An Integrity-guaranteed Timeout Threshold Algorithm for UDP Flow Identification

Xiaoguo Zhang[1,2], Yanjun Su[1], Wei Ding[1]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 211189, China
[2]School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China
e-mail: {xgzhang, yjsu, wding}@njnet.edu.cn

***Abstract*—Network flow records are the foundational and key data of network flow technology. As network flow technology is widely applied, the correctness of network flow record becomes more and more important. For UDP flow identification, timeout strategy is accepted. However, there is no research to analyze the principle and reasonableness of timeout strategy, so we propose a concept of attribute recognition degree that can make a quantitative analysis of the contribution degree of flow attribute for flow identification, and based on this concept we analyze the principle and reasonableness of timeout strategy and the analysis results prove that timeout strategy is suitable for UDP flow identification. Then, on the basis of the rate of number change of UDP flow records, we propose a timeout threshold algorithm for UDP flow identification in order to improve the accuracy of UDP flow identification with timeout strategy. A lot of analyses and experimental results demonstrate that our timeout threshold algorithm can guarantee the integrity and correctness of UDP flows.**

Keywords-flow identification; timeout threshold; UDP flow; flow integrity; attribute recognition degree

I. INTRODUCTION

As network flow technology is widely applied in network security management, performance management, accounting management, traffic classification, traffic engineering, and so on [1-6], the accuracy and efficiency of network flow identifying algorithm become more and more important. It is always one of the important research issues of network flow technology to improve the accuracy and efficiency of network flow identifying algorithm. The core of network flow identifying algorithm is flow termination strategy, it usually refers to flow timeout strategy and directly influences the efficiency and accuracy of flow identifying algorithm. In a general way, accuracy is the opposite of efficiency, and flow termination strategy is aimed to look for a balance between efficiency and accuracy. Although there are many excellent researches and each of them has its own advantages in performance, all of them fail to achieve the best balance between efficiency and accuracy [7-11].

The network flow records are the result of network flow identification. For the applications based on network flow, these network flow records are the foundational and key data. The correctness of network flow records directly affects the validity and the accuracy of these applications. Furthermore, the correctness of network flow records also directly affects the accuracy of flow characteristics, and this will bring about errors in analyzing of flow lifetime, flow volume, flow rate, flow life stage, flow protocol, flow port, and so on. Therefore, focus on the correctness of network flow records, this paper proposes an integrity-guaranteed timeout threshold algorithm for UDP flow identification.

The main contributions of our work are as follows: 1) We define the recognition degree of flow attribute that can quantify the contribution of flow attribute for flow identification. Based on this definition we give its detailed calculation method, and then analyze recognition degrees of many flow attributes and prove that timeout strategy is suitable for UDP flow identification. 2) Based on the rate of number change of UDP flow records, using its mutation point and reasonable point we propose a timeout threshold algorithm for UDP flow identification. Then we do a lot of experiments and analyses, and the results show that the algorithm can guarantee the accuracy of UDP flow identification under the premise of ensuring efficiency.

The remainder of this paper is organized as follows. Section II analyzes existing flow identifying strategies, and points out their advantages and disadvantages. Section III defines and calculates attribute recognition degree, and analyzes the reasonableness of timeout strategy for UDP flow identification. Section IV describes our timeout threshold algorithm based on the rate of flow number change in detail. Finally, we conclude the paper in section V.

II. RELATED WORK

Network flow termination strategy is a very important research content of network flow technologies. From the principle scope, existing network flow termination strategies can be classified as three categories. The first category is the timeout strategy that uses timeout threshold to identify all the flows [7-11]. The second category is the forced termination strategy that directly terminates the flows based on memory limit or time limit [12]. The third category is the termination strategy based on special ending flag and symbols [12]. Because UDP flows have no ending signs and symbols, the third category is not suitable for UDP flows. And the second category sacrifices accuracy in exchange for efficiency, so the second category is also not in conformity with the purpose of our study.

