Old Habits Die Hard: Fingerprinting Websites On The Cloud

Xudong Zeng[1,2], Cuicui Kang[1,2], Junzheng Shi[1,2], Zhen Li[1,2](✉), and Gang Xiong[1,2]

[1] Institute of Information Engineering
Chinese Academy of Sciences
{zengxudong,kangcuicui,shijunzheng,lizhen,xionggang}@iie.ac.cn
[2] School of Cyber Security
University of Chinese Academy of Sciences

Abstract. To detect malicious websites on the cloud where a variety of network traffic mixed together, precise detection method is needed. Such method ought to classify websites over composite network traffic and fit to the practical problems like unidirectional flows in ISP gateways. In this work, we investigate the website fingerprinting methods and propose a novel model to classify websites on the cloud. The proposed model can recognize websites from traffic collected with multi-tab setting and performs better than the state of the art method. Furthermore, the method keeps excellent performances with unidirectional flows and real world traffic by utilizing features only extracted from the request side.

Keywords: Website fingerprinting, traffic analysis, cloud platform

1 Introduction

Cyber security has become the focus of governments and public after the PRISM[3]. In the same time, encrypted communication protocols like SSL/TLS are becoming ever more popular. The percentage of encrypted requests to the google services is up to 75% [6] and still keeps increasing with the rapid development of cloud services. Nowadays, almost every cloud platform can provide users with free certificates to apply SSL/TLS and encrypt their communications. However, malicious websites are spreading among the cloud and serving with other normal websites on the same IP address [18] because of the techniques used by cloud platforms. A Chinese report [4] shows that almost 74% handled phishing websites are detected by public tip-off which embodies the need for precisely detection method.

To solve the problem, the usual ways are domain or IP blacklist. However, domain blacklist only solve the problem partially as discovering all variants isn't easy. For example, pornographic websites or online gambling websites have applied in many domains that domain blacklist could hardly gather all of them.

[3] https://en.wikipedia.org/wiki/PRISM_(surveillance_program)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_45

And the IP blacklist is likely to block normal websites serving on the same IP address. Thus, traffic analysis is suggested since the superficial information is disappeared in encrypted traffic. In this field, the website fingerprinting (WF-P) is popularly used in classifying websites over encrypted traffic, which utilize time, packet length, distribution and other statistical meta-data as a feature set.

In this paper, we take advantage of WFP to classify encrypted websites over HTTPS and propose a novel composite feature, termed as Request-Respond-Tuples (RRT, see Section 4.2). RRT is unlike the packet direction based bursts. It's defined based on a client request and the corresponding server response. After compared the performance of several machine learning algorithms, the proposed model is implemented with Random Forest [1]. And then, the proposed model is experimented with challenging scenarios where traffic is collected with simulated user behaviors. Finally, the proposed model achieves more than 20% true positive rate than the state of the art WFP model k-fingerprinting [7] in classifying websites on Content Delivery Network (CDN). And also shows promising results when being applied to real world problems.

The key contributions of this paper are listed as follows:

- This paper presents a three-phase domain collecting method, which gathers websites in the same IP address and is suitable for several cloud platforms.
- Our data sets are collected with simulated user behaviors which is more complex than previous studies.
- This paper proposed a novel WFP model and achieve a better true positive rate than k-fingerprinting [7] in recognizing websites on cloud platforms.
- The proposed model performs well with unidirectional flows and achieve 84.72% TPR with 200 positive classes.
- To the best of our knowledge, we are the first who apply WFP on traffic from ISP gateways.

The remainder of this paper is organized as follows: Section 2 shows the related work of WFP and Section 3 provides insight into the data sets. Then, data processing and RRT feature are introduced in detail in Section 4. The experiments are conducted in Section 5, where the proposed model is constructed with several the state of the art algorithms, including Naïve Bayes (NB) [22], Decision Tree (DT) [14], SVM [17] and Random Forest (RF) [1]. Furthermore, the section also discusses how to apply the proposed WFP model with realistic problems in ISP gateways. Finally, Section 6 concludes this paper.

