SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM

TNC TRANSACTIONS ON NETWORKS AND COMMUNICATIONS

VOLUME 5, NO. 6
ISSN: 2054 -7420

Hierarchy Website Fingerprint Using N-gram Byte Distribution

[1]Mohammed Aldarwbi, [2]Essa Shahra
[1,2] Computer Engineering, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia
m.aldarwbi@gmail.com; eissa.qassim@gmail.com

ABSTRACT

According to www.internetlivestats.com, there are over one billion websites on the world wide web (WWW) today while in 1991, there were only one single website. Websites classification based on traffic analysis has become a difficult problem due to the large number of websites within the internet. All the proposed approaches in the literature could not classify more than 100 websites which is a very trivial number compared to the total number of websites over the internet. In this paper, a two-level websites' classification technique is proposed. At the first level, the traffic is classified to a general category such as sports, news, social, healthy, education, etc. Then, for further information the packet could be classified within the same category to identify from which websites the packet came.

Keywords: Website fingerprinting; Traffic analysis; N-gram byte distribution.

1 Introduction

Digital forensics is considering as extremely youthful science as the number digital crimes have been increased dramatically. The new emerging digital forensics issues needs creative solutions to be utilized by the investigators to achieve their work in the optimal way. Network forensics issues are considered as the most difficult issues in digital forensics as investigator endeavors to recreate or comprehend occasions from the data observed in the network (network traffic). Network forensics enables us to make measurable decisions in view of the captured traffic, which might be significant over the span of an investigation [1].

Network forensics defined in DFRWS as "use of scientifically proven techniques to collect, use, identify, examine, correlate, analyze, and document digital evidence from multiple, actively processing and transmitting digital sources for the purpose of uncovering facts related to the planned intent, or measured success of unauthorized activities meant to disrupt, corrupt, and or compromise system components as well as providing information to assist in response to or recovery from these activities.". The point of the analysis is normally to build up abnormal facts of truths, for example, attribution, aim, personality, timetables and other data which might be important to the study case network forensic.

Network administrators use network forensic analysis tools (NFAT) to monitor network, capture network traffic, play main role in network crime investigation to assist and help in generating appropriate decision of an incident. In addition, NFATs help in investigating the insider burglary and abuse of assets, anticipate attacker goals in near future, perform risk estimation, assess network achievement, and help to secure

DOI: 10.14738/tnc.56.3767
Publication Date: 09th November 2017
URL: http://dx.doi.org/10.14738/tnc.56.3767

intellectual proprietary. NFATs collect the entire traffic of network, provide users the ability to analyze network traffic according to their need and try to discover and find significant features about the traffic [2].

From a forensic perspective, we are regularly concern more about high level information, than network protocol information. For example, in a normal forgery case we might be concerned in the content of the packet itself sent through the network rather than how the packet was sent (e.g. using instant messenger, email or web pages) [3]. Website fingerprint is an attack of traffic analysis running by a local eavesdropper, its goal is to infer information about the visited website by user by defining a feature of data flow. The attacker use meta information, such as traffic direction, number of packets, packet size or the content of packet as we did for website Fingerprint [4].

In this paper, a two-level website classification technique is presented. The first level is classifying the traffic to a general category of websites like sport, news, social, etc. The second level of classification can be used for further information; the packet could be classified within the same category to identify from which websites the packet came. We build our own dataset by controlling google chrome browser automatically and visiting a set of selected websites in each category. We utilize the power of Selenium python library in the process of collecting the traffic. N-gram analysis is used in the classification. Unlike the literature, we based on the payload of the packets not the header.

The rest of the paper is organized as follows. The literature review is mentioned in section 2. Our scheme is described in section 3. Building data set is presented in section 4. Experiments and analysis is presented in section 5. finally, the conclusion of our work.

2 Related Work

In [4], the authors imply that when the encrypted packets traverse the tunnel in the uplink direction and in the absence of clients' activity information, the attacker can detect the packet timestamp easily by exploiting the packet timestamp.

The attacker can with high probability guess the websites that the client visit. To classify the timestamp sequences, they use K-Nearest neighbors and Naive Bayes Classification. The proposed work is timing-only attack while the others focus on the packet count and packet size information. The work in [5] focus on the network traffic of five popular websites namely, YouTube, Gmail, Skype, Facebook and Gmail video chat. The traffic features that have been extracted are bandwidth, inter-arrival time, average packets sent/received per second, and packet length. During their investigation, they noticed that each website has different traffic with respect to different web browser. They use as more features as possible. They prove that each website has different traffic with respect to different web browser.

