This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE Globecom 2010 proceedings.

An Novel Hybrid Method for Effectively Classifying Encrypted Traffic

*Guang-Lu Sun[1, 4], Yibo Xue[1, 2], Yingfei Dong[3], Dongsheng Wang[1, 2] and Chenglong Li[1]
[1]Research Institute of Information Technology, Tsinghua University, 100084, Beijing, China.
[2]Tsinghua National Lab for Information Science and Technology, Beijing, 100084, China.
[3] Dept of Electrical Engineering, University of Hawaii, Honolulu, HI, 96822, USA.
[4] School of Computer Science and Technology, Harbin University of Science and Technology, China

Abstract—**Classifying encrypted traffic is critical to effective network analysis and management. While traditional payload-based methods are powerless to deal with encrypted traffic, machine learning methods have been proposed to address this issue. However, these methods often bring heavy overhead into the system. In this paper, we propose a hybrid method that combines signature-based methods and statistical analysis methods to address this issue. We first identify SSL/TLS traffic with signature matching methods, and then apply statistical analysis to determine concrete application protocols. Our experimental results show that the proposed method is able to recognize over 99% of SSL/TLS traffic and achieve 94.52% in F-score for protocols identification.**

Keywords- Traffic classification, Encrypted protocol, Hybrid method, Statistical analysis

I. INTRODUCTION

Real-time traffic classification is critical for many network management tasks, such as adaptive Quality of Service (QoS), security, dynamic access control, and intrusion detection systems (IDSs) [1]. Internet service providers (ISPs) mostly use port-based methods or payload-based methods. Port-based classification cannot deal with applications with dynamic ports [2]. Payload-based classification usually compare packet payload with known signatures, and it does not work when packet payloads are encrypted [3].

Recently, traffic classification approaches based on machine learning (ML) methods have been developed to address the limitations of the above two methods [4]. We can use flow statistics (e.g., flow duration, mean packet size) to build profile patterns to associate flows with application protocols. Classification models, such as Naïve Bayes, Support Vector Machine, and Expectation Maximization, are built for flow classification [5]. Although these methods address the limitations of port-based and payload-based methods, their efficiency, real-time capability, and accuracy are still fallen behind our requirements. Furthermore, these methods classify traffic at traffic points, and are difficult to evaluate and obtain comprehensive and stable results [6].

In this paper, we propose a hybrid approach to identify application protocols that encrypt payloads with Secure Socket Layer protocol (SSL) or Transport Layer Security protocol (TLS). For example, HTTPS and TOR both use SSL/TLS to encrypt their packets. First, we extract flows that match a series of SSL/TLS rules. We can achieve this with good accuracy and high processing speed. We then perform statistical analysis to figure out which application protocols are using these SSL/TLS connections. We train our Bayesian classification model with real data to acquire statistical information for identifying different application protocols. Application flows based on SSL/TLS are identified through their statistical information. Our experimental results show that the proposed method achieves a satisfied performance, and overcome the issues of solely applying signature-based methods or ML-based methods.

The remainder of this paper is structured as follows. Section II discusses related work. Section III introduces signature-based method and statistical analysis methods used in our work. We present the hybrid method and discuss signatures, feature templates and merging strategy in Section IV. We introduce our data sets experimental results in Section V. We conclude this paper and discuss future work in Section VI.

II. RELATED WORK

Port-based methods directly inspect the 5-tuple of packets and well-known port numbers, because many traditional application protocols utilize fixed port numbers assigned by IANA. To avoid completely relying on port numbers, payload-based methods are deployed. Packet payloads are matched to the known characteristics of application protocols [7]. These application signatures are extracted by analyzing available documentations and packet-level traces, and then used in online identifiers to track application traffic.