2 Related work

Deep packet inspection (DPI) [15] is a famous technique in traffic classification. But with the development of encryption, DPI faced with the great challenge that traffic analysis based on machine learning becoming another alternate method. WFP is a type of traffic analysis and it has been researched for a long time. The very first WFP models are aiming to identify pages over SSL [3, 10], but more following WFP studies [**?**, 2, 5, 7, 8, 11–13, 20] turn to focus on applying WFP in

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_45

anonymous communication networks. However, people generally don't use any proxies that most WFPs appealed to a restricted scenario. Recently, Hayes et al. [7] made well performed WFP by using the predicted results of each decision tree in a trained RF, called k-fingerprinting (KFP). KFP classified 55 webpages over 7000 webpages from Alexa top 20000 and achieved 95% true positive rate.

Liberatore et al. [**?**], Herrmann et al. [8], Dyer et al. [5] and many early studies adopt Naïve Bayes to learn the patterns of website. But Naïve Bayes amuse features are independent of each other which is difficult to achieve in reality. Other researchers [2, 11–13] mainly focus on SVM. Panchenko et al. in 2011 [13] proposed some useful statistical features, such as burst and HTML Marker, processed WDP with SVM. And later in 2016 [11], they showed another WFP model named CUMUL with interpolants of cumulative packet lengths curve. Although previous studies obtained excellent results, WFP still need to try more machine learning algorithms and compared the difference. Juarez et al. in 2014 [9] argued that WFP studies made some naive settings such as no background traffic and ignore the difference between browser versions, user behaviors and regions. Regardless of the fact that it's hard to obtain a excellent result in complex settings, researchers still made some achievement. Wang et al. [21] applied WFP to Tor with background traffic and Panchenko et al. [12] compared results of classifying Tor hidden service between different Tor Browser Bundles.

To sum up, WFP has obtained a lot of achievement on classifying websites over anonymous communication network, but real world network is more complicated and WFP still need more breakthrough to deal with background traffic and other practical problems.

3 Data set

We collected two data sets from different kinds of cloud platforms: App Engine and CDN. As Baidu APP Engine(BAE)[4] and Tencent CDN[5] (TCDN) are popular in China, and most of websites on them can be accessed by HTTPS. Besides the platform difference, some websites in data set **TCDN** were similar or owned by the same large website, but data set **BAE** excludes those websites manually.

3.1 Domains collection

As we haven't found any public websites list for those two platforms. So, we come forward with a three-phased method to gather domains on cloud. First, we can build a seed list of domains which serve on specific cloud platforms by Internet search engines. Second, performing active requests to extract IP addresses for each domain in the seed list. Finally, gather domains from dns.aizhan.com or other tools which can reverse search domain by IP address. Although this three-phase method can't provide a perfect website list, it still good enough to collect numerous domains which were available on the same IP address recently.

[4] https://cloud.baidu.com/product/bae.html
[5] https://cloud.tencent.com/act/event/cdn_brief.html

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_45

3.2 Collection Settings

Based on the three-phase method, the data sets are collected by utilizing Selenium[6] to take control of Firefox and assess domains of BAE and TCDN repeatedly. The raw traffic is recorded into PCAPs[7] by Tshark[8]. For each domain, we first get its index page and then randomly perform several clicks to simulate user behaviors. Therefore, traffic is collected with multi-tab and the user behaviors are varied in each collection.

By performing random clicks in traffic collection, the generated traffic is more complicated than others with on interactions. For example, first we open the homepage of website A, and then we randomly click a hyperlink which will open a webpage of A or other websites. After performing these behaviors several times, the generated traffic behaviors will be different in each time. As a result, the target website to classify is multi-modal [19] which means the fingerprint of the target isn't stable and the classification is become more challenging.

