Role Recognition Method Based on Network Flow Characteristics

Tao YU\textsuperscript{1} and Rui-qi YUE\textsuperscript{2}

\textsuperscript{1}Beijing Institute of Tracking and Telecommunications Technology, Beijing, China
\textsuperscript{2}Science and Technology on Information Systems Engineering Laboratory, National University of DT, Changsha, China

Keywords: Intranet security, Network flow, Organizational role, Random forest.

Abstract. Network user behavior analysis is very important for the security of large organizations' intranet. In order to analyze the behavior of Intranet users, it is necessary to identify the roles of Intranet users. In this paper, a role recognition method based on network flow characteristics is proposed for organizational intranet, which is a specific network space. Firstly, the interactive distribution characteristics of network flow based on information entropy are used to describe the distribution characteristics of different network nodes. Secondly, traffic characteristics are used to analyze the interaction flow between nodes. Finally, based on these feature sets, Stochastic Forest method is applied to recognize the role of intranet. Experiments on real intranet data verify the effectiveness of the proposed method.

Introduction

Private Intranet is an important basis for daily communication and operation of organizations or enterprises. Intranet attacks may cause huge financial losses and reputation damage. Therefore, data-driven Intranet security tools and analysis methods have attracted great attention in recent years. In many data, network flow is very important for intranet security monitoring. A typical Netflow data includes source IP (srcIP), source port (srcPort), destination IP (dstIP), destination port (dstPort), protocol, bit number and timestamp, etc. Because it is very difficult to analyze the total traffic of the network, Cisco has developed Netflow protocol [1], which collects the complete traffic data as a sample stream, so that network administrators can use Netflow and analysis tools to identify malicious behavior of users. However, the current features provided by Netflow tools are only applicable to simple statistical analysis, which is not enough to support analysis such as role recognition and discovery of new attacks.

Network role recognition is to automatically divide users, hosts and services based on network interaction into different groups, each group has different behavior patterns. Role classification has been widely studied in the field of social networks [2]. Because intranets can be regarded as some kind of social networks, some indicators or methods commonly used in the study of social networks can be applied to the analysis of intranets interaction, such as measuring the degree of importance of nodes, outcomes and so on. The basis of role classification is traffic classification, which can be divided into four categories: port-based classification, payload-based classification, statistical classification and behavior classification [3]. Although there are still some problems in Netflow, such as data packet loss due to router flooding by massive data records, it is still a very useful resource for role recognition [4].

For organizations, identifying roles in the Intranet can help managers understand the workload and efficiency of personnel, as well as discover anomalies. At present, there are some studies on intranet security models or methods, such as Yu et al. [5] grouping intranet user behavior using improved regular expression rules. AsSadhan et al. [6] proposed a long-range Dependence (LRD) behavior analysis method based on time series analysis, and Dong et al. [7] established a method. An internal network user behavior control system helps managers monitor employees' behavior. Xu et al. [8] proposed a behavioral classification framework based on information theory. The index used was information entropy (Entropy) proposed by Shannon. Information entropy has been widely used in
Characteristic Construction of Network Flow

Appropriate network flow characteristics can effectively characterize the user's behavior characteristics. Network flow features based on simple statistics contain redundant information. In order to compress information and extract more compact network flow features, this paper uses entropy to quantify the uncertainty contained in the data.

For a random variable $X$, its possible number of values is $N_X$, assuming that take sampling $m$ times, a random distribution on $X$ is obtained, $p(x_i) = m_i / m$, $x_i \in X$, in which $m_i$ denotes the number of times $x_i$ was observed, then the Shannon entropy of $X$ is defined as [13]:

$$H(X) = - \sum_{x \in X} p(x_i) \log p(x_i)$$

(1)

Formula (1) shows that the more the number of possible values of $X$, the more average the probability distribution and the larger the entropy value. It is noted that the maximum number of different values of $X$ that may be observed in $m$ times observations is $\min\{N_x, m\}$, thus the maximum possible value of entropy is $H_{\text{max}}(X) = \log \min\{N_x, m\}$, i.e., $0 \leq H(X) \leq \log \min\{N_x, m\}$.

