TCPI: A Novel Method of Encrypted Page Identification

Wei Xia
School of Control and Computer Engineering, North China Electric Power University
Beijing, 102206, China
xiawei@ncepu.edu.cn

Yan Ren
CNCERT/CC China
Beijing, 100029, China
renyan@cert.org.cn

Zhenlong Yuan
Department of Automation, Tsinghua University,
Beijing, 100084, China
yuanzl11@mails.tsinghua.edu.cn

YiboXue*
TsinghuaNationalLaboratoryforInformationSciceandTechnology,
Beijing,100084,China
yiboxue@tsinghua.ed

***Abstract*—Encrypted Page Identification (EPI) has been increasingly attracted much attention in recent years. Traditional methods now face challenges due to the emergence and development of HTTPS and Cloud Computing. To address this issue, we analyze HTTPS communication mechanism and present algorithm of Calculation of Timing Characteristics (CTC) to extract the Timing Characteristics(TC) as identification signature of encrypted page from time feature of traffic. Based on the CTC, a novel framework called Timing Characteristics-based Page Identification (TCPI) for EPI is proposed in this paper. Additionally, a complete methodology to verify the approachfor precise EPI is designed based on real large-scale datasets. The experimentalresults demonstrate the efficacy and accuracy of the identification method.**

Keywords-Encrypted Page Identification; Timing Character -istics;Calculation of Timing Characteristics;

I. INTRODUCTION

The World Wide Web has been the most important provision of services for information retrieval and resources sharing. Billions of people are intending to browse the Web due to its easy-to-use and flexible connection in order to get information and communicate with others. With the rapid development and extensive adoption of the Web, it is increasingly critical to guarantee the quality and legality of information on the Web. Users especially for the teenagers, who may be vulnerable to harmful information, malicious attack and privacy leakage, call for supervision and protection in web browsing. Additionally, as the Web browsing traffic is ever-growing, network management calls for page identification techniques to optimize resource allocation and advance the quality of service. Consequently, our study focuses on finding a new method of addressing the issue of encrypted page identification.

Although traditional page identification methods have ever achieved certain successes in page identification, they are no longer effective due to two major challenges brought by the emergence of HTTPS and cloud computing.

- **HTTPS:**The extensive use of Security Socket Layer, and its successor Transport Layer Security [4],which devotethemselves to guarantee the data security and information integrityby encrypting the payload by the cipher algorithms, is the most important challenge for page identifying. The plaintext related to pages is invisible owing to the encryption. Consequently, we cannot utilize the method of analyzing the content of payload to get the detail information unless using the brute force attack which is very time-consuming.
- **Cloud Computing:**The advent of cloud computing has set off a revolution. As a result, not one but a group of IP addressesmay provide the same Web service or responseidentical request for the Web site. Meanwhile, one IP address may simultaneously support amount of pages and resources. Hence it isnot realistic to separatespecific pages from the others simply by the IP address.

In order to address the issues above, we propose a novel method called Timing Characteristics-based Page Identifi -cation (TCPI), which adopts time feature of trafficand extractsTiming Characteristics as identification signature of encrypted Web page.

The highlights of the paper include:

1) Based on the recovery of time-related self-similarity of one encrypted page, as well as the discrepancy of different pages, we analyze the mechanism of communication between the server and client. We find out certain heuristic features and proposean algorithm named Calculation of Timing Characteristics (CTC) to obtain identification signature for identifying encrypted pages.

2) On the basis of presentation of the Algorithm, We further give the complete description of the framework called Timing Characteristics-based Page Identification (TCPI), in order to achieve the goal of EPI.

3) We verify and evaluate the proposed method by groups of experiment. The results demonstrate that the approach has achieved high efficacy and accuracy via the metric of accuracy and time. Our approach works well on identifying encrypted pages.

The rest of this paper is organized as follows. Section IIpresents the related work on page identification. Section III analyzes the mechanism of the encrypted web browsing and proposes the core algorithm.Section IVputs forward the full view of the framework based on the algorithm. Section Vevaluates our approach experimentally. At last, we conclude this paper and discuss the future work.