The first category is the most extensively studied flow termination strategy and its basic idea is that if the inactive time of a flow exceeds a certain threshold, the flow is

978-1-5386-2030-4/17/$31.00 ©2017 IEEE

terminated. The representative timeout strategies mainly include Fixed Timeout strategy (denoted as FT), Measurement-based Binary Exponential Timeout strategy (denoted as MBET), Probability-Guaranteed Adaptive Timeout strategy (denoted as PGAT), Two-level Self-Adaptive Timeout strategy (denoted as TSAT) and Dynamical Timeout Strategy (denoted as DToS) [7-11]. Moreover, there are some dedicated flows identifying strategies that are only for UDP traffic, such as multiclass support vector machines timeout strategy [13], differentiated updating timeout strategy [14], and so on.

FT strategy evaluates the reasonability of its timeout threshold based on the number of newly created flows, the number of active flows and the number of repeatedly-created flows, so FT strategy pays more attention to efficiency on memory and CPU. MBET strategy chooses timeout threshold based on flow throughput, so MBET may lose accuracy because of the fluctuation of flow throughput, and MBET can gain high efficiency on memory. PGAT strategy chooses timeout threshold based on application type of a flow, flow size and guaranteed probability, and this strategy also aims to gain high efficiency on memory. TSAT and DToS strategies employ MBET strategy to identify UDP flows, so they also gain high efficiency on memory at the sacrifice of accuracy. Moreover, multiclass support vector machines timeout strategy and differentiated updating timeout strategy also aims to gain high efficiency, and they directly use the experiential timeout thresholds.

It is thus clear that the existing timeout strategies for UDP flows pay more attention to efficiency on memory and CPU than accuracy. Furthermore, these strategies also do not analyze the principle and reasonableness of flow timeout strategies. Therefore, this paper proposes a concept of attribute recognition degree to analyze the principle of timeout strategy, and proposes a timeout threshold algorithm that aims to guarantee high accuracy of UDP flow identification.

III. TIMEOUT STRATEGY ANALYSIS

For a network flow, it is composed of a number of packets; each packet has an arrival time relative to the observation point, so that there is an arrival time interval of two consecutive packets. If there are n packets in a flow, there are n-1 arrival time intervals, and the maximum of arrival time interval is called Maximum Packet Arrival Interval (denoted as MPAI). The timeout strategy predicts the termination of a flow just based on the arrival time interval between two consecutive packets of a network flow; timeout strategy sets a timeout threshold, when the timeout threshold is less than the time that a network flow waits for the next arrival packet, the network flow is terminated. It is not hard to find, for timeout strategy, when a network flow is ended, it needs to be preserved in the memory for a period of time, and this period of time is the timeout threshold. If the timeout threshold is smaller, a network flow may be truncated into two or more network flows, and network flow correctness is reduced. If the timeout threshold is larger, it is a long additional time that a network flow is kept in memory, so memory consumption is more and memory utilization is lower. For network flow correctness, MPAI is the most ideal timeout threshold. However, for memory utilization, the most ideal condition is that a network flow should be exported out of memory immediately when it is ended; i.e., it is zero the additional time that a network flow is kept in memory.

It is not hard to find the principle of timeout strategy from the above descriptions. Although the existing flow identification strategies mostly identify a network flow based on Packet Arrival time Interval (denoted as PAI), these strategies do not analyze the recognition ability of PAI for flow termination. And furthermore, there is also no research to analyze the recognition ability of other attributes, such as packet size, port number, and so on. Therefore, we will quantify the recognition ability of general flow attributes based on the relevant theories of information theory in order to analyze the reasonableness of timeout strategies.

A. Attribute Correlation Quantification

For a random variable X, if its possible values are $\{x_1, x_2, ..., x_n\}$, and the corresponding probabilities of these possible values are $\{P(x_1), P(x_2), ..., P(x_n)\}$, then the uncertainty of this available can be quantified by information entropy $H(X)$.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \quad (1)$$

If in the circumstance that the possible values $\{y_1, y_2, ..., y_m\}$ and their corresponding probability $\{P(y_1), P(y_2), ..., P(y_m)\}$ of a random variable Y are known, then the uncertainty of a random variable X can be quantified by conditional entropy $H(X|Y)$.

$$H(X|Y) = -\sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} P(x_i|y_j) \log P(x_i|y_j) \quad (2)$$

