTSOS M. Long-range dependence: Ten years of Internet traffic modeling[J]. IEEE Internet Computing, 2004, 8: 57-64.
[24] RIEDI R H, VEHEL J L. Multifractal properties of TCP traffic: A numberical study[R]. Technical Report RR-3129, INRIA, Rocquencourt, 1997.
[25] MAULIK K, RESNICK S. The self-similar and multifractal nature of a network traffic model[J]. Stochastic Models, 2003, 19: 549-577.
[26] MASUGI M. Multi-fractal analysis of IP-network traffic based on a hierarchical clustering approach[J]. Communications in Nonlinear Science and Numerical Simulation, 2007, 19: 1316-1325.
[27] TARTAKOVSKY A.G. Asymptotic Properties of CUSUM and Shiryaev's Procedures for Detecting a Change in a Nonh-omogeous Gaussian Proces[J]. Mathematical Methods of Statistics, 1995, 4: 389-404.
[28] TARTAKOVSKY A.G, Rozovskii B, Shah K. A Nonparametric Multichart CUSUM Test for Rapid Intrusion Detection[C]//Pro-ceedings of Joint Statistical Meetings. Minneapolis, MN, 2005.
[29] BAUER m. Approximation Algorithms and Decision Making in the Dempster-Shafer Theory of Evidence—an empirical study[C]//12th Conference on Uncertainty in Artificial Intelligence (UAI 96), 1997: 217-237.
[30] YAGER R, LIU L. Classic Works of the Dempster-Shafer Theory of Belief Functions[M]. Springer-Verlag Berlin, 2008.
[31] SEN S, SPATSCHECK O, WANG Dongmei. Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures[C]// Proceedings of the 13th international conference on World Wide Web. New York, USA: ACM, 2004: 512-521.
[32] KASERA S, PINHEIRO J, Loader C. Fast and robust signaling overload control[C]//Proceedings of 9th International Conference on Network Protocols. Riverside, USA: IEEE, 2001: 323-331.
[33] STEGGINK M, IDZIEJCZAK I. Detection of Peer-to-Peer Botnets[EB/OL]. http://staff.science.uva.nl/~delaat/sne2007-2008/p22/report.pdf.

Biographies

SONG Yuanzhang, is currently an Assistant Research Fellow in Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His research interests mainly focus on network security and distributed computing system. E-mail: songyuanzhang@163.com.