II. RELATED WORK

As the encrypted Web browsing traffic takes up increasingly significant portion of the whole network traffic, the researchers have made great efforts to solve the problem through different approaches.

*Corresponding author. E-mail: yiboxue@tsinghua.edu.cn

One way of achieving EPI is traffic analysis attack [5-8]. Felten et al. [5] put forward an attack via the browser caching features of the user's. In the work of Gilbert Wondracek et al. [6], for tracking users on social network, they present a novel de-anonymization attack to server for stealing web browsing history. Cheng et al. [7] and Mistry et al. [8] demonstrated that Web pages browsed using SSL could be identified using simple statistical characteristics. But the methods are feasible only on the premise of using specific Web server accessed by the users.

Another way is to take the length information of packets as a key feature to identify the Web page [10-14]. Sun et al. [10] deeply investigated the identification ability of encrypted Web-browsing traffic based on HTTP object count and sizes. Bissias etal. [11] presented a straightforward identification by creating a profile with the statistical characteristics of Web requests. Lu etal. [12] pointed out that packet ordering information could be utilized to enhance website fingerprinting as well. Danezisetal. [13] not merely took advantage of the data length information leaked by HTTP transactions over TLS protocol but also introduced the Hidden Markov Model to analyze the sequences of user requests and then found the most plausible resources accessed. Yu et al. [14] also established a specific hidden Markov model, then used this model to identify the optimal sequence of the accessed pages.

Different from them, this paper presents a page identification method which uses time feature after making a comprehensive analysis of the mechanism of communication between the server and client connected through HTTPS.

III. OBSERVATION, ANALYSIS AND ALGORITHM

As mentioned above, our work is proposed based on the time feature of the traffic. Our study draws the following conclusions via series of observation and analysis.

A. *Experimental Observation*

Although the pages are encrypted for concealing the payload, it still exists special attributions among them. For instance, as shown in Fig. 1(a) and Fig. 1(b), the curves may depict the tendency distinction in the case of visiting twoencrypted pages. The horizontal axis denotes packet numbers

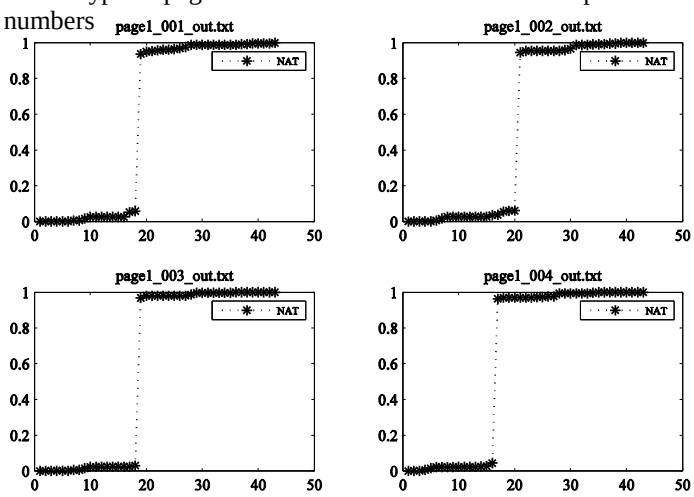

Figure 1. (a)Timing Tendency of Normalized Arrival Time of Page 1.

passing on the network, while the vertical axis represents normalized arrival time. We have found that within each side of the two pages, the tendency of each curve keeps similar with the others and relatively constant, which means there exists self- similarity in accessing the same page repeatedly. In addition, significant difference between the two pages also can be found via the different curve tendency in the figures. With assumption that the network environment is reliable and without malicious attack, we figure out that it would be feasible to identify the encrypted pages by certain identifi -cation signature based on the time sequence.