Because $H(X)$ is the uncertainty of X before all the possible values of Y are known and $H(X|Y)$ is the uncertainty of X after all the possible values of Y are known, $H(X)$-$H(X|Y)$ is the amount of information that random variable Y provides for variable X, namely the mutual information between X and Y, denoted as $I(X;Y)$.

$$I(X;Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i y_j) \log \frac{P(x_i|y_j)}{P(x_i)} = H(X) - H(X|Y) \quad (3)$$

$I(X;Y)$ represents statistical constraint degree between random variables X and Y. If variables X and Y are not correlative, $I(X;Y)=0$; Otherwise $I(X;Y)>0$, and the greater the value of $I(X;Y)$ the stronger the correlation of X and Y. However, mutual information can only reflect the reduction of the uncertainty and its value is affected by the information entropy of variables. Therefore, mutual information should be normalized and Symmetrical Uncertainty (denoted as SU) usually be adopted to normalize the mutual information $I(X;Y)$ [15].

$$\mathrm{SU}(X;Y) = \mathrm{SU}(Y;X) = \frac{2I(X;Y)}{H(X)+H(Y)} \quad (4)$$

The value of SU(X;Y) is between 0 and 1, the greater the value of SU(X;Y) the stronger the correlation degree between the two variables. When the value of SU(X;Y) is 1, one variable can perfectly predict the value of the other variable, namely completely correlative and when the value of SU(X;Y) is 0, two variables are completely independent.

B. Attribute Recognition Degree

Definition 1. The amount of information that an attribute can provide for the termination state of a network flow is called the attribute recognition degree of the attribute.

The mutual information $I(Y;X)$ is the amount of information that random variable X provided for Y. If random variable X is a flow recognition attribute, such as source port, destination port, packet size, PAI, and so on; all possible values of X are $\{x_1, x_2,..., x_n\}$ and the corresponding probabilities of X are $\{P(x_1), P(x_2),..., P(x_n)\}$. Meanwhile, Y is the termination state of a network flow and there are two values of Y (1 and 0, 1 represents flow termination, 0 indicates that flow is not ended); all possible values of Y are $\{y_1=1, y_2=0\}$ and the corresponding probabilities of Y are $\{P(y_1=1), P(y_2=0)\}$. Therefore, the value of $I(Y;X)$ is the attribute recognition degree of X. For a fair comparison of each attribute recognition degree, we adopt symmetrical uncertainty SU to normalize each attribute recognition degree.

In this paper, we use a random variable X to represent an attribute of a network flow and employ random variable Y to denote the termination state of a network flow. Firstly, we obtain all possible values and their corresponding probabilities of X and Y, the conditional probabilities of X taking Y as given. Secondly, we calculate and normalize the attribute recognition degree $I(Y;X)$ as following steps: 1) Calculate information entropy $H(X)$ and $H(Y)$ by (1); 2) Calculate conditional entropy $H(X|Y)$ by (2); 3) Calculate mutual information $I(X;Y)$ by (3), and then obtain $I(Y;X)$ by symmetry $I(Y;X)=I(X;Y)$; 4) Normalize $I(Y;X)$ to get SU($Y;X$) by (4).

TABLE I. THE BASIC INFORMATION OF TRACES

Trace ID	Time	Packets number	UDP ratio	Others ratio
1	9/3/2015 8:00-10:00	1.5E+09	18.79%	81.21%
2	9/3/2015 14:00-16:00	2.1E+09	30.82%	69.18%
3	9/3/2015 20:00-22:00	2.7E+09	37.53%	62.47%
4	10/18/2016 8:00-10:00	1.7E+09	25.46%	74.54%
5	10/18/2016 14:00-16:00	2.4E+09	30.65%	69.35%
6	10/18/2016 20:00-22:00	2.9E+09	37.70%	62.30%

Definition 2. A flow is defined as a unidirectional stream of packets subject to a specification that all the packets have same five-tuple (source IP address, destination IP address, source port number, destination port number, layer 3 protocol type) and comply with a certain termination constraint. When the layer 3 protocol of a flow is UDP, the flow was called a UDP flow.