It is known that the maximum value of $H(X)$ is affected by $m$ and $N_x$. In order to measure the uncertainty of variable $X$ under the condition that $m$ observations have been obtained, $H(X)$ should be standardized. Therefore, Shannon entropy should be replaced by Relative Uncertainty ($RU$), then we have:

$$RU(X) = \frac{H(X)}{H_{\text{max}}(X)} = \frac{H(X)}{\log \min\{N_x, m\}}$$

(2)

$RU(X)$ takes values from 0 to 1. When $RU(X) = 0$, it means that the observed $m$ values are the same, that is, there exists a certain value $x \in X$, so that $p(x) = 1$. In this case, there is no uncertainty in the observation set. Let $A$ denotes the set of observations of $X$, under the condition of $m \leq N_x$, if and only if $|A| = m$ and for any $x_i \in A$, $p(x_i) = 1/m$, then $RU(X) = 1$. That is to say, $m$ observations are different from each other, and the observation set shows the greatest uncertainty. At this time, $RU(X)$ measures the stochastic degree of $X$. In this case of $m > N_x$, if and only if $m_i = m / N_x$, i.e., for any $x_i \in A$, $p(x_i) = 1/N_x$, then $RU(X) = 1$. In this case, $RU(X)$ is used to measure the uniformity of the random distribution. In this article, $m \ll N_x$.

For a dimension $X$ (such as $srcIP$) and a time period $T$ in Netflow Quaternion ($srcIP$, $dstIP$, $srcPort$, $dstPort$), $m$ is used to represent the number of all network flows generated in $T$, and the set $A = \{a_1, a_2, \ldots, a_n\}$, $n \geq 2$ is used to represents the values of $X$ dimension in these network flows, $m_i$ denotes the number of network flows whose value is $a_i$. Then the probability distribution on $X$ can be expressed as $p_i = P_A(a_i) = m_i / m$. Then $RU(P_A)$ measures the degree of randomness of dimension $X$ in time period $T$. For the network flow in time $T$ with source $IP = srcIP$, we calculate the relative uncertainty of the other three dimensions in the network flow set ($RU_{dstIP}, RU_{srcIP}, RU_{dstPort}$). This triple describes the characteristics of the interaction behavior distribution of the intranet individual $per_i$ in time $T$, as shown in Table 1.
Table 1. Relationship between relative uncertainty and distribution of interaction behaviour.

| RU_dstIP          | RU_srcPort      | RU_dstPort      |
|-------------------|-----------------|-----------------|
| more fixed interactive objects | more fixed sending ports | more fixed traffic |
| Interactive Objects are more random and diverse | random and diverse | Use multiple ports to send traffic to multiple ports |

In order to characterize the behavior patterns of individual Intranet in the morning, afternoon and evening of a day, the sampling time periods are 8:00 to 12:00 ($T_1$), 12:00 to 18:00 ($T_2$), 18:00 to 24:00 ($T_3$), respectively. For individuals $per_i$ with source IP = $srcIP$, the relative uncertainties in three periods are calculated as their network flows. Connection Distribution Features ($CDF$) is showed as:

$$CDF_{per_i} = (T_j RU_{dstIP} , T_j RU_{srcPort} , T_j RU_{dstPort})$$  \hspace{1cm} (3)

The calculation of the above-mentioned entropy only takes into account the number of interactions between nodes, but does not take into account the flow of network flows. For a network individual $per_i$ with source IP = $srcIP$, the total network traffic in time $T$ is defined as $\sum_{per_i,T} Fs$, the number of netflows of $per_i$ in time $T$ is $|netflow_{per_i,T}|$, thus the average network traffic of $per_i$ is defined as:

$$Avg\left(Fs\right)_{per_i,T} = \frac{\sum_{per_i,T} Fs}{|netflow_{per_i,T}|}$$ \hspace{1cm} (4)

Taking $T_1$, $T_2$, $T_3$ as the three time periods of morning, afternoon and evening, the network flow characteristics of network individual $per_i$ are recorded as follows:

$$Avg\left(Fs\right)_{per_i} = \left\{Avg\left(Fs\right)_{per_i,T_1}, Avg\left(Fs\right)_{per_i,T_2}, Avg\left(Fs\right)_{per_i,T_3}\right\}$$ \hspace{1cm} (5)

$Avg\left(Fs\right)_{per_i}$ describes the traffic characteristics of network individuals in one day.