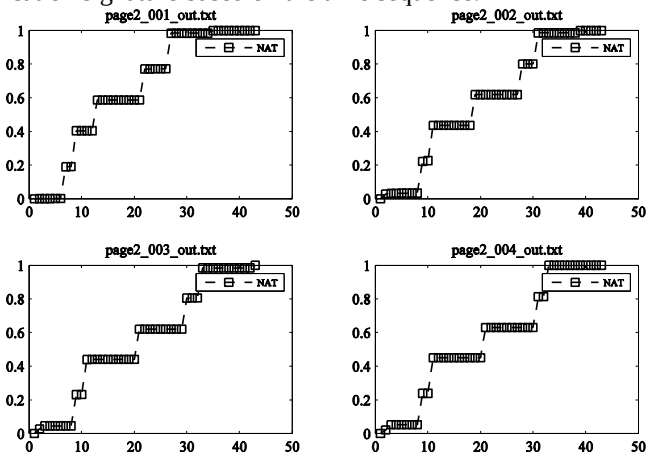

Figure 1. (b)Timing Tendency of Normalized Arrival Time of Page 2.

B. *HeuristicAnalysis*

After having done abundant of analyses and experiments, we give hypothesis by analyzing the mechanism of the process of encrypted page browsing between the server and client.

Taking account of communication between the server and client, we figure out the reason for the discrepancies of pages. Despite of encryption that results in disability to decode in the application protocol layer, we could start with the procedure of request-response of HTTP protocol. First, the client submits an initial request for the page. Then the server which provides resources returns an initial response for the initial request containing the HTML document of the page what we call the root page. The HTML document may consists of other content such as JavaScripts, CSSs or images. All of these embedded objects would be requested by the browser during receiving and parsing the initial response HTML. Suppose that the HTML codes of varieties of pages differentiate from others, so that the root page and follow up objectsin initial response would be different. Consequently, these differences result in different arrival time sequences when we focus on the timing feature of aggregated flow from the server to the client. Coincidentally, although the application data that encrypted and encapsulated by the SSL or TLS protocol, the conclusion from the analysis is still applicable

The aforementioned is the analytical basis of our heuristic work. Since there are similarities among every-time accessing to the same page, we pay attention to several characteristics as mentioned in the work of More of the encrypted traffic and try our best to find out certain typical characteristic experimentally. Fortunately, we find out that

the inter-arrival time could meet our demand. In fact, the network latency may be under constraint of the speed or bandwidth in the different network environment. Taking this issue into account, we convert the inter-arrival time sequence into its change rate (also called gradient) to eliminate the impact of the external factor. We employ gradient sequence as our Timing Characteristics (TC) to describe the relation which is specifying arrival time in terms of packets. That is, TC turns into representation for the patterns of pages.

To indicate the relevance between the two pages, measuring the similarity is reasonable. In light of the fact that the gradient sequence can be regarded as the trait of a page, we can use them to measure similarity of two pages. As for the gradient is time-related, the similarity can be depicted by certain distance metric which is based on the time sequences.

C. *CTC Algorithm*

- Basic Idea

Assume that the time sequence $\boldsymbol{T}=(t_1,t_2,\ldots,t_i,\ldots,t_L)$ represents for $\boldsymbol{L}$ packets in the aggregated flow from the server to client when client is visiting anencrypted Web page, where t_i stands for the arrival time of the *ith* packet. Then we set every $\boldsymbol{C}(1<C\leqslant\lfloor(L-1)/2\rfloor+1)$packets to as the time interval.

