ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 616-620
doi: 10.18178/wcse.2017.06.107

Fingerprinting Attack Based on SSH Anonymous Websites

Lu Han [1], Zhengmin Li [2 +] and Zhengping Jin [3]

[1] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China

[2] National Computer Network Emergency Response Technical Team/Coordination Center of China (CNCERT/CC)

[3] Institute of Information Engineering, Chinese Academy of Sciences (IIE, CAS)

Abstract. Internet criminals can carry out illegal and criminal activities through SSH anonymous communication system. However, the highest accuracy rate of the fingerprinting attack method based on SSH anonymous website is about 93%, which makes the perpetrators still have the opportunity to take advantage of this shortage which can be used to make a serious threat to the Internet. On the above issues, this paper proposed an efficient fingerprinting attack method, making the fingerprint attack accuracy rate reached 96.4%. This method is based on analyzing upstream traffic which over the SSH anonymous communication system with the use of the random forests classifier, the number of TCP requests, the number of packets, the size of packet, the packet sequence and so on as the classified characteristics, we have given sufficient evidence to verify the effectiveness of this fingerprinting attack method.

Keywords: traffic analysis, anonymous communication, random forests, fingerprinting attack.

1. Introduction

Although anonymous communication systems (such as SSH, Tor and so on.) can encrypt the data passing through the encrypted channel when the user on the Internet, it cannot hide the traffic information which are transmitted between the client and the server, such as packet size, direction, delay and so on, these traffic characteristics make the attacker have the opportunity to use traffic detection and analysis and other means to obtain the location, identity and other privacy information of the both sides of the communications. The disclosure of this private information will cause serious impact on the individuals, businesses and even the military. On the other hand, although fingerprinting attacks about the anonymous site may be defined as an immoral or even illegal behaviour that steals user privacy, Internet criminals may also use anonymous communication system to institute some criminal activities. Thus, network regulators can take full advantage of fingerprinting attack technology to effective combat illegal and criminal activities on the Internet. However, the existing fingerprint attack technology used only a small amount of traffic characteristics, and the construction of the attack model is not very robust. So the choice of traffic characteristics and the structure of attack models need to be further studied.

In this paper, the above problems are discussed in detail, and a novel fingerprinting attack technology based on random forests algorithm is proposed. The main research results are as follows: (1) Analyzing and understanding the HTTP data flow in anonymous communication system based on SSH protocol, extracting and further processing of high-distinguishing traffic characteristics. (2) Using the random forests algorithm to build the attack model in the closed environment of the upstream traffic, and the accuracy achieved 96.4% which higher than all existing attack models, therefor proves the high efficiency of this fingerprinting attack method. The structure of this paper is as follows: In Section 2 we mainly introduce the related work of

[+] Corresponding author. Tel.: 15600931143; fax: +86-010-62283192.
E-mail address: 870810588@qq.com.

fingerprinting attack. The Section 3 we introduce the fingerprint attack model of this subject. In Section 4 we prove the high efficiency of the fingerprinting attack model by the relevant experimental evaluation. After are the summary of this paper and the next step in the research.

2. Related Work

Fingerprinting attack technology of websites continuous development with the advent of anonymous technology, we focus on analysis of fingerprinting attack technology development process in this section.

Hintz et al.[1] certificated the feasibility of the anonymous website fingerprinting attack in theory. Sun et al.[2] suggested that the number of TCP connections and the total length of the packet as features, using Jaccard coefficient method to match fingerprint's similarity. Bissias et al.[3] used the length and interval time of the packet as features, and the cross-correlation formula to assess the relevant degree between the data. Liberatore et al.[4] regarded the length distribution of the packet as the feature and used naive Bayes for fingerprinting attack, the average accuracy of the attack reached 73%. Herrmann et al.[5] used the weight calculating technique in text mining to standardize the feature vector, and used the Multinomial Na ïve Bayes as the classifier. Compared to most researchers who focused only on the length of the packet, Lu et al.[6] suggested that the order of the packet generated when accessing the anonymous website can also be used as a feature. Ling et al.[7] argued that RTT statistics (the RTT mean and variance) which were generated by the anonymous website could also be used as features in addition to the size and order of the packet. Then Gu et al.[8] proposed that the length, order and total length of the packet and the number of TCP connections can be token as the classified features, using the longest common sub-string algorithm to match the data, the attack effect is greatly improved than before, but this method is very sensitive to sequence of the packet.

In this paper, we use the openSSH2000 data sets which is belong to Liberatore et al.[4] which have high degree of public confidence, but we have adopted a more innovative and powerful random forests algorithm, and we have extracted characteristics of significant discrimination in traffic more than the existing work, the final fingerprinting attack accuracy is higher than their predecessors.