As the increasing deployment of many encrypted protocols, payload-based methods become less attractive while ML-based methods gain more attention. McGregor et al. firstly used unsupervised machine learning techniques to cluster traffic flows [8]. Moore and Zuev applied a supervised Naive Bayes estimator to classify application protocols and further improved the accuracy of refined variants [4]. They used manually-classified data corresponding to the actual category of flows as the training data set. Statistical patterns are abstracted from the flows as model features. After the adjustment of estimation parameters in the training phase, the training model can be used to classify other flows by computing statistical information of their patterns. Following the above algorithms, a lot of machine learning models were applied to traffic classification, such as

* Contact Author: Dr. Guang-Lu Sun, e-mail: guanglu.sun@gmail.com .
This work was supported in part by China National Nature Science Grant (60903083), China Postdoctoral Science foundation (20090450390), Program for New Century Excellent Talents In Heilongjiang Provincial University and The 973 National Basic Research Program of China (2007CB311102).

978-1-4244-5638-3/10/$26.00 ©2010 IEEE

simple K-Means, Nearest Neighbor, Decision Tree, and Bayesian Network [6]. Performance comparison of these algorithms are explained and reported in [9], from the aspects of classification accuracy and computational performance.

SSL/TLS protocol [10] is between the TCP/IP protocol and an application protocol. Its primary goal is to provide security, privacy and data integrity between two applications. It is composed of two layers: the Record Protocol and the Handshake Protocol. At the lower level, the Record Protocol that is layered on top of some reliable transport protocol. The Handshake Protocol helps a server and a client authenticate each other and negotiate an encryption algorithm and keys, before an application protocol transmits its data.

III. SIGNATURE-BASED METHOD AND STATISTICAL ANALYSIS METHOD

In this section, the two empirical methods used in this paper are introduced briefly. Firstly, we describe signature-based method and its pattern matches. The statistical analysis method based on Bayesian theory is presented as follows.

A. Signature-based Method

A signature-based method is based on the matches between the identification signatures and the actual information in TCP/UCP payloads. The signatures include the universal characteristics of an application protocol, and built via two sources: from the standard protocol specifications and documentations, or from manual observation and analysis. Fixed strings in payloads and special behaviors in the transmission procession are considered as signatures. For example, "<a character(1 byte)><a string (19 byte)>" represents the BitTorrent header format of the handshake messages.

From the aspect of matching spectrum, signature-based methods are divided as single-packet matching and multi-packets matching. From the aspect of matching module, signature-based methods are divided as fixed offset matching and variable offset matching [3]. In practice, different strategies are chosen based on trade-offs in terms of the level of accuracy, scalability and robustness.

B. Statistical Analysis Methods

For identifying encrypted protocols and automatic analysis, statistical analysis methods are brought into traffic analysis. These methods assume that the application protocols generate stable transmission patterns when they transmit data in the network. These patterns represent protocol behaviors and used as special information to identify the protocols. The numerical characteristics of these patterns are statistical information such as flow length and duration. By combining sufficient statistical information, these classifiers can identify protocols accurately.

We use widely-accepted Bayesian methods in this paper. Bayesian methods are popular parameter estimation methods. Consider a sample set $X = \{x_1, x_2, ..., x_n\}$ belongs to a class set $C = \{c_1, c_2, ..., c_l\}$. For each sample x_i, its characteristics (named as features) are described as $F_i = \{f_{i1}, f_{i2}, ..., f_{im}\}$, which are numeric or discrete values. $FT = \{ft_1, ft_2, ..., ft_m\}$ is a feature template based on which features are abstracted. For a pending sample x^*, Bayesian methods compute the conditional probability $p(c_i \mid x^*)$ based on Formula (1).

$$p(c_i \mid x^*) = \frac{p(x^* \mid c_i)p(c_i)}{\sum_i p(x^* \mid c_i)p(c_i)} \quad (1)$$

$\sum_i p(x^* \mid c_i)p(c_i)$ is the normalization factor. $p(c_i)$ denotes prior distribution which can be compute based on Formula (2).

$$p(c_i) = \frac{n_{c_i}}{n} \quad (2)$$

n_{c_i} is the number of samples belonging to class c_i.

Based on the split methods of $p(c_i \mid x^*)$, Bayesian methods can be sorted to independent estimation and dependent estimation. Based on the computation methods of $p(x^* \mid c_i)$, Bayesian methods can be sorted to parameter estimation (Naïve Bayes) and non-parameter estimation (Kernel Estimation).