With the time interval of every $\boldsymbol{C}$packets calculated, the $\Delta\boldsymbol{T}=(\Delta t_1,\Delta t_2,\ldots,\Delta t_j,\ldots,\Delta t_N)$ means the inter-arrival time came from the$\boldsymbol{T}$, where the length $\boldsymbol{N}$can be expressed by$\boldsymbol{L}$and $\boldsymbol{C}$as follows:

$$\boldsymbol{N}=\lfloor(\boldsymbol{L}-1)/(\boldsymbol{C}-1)\rfloor. \quad (1)$$

In the sequence $\Delta\boldsymbol{T}$,Δt_ican be expressed by $\boldsymbol{T}=(t_{f(1)},t_{f(2)},\ldots,t_{f(i)},\ldots,t_{f(L)})$as Eq. (2) shows:

$$\Delta t_j=t_{(C-1)*(j+1)-(C-2)}-t_{(C-1)*j-(C-2)},\ j\in 1,2,\ldots,N \quad (2)$$

And the gradient $\boldsymbol{M}=(m_1,m_2,\ldots,m_k,\ldots,m_{N-1})$, which represents the change rate of inter-arrival time,can be expressed as Eq. (3):

$$m_k=(\Delta t_{k+1}-\Delta t_k)/(k+1-k)=\Delta t_{k+1}-\Delta t_k, k\in 1,2,\ldots,N-1 \quad (3)$$

Then we use the sequence $\boldsymbol{M}=(m_1,m_2,\ldots,m_i,\ldots,m_{N-1})$ as the signature to represent the visited page.

- Introduction to the Algorithm

At the first, the terminology is given as TABLE I shows.

TABLE I. TERMINOLOGY

Terminology	
Name	**Remark**
Parameter $\boldsymbol{P}$	$\boldsymbol{P}$ means how many encrypted pages we focus on
Parameter $\boldsymbol{L}$	$\boldsymbol{L}$ means how many packets we concern about.
Parameter $\boldsymbol{C}$	$\boldsymbol{C}$ means the packet number of interval for timing.
Parameter $\boldsymbol{N}$	$\boldsymbol{N}$ means the length of TC sequence which is determined by the $\boldsymbol{L}$ and $\boldsymbol{C}$
Time Sequence $\boldsymbol{T}$	$\boldsymbol{T}$ means the arrival time sequence of traffic $\boldsymbol{T}$ can be expressed as $\boldsymbol{T}=(t_1,t_2,\ldots,t_i,\ldots,t_L)$
Time Sequence $\Delta\boldsymbol{T}$	$\Delta\boldsymbol{T}$ means the inter-arrival time sequence $\Delta\boldsymbol{T}$ can be expressed as $\Delta\boldsymbol{T}=(\Delta t_1,\Delta t_2,\ldots,\Delta t_i,\ldots,\Delta t_N)$
Sequence $\boldsymbol{M}$	$\boldsymbol{M}$ means the gradient of $\Delta\boldsymbol{T}$ $\boldsymbol{M}$ can be expressed as $\boldsymbol{M}=(m_1,m_2,\ldots,m_i,\ldots,m_{N-1})$
Pattern Set $\boldsymbol{PS}$	$\boldsymbol{PS}$ means the pattern set of process of identification PS can be expressed as PS=$(M_1,M_2,\ldots,M_j,\ldots,M_P)$

We propose an algorithm named Calculation of Timing Characteristics (CTC) to calculate Timing Characteristics (TC) for generating identification signature. The algorithm is shown in Algorithm 1 in TABLE II.

TABLE II. ALGORITHM OF CTC

Algorithm 1 Algorithm of CTC

Input:
Packet number :$\boldsymbol{L}$
Packet interval:$\boldsymbol{C}$
Arrival time sequence of encrypted pages: $\boldsymbol{T}=(t_1,t_2,\ldots,t_i,\ldots,t_L)$
Output:
1. Calculate the length of inter-arrival time sequence
$\boldsymbol{N}=\lfloor(L-1)/(C-1)\rfloor$
2. Normalize the arrival time sequence$\boldsymbol{T}$.
3. Calculate the first inter-arrival time $\Delta t_1=t_C-t_1$
4.**for**i = 2:N
5. $\Delta t_i=t_{(C-1)*(i+1)-(C-2)}-t_{(C-1)*i-(C-2)}$
6. $m_{i-1}=\Delta t_i-\Delta t_{i-1}$
7.**end for**
8.return$\boldsymbol{M}$

From CTCalgorithm, we can automatically obtain Timing Characteristics(represented bygradient sequence) of various different pages, then they can be used as signatures for identifying.