In this part, we will calculate the attribute recognition degrees of some important flow attributes based on IP traces according to our method. We use 64 seconds fixed timeout strategy to identify network flows and choose source port, destination port, packet size and PAI as flow attribute X.

TABLE II. ATTRIBUTE RECOGNITION DEGREE BASED ON TRACE 1~6

Trace ID	PAI	Packet size	Source port	Destination port
1	0.3940	0.1169	0.0237	0.0164
2	0.3937	0.0882	0.0028	0.0018
3	0.3685	0.1145	0.0028	0.0027
4	0.3825	0.1038	0.0020	0.0010
5	0.3781	0.0961	0.0019	0.0008
6	0.4278	0.1125	0.0131	0.0078

When source port, destination port and packet size is attribute X respectively, the possible values of X is respectively {0, 1, 2, ..., 65535}, {0, 1, 2, ..., 65535} and {28, 29, 30, ..., 1500}. Before we calculate the recognition degrees of attributes, we firstly must identify all flows from IP traces and mark the value of Y for every packet (mark 1 for the terminated packet of a flow and mark 0 for others). Secondly, we obtain the distributions of attribute X after Y is known (the value of Y is marked 0 or 1) and the distributions are respectively denoted as D_0 corresponding to Y=0 and D_1 corresponding to Y=1. Based on D_0 and D_1 we can calculate the conditional probabilities of X as Y is known. Then we can calculate the total numbers of Y=0 and Y=1 and denote them as SUM_0 corresponding to Y=0 and SUM_1 corresponding to Y=1. Based on SUM_0 and SUM_1 we can get probabilities of all possible values of Y. Similarly, we can get the distribution of all possible values of X by adding up D_0 and D_1 and denote the distribution as D_2. Based on D_2 we can calculate probabilities of all possible values of X. When we get all these probabilities, we can calculate attribute recognition degrees by our method. However, it is noteworthy that when PAI is attribute X, because 64 seconds fixed timeout strategy is used, all the possible values of X are {0, 1, 2, ..., 64}. For a flow that has n packets, it has n-1 PAIs. We ceil the value of PAI, when n=1 the flow is a single packet flow and it has a PAI=0. Because the best timeout threshold is the maximum PAI of a flow, we make a mark 1 on the maximum PAI (this PAI is the terminated PAI) and the others are made mark 0 (these PAIs are nonterminal PAIs). The values of PAIs and the marks can be obtained in the course of flow identification and then we can calculate the attribute recognition degree of PAI as other attributes.

All IP traces used in this paper are collected from the main channel in the CERNET with 1/4 flow sampling, and this main channel covers over 100 universities and high schools, and its bandwidth is 10G bps [16]. Table 1 lists the basic information of 6 traces that are used in the paper.

In order to accurately quantify identifying capability of each attribute, we calculate the recognition degrees of

attributes respectively on trace 1~6. Table 2 lists the attribute recognition degree measurement results of trace 1~6. From the results we can find that the measurement results of different traces are roughly the same. The attribute recognition degree of PAI is the highest. Packet size also contributes to flow identification but its recognition degree is lower. However, the attribute recognition degrees of other attributes are very low and these attributes can be viewed as uncorrelated elements with flow identification. UDP flows have no ending signs and symbols, and the amount of its attributes is fewer than TCP flows. Among the attributes of UDP flows, PAI has the largest relativity with flow termination state, and the core idea of timeout strategy is based on PAI, so timeout strategy is suitable for UDP flow identification.

IV. TIMEOUT THRESHOLD ALGORITHM

Timeout strategy is accepted by UDP flow identifying algorithms. The key issue of timeout strategy is timeout threshold algorithm. In order to ensure the high correctness of UDP flows, we propose a timeout threshold algorithm that pays more attention to accuracy than efficiency. We choose timeout threshold based on change percentage of flows number. The Change Percentage of Flows Number is denoted as CPFN. For a trace, when timeout threshold is t, the CPFN is denoted as C_t, and it can calculate by (5).

$$C_t = \frac{|F_t - F_{t-4}|}{F_{t-4}} \tag{5}$$

Where, F_t denotes the flows number when timeout threshold is t, and F_{t-4} denotes the flows number when timeout threshold is t-4.