3. Fingerprinting Attack Model

The fingerprinting attack model is shown in Figure 1, and the fingerprinting attack method proposed in this paper is mainly playing a role in the sniffer. At first extract feature data and store it in the database, afterwards select the efficient machine learning algorithm for classification, use training data sets to build a robust classified model, and testing data set to be classified to detect the efficiency of the fingerprinting attack model.

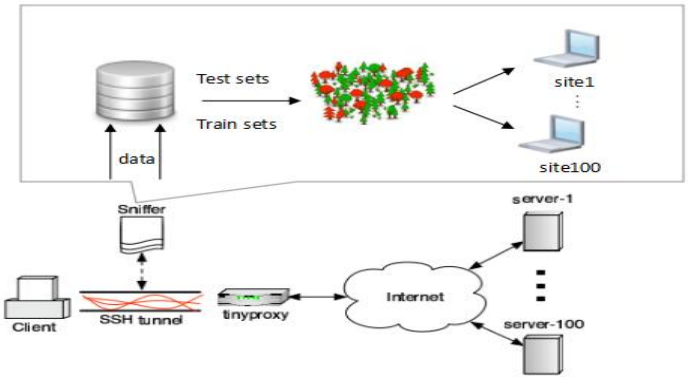

Fig. 1: Fingerprinting attack model [3]

In above fingerprinting attack method, how to select the efficient classified method which are suitable for high-dimensional data and the high-distinguishing characteristics is the focus of this paper.

3.1. Attacking method

We choose the random forests algorithm as classified method of fingerprinting attack model in many machine learning methods. Because it is a kind of more suitable for high-dimensional and large data set

classified method than other means, and has shown the effectiveness in the Tor[10]. The random forests have some prominent advantages relative to the previous classified method in dealing with data:

- Good performance on large data sets. Because of the large amount of data are extracted from the traffic flow, hundreds of decision trees are very powerful compared to the simple decision-making ability of single decision tree, and multiple random sampling makes the random forests do not need K-fold cross-validation to test the performance of the classifier.
- Anti-noise ability. There is a large part of the dirty data in the classified sample due to network, communication, hardware and other reasons, the random forests can still judge and estimate the missing data and maintain a high classified accuracy.
- Able to handle high-dimensional data. From the previous work, The high-dimensional data processing capability of the classifier selected by the previous work was limited, so that the accuracy of the attack is greatly restricted. The random forest method in the high-dimensional data processing effect is better than any other classifier.

3.2. Feature-extraction

This paper extracts the number of TCP request connections, the total number of TCP connections, the size of the upstream packets, the total size of the packets, the number of upstream packets, the total number of packets, the information of RTT, packet order etc. as the data characteristics of the classified model, feature's dimensions which we used is 7-8 times of the predecessor, and many of the high-distinguishing traffic characteristics, such as the number of upstream packets, RTT time information, etc., are not appeared in the attack model of the predecessors, resulting in the accuracy of fingerprinting attacks unsatisfactory.

In order to stabilize the dimension of the feature, the RTT is converted to RTT average and variance, the order of the packet is converted into a website number which best matches the order of the packets, the other statistical information derived from the existing feature as some features is also an important step in feature-extraction.

4. Experiment Evaluation

This paper uses public openSSH2000 data set for experimental verification. Experiments demonstrate the high-distinguishing of the selected features and the efficiency of the methods used by comparing with the existing polynomials such as naive Bayesian, KNN, SVM, and other existing integrated classifiers such as AdaBoost and ExtraTrees. One consensus is that the accuracy of classifier which used by the fingerprinting attack method will be regarded as the the accuracy of the fingerprinting attack method.

4.1. Selecting experimental parameters

The two main steps in the classified process are to use the prepared training samples to train the classifier in order to build a classified model, and then use the testing samples for testing. Various parameters in the evaluation will have an impact on the experimental results, including the size of the testing sample set n_{test} (the number of samples per website), the size of the training sample set n_{train}, the time interval of the training sample and the test sample Δ t, and the size of the training set N_{train} (the number of sites) and the testing set size N_{test} and so on. In order to select the appropriate experimental parameters, we test some of the sample set after fixing certain parameters, observe the accuracy of the fingerprinting attack by modifying the candidate parameters that need to be substituted, and select the appropriate experimental parameters according to the accuracy. In the case of $N_{train} = N_{test} = 100$, $n_{test} = 4$, Δ t = 0.5, as shown in Figure 2, we can find that the accuracy of the attack increases with the increase of the training sample set. When $n_{train} = 4$, the accuracy of the attack is 96.4%. When the increase continues, the accuracy increases but the increase is less, so we set n_{train} to 4 in the subsequent experiments in order to reduce the computational loss of the training ample set and to facilitate the comparison with the existing research work. In the case of $n_{train} = 4$, $n_{test} = 4$ as shown in Figure 3, the accuracy of fingerprinting attack decreases gradually with the increase of interval time Δ t. Even if there is a large time delay between the train set and the test set, the influence on the

accuracy of the fingerprinting attack is relatively small, When Δ t=16, the accuracy can still be maintained at 87.3%, Therefore, the fingerprinting attack model proposed in this paper has high robustness to the time interval, and choose Δ t=0.5 can verify the efficiency of fingerprinting attack method is better.

