Website Fingerprinting Attack on SSH with Random Forests Classifier

Yong-Jun Wei[1, a], Su-Juan Qin[1, b]

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

[a]13141205240@163.com, [b]qsujuan@bupt.edu.cn

Keywords: anonymity network、website fingerprinting、traffic analysis、random forests

Abstract: In this paper, we implement a special websites fingerprinting attack based on SSH anonymous communication system with random decision forests classifier. When users access to the Internet through anonymous communication network, website fingerprinting attack can listen in the client passively to determine the source address to which user access, so as to achieve the goal of regulation and monitoring anonymous communication system. However, there are plenty of differences between our work and the previous research. We pay more attention to outgoing traffic which is sent by the servers to the client and achieving a well performance by extracting our own features with some specific strategies. According to the experiments, we have proven that the random decision forests classifier has better performance than other classifiers in the field of outgoing traffic.

Introduction

With the increasing use of Internet in personal and business communication, users have higher requirements for privacy protection when using the Internet in specific fields, such as bank transactions, online shopping and medical care. If the users were surfing the Internet without anonymous communication system directly, the attackers can steal users' personal information by monitoring the activities of users' web communication in the HTTP request, and then obtain the content of the communication and the destination address of the traffic. In this case, SSH anonymous communication system with single proxy, Tor anonymous communication system with three hops, and IPSec tunnels to a remote VPN concentrator are presented to protect the privacy of users. Although anonymous communication system is designed to protect user privacy, however, some cyber criminals used it to cover up illegal and criminal behaviors, such as spreading the anti-social speeches, rumors and blackmail, etc. In order to fight against the cybercrime activities effectively, and guarantee the effectiveness of the network supervision, The techniques of website fingerprinting attack based on a variety of anonymous communication system were proposed decades ago.

With the generate of anonymous communication system, the website fingerprinting attacks has been developed. In 2002, Hintz [1] point at the hidden trouble in security encryption agent, the concept of website fingerprinting attacks was proposed firstly and proved the availability of the attack method using total data sent over SSL connections. Sun, et al. [2] also found the loopholes in the anonymous system. They took the number of TCP connection and the total length of packets as features, and made use of Jaccard's coefficient to finish the experiments of fingerprinting.

With the widely use of HTTP/1.1, web browser use persistent connections and pipeline technology when they surfing the Internet. What's more, a variety of single hop or multi-hop

anonymous communication system has a wide range of applications, which make the features be invalid.

In 2006, Liberatore et al [4], firstly presented naïve Bayes classifier method with packet size distribution as feature and reached a 73% accuracy rate through the experiment. At the same time, Liberatore use Jaccard's coefficient to achieve over 60% accuracy on SSH. In 2009, Herrmann et al [5] used text mining techniques to improve Liberatore's features, and took advantage of Multinomial Naïve Bayes (MNB) classifier to achieve the accuracy range from 73% to 89%. In 2012, Ling et al [6] extracted the sample mean and sample variance of RTT as features to complete website fingerprinting attack. In 2014, Gu et al [7] used the difference between incoming traffic and outgoing traffic, and took TCP connection count, total size of packets, as well as the length and sequence of packets as characteristics, through the longest common substring algorithm to obtain the highest accuracy rate of 93.7%. However, for the outgoing packets, only length distribution and Naive Bayes classifier were used to complete the website identify, which would cost a lot of bandwidth with an uncertain effect.

All the above methods on SSH ignored the outgoing traffic features and didn't analyze the integrated traffic. Therefore, we presented a new method of website fingerprinting attack in this paper. We analyze mainly the outgoing traffic, and classify the websites by using total transmission size, total transmission time, the count of TCP response, the total size and number of the outgoing packets as well as the size distribution of the outgoing packets as features. Moreover, the random forests algorithm is used to take related experiments in openSSH2000 dataset.

The paper is organized as follows. Firstly, we briefly introduce the background and the current research status of website fingerprinting attack on SSH, and then, we analyses the attack model and the protocol of SSH. The introduce of the random forests classifier was followed by analyses the protocol of SSH, in this part, we analysis the advantages of random forests classifier with high dimension features and abnormal data theoretically. In the experimental evaluation module, we compare the accuracy of our method with several common classifiers [3, 5, 8]. The last part of the paper, we summarize the paper and points out the next step research direction.

Attack Model and Protocol Analysis

SSH protocol aims to provide secure remote login and other secure network services on insecure network by using a variety of encryption and authentication. And provides encryption channel between the client and the server to solve privacy problem base on plaintext transport.