3.3 Summary

In the data sets, **BAE** data set consists of 40 websites each with 100 instances, while the **TCDN** data set consists of 30 instances each of 200 websites and 10 instances each of 1000 websites. The summaries for both data sets are presented in Table 1. The column of different IPs describes the number of unique IP addresses. It's obvious that every IP holds more than one website in average and **TCDN** is more centralized than **BAE**. Due to the relations between websites and IP addresses are changing over time. We assume all websites of each data set serve on the same IP address.

Table 1: Summaries of BAE and TCDN

Data set	Total instances	TCP Flows	Different IPs	Different SNIs
BAE	4000	52.92K	17	190
TCDN	16000	134.88K	68	2213

4 Data processing

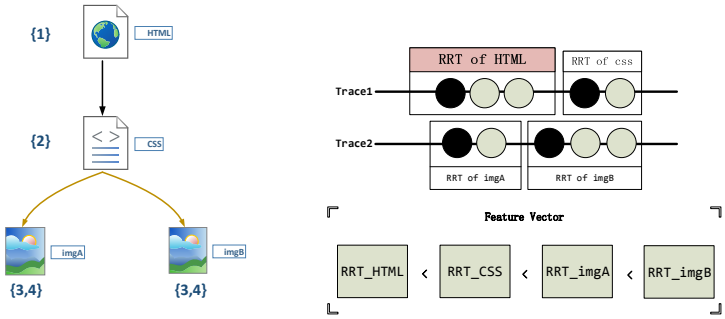
4.1 Preprocessing

For each trace T in the data sets, we assign $T = (SNI, P)$ where SNI is a field of SSL/TLS protocol and $P = (t, l, s)$ which stands for timestamp, packet length

[6] https://www.seleniumhq.org
[7] https://en.wikipedia.org/wiki/Pcap
[8] https://www.wireshark.org/docs/man-pages/tshark.html

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_45

and packet states. In particular, when $l > 0$ indicates an incoming packets and $l < 0$ indicates an outgoing packet. And the states can be used to filter packets without payload and unused connection.



(a) Loading process (b) Data processing

Fig. 1: In figure(a), browser first gets a CSS stylesheet and then requests two images and figure(b) demonstrates the data process where the black circles represent outgoing packets and the white are incoming.

4.2 Features

Based on the previous studies, some useful statistical features have been extracted in a new perspective. We assume that website can be identified by a set of loading patterns which consist of a partial order relation (POR) set and RRTs. In particular, the RRT is a kind of ensemble feature which describes the statistical features of the request packet(s) and the following incoming burst. And the POR in our study is based on the timing relationship that describes the start order between two RRTs. Therefore, PORs of a certain website can remain intact with background traffic. To illustrate the work more specifically, Fig. 1(a) shows the loading process of a simple webpage and the data processing is displayed in Fig. 1(b). From the figure, RRTs are slightly independent with each other which means RRTs are focused and immune to background noise. Finally, the top K RRTs are chosen to represent the loading patterns of websites. The feature set of RRT is listed as follows:

- ***Time*** indicates the status of server and network. So that we extract start time, time cost, distribution of packet arrival time and inter arrival time to embody this field.
- ***Packet length*** is the "shape" of the encrypted content. Thus, we extracted request packet length, total response packet length, last packet length and other statistical features to describe it.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_45

- ***Proportion*** is the overall distribution of packets/bytes/time. According to the proportion, the local statistical feature can associated with the overall information which imply the profile of website.

5 Experiment

5.1 Experiment setup

The websites in the experiments are recognized by traces aggregated by SNI which named as TASNI for short. However, most previous WFP models are applied on all traces belong to the same domain which is hard to achieve in reality. Furthermore, websites in WFP studies ether be regarded as monitored or unmonitored, similar to interested in or not. But we recognize websites by TASNI, that an instance may consist of several TASNIs. Thus, we assume only TASNI whose SNI equal to the domain can be regarded as monitored, because most unique resources are more likely belong to it.