Gong in [6] proposed work is trying to prove that the remote traffic analysis could be used by eavesdroppers. The adversary can identify the websites that a remote user is accessing by knowing his/her IP address. Their classification is performed using Dynamic Time Warping (DTW), which is a method used to find an optimal alignment between two temporal sequences (time-dependent). Then, the process of matching the user's traffic to the previously collected traffic is done by the k-nearest neighbor (k-NN) algorithm. Instead of monitoring and analyzing the victim traffic patterns by capturing the traffic from the same LAN, this work carries it out remotely by exploiting the queuing side channel in the routers.

In [7] studied an attack is based on forming profiles for the most visited websites and matching the traffic against these profiles. They collect the traffic of the most visited websites by their department users (24 volunteers for 214 days). The features that composed the profile are the inter-arrival time distribution and packet size.

Lu and Chan in [8] handle two approaches, classification and detection. The first scenario for given dataset which is known to be a visited website and its objectives is to recognize the site. This is called Classifications. The second situation: for given dataset, decide if it is a visit or not if visited to a site and recognize the site that was visited. they propose an effective strategy that uses packet requesting data for site fingerprinting. They utilize noisy requesting data instead of simply the distribution of packet size.

In [9] they assess traffic analysis approaches that derive the wellspring of a site page recovered under the cover of an encipher passage and their approaches distinguish sources by differentiate experimental traffic with profiles of known website made from packet lengths, and are referred to as profiling attack. In [10] they propose new approach utilizing the aggregated whole of packet sizes as the abstract representation furthermore, to test a constant number n of extra features from this implementation and analyze one type of attack detection (remote traffic) that operates against local network clients. By observing the queue delay of request packet, they observe that it is available to extract the router's queue state. The attacker can estimate the packet size, time arrival, and several packets delivering at the router. They utilize the total of packet sizes in the queue instead of the size of the packet itself. Utilize diverse condition for gathering the data set.

2.1 N-gram Distribution

In this section, we describe the N-gram distribution approach used to classify the visited websites. n- gram is a series of contiguous items from a given streams. N-gram distributions have been applied in different applications, and it is easy to understand and implement, and get more accuracy. For each packet n-gram is computed by extracting the content of the paced (payload) and counting the number of appearance for each gram for example using 1-gram one byte from the packet will counts its occurrences and so on for each gram. Determining the size of gram depends on the used applications, the complexity of computation is increased exponentially as the size of gram increase [11].

Our approach worked rely on computing and comparing n-gram frequencies profiles. First, we use the n-gram distribution to compute websites profiles form training dataset which represent different websites category such as healthy, news, and social websites. Then the system calculates a profile for each website that needs to be classified. Finally, the system calculates a distance between web packet and each of profiles category. The system selects the category with smallest distance to website.

3 Data Collection

We have conducted our experiments using three groups of different datasets. The data collected using our own code that visit the website automatically in which for each cycle it visits all websites in our lists, for each visit several packets are collected through network using ***tshark*** tool. The following figure (1) represent the algorithm of data collection.

4 Implementations

According to www.internetlivestats.com there are over 1 billion websites on the world wide web today while in 1991 there were only single website. Websites classification by analyzing the traffic become a hard problem due to the large number of website. All the proposed approaches in the literature could not classify more 100 websites which is nothing comparing to the total number of websites over the internet.

We propose a two-level classification technique. At the first level, the traffic is classified to a general

```
For each round
    For each website
        For each visit
            - Open new google chrome( no cache)
            - Open web page
            - Run tshark for traffic capturing
            - wait  20 second
            - Stop tshark
            - close chrome
        End
    End
End
```

Figure 1: Data collection algorithm

category like sports, news, social, healthy, education, etc. Then, for further information the packet could be classified within the same category to identify from which websites the packet came.

4.1.1 First level classification

We collect traffic for three distinct categories which are healthy, news, and social websites. The first category is for the most visited health website which is used in figure (6) which are asthmacare.ie, kingfisherclub.com, whitefeatherhealing.com, psychotherapy.com, and hse.ie respectively. The second category is for the most visited news websites which are www.cnn.com,www.foxnews.com, www.reuters.com, www.cnbc.com, www.cbc.ca according to www.alexa.com. The third category is for the well-known social websites which are (www.facebook.com, www.twitter.com, and www.instgram.com) according to www.alexa.com. Then, the collected traffic has been divided into two parts which are training and test. About 80 of the traffic is taken as the training. The first level classification model is presented in figure (2). Then the BFD is used to build the websites category profile as it is shown