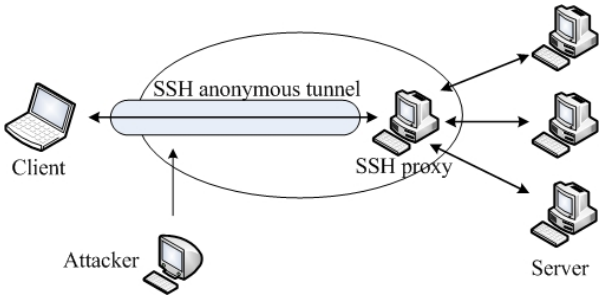

Figure1: the model of website fingerprinting attack

We listen the traffic passively in the tunnel between SSH proxy and client to complete the website fingerprinting attack. Attack model is shown in Fig. 1.

In Fig. 1, the client sends incoming traffic through the encrypted tunnel, and the outgoing traffic is send back to client by the encrypted tunnel. Attacker can only check the encrypted communication, and then create a log of package length and interval time. The observer can store these logs without limit. Like Sun et al [2], we assume the observer could not distinguish the single object.

In order to improve the feasibility of our method, we adopt the same assumption environment with previous scheme [6].

Firstly, attacker can configure the similar network environment with criminal, and has a list of blacklist. Meanwhile, we can use the same SSH proxy for website access, so as to establish a fingerprinting database.

Secondly, attacker has the same configure of browser, such as prohibit the cache, use the same protocol. Moreover, attacker can extract packets from the same website by using the information about interval time.

In respect of feature extract, we only analyze the outgoing traffic. With the widely use of HTTP / 1.1, the number of object in one page and the volume of each connection lose its' effect. At the same time, multichannel parallel transmission makes the outgoing packet is out-of-order. So the longest common substring algorithm for packet sequence in Ref. [7] can't perform well.

Normally, outgoing packets number is larger and most of the length equal MTU. Considering these characters, we calculate the statistics of packets quantity with [0-200], [200-1000], [1000-1448]. Then we use statistics of packets quantity, the total size, and the total numbers as the features of outgoing packets. At the same time, we put total size of packets and website access time in feature set. Furthermore, when we parse the original data of openSSH2000, the control packet will be discarded, which is influenced by the network environment significantly.

Attack Method

Our attack use random decision forests [9，10] which is an ensemble method using multiple decision trees. Different with other classifier [3,5,7], random forests classifier can analyze the contribution of features，which can be used to evaluate the importance of characteristics. Furthermore, in the face of dimension reduction, data leakage and data exception, it still can maintain a good performance.

The dimension in our characteristics is 3 ~ 4 times of previous. Due to influence of network and time, there are some empty data and abnormal data which potentially reduce the classifiers' effect. It's one of the reason that Support Vector Machine (SVM), K Nearest Neighbor (KNN), Multinomial Naïve Bayes (MNB), Gaussian Naïve Bayes (GNB) etc classifiers perform poorly.

Random forests achieve the estimate of the missing data, even if a big part of the data was loss can still remain the accuracy, and make the error of abnormal data balance. With sample many times randomly with reset, we do not need use k-fold crosses validation to measure random forests performance, it is estimated through the unused training samples on each tree, and the callback mechanism improves the accuracy. What's more, fast learning and detect the feature interactions make random forests perform better.

In random forests, differ from only one tree in the CART model, many decision trees will be generated. Each tree in the forests uses a fixed length feature vector for training and through select training samples randomly to prevent over-fitting [11]. When an object is classified, every tree in the random forests tree will give their selection, and the output result of a whole forest is the website with the most votes.

The "plant" and "growth" rules of every decision tree are shown as follows:

• Suppose that we set the sample number of training set as N, and then we get the N samples by repeated sampling with reset, so that the sampling results will be used as the training set of the decision tree;

• If there are M input variables, each node will be randomly selected m (m<M) specific variable, and then use the m variables to determine the best splitting point. In the process of generate the decision tree, the value of m is kept constant;

• Each decision tree is most likely to grow without pruning;

• Predict the new result through compute the selections of the entire decision tree.

Experimental Evaluation

In this paper, we use the public dataset in Ref.[4] to finish the experimental verification. By comparing with SVM, KNN, MNB and GNB classifiers, we prove the high accuracy of our method.

The dataset is collected by Liberatore et al in 2006 and each sample consists of the 2,000 most-visited websites of traffic. They created a new sample once every six hours for a period of two months, and total get 480,000 samples [4].