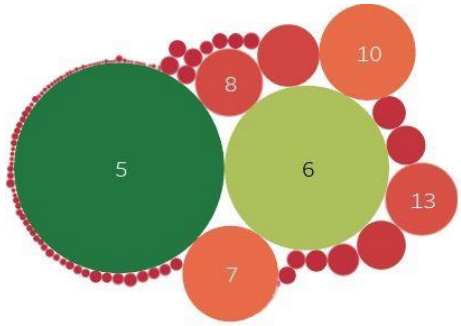


Fig. 2: The distribution of TASNI's RRT count.

It should be noted that we only conduct experiments with open-world setting [13] where the test set owns some cases that train set don't have. In order to evaluate the performance of the proposed method, True Positive Rate and False Positive Rate are used as the major evaluation metric in the experiment, which are defined as follows:

- ***True Positive Rate (TPR)*** is the probability that a monitored website is classified as the correct monitored page. In the following equation, $|test_{mon}|$ means the number of monitored instances for testing.

$$TPR = \frac{\sum_{i=0}^{|test_{mon}|}(Predict_i = Label_i)}{|test_{mon}|} \tag{1}$$

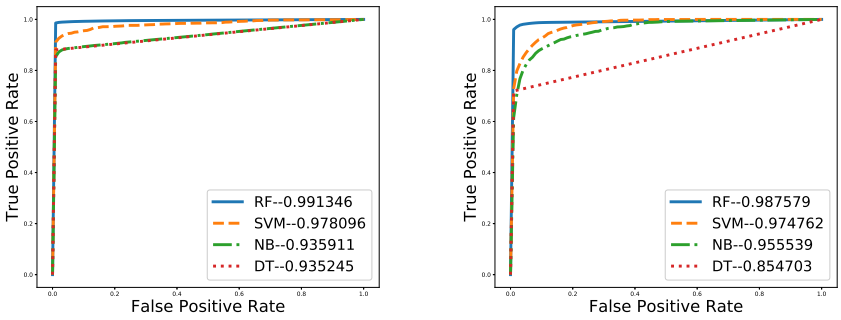
ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_45

- ***False Positive Rate (FPR)*** is probability that an unmonitored page is incorrectly classified as a monitored page. In the following equation, $|test_{unmon}|$ means the number of unmonitored instances for testing and NOISE is constant for every unmonitored instance.

$$FPR = \frac{\sum_{i=0}^{|test_{unmon}|}(Predict_i \neq NOISE)}{|test_{unmon}|} \tag{2}$$

5.2 Model

With features extracted, several supervised machine learning algorithm are compared. As most of TASNIs in Fig. 2 have RRTs less than 6, that top 6 RRTs (K=6) is selected. For each data set we randomly chose 60 percent of monitored TASNIs and 1000 unmonitored TASNIs for training. The rest of monitored TASNIs and 2000 randomly selected unmonitored TASNIs are used for testing. The results are displayed in Fig. 3 and Table 2, from which it can be found that the Random Forest performed much better than the others.



(a) BAE (b) CDN

Fig. 3: The ROC curve of four classifiers on two data sets and the AUC scores of each algorithm are listed in legends. The difference between algorithms is more clear in CDN data set, as the positive kinds of CDN data set are more than BAE data set.

From the result of the comparison, the NB not perform very well. As we know, the NB algorithm assumes the features in the feature set are independent of each other. The assumption is much suitable for RRT, but it isn't very suitable for the features of RRT. The DT algorithm and SVM algorithm also not perform as well as RF. In our opinion, we assume the reason is because of the multi-modal. As Wang Tao [19] hold that WFP is a kind of multi-modal classification whose class is consisted of several subclasses, where the instances of the same

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_45

subclass are similar, but not similar in different subclasses. In our data set, all the traffic collected with random interaction which caused many subclasses. And the DT and SVM take all instances in a class into consideration, that the performances have been influenced. However, RF use several randomly sampled data set to build an ensemble classifier, where the influence of multi-modal would be weakened.

Table 2: Precision, Recall, F1-score of four classifiers.