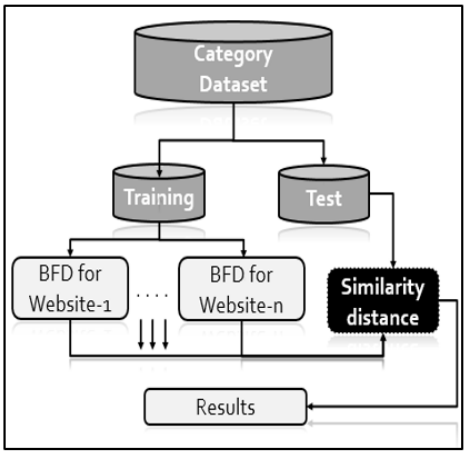

Figure 1: Second Level Website classification

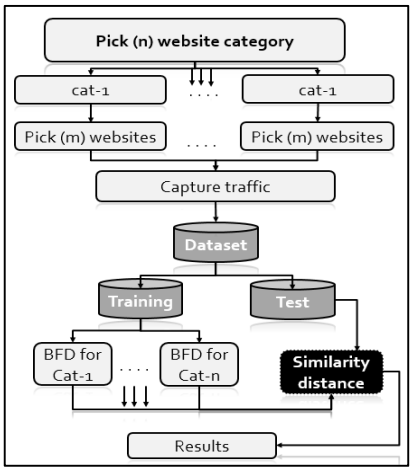

Figure 2: First level model

in figure (3). The BFD presented in table (1) is a result of 3-gram analysis. The last 20% of the dataset is used for prediction, we find similarity distance between the tested packet and the website category profiles to find to which website category profile it belongs. As it is shown in table 1, 3-gram analysis provide high accuracy.

Table 1: Websites category accuracy

Group	Accuracy
Health	94.58 %
News	90.3 %
social	99.89

4.1.2 Second level classification (Http)

After the category, has been identified, the packet could be classified within the category itself. Each category has number of websites, n-gram analysis is used to classify the packet to which website it belongs within the same category. The classification model for this level is presented in figure (3). We propose two approaches in classifying the packet, the first one is without filtering the packets. The second one is after we eliminate the images and videos from the traffic, the two approaches will be presented in the following section.

Website classification with images

The model of this approach is shown in figure 5. We take the whole packet payload in building the website profile. We do the training and test for each category websites separately. For the training dataset, we build BFD for each website. Similarly, we build the BFD for each website in news category. Once the training phase for each category websites complete, the test phase is beginning. The following tables present the results of websites classification within each category.

a. Healthy websites results

We implement 1-gram, 2-gram, and 3-gram in search for the highest accuracy that n-gram analysis can provide. We end up with that, 3-gram provide the highest accuracy as it is show in table 2. The results in Table (2). shows the accuracy of healthy websites within three different n-gram. As can be noticed from this table, the accuracy is increased as the gram increased significantly. However, the results still not good

enough even with 3-Gram. The results of 3-gram, in the Table (2), shows that some websites classification accuracy are excellent (93%) while some of them are not relatively good (39%). The reason behind this fluctuating is due to the variation in the websites content. Most of these websites has images which affect the profiling process.

Table 2: Healthy websites accuracy with images

Websites	1- gram analysis	2-gram Analysis	3-Gram analysis
asthmacare.ie	20%	66%	74%
kingfisherclub.com	84%	93%	93%
Whitefeatherhealing.com	34%	58%	51%
psychotherapy.com	20%	58%	66%
Hse.ie	2%	29%	39%

In the next section we tried to eliminate all packets that include images and the accuracy of website classification is increased dramatically after. As in the table (2) 3-gram analysis provide the best result in identifying the websites.

a. News websites results

In this section, 1-gram, 2-gram, and 3-gram is implemented in search news websites as it is shown in the following table (3). It can be noticed from Table 3 that the accuracy is significantly increases as going from low gram to higher one. However, the results still not good enough even with 3-gram. This is because the content of the packets is compressed which make the distribution looks random

Table 3: News websites accuracy with images.

Websites	1- gram analysis	2-gram Analysis	3-Gram analysis
Cnn.com.	1%	14%	5%
foxnews.com	34%	33%	76%
retuters.com	37%	40%	38%
cnbc.com	72%	65%	59%
cbc.ca	2%	1%	2%

Website classification without image

Regarding the classification of websites without imaging, it was noticed that the results of identifying not good enough since most of the websites has a lot of image which makes the decision in website classification looks random. To solve this problem, we proposed another approach to overcome the decision randomness. The main idea behind the proposed approach is by filtering the images from the traffic.

a. Healthy websites results

The results in Table 4 shows the accuracy of healthy websites identification within three different n-grams. It can be noticed in table (4) how the accuracy is increased after eliminating the images which provides excellent accuracy with 3-gram which gets 60% of the websites with 100% accuracy, while the lower accuracy was 61%.