In independent estimation, $p(c_i \mid x^*)$ can be decomposed based on the assumption of independence:

$$p(c_i \mid x^*) = \prod_m p(c_i \mid f_{*m}) \quad (3)$$

For each $p(c_i \mid f_{*m})$, Naïve Bayes method gives the estimation based on the experimental distribution like the normal distribution. In a dependent estimation, $p(c_i \mid x^*)$ cannot be decomposed. A Bayesian model computes $p(x^* \mid c_i)$ by a covariance matrix in the normal distribution. Different Bayesian methods are discussed in [11].

The advantages of Bayesian methods include the following four points: 1) They consider the numerical characteristics of flow. 2) Their probability models can give the confidential value of results. 3) The methods have performed well in many classify tasks. 4) Naïve Bayes method needs little training and testing time. Kernel estimation provides high classification performance.

IV. HYBRID METHOD TO ENCRYPTED TRAFFIC CLASSIFICATION

A. The Framework of Encrypted Traffic Classification

As a network monitor needs to classify all traffic at a link, it needs to address two issues: the fuzzy descriptions of network traffic and the challenge of accurate evaluation. In this paper, we propose to identify a particular class of traffic and develop effective methods to achieve this goal.

SSL/TLS traffic is arguably the most important class for security and privacy in the network. Many application protocols like TOR and HTTPS are based on it. For controlling

978-1-4244-5638-3/10/$26.00 ©2010 IEEE

these types of traffic and avoiding vicious accesses and intrusions, it is essential to find traffic and identify their concrete protocols.

The advantages of signature-based methods are accurate and fast, but they cannot deal with encrypted protocols and demand manual analysis. While statistical analysis methods are able to address these issues, they usually have lower accuracy for special traffic classification and are difficult to model all traffic. Therefore, we proposed a hybrid method to identify the encrypted protocols. Our framework is shown in Fig. 1.

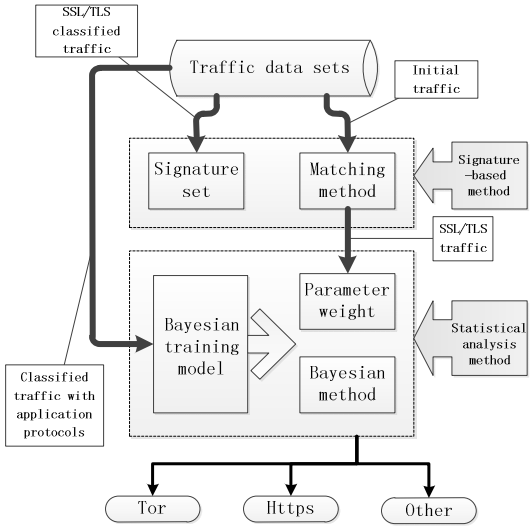

Figure 1. The framework of the hybrid method for protocols identification

There are several key phases in the framework. In the processing phase of signature-based method, we use a small amount of SSL/TLS traffic to analyze its signatures. We also analysis of SSL/TLS documents and SSL/TLS payloads, to generate a signature sets of SSL/TLS protocols. In the training phase of statistical analysis method, we use a large amount of classified flows of encrypted protocols to train our Bayesian model. The model then generates the features with the parameter weights.

In testing our method, we first put testing flows with no classified tag into the signature-based processing phase. Once matched with SSL/TLS signatures, applications flows on top of SSL/TLS sessions are separated from other testing flows. We then input these flows to the statistical analysis method and identify the flows belong to which application based on our Bayesian model.

B. Signature-based Method for SSL/TLS Protocols

In the class of SSL/TLS protocols, the most frequent used version is SSL3.0 and TLS1.0. Their basic protocol is the Record Protocol on which other operational protocols are based. Fig. 2 gives the general formation of Record Protocol.

We use the characteristics of the Record Protocol as signatures to classify flows as SSL/TLS flows. The fixed single packet matching method and the fixed multi-packets matching method mentioned in Section III are used to realize the signature identifying algorithm. The algorithm is as follows:

For each packet of a flow,

1) Match one of the type signatures (0x14、0x15、0x16、0x17) in the first byte of the TCP payloads.