We employ the mean of CPFN as a reference value, and the mean of CPFN is denoted as MC. For timeout threshold t, if C_{t+4}>MC, threshold t is not a reasonable timeout threshold. Because it will lead to a large change of flows number to increase the timeout threshold. This kind of situation is called mutation of flow number, and C_{t+4} is called mutation point, MC is called mutation threshold. If C_{t+4}<MC, it does not lead to mutation of flow number. Furthermore, we use RT to denote reasonable threshold, and RT=2*MC/5. If C_{t+4}<RT, C_{t+4} is considered to be candidate value because it only leads to tiny change of flows number to increase the timeout threshold. It is worth noting that, even if C_t<MC, there is no guarantee that for any timeout threshold τ, if τ>t, C_τ<MC. Therefore, a reasonable timeout threshold must satisfy two conditions as follow. Firstly, the timeout threshold must be larger than all the mutation points. Mutation point means that there is large change of flows number in the process of increasing timeout threshold, so the reasonable timeout threshold should avoid this irrationality situation. Secondly, timeout threshold should be as small as possible of the candidate values; although this paper focuses on the authenticity of flows, we also need consider the efficiency on the premise of ensuring authenticity of flows.

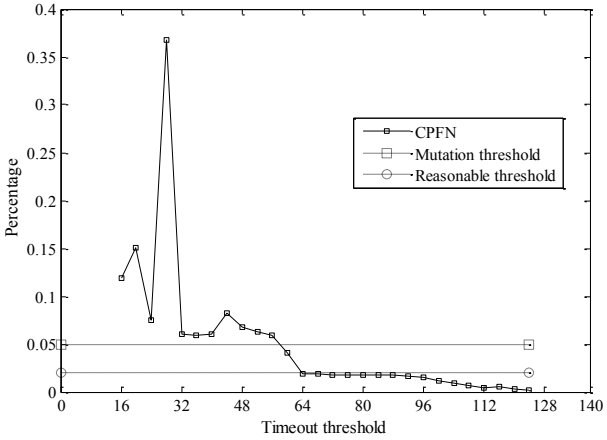

Figure 1. The change percentages of UDP flows number in trace 1

We calculate the timeout threshold respectively on trace 1~6 for UDP flows. Fig. 1 depicts the CPFNs under different timeout threshold of trace 1; from the results we can find when the timeout threshold is 64s the CPFN is reasonable for the first time, and the CPFN decreases with the increase of timeout threshold. For trace 2~6, the CPFNs under different timeout thresholds are similar to trace 1, so Fig. 2 depicts the average CPFNs under different timeout threshold of trace 1~6, and we can find the universal features of CPFNs under different timeout thresholds. It is thus clear that, 64s is always rational timeout threshold for these trace data, and the CPFNs of UDP flows have certain universality not only in different time of a day, but also in different date.

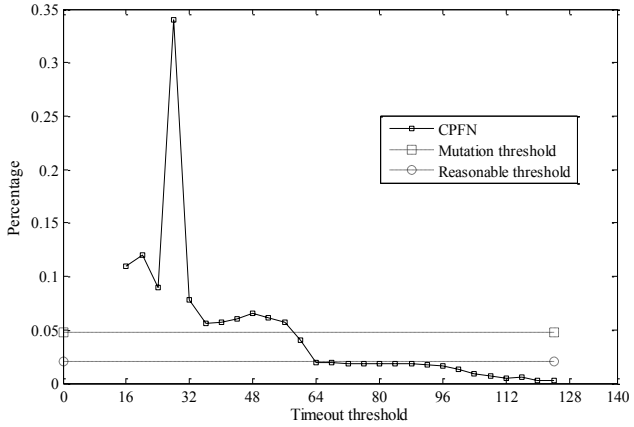

Figure 2. The average CPFN of UDP flows in trace 1~6

V. CONCLUSIONS

In order to gain high accuracy of UDP flow identification, we analyze the existing flow identifying algorithms and the characteristics of UDP flows, and then we use the recognition degrees of attributes to demonstrate the reasonability of timeout strategy for UDP flows. And then we propose a timeout threshold algorithm for UDP flows based on CPFN. We expound the principle of our timeout threshold algorithm and do a lot of experiments on IP traces. The results of experiments show that, 64s is always rational

timeout threshold, and our algorithm is suitable for UDP flow identification that can guarantee the integrity and correctness of UDP flows.