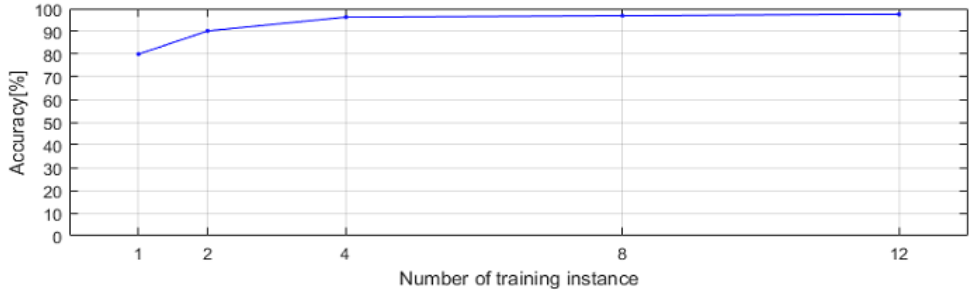

Fig. 2: The influence of different training sample set size on accuracy

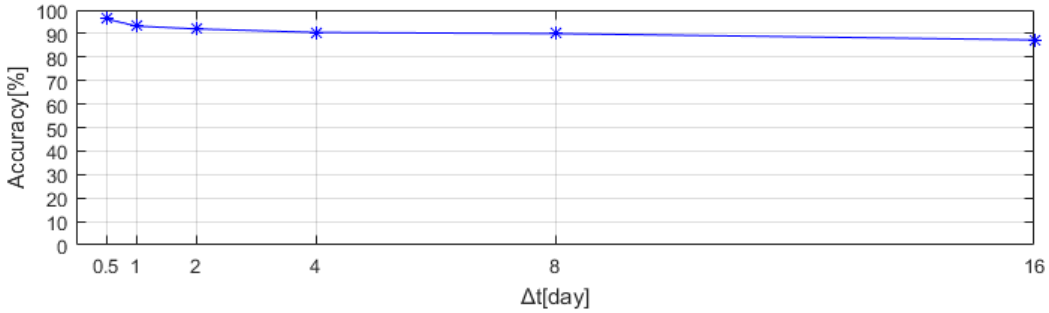

Fig. 3: The influence of different △t on accuracy

4.2. Experimental results

Figure 4 shows the change in the accuracy of the fingerprinting attack when the training sets and the testing sets are increased synchronously in the case of $n_{train} = 4$, $n_{test} = 4$, and Δ t=1. Figure 4 compared with random forests classifier and several other commonly used classifier, we can see from the figure, although the attacked accuracy of categories with the increase in the number of sites has a certain degree of reduction, the random forest algorithm has always maintained the accuracy at the highest position, however, other methods used in the previous fingerprinting attack, such as SVM, KNN, GNB and so on, did not achieve the desired results. When $N_{train} = N_{test} = 600$, the accuracy can still be kept at 85%, which further proves that the random forest method has obvious advantages in the efficiency and stability of attack accuracy with other competitive algorithms.