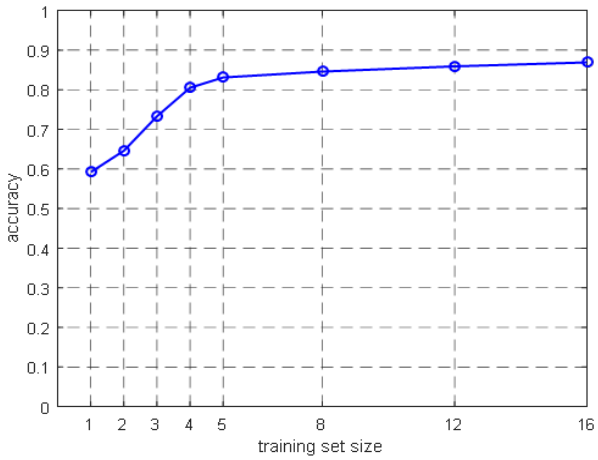

Figure 2: Effect upon accuracy of varying training set size

Website fingerprinting attack mainly includes two stages: known sample training and unknown sample testing. The accuracy of the test results is the criterion for judging the effect of attack. Parameters setting in the experimental process will have an impact on the accuracy of the test results which include the training set size, the size of the testing set and interval time. In order to ensure the credibility of the results and compare with the previous experiments, we use the top 100 websites of dataset as our target websites.

As shown in Fig.2, we take the experiment of training set (n_{train}) and compute the accuracy when the interval time of $\Delta t=1$ and the testing set size of $n_{test}=4$. After repeated experiments many times, we calculated the average value of each point. Compared with Liberatore et al [4] and Gu et al [6], our accuracy is increased by nearly 7% form $n_{train}=4$ to $n_{train}=16$ and their result tend to be stable. With the increase of the training set, more data can be used to "plant" and "growth", which makes random forests has more samples to generate training set of decision tree, so the accuracy has been significantly improved. Considering the consistency with previous work, we still set the $n_{train}=4$.

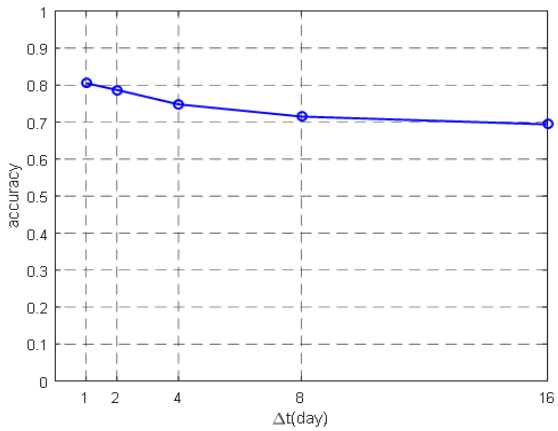

Figure 3: Effect upon accuracy of varying interval time

Fig. 3 shows the effect of interval time on the accuracy when $n_{train}=4$ and $n_{test}=4$. From the Fig.3, We can know, the accuracy appeared different degree of decline when interval time increasing. The reason is that the outgoing traffic is used to transfer the object, have a large impact on the total size and the total number of outgoing packets. And then affect the distribution of packet size in different ranges, which led to the decline in accuracy. However, accuracy remains at around 70% even interval time up to 15 days, this shows that our method has a high robustness to time delay. Without update the training set frequently, the method can still ensure the effectiveness of attack. In order to facilitate the comparison with previous, we

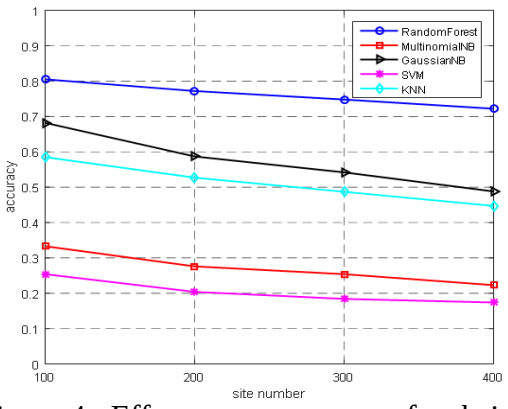

Figure 4：Effect upon accuracy of websites number

set the time interval $\Delta t=1$.

Then we verify the relationship between the numbers of websites and the accuracy through experiments. Through theoretical analysis we can know that with the increase of websites numbers, more similar data appeared in the training set and the testing set, which makes the accuracy rate appear to varying degrees of decline.

Fig. 4 shows the influence of website numbers on the accuracy by comparing with several commonly used classifiers, such as MNB, GNB, SVM, and KNN. The experimental parameters are $n_{train}=4$, $n_{test}=4$, $\Delta t=1$. We can see from Fig.4, the accuracy of all classifiers appeared different degree of decline with the increase of the websites numbers. When the numbers of sites reach 400, our method can still keep 72.2% accuracy, compared to other methods, the accuracy of our method decreases slowly. The result shows that our method can keep good effect in big data.