Role Recognition Based on Random Forest

Random Forest [12] is used to classify unknown network individuals based on known role information and network flow characteristics. Random forest model has been widely used in many fields due to its advantages of less parameter adjustment, less over-fitting and providing importance ranking of classification features.

The specific process of role recognition of network individuals based on network flow characteristics using stochastic forest model is as follows:

1. Establish the eigenvector model, i.e., the network flow feature set $F_c(per_i) = (CDF_{per_i}, Avg(Fs_{per_i}))$, and use $(F_c(p_i), r_i)$ to represent all training examples, where $r_i$ represents the role label.

2. Assuming that there are $N$ samples in the training set $S$, $N$ training samples are extracted from $S$ by using the bagging sampling method with playback as the training set $S_i$ of the $i$th tree, and the $i$th decision tree $t_i$ is constructed based on $S_i$.

3. The construction method of decision tree $t_i$: If the number of features in the network flow feature set $F_c$ is $m$, the dimension features $m_{try}$ are randomly selected, and the best classification features $a$ are selected as the splitting attributes of the current node. According to $a$, splitting the node into two branches, the splitting process of the node is recursively called in each branch. The training set $S_i$ can be classified accurately by the decision tree constructed, or all the features have been used.
up. In the process of constructing each decision tree, $m_{ry}$ is a fixed value, usually $m_{ry} = \text{int}(\log_2 m + 1)$.

4) Repeat steps (2) and (3) until $k$ decision trees are established to complete the construction of random forests.

5) Random forests are used to classify the roles of network individuals with unknown roles. The role labels are obtained by considering the classification results of all decision trees. The common methods are voting and probability averaging. That is, when the network flow characteristics $F_c(p_i)$ of the network individual with unknown roles are obtained, the judgment of the role $r_p$ of the network individual is as follows:

$$r_p = \text{arg max}_c \left\{ \frac{1}{N} \sum_{i=1}^{N} I \left( \frac{n_{h_i,c}}{n_{h_i}} \right) \right\},$$

(6)

Probability averaging, $r_p = \text{arg max}_c \left\{ \frac{1}{N} \sum_{i=1}^{N} \omega_i \frac{n_{h_i,c}}{n_{h_i}} \right\}$,

(7)

In which, $N$ is the total number of decision trees in random forests, $I(*)$ is the indicative function, $n_{h,c}$ is the classification result of decision trees for class $C$, $n_{h_i}$ is the number of leaf nodes $h_i$.

Experimental Analysis

This paper studies the internal network environment of a unit, and the network structure is connected with the external Internet. The whole network is divided into seven sub-networks. All network data packets will pass through the central switch, and aggregate the data packets into network flows in the core switch. Seven subnets correspond to six functional departments and one server. If only personal computers are considered in each department, the set of roles $R=\{\text{department 1, Department 2, Department 3, Department 4, Department 5, Department 6, server}\}$. The IP address of the server is fixed, that is, there is only a single element in $\text{per.ip}$. The IP address of the PC is dynamically allocated, that is, there are many IP addresses in $\text{per.ip}$, but only one IP address at the same time. By collecting the data from the core switch, the network flow data generated by 661 hosts in the intranet during the 17 days from November 11 to November 27, 2016 are obtained. There are 195,037,491 data records in total, and the data size is 29.63GB.