Algorithm	BAE			CDN		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Random Forest	98.67%	94.64%	96.32%	89.23%	88.21%	88.2%
SVM	89.6%	88.01%	88.4%	57.27%	56.53%	55.69%
Natïve Bayes	94.1%	81.64%	85.91%	41.59%	35.89%	35.3%
Decision Tree	92.52%	90.41%	90.83%	73.79%	71.79%	71.63%

Therefore, Random Forest is chosen as our classifier, named as RRT+RF shortly. To elevate RRT+RF, we set up an experiment for varying the sample amount of RF, N, from 1 to 200 on **TCDN** data set. From this Fig. 4, we can see that the TPR and FPR change quickly when N is low, and become convergence when N is greater than 100. Finally, N is set to 150 where the model can achieve a nice result and training quickly as well.

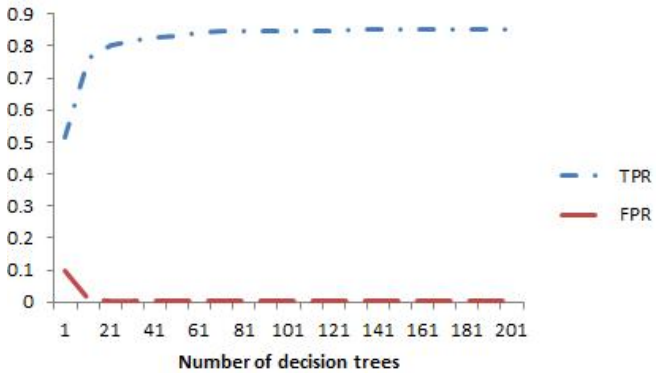


Fig. 4: TPR and FPR in different number of decision trees.

However, the number of RRTs in previous experiments is fixed. Thus, we'd like to vary the K which stands for the number of RRTs used in the model, and

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_45

analyze the influence. Experiments are based on **TCDN** data set and K is varied from 1 to 10. The results of the experiment are shown in Fig. 5. As a result, we find the TPR is nearly the best when K is set to 6 which means the current WFP suffered a lot influence on default values. And when K is set to 1, the feature set degenerate into HTML Marker [13]. The proposed WFP still keeps an excellent score that means HTML document is useful for classifying websites. However, based on the results, non-HTML RRTs seems useless. Therefore, we set up our model with the K-th RRT for only and vary K from 1 to 10. The results are shown in Figure 4 as well. From the figure, RRTs from the 2nd to 4th keep similar importance as the 1st and the performance get worse slightly when K larger than 4. The phenomenon is because of the order for the latter RRTs may change frequently as the multicore processors are used. And another reason is due to more RRTs are filled with default values when K grows up.

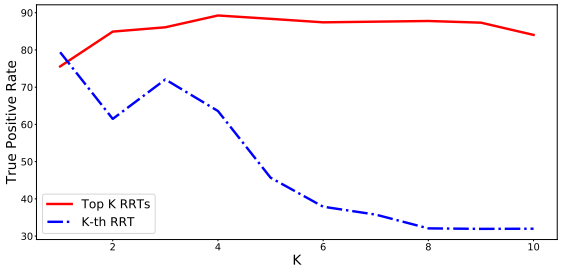


Fig. 5: TPRs of model with the top K RRTs or the K-th RRT.

In order to evaluate RRT+RF, an experiment is set to contrast RRT+RF with the state of the art WFP model. As k-fingerprinting(KFP) [7] has tested on traffic from standard browser which is similar to us and built model by RF as well. So that KFP is chosen to make the comparison, and KFP is implemented by source code shared in Github[9] [7]. However, in default, KFP doesn't take advantage of length related features as all packet length of traffic from Tor are fixed. Furthermore, ACK packets which may cause noise [20]. As a result, KFP is implemented with all features without ACK packets.

The experiments performed in both data sets and the results are listed in Table 3. According to the results, RRT+RF achieves a higher TPR for BAE data set and performs much better than KFP for **TCDN** data set. To summarize, the RRT+RF achieves a better result than KFP and performs excellent in multi-class classification with cloud platforms whose servers often carry numerous websites.