IV. TCPI FRAMEWORK

In this section, we propose the Timing Characteristics -based Page Identification (TCPI) framework as shown in Fig.2.

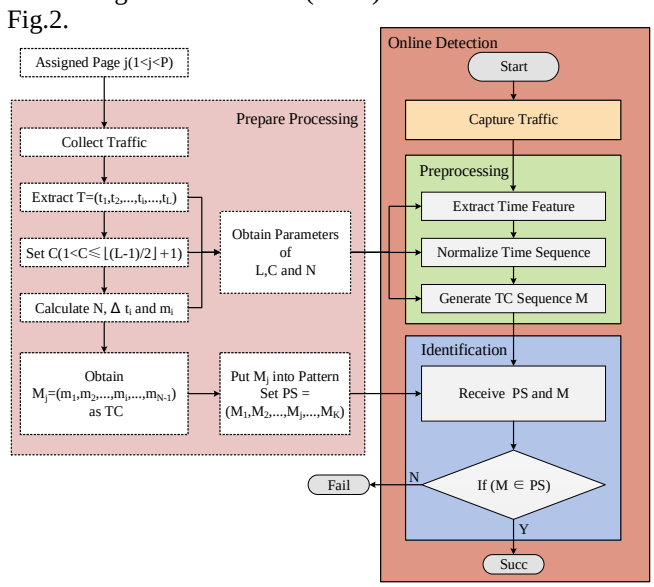

Figure 2. Overview of TCPI Framework

The whole process includes two parts: Prepare Process -ing and Online Detection. The former is offline, it mainly generates the time characteristics signature for various differ -ent pages; while the latter is for real-time *Online Detection*.

A. *Prepare Processing*

This part deals with preparing parameters and generatingpattern set for the online identification.

First of all, we select $\boldsymbol{P}$ pages to be identified as assigned pages. Then we collect traffic by visiting every *Assigned*

Pagej. Then get $\boldsymbol{L}$ packets and extract the time sequence from the traffic. After that, Set $\boldsymbol{C}$, then calculates $\boldsymbol{N}$.

These three parameters determine Δt_i and m_i. The Obtained Parametersare passedto *Online Detection* part.

Meanwhile, the *Obtaining*$\boldsymbol{M_j}$ module gets the three parameters and generates the TC sequence $\boldsymbol{M_j}$ for the *Assigned Page j* by Algorithm CTC. Then $\boldsymbol{M_j}$ will be put into the pattern set $\boldsymbol{PS}$. With the $\boldsymbol{P}$ pages performed, we can obtain a pattern set $\boldsymbol{PS}$= ($\boldsymbol{M_1}$,$\boldsymbol{M_2}$,…,$\boldsymbol{M_j}$,…,$\boldsymbol{M_p}$), which is used to be the identification signature for identifying.

B. *Online Detection*

This part is designed for the real time online page identification, including three modules.

Capture Traffic is used for capturing and filtering the online traffic.

Preprocessing firstly receives the parameters $\boldsymbol{L}$, $\boldsymbol{C}$ and $\boldsymbol{N}$ from the *Prepare Processing* part, then extracts and normalizes time sequence, lastly generates TC sequence $\boldsymbol{M}$ by CTC algorithm.

Identification receives the pattern set $\boldsymbol{PS}$ from *Prepare Processing* and the TC sequence $\boldsymbol{M}$ from the previous module. Then match $\boldsymbol{M}$ with PS. If it is yes, the identification successes, otherwise it fails.

V. EXPERIMENTS AND EVALUATIONS

A. *Experiments Description*

To verify the proposed framework, we have set up several groups of experiments. We choose 100 encrypted pages randomly from Wikipedia by its URL: *https://en.wikipedia.org/wiki/Special: Random*.

Besides, we have conducted the experimentson several groups with different combination of parameters $\boldsymbol{L}$and $\boldsymbol{C}$ in order to find out the optimal for the framework.