By calculating each network flow of $\text{srcIP}$, a total of 2 390 CDFs are obtained. Each CDF corresponds to the random degree characteristics of an intranet individual's interaction behavior on a given day. Three dimensions $(\text{RU}_{\text{dstIP}}, \text{RU}_{\text{srcIP}}, \text{RU}_{\text{dstPort}})$ of the CDF are scattered in three time periods as shown in Figure 3. Three time periods are drawn in three diagonal squares (morning, afternoon, evening). In the other six squares, the dispersion relations between each two time periods are given.
From Figure 1, we can see that the relative uncertainties of the three dimensions are distributed in low-value areas, which shows that most of the intranet objects studied have relatively fixed patterns of interaction behavior, and only a small number of individuals will switch or use more ports between multiple interactive objects, that is, individuals are "uncertain". In the morning, afternoon and evening, it can be seen from the scatter graph matrix $T_j RU_{\text{dstIP}}$ that the scatter points in the scatter graph matrix are mainly distributed on the diagonal line, indicating that the uncertainty of most individuals in choosing interactive objects does not change with the time period; in the scatter graph matrix $T_j RU_{\text{srcPort}}$ and $T_j RU_{\text{dstPort}}$, although the overall distribution of relative uncertainties in the three intervals are similar, but the distribution of walking points is scattered, which indicates that the uncertainty of using interactive ports varies with time.

According to the different roles and departments of Intranet individuals, we can divide the involved intranet individuals ($srcIP$) into seven categories, including the intranet individuals belonging to six different departments (represented by departments 1 to 6) and the intranet individuals belonging to server roles. Accordingly, in a total of 2390 CDFs, the number of CDFs calculated by individuals belonging to seven roles as benchmarks is shown in Table 2. In order to compare the behavior patterns of Intranet individuals of these roles, the distribution histograms of these roles on $(RU_{\text{dstIP}}, RU_{\text{srcIP}}, RU_{\text{dstPort}})$ are drawn, as shown in Table 3. Since most uncertainties do not change much over time, in order to compare the statistical differences among individuals with different roles, the three periods are regarded as equal sampling intervals without considering the distinction of time periods.

In terms of interactive objects, it can be seen from the distribution histogram of $RU_{\text{dstIP}}$ that the other five departments are mainly distributed in the low-value areas, except for the middle-high value areas ($>0.3$) of the 6th departments. The results show that only 6 departments have relatively average distribution of interaction objects and have greater uncertainty. The vast majority of individuals in the remaining 5 departments have smaller range of interaction objects and lower uncertainty. $RU_{\text{dstIP}}$ of Department 1 has a maximum of 0.6 and little distribution above 0.3, which indicates that Department 1 has the lowest uncertainty in terms of interactive objects among the six Departments. $RU_{\text{dstIP}}$ of Department 2 has a part of the high value range from 0.7 to 0.8, which indicates that although most individuals in Department 2 show low uncertainty in the interaction object, there are still a small number of individuals with high uncertainty. The main distribution range of $RU_{\text{dstIP}}$ of servers is similar to other roles, and mainly concentrates on the low and middle value areas, but it shows a more median value (0.3-0.6). This reflects that the server has a higher uncertainty in the interactive objects compared with other types of individuals. This is also due to the working characteristics of the server, that is, the need to ring. It should be decided at the request of different individuals.

In terms of using ports, we can see that the $RU_{\text{srcPort}}$ of server is lower, which reflects that the same server generally uses fixed ports for service response and response, while $RU_{\text{srcPort}}$ in both low-value areas (0-0.3) and (0.7-0.9) are distributed, indicating that due to the different functions of the server,
some servers only go to more fixed individuals and end-to-end. Ports send reply messages, while some servers send reply messages to different ports of different individuals at the same time. $RU_{srcPort}$ of Departments 1 and 4 are mostly concentrated in high-value areas, that is, the ports used to send messages are more uncertain, which reflects the working characteristics of these two departments through multi-ports for a variety of services. The $RU_{srcPort}$ and $RU_{dstPort}$ of department 6 are concentrated in very low areas, which indicates that Department 6 has very low uncertainty in port usage, that is, the network interaction is concentrated on fewer ports, reflecting the scarcity of business types processed through intranet. The other three departments, Department 2, Department 3 and Department 5, have similar $RU_{srcPort}$ and $RU_{dstPort}$ value distribution patterns respectively, which indicates that they have similar behavior rules in the use of Intranet ports.