[9] https://github.com/jhayes14/k-FP

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_45

Table 3: Results for RRT+RF and KFP.

Data set	TPR		FPR	
	RF+RTT	KFP	RF+RTT	KFP
BAE	98.62%	94.24%	0.95%	0.65%
TCDN	84.85%	62.89%	0.45%	5.52%

5.3 Unidirectional flows

In an ideal situation, the WFP only needs to deal with the problems bring with encrypted technique. However, in the real world, applying WFP still has to overcome many practical difficulties. Shi et al [16] thought perform traffic analysis will be more and more harder, with the depth of networks because of traffic aggregation. In this section, we'd like to discuss about a practical problem for applying WFP to ISP gateway. The unidirectional flow is common in ISP gateways. Its outgoing and incoming packets appear in different gateways. Faced with such problem, the WFP needs well worked with only request side or server side information as it's hard to match an outgoing trace with another incoming trace exactly. As a result, many features can't be extracted without half information. However, the response data size can be extracted base on the difference between acknowledgement number[10] of packets in the request side like Fig. 6. Finally, we select features which can be extracted in the request side, such as total outgoing bytes and total bytes.

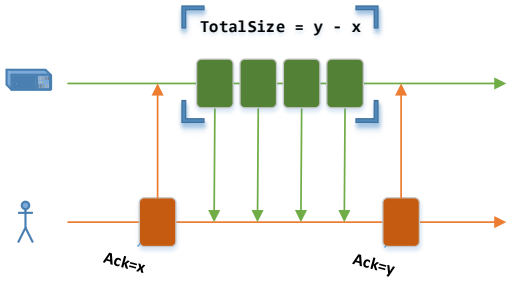


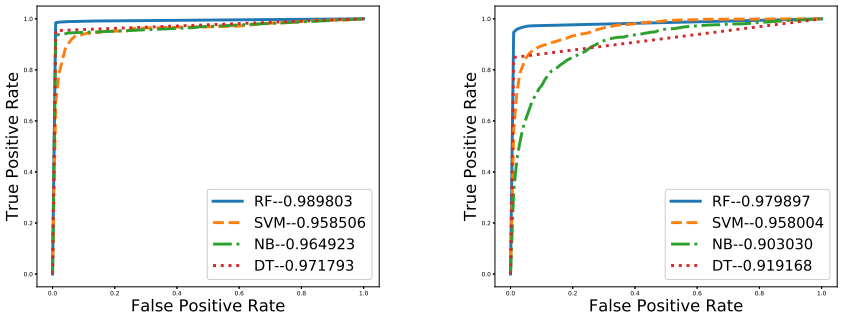
Fig. 6: Evaluate opposite consecutive packet size by acknowledgement number

Experiments conducted with traffic of request side and the results are shown in Table 4 and Fig. 7. From the results, the performance of RF+RRT is still promising on such practical problem. And the performance of The classifiers

[10] Acknowledgement number is a field of TCP protocol and used for tracing transmitted traffic volume.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_45

based on SVM and DT have enhanced a bit, which may because of the number of features has been decreased that reduce the difference between subclasses. According to these results, we found we can only focus on the request side traffic to implement a promising WFP when applying WFP with unidirectional flows, so that we can try to apply WFP the real world.



(a) BAE (b) CDN

Fig. 7: The ROC curve of four classifiers on two data sets only with traffic of the request side.

Table 4: Precision, Recall, F1-score of four classifiers in unidirectional flows.