2) Match one of the version signatures (0x03 00、0x03 01、0x03 02、0x03 03) in the second and third bytes.

3) Compute the length of SSL/TLS structure denoted in the fourth and fifth bytes.

4) Detect the subsequent bytes of current packets, if the subsequent length is longer than the length of structure, go to 5; else go to 6.

5) If procedure 1-4 is not executed three times, repeat procedure 1-4; else the flow belongs to SSL/TLS protocol.

6) If procedure 1-4 is not executed three times, read next packet of the flow, repeat procedure 1-4; else the flow belongs to SSL/TLS protocol.

+	Byte + 0	Byte + 1	Byte + 2	Byte + 3
Byte 0	Content type			
Bytes 1 - 4	Version		Length	
	(Major)	(Minor)	(bits 15..8)	(bits 7..0)
Bytes 5 - (m-1)	Protocol message(s)			
Bytes m - (p-1)	MAC (optional)			
Bytes p - (q-1)	Padding (block ciphers only)			

Figure 2. The general formation of Record Protocol

C. Bayesian Method for Encrypted Application Protocols

A Bayesian method is a typical classifier model. We use the Naïve Bayes model because of its lower training and testing cost in computation and storage. The flow patterns of encrypted application protocols can be represented by features in the Naïve Bayes model. The features are draw-out based on features templates. The definition of features templates is shown in Table I.

TABLE I. FEATURE TEMPLATES OF NAÏVE BAYES MODEL

Feature Template	**Features Description**
Feature template 1	Mean packet length
Feature template 2	Maximum and Minimum packet length
Feature template 3	Mean inter-arrival time of packet
Feature template 4	Maximum and Minimum inter-arrival time of packet
Feature template 5	Flow duration
Feature template 6	Packet-count of flow

In the training phase of Naïve Bayes model, the classified flows enter into the training model. The prior probabilities of classes are computed based on Formula (2). The mean value and the variance of the distribution function of each class are defined as Formula (4) and (5).

$$\hat{\mu}_i = \sum_{x_k : C(x_k)=c_i} \frac{x_k}{n_{c_i}} \tag{4}$$

$$\hat{\sigma}_i^{\,2} = \sum_{x_k : C(x_k)=c_i} \frac{(x_k - \hat{\mu}_i)^2}{n_{c_i} - 1} \tag{5}$$

In the testing phase, if consider the features independent from each other and the feature distribution satisfying the formal distribution, the distribution function is defined as:

$$p(x^* \mid c_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\{-\frac{(x^* - \hat{\mu}_i)^2}{2\sigma_i^2}\} \tag{6}$$

Through the numerical features of unknown flows and the distributions of features corresponding to each class, the posterior probability of a new flow is computed based on Formula (1), (3) and (6).

The new flow is identified to protocol $c^{'}$, if $p(c^{'} \mid x^*)$ has the highest value in all $p(c_i \mid x^*)$.

V. EXPERIMENTAL RESULT AND DISCUSSIONS

We first introduce the data sets and metrics and then discuss the experimental results.

A. Data sets and Evaluation Metrics

We first present how to compose the data sets for training and testing with the signature-based method and the statistical analysis method. To evaluate the performance of different methods, we choose two types of data sets.

For the purpose of evaluating the signature-based method, flows based on SSL/TLS protocols are obtained from a network monitor on the sender side. Table II contains information about the structure of the data set of SSL/TLS, including four known application protocols and one unknown class (*Other*) based on SSL/TLS protocol. The table shows the protocol types, data set size, the number of packets, and the number of flows of each type.

Background flows are obtained from the monitor of edge network and other public data sets. Table III contains information about the structure of the background traffic. Seven types of data sets are obtained as background traffic. The DARPA data set is public trace which contains over 20 types of protocols such as HTTP, SSH, SMTP, SNMP, Route/u, TELNET, FTP, POP3. The other six types of data are public traffic acquired from the sender of our network. Thunder is the most popular P2P application in China. Table III also shows the protocol types, data set size, the number of packets, and the number of flows of each type of application protocols.