In order to analyze the size and distribution of network traffic in the whole intranet, a scatter plot of network traffic is drawn from the collected network flow data, as shown in Figure 2. Among them, the horizontal axis represents the size of network traffic, in megabytes (MB) as a unit, and the vertical axis represents the number of network traffic corresponding to the size of traffic in that period, where logarithmic coordinates are used. It can be seen that traffic distribution charts show the characteristics of heavy-tailed distribution, that is, most network traffic only transmits a small amount of traffic, and most traffic is concentrated on a small amount of data flow.

In order to compare the difference of network traffic characteristics among different roles of Intranet individuals, the average traffic is divided according to the roles of Intranet individuals, and the average and median values of each role are calculated. The results are shown in Figure 3. The bar chart above shows the average traffic of individuals with different roles in the morning, afternoon and evening, while the bar chart below shows the median. On average, Section 5 has the highest average traffic, followed by servers, while Section 6 has the highest average traffic in the evening, although the average traffic in the day is low. In addition, the average traffic in the day is significantly higher than that in the night except Department 6. From the median point of view, the average traffic of department 6 in the day is higher than that in the night. Considering the heavy tail of traffic distribution, it can be inferred that this difference is due to the significant increase of the "tail" traffic of department 6, i.e. a small number of individuals in the evening.

Figure 2. Overall distribution of network flow size in Intranet.
Summary

In view of the special network environment of organizational intranet, this paper classifies the roles of different individuals in the intranet by using network flow data and constructing a network flow feature set that can describe the characteristics of interaction distribution and interaction flow. Experiments show that the network flow feature set proposed in this paper can better characterize the interaction behavior characteristics of Intranet individuals, and based on random forest, it can classify the roles according to the network flow data generated by intranet individuals, thus providing a basis for anomaly detection of network users. In the future, more effective features will be proposed to describe Netflow, and other classification methods will be explored to achieve better role recognition results.

References

[1] C. Estan, K. Keys, D. Moore and G. Varghese, “Building a better Netflow,” Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, USA, pp. 245-256, August/September 2004.

[2] S. Zhang, X. Liang and J. Qi, “A review on role identification methods in social networks,” Chinese Journal of Computers, vol. 40, pp. 649-673, 2017.

[3] N. A. Khater and R. E. Overill, “Network traffic classification techniques and challenges,” Tenth International Conference on Digital Information Management, South Korea, pp. 43-48, October 2015.

[4] N. G. Duffield and C. Lund, “Predicting resource usage and estimation accuracy in an IP flow measurement collection infrastructure,” 2003 Internet Measurement Conference, USA, pp. 179-191, October 2003.

[5] Y. Yu, Y. Fu and X. Wu, “Audit scheme for intranet behavior based on improved regular expression rule grouping,” Journal of Computer Applications, vol. 36, pp. 2241-2245, 2016.
[6] B. AsSadhan, K. Zeb, J. Al-Muhtadi and S. Alshebeili, “Anomaly detection based on LRD behavior analysis of decomposed control and data planes network traffic using SOSS and FARIMA models,” IEEE Access, vol. 5, pp. 13501-13519, 2017.

[7] Q. Dong, Q. He, H. Chai, Y. Wang and S. He, “PM-IUBC: A P2P and MongoDB based intranet user behavior control system,” Journal of Communications, vol. 10, pp. 1027-1036, 2015.

[8] K. Xu, Z. Zhang and S. Bhattacharyya, “Profiling internet backbone traffic: behavior models and applications,” Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, USA, pp. 169-180, August 2005.

[9] P. Tang and Z. Zhang, “Multi-agent DDOS attack detection based on information entropy,” Computer Science, vol. 35, pp. 292-295, 2008.

[10] J. Liang and Z. Shi, “The information entropy, rough entropy and knowledge granulation in rough set theory,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 12, pp. 37-46, 2008.

[11] C. Peng, H. Ding, Y. Zhu, Y. Tian and Z. Fu, “Information Entropy Models and Privacy Metrics Methods for Privacy Protection,” Journal of Software, vol. 27, pp. 1891-1903, 2016.

[12] A. Cutler, D. R. Cutler, J. R. Stevens. “Random Forests”, Machine Learning, vol. 45, No. 1, pp. 157-176, 2004.

[13] C. E. Shannon. “A mathematical theory of communication”, Bell System Technical Journal, vol. 27, pp. 379-423, 1948.