Algorithm	BAE			CDN		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Random Forest	99.47%	95.09%	97.15%	88.25%	86.8%	86.8%
SVM	97.47%	82.52%	77.76%	64.55%	66.04%	62.71%
Natïve Bayes	93.6%	89.43%	90.7%	15.5%	17.6%	13.51%
Decision Tree	97.1%	95.38%	96.16%	79.35%	85.73%	81.28%

5.4 Real world experiment

Base on the previous experiments, we conduct an experiment on the traffic from several gateways of CSTNet. The traffic on the ISP gateways is unlike the traffic on the local network. First, almost all of the traffic passing by the ISP gateways is unidirectional. Second, the traffic from the same client generally pass by several gateways. Finally, the situation is more open as the traffic may come from arbitrary websites. In such condition, we build a simple webpage on Amazon and collect traffic on a gateway of CSTNet.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_45

To label ISP traffic and evaluate our method, we access our simple webpage for serval times by Selenium and record the traffic as well. With the traffic collected on the client, the Client Random String (CRS)can be extracted in the SSL/TLS handshakes. As the CRS is consisted of 28 bytes, so that it's not easy to find two equal CRS. We labeled the ISP traffic's TASNI whose CRSs have occured in client traffic, and divide all of the TASNI into timeslices for every 10 seconds. Finally, we get a labeled real world data set.

The traffic has been transformed into almost 20000 TASNIs, and about 5% of them are labeled as positive. The RF+RRT first trained with client collected data set, and then test with the ISP data set. The TPR and FPR of the purposed method are 82.88% and 4.9762%, while the ROC curve is displayed in Fig. 8. According to the results, the FPR of the purposed method isn't well, but the overall results are still promising as we are the first to apply WFP on ISP traffic.

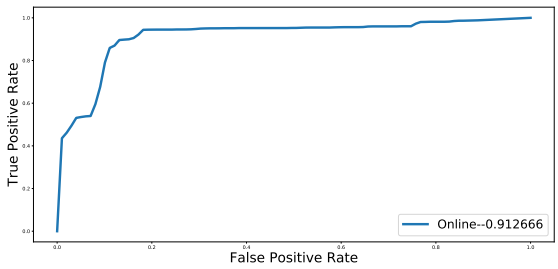


Fig. 8: The ROC curve for the RF+RRT tested on the ISP data set

6 Conclusion

In this paper, we recognize websites with traffic aggregated by SNI and propose a novel WFP model to precisely classify websites on the cloud. The proposed model consists of two parts. One is RRT, a unit to describe a pair of HTTPS request and respond. The other is Random Forest which is a well known machine learning algorithm. From the comparison of the state of the art algorithm, we discovered that a sequence of RRTs can characterize websites better than extract features from the entire traffic roughly. Finally, the purposed method deal with unidirectional flows and applied to ISP traffic, that the proposed method shows a promising performance. In the future, we will improve the feature set of RRT and practically apply to the real world.

Acknowledgment

Ths work is supported by The National Key Research and Development Program of China (No. 2016QY05X1000, No. 2016YFB0801200), The National Natural

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_45

Science Foundation of China (No. 61702501), The CAS/SAFEA International Partnership Program for Creative Research Teams and IIE, CAS International Cooperation Project.