Table 1: Healthy websites (without images)

Websites	1- gram analysis	2-gram Analysis	3-Gram analysis
asthmacare.ie	100%	100%	100%
kingfisherclub.com	100%	100%	100%
Whitefeatherhealing.com	100%	100%	100%
psychotherapy.com	%0	52%	61%
Hse.ie	56%	71%	80%

b. News websites results

In this part of the simulation, we implemented 1-gram, 2-gram, and 3-gram in search news websites as it is shown in the following table (5). The results presented in table 5. shows the accuracy of news websites within three different n-gram. It can be noticed that filtering the packets from images and videos enhanced the classification performance significantly. As seen in table (5), high accuracy with 3-gram with 100% is provided for best case and 34% for worse case.

Table 2:News websites (without images)

Websites	1- gram analysis	2-gram Analysis	3-Gram analysis
Cnn.com.	91%	100%	100%
foxnews.com	0%	14%	34%
retuters.com	0%	45%	45%
cnbc.com	19%	57%	64%
Cbc.ca	85%	86%	86%

Second level (HTTPS) websites fingerprint

All the results in previous sections was with Http traffic. However, in this section we used our approach with Https traffic and the results are showed in table (6). The results in table 6 shows a good result for Https traffic using 3-gram, unless for google.com this is due to the natural design of google web page that has a less content.

Table 6:Https websites accuracy

Websites	1- gram analysis	2-gram Analysis	3-Gram analysis
google	19%	2%	3%
Facebook	42%	52%	54%
amazon	54%	58%	59%
instagram	30%	40%	42%
cbc.ca	64%	65%	64%

5 Conclusion

In this paper, we presented two level of website classifications, the first level is classifying the traffic to a general category and the second level of classification can be used for further information; the packet could be classified within the same category to identify from which websites the packet came. The results showed that byte frequency distributions can be used to classify the website with a high accuracy in different types of n-gram size. The results showed that 3-gram provides more accuracy for both level of classification; category and websites.

REFERENCES

[1] M. Cohen, "PyFlag–An advanced network forensic framework," Digital investigation, vol. 5, pp. S112-S120, 2008.

[2] E. S. Pilli, R. C. Joshi, and R. Niyogi, "Network forensic frameworks: Survey and research challenges," digital investigation, vol. 7, pp. 14-27, 2010.

[3] K. Karampidis and G. Papadourakis, "File Type Identification for Digital Forensics," in International Conference on Advanced Information Systems Engineering, 2016, pp. 266-274.

[4] S. Feghhi and D. J. Leith, "A Web Traffic Analysis Attack Using Only Timing Information," IEEE Transactions on Information Forensics and Security, vol. 11, pp. 1747-1759, 2016.

[5] S. S. Kowsalya, "Website Fingerprinting using Traffic Analysis Attacks."

[6] X. Gong, N. Borisov, N. Kiyavash, and N. Schear, "Website detection using remote traffic analysis," in International Symposium on Privacy Enhancing Technologies Symposium, 2012, pp. 58-78.

[7] G. D. Bissias, M. Liberatore, D. Jensen, and B. N. Levine, "Privacy vulnerabilities in encrypted HTTP streams," in International Workshop on Privacy Enhancing Technologies, 2005, pp. 1-11.

[8] L. Lu, E.-C. Chang, and M. C. Chan, "Website fingerprinting and identification using ordered feature sequences," in European Symposium on Research in Computer Security, 2010, pp. 199-214.

[9] M. Liberatore and B. N. Levine, "Inferring the source of encrypted HTTP connections," in Proceedings of the 13th ACM conference on Computer and communications security, 2006, pp. 255-263.

[10] Panchenko, F. Lanze, A. Zinnen, M. Henze, J. Pennekamp, K. Wehrle, et al., "Website fingerprinting at internet scale," in Network & Distributed System Security Symposium (NDSS). IEEE Computer Society, 2016.

[11] W.-J. Li, K. Wang, S. J. Stolfo, and B. Herzog, "Fileprints: Identifying file types by n-gram analysis," in Information Assurance Workshop, 2005. IAW'05. Proceedings from the Sixth Annual IEEE SMC, 2005, pp. 64-71.