Because "the arriving order" of the flows has no effect on traffic classification, we randomly input flows into the classifier in the training or testing phases.

Besides the common metrics, we also use the following metrics for evaluation: We use Accuracy to represent the number of correctly classified examples. We use Recall (R) to represent the fraction of positive instances that is positively predicted. We use Precision (P) to represent the fraction of predicted instances that is positively predicted. It is equivalent to True Positives. F-score (F) is the harmonic mean of precision (P) and recall (R). It is computed as follows.

$$F = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R} \tag{7}$$

where β is usually defined as 1.

TABLE II. THE TYPES OF APPLICATION DATA SETS BASED ON SSL/TLS

Protocol Type	Size	Number of Packets	Number of Flows
HTTPS	60.3M	565298	2256
TOR	572.7M	569104	1696
UPDATE	637.2K	5777	104
OSCAR	1.2M	11663	14
Other	12M	114262	4514
Total	**646.8M**	**1266104**	**8584**

TABLE III. THE TYPES OF APPLICATION DATA SETS AS BACKGROUND TRAFFIC

Protocol Type	Size	Number of Packets	Number of Flows
DARPA	63.0M	346672	69268
BitTorrent	79.1M	119268	1288
Edonkey	146.9M	258591	3373
HTTP	141.1M	127495	4065
FTP	6.4M	152868	1081
Thunder	88.9M	150830	1364
GRE	30.5M	69591	22
Total	**555.9M**	**1225315**	**80461**

B. Experimental Results and Discussions

The hybrid method includes two main steps. The first step classifies SSL/TLS traffic from the background traffic using the signature-based method. The second step identifies application protocols using the statistical analysis method.

We will first report the experimental results of the signature-based method. Because it is difficult to acquire standard testing data sets with background traffic specially for testing SSL/TLS protocols, we design the mixture of data sets to validate the performance of the signature-based method, by combining the data sets described in Table II and Table III.

Table IV shows the performance of the signature-based method for SSL/TLS classification. Almost all the SSL/TLS traffic can be classified from background traffic. It confirms that our signature-based method is effective to recognize SSL/TLS-based protocols and achieves the precision upwards to 99% accuracy. Furthermore, other background traffic like BitTorrent, HTTP and Thunder are completely not classified into the class of SSL/TLS.

Then, we will report the experimental results of the statistical analysis method. Because it needs more information for training our Bayesian model, we use TOR and HTTPS flows in the data sets to validate the method. Other three types

of flows are not suitable because they have only a small numbers of flows. Table V shows the performance of identifying TOR and HTTPS flows with the independent assumption based on the Naïve Bayes model.

TABLE IV. THE PERFORMANCE OF SSL/TLS CLASSIFICATION

Protocol Type		Number of Packets		Number of Packets		Per-Packet Accuracy
		Original	Identified	Original	Identified	
SSL/TLS	HTTPS	565298	560589	2256	2223	99.16%
	TOR	569104	568493	1696	1591	99.89%
	UPDATE	5777	5774	104	103	100.00%
	OSCAR	11663	11663	14	14	100.00%
	Other	114262	113094	4514	4321	98.98%
	Total SSL	**1266104**	**1259613**	**8584**	**8252**	**99.49%**
Background Traffic	Darpa	346672	0	1288	0	0%
	Bittorrent	119268	0	1364	0	0%
	Edonkey	258591	0	3373	0	0%
	HTTP	127495	0	1081	0	0%
	FTP	152868	0	69268	0	0%
	Thunder	150830	0	4065	0	0%
	GRE	69591	0	22	0	0%
	Total Background	**1225315**	**0**	**1288**	**0**	**0%**

TABLE V. THE PERFORMANCE OF TOR/HTTPS IDENTIFICATION WITH INDEPENDENT ASSUMPTION BASED ON NAÏVE BAYES MODEL (PER FLOW)

Protocol	Training Data Set	Test Data Set	Precision (P %)	Recall (R %)	F-score (F %)
HTTPS	1200	155	0.9313	0.9613	0.9460
TOR	1500	155	