References

1. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
2. Cai, X., Zhang, X.C., Joshi, B., Johnson, R.: Touching from a distance: website fingerprinting attacks and defenses. In: the ACM Conference on Computer and Communications Security, CCS'12, Raleigh, NC, USA, October 16-18, 2012. pp. 605–616 (2012)
3. Cheng, H., Avnur, R.: Traffic analysis of ssl encrypted web browsing. `http://www.cs.berkeley.edu/~daw/teaching/cs261-f98/projects/final-reports/ronathan-heyning.ps`, project paper, University of Berkeley, 1998
4. CNNIC: Global chinese phishing sites report. `http://www.cnnic.cn/gywm/xwzx/rdxw/20172017/201706/P020170609490614069178.pdf`, june, 2016
5. Dyer, K.P., Coull, S.E., Ristenpart, T., Shrimpton, T.: Peek-a-boo, I still see you: Why efficient traffic analysis countermeasures fail. In: IEEE Symposium on Security and Privacy, SP 2012, 21-23 May 2012, San Francisco, California, USA. pp. 332–346 (2012)
6. Google: Google transparency report. `https://transparencyreport.google.com/https/overview`, accessed September 30, 2017
7. Hayes, J., Danezis, G.: k-fingerprinting: A robust scalable website fingerprinting technique. In: 25th USENIX Security Symposium, USENIX Security 16, Austin, TX, USA, August 10-12, 2016. pp. 1187–1203 (2016), `https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/hayes`
8. Herrmann, D., Wendolsky, R., Federrath, H.: Website fingerprinting: attacking popular privacy enhancing technologies with the multinomial naïve-bayes classifier. In: Proceedings of the first ACM Cloud Computing Security Workshop, CCSW 2009, Chicago, IL, USA, November 13, 2009. pp. 31–42 (2009)
9. Juárez, M., Afroz, S., Acar, G., Díaz, C., Greenstadt, R.: A critical evaluation of website fingerprinting attacks. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014. pp. 263–274 (2014)
10. Mistry, S., Raman, B.: Quantifying traffic analysis of encrypted web-browsing. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.10.5823\&rep=rep1\&type=pdf`, project paper, University of Berkeley, 1998
11. Panchenko, A., Lanze, F., Pennekamp, J., Engel, T., Zinnen, A., Henze, M., Wehrle, K.: Website fingerprinting at internet scale. In: 23rd Annual Network and Distributed System Security Symposium, NDSS 2016, San Diego, California, USA, February 21-24, 2016 (2016), `http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2017/09/website-fingerprinting-internet-scale.pdf`
12. Panchenko, A., Mitseva, A., Henze, M., Lanze, F., Wehrle, K., Engel, T.: Analysis of fingerprinting techniques for tor hidden services. In: Proceedings of the 2017 on Workshop on Privacy in the Electronic Society, Dallas, TX, USA, October 30 - November 3, 2017. pp. 165–175 (2017)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_45

13. Panchenko, A., Niessen, L., Zinnen, A., Engel, T.: Website fingerprinting in onion routing based anonymization networks. In: Proceedings of the 10th annual ACM workshop on Privacy in the electronic society, WPES 2011, Chicago, IL, USA, October 17, 2011. pp. 103–114 (2011)
14. Quinlan, J.R.: Induction on decision tree. Machine Learning **1**(1), 81–106 (1986)
15. Sherry, J., Lan, C., Popa, R.A., Ratnasamy, S.: Blindbox: Deep packet inspection over encrypted traffic. Computer Communication Review **45**(5), 213–226 (2015)
16. Shi, Y., Biswas, S.: Website fingerprinting using traffic analysis of dynamic webpages. In: IEEE Global Communications Conference, GLOBECOM 2014, Austin, TX, USA, December 8-12, 2014. pp. 557–563 (2014)
17. Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural Processing Letters **9**(3), 293–300 (1999)
18. Trevisan, M., Drago, I., Mellia, M., Munafò, M.M.: Towards web service classification using addresses and DNS. In: 2016 International Wireless Communications and Mobile Computing Conference (IWCMC), Paphos, Cyprus, September 5-9, 2016. pp. 38–43 (2016)
19. Wang, T.: Website Fingerprinting: Attacks and Defenses. Ph.D. thesis, University of Waterloo, Ontario, Canada (2015), `http://hdl.handle.net/10012/10123`
20. Wang, T., Cai, X., Nithyanand, R., Johnson, R., Goldberg, I.: Effective attacks and provable defenses for website fingerprinting. In: Proceedings of the 23rd USENIX Security Symposium, San Diego, CA, USA, August 20-22, 2014. pp. 143–157 (2014), `https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/wang_tao`
21. Wang, T., Goldberg, I.: On realistically attacking tor with website fingerprinting. PoPETs **2016**(4), 21–36 (2016)
22. Zolnierek, A., Rubacha, B.: The empirical study of the naive bayes classifier in the case of markov chain recognition task. In: Computer Recognition Systems, Proceedings of the 4th International Conference on Computer Recognition Systems, CORES'05, May 22-25, 2005, Rydzyna Castle, Poland. pp. 329–336 (2005)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_45