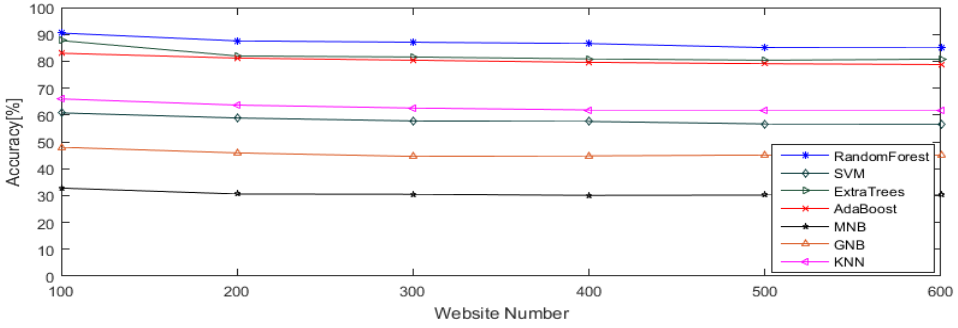

Fig. 4: The influence of different number of websites on accuracy

Figure 5 shows the accuracy of fingerprinting attacks associated with seven classifiers such as MNB, SVM, GNB, ExtraTrees and so on which be used in the most of fingerprinting attack. In the case of $n_{train} = 4$, $n_{test} = 4$, and Δ t=0.5, it can be seen from the figure that the accuracy of the fingerprinting attack method which use the random forest as the classifier has achieved 96.4%, which higher than other classifiers in dealing with the fingerprinting attack of anonymous websites such as SVM, KNN and so on, and the most invalid method-MNB whose accuracy only achieve about 50%, Which is 46% lower than the fingerprinting attack method proposed in this paper. it coincides with the previous theoretical analysis anonymous websites,

which proves the correctness of selecting the random forest algorithm to construct the fingerprinting attack model.

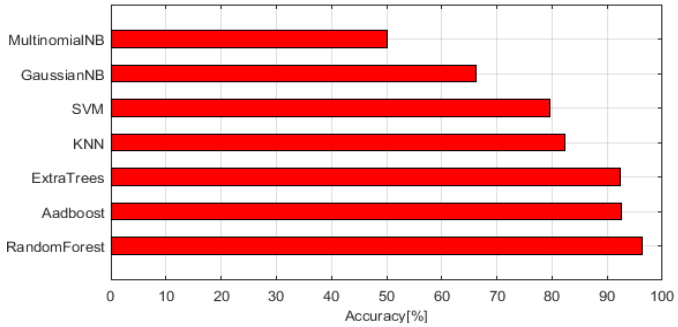

Fig. 5: Comparison of the accuracy of seven attack models

5. Summary

In this paper, the experimental results fully indicate that the fingerprinting attack model which be proposed in this paper has high robustness at the time interval of the sample, while maintaining the stability of the accuracy rate in the case of increasing the number of websites, and the highest accuracy rate is 96.4%, which prove the efficiency of this model. All in all, the proposed method which use random forests as the classifier in this paper is advantageous compared with competitive existing methods which use the SVM, KNN and son on as classifier, not only the robustness but also the efficiency.

These Fingerprinting attack experiments have been carried out in a closed environment, the effect of the method in the real environment has yet to be inspected.

6. Acknowledgements

This work is supported by NSFC (Grant No. 61502044), the Fundamental Research Funds for the Central Universities (Grant No. 2015RC23).

7. References

[1] Hintz, A.: Fingerprinting Websites using Traffic Analysis. In: Dingledine, R., Syverson, P.F. (eds.) PET 2002. LNCS, vol. 2482, pp. 229–233. Springer, Heidelberg(2003)

[2] Sun, Q., Simon, D.R., Wang, Y.M., Russell, W., Padmanabhan, V.N., Qiu, L.:Statistical Identification of Encrypted Web Browsing Traffic. In: IEEE S&P 2002, pp. 19–30 (2002)

[3] Bissias, G.D., Liberatore, M., Jensen, D., Levine, B.N.: Privacy Vulnerabilities in Encrypted HTTP Streams. In: Danezis, G., Martin, D. (eds.) PET 2005. LNCS, vol. 3856, pp. 1–11. Springer, Heidelberg (2006)

[4] M. Liberatore and B. Levine.Inferring the Source of Encrypted HTTP Connections. In Proceedingsof the 13th ACM Conference on Computer andCommunications Security, pages 255–263, 2006.

[5] D. Herrmann, R. Wendolsky, and H. Federrath, “Website fingerprinting: attacking popular privacy enhancing technologies with the multinomial na¨ıve-bayes classifier,” in Proceedings of the 2009 ACM workshop on Cloud computing security (CCSW), September 2009.

[6] L. Lu, E.-C. Chang, and M. C. Chan, “Website fingerprinting and identification using ordered feature sequences,” in Proceedings of the 15th European conference on Research in Computer Security (ESORICS), September 2010.

[7] Ling Z;Luo JZ;Zhang Y A novel network delay based side-channel attack:Modeling and defense 2012

[8] Xiaodan Gu, Ming Yang, and Junzhou Luo. ”A novel Website Fingerprinting attack against multitab browsing behavior”. In 19th IEEE International Conference on Computer Supported Cooperative Work in Design, CSCWD, pages 234–239, 2015.

[9] Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22.

[10] Hayes and G. Danezis. k-fingerprinting: a Robust Scalable Website Fingerprinting Technique. In USENIX Security Symposium. USENIX Association, 2016.

[11] Hayes J, Danezis G. Better open-world website fingerprinting[J]. arXiv preprint arXiv:1509.00789, 2015.