ACKNOWLEDGMENT

We thank the anonymous reviewers for their valuable comments. This work is supported by the National Key Technology Support Program of China under Grant No. 2008BAH37B04 and the State Key Development Program for Basic Research of China under Grant No. 2009CB320505.

REFERENCES

[1] Q. Huang, and P. C. Lee, "LD-Sketch: A distributed sketching design for accurate and scalable anomaly detection in network data streams," IEEE Conference on Computer Communications, Toronto, Canada, 2014, pp. 1420-1428.

[2] X. L. Jiang, J. G. Yang, G. Jin, and W. Wei, "RED-FT: A scalable random early detection scheme with flow trust against DoS attacks," IEEE Communications Letters, Vol. 17, No. 5, pp. 1032-1035, May, 2013.

[3] H. T. Zhu, W. Ding, and L. H. Miao, "Effect of UDP traffic on TCP's round-trip delay," Journal on Communications, Vol. 34, No. 1, pp. 19-29, January, 2013.

[4] H. S. Lee, "Implementing effective and reliable framework for usage-based billing". In: Proc. of NOMS 2004. Piscataway: IEEE, 2004, 889-890.

[5] B. D. Li, J. Springer, G. Bebis, and M. H. Gunes, "A survey of network flow applications," Journal of Network and Computer Applications, Vol. 36, No. 2, pp. 567-581, March, 2013.

[6] Q. Zhao, Z. H. Ge, J. Wang, and J. Xu, "Robust traffic matrix estimation with imperfect information: Making use of multiple data sources". In Proc. of SIGMETRICS 2006, Saint Malo, 2006, 133-144.

[7] K. C. Claffy, H. W. Braun, and G. C. Polyzos, "A Parameterizable Methodology for Internet Traffic Flow Profiling," IEEE Journal on Selected Areas in Communications, Vol. 13, No. 8, pp. 1481-1494, October, 1995.

[8] B. Ryu, D. Cheney, and H. W. Braun, "Internet Flow Characterization: Adaptive Timeout Strategy and Statistical Modeling," PAM 2001 Workshop, Amsterdam, Netherlands, 2001, pp. 95-105.

[9] J. F. Wang, L. Li, F. C. Sun, and M. T. Zhou, "A probability-guaranteed adaptive timeout algorithm for high-speed network flow detection," Computer Networks, Vol. 48, No. 2, pp. 215-233, June, 2005.

[10] M. Z. Zhou, J. Gong, and W. Ding, "Study of network flow timeout strategy," Journal on Communications, Vol. 26, No. 4, pp. 88-93, April, 2005, in Chinese.

[11] M. Z. Zhou, J. Gong, and W. Ding, "High-speed network flows' dynamical timeout strategy based on flow rate metrics," Journal of Software, Vol. 17, No. 10, pp. 2141-2151, October, 2006, in Chinese.

[12] Cisco, "NetFlow Services Solutions Guide," http://www.cisco.com.

[13] J. Cai, Z. B. Zhang, P. Zhang, and X. B. Song, "An adaptive timeout strategy for UDP flows using SVMs," International Conference on Parallel and Distributed Computing, Applications and Technologies, Wuhan, China, 2010, pp. 118-127.

[14] X. H. Zhao, J. B. Xia, and C. H. Zhu, "Research on UDP flow timeout strategy in high-speed network," Journal of Hefei University of Technology (Natural Science), Vol. 36, No. 2, pp. 176-180, February, 2013, in Chinese.

[15] H. L. Zhang, G. Lu, M. T. Qassrawi, Y. Zhang, and X. Z. Yu, "Feature selection for optimizing traffic classification," Computer Communications, Vol. 35, No. 12, pp. 1457-1471, July, 2012.

[16] Jiangsu Key Laboratory of Computer Networking Technology, http://iptas.edu.cn/src/system.php.