B. *Experiments and Evaluation*

- Accuracy

We have launched multiple groups of experiments with different combination of parameters. Here we give the average accuracy of online identifying in Fig. 3. The average accuracy of 500 pages is more than 90%. It demonstrates that TCPI can achieve high effectiveness.

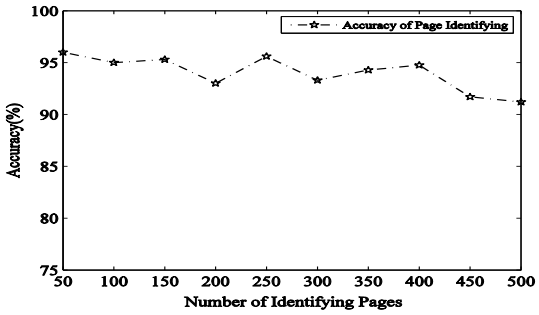

Figure 3. Accuracy of Online Identifying

- Time

To assess the speed of page identification, we test the timing consumption. It can reach 300MbpsunderPentium 4 3.0GHz and 1G DDR. It can be deployed online.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel method TCPI which focus on the timing characteristics in form of complete framework. We verify and evaluate our method for page identification with high accuracy and speed. According to the results, we would focus on the optimization of generating TC sequence and automated pattern update which can solve the issues of efficiency advance and pages variation.

ACKNOWLEDGMENT

This work was supported by the National Key TechnologyR&D Program of China under Grant No.2012BAH46B04.

We would like to thank the reviewers for their insightful comments.

REFERENCES

[1] World Wide Web. [Online]."http://en.wikipedia.org/wiki/World_Wide_Web"

[2] Jeffrey Erman, Martin Arlitt and Anirban Mahanti, "Traffic classification using clustering algorithm," MineNet ’06 Proceedings of the 2006 SIGCOMM workshop on Mining network data, pp.281-283.

[3] Wikipedia. [Online]. "http://en.wikipedia.org/wiki/Wikipedia:about"

[4] HTTPS. [Online]. "http://en.wikipedia.org/wiki/HTTP_Secure"

[5] Felten, Edward W., and Michael A. Schneider. "Timing attacks on web privacy." Proceedings of the 7th ACM conference on Computer and communications security. ACM, 2000.

[6] Wondracek, Gilbert, et al. "A practical attack to de-anonymize social network users." Security and Privacy (SP), 2010 IEEE Symposium on. IEEE, 2010.

[7] Cheng, Heyning, and Ron Avnur. "Traffic analysis of ssl encrypted web browsing." URL citeseer. ist. psu. edu/656522. html (1998).

[8] Mistry, Shailen, and Bhaskaran Raman. "Quantifying Traffic Analysis of Encrypted Web-Browsing." (1998).

[9] Labovitz, Craig, et al. "Internet inter-domain traffic." ACM SIGCOMM Computer Communication Review. Vol. 40. No. 4. ACM, 2010.

[10] Sun, Q., Simon, D. R., Wang, Y. M., Russell, W., Padmanabhan, V. N., & Qiu, L. (2002). "Statistical identification of encrypted web browsing traffic." In Security and Privacy, 2002. Proceedings. 2002 IEEE Symposium on (pp. 19-30). IEEE.

[11] Bissias, George Dean, et al. "Privacy vulnerabilities in encrypted http streams." Privacy Enhancing Technologies. Springer Berlin Heidelberg, 2006.

[12] Lu, L., Chang, E. C., & Chan, M. C. (2010). Website fingerprinting and identification using ordered feature sequences. In Computer Security–ESORICS 2010 (pp. 199-214). Springer Berlin Heidelberg.

[13] Danezis, George. "Traffic Analysis of the HTTP Protocol over TLS." Unpublished draft (2009).

[14] Yu, Shui, et al. "Attacking anonymous web browsing at local area networks through browsing dynamics." The Computer Journal 55.4 (2012): 410-421.