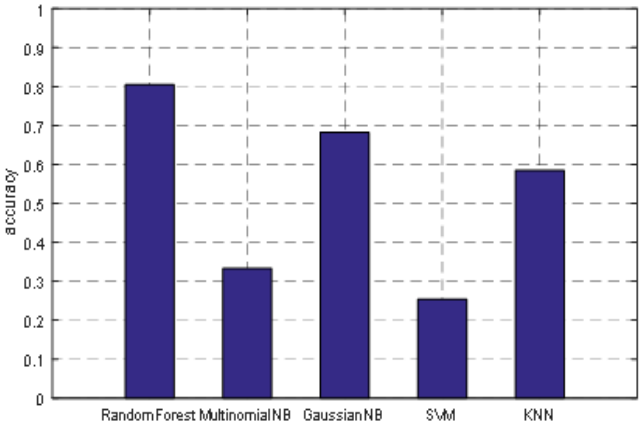

Figure 5：Effect upon accuracy of five different methods

Fig. 5 shows the accuracy of five different methods in $n_{train}=4$, $n_{test}=4$ and $\Delta t=1$. The experimental results show that random forests classifier perform better for website classification in the face of high dimension features and abnormal data.

Conclusion

In this paper, we aim at monitor the illegal use of SSH. Random forests is put forward on outgoing traffic and compared with common methods. It is proved that only rough outgoing traffic analysis can also get a good effect. The experiments result shows even in big dataset and larger intervals, our method can also maintain a good accuracy. Moreover, our method can ensure the stability of the attack model without update training set frequently.

In the future, we will pay more attention to the extract of features and find a kind of effective defenses to the features we used. At the same time, the existing methods were always verified in closed environment, which affect the practicability of attack methods. How to guarantee the accuracy in open experimental environment is an urgent problem.

Acknowledgments

This work is supported by NSFC (Grant Nos. 61300181, 61502044), the Fundamental Research Funds for the Central Universities (Grant No. 2015RC23).

References

[1] A. Hintz. Fingerprinting Websites Using Traffic Analysis. In Privacy Enhancing Technologies (PETs), pages 171–178. Springer, 2003.

[2] Qixiang Sun, Daniel R. Simon, Yi-Min Wang, Wilf Russell, Venkata N. Padmanabhan, and LiliQiu. Statistical identification of encrypted web browsing traffic. In Proceedings of the 2002 IEEE Symposium onSecurity and Privacy, Berkeley, California, May 2002.
[3] Tao Wang, Xiang Cai, RishabNithyanand, Rob Johnson, and IanGoldberg. Effective Attacks and Provable Defenses for Website Fingerprinting. In Proceedings of the 23rd USENIX Security Symposium,San Diego, CA, USA, August 20-22, 2014., pages 143–157, 2014.
[4] M. Liberatore and B. Levine. Inferring the Source of Encrypted HTTP Connections. In Proceedings of the 13th ACM Conference on Computer and Communications Security, pages 255–263, 2006.
[5] D. Herrmann, R. Wendolsky, and H. Federrath. Website Fingerprinting: Attacking Popular Privacy Enhancing Technologies with the Multinomial Naive-Bayes Classifier. In Proceedings of the2009 ACM Workshop on Cloud Computing Security, pages 31–42, 2009
[6] Ling Z, Luo J, Zhang Y, et al. A novel network delay based side-channel attack: modeling and defense[C]//INFOCOM, 2012 Proceedings IEEE. IEEE, 2012: 2390-2398.
[7] GU Xiao-Dan, YangMing, LUOJun-Zhou, JIANGPing, Website Fingerprinting Attack Based on Hyperlink Relations, Chinese Journal of Computer, v 38, n 4, p 833-845, April 1, 2015
[8] Andriy Panchenko, Lukas Niessen, Andreas Zinnen, and Thomas Engel. Website fingerprinting in onion routing based anonymization networks. In Proceedings of the 10th annual ACM workshop on Privacy in theelectronic society, WPES 2011, Chicago, IL, USA, October 17, 2011, pages 103–114, 2011.
[9] Breiman L. Random forests [J]. Machine learning, 2001, 45(1): 5-32.
[10] Pall Oskar Gislason, Jon AtliBenediktsson, and Johannes R. Sveinsson. Random forests for land cover classification. Pattern Recogn.Lett, 27(4):294–300, March 2006
[11] Hayes J, Danezis G. Better open-world website fingerprinting [J]. arXiv preprint arXiv:1509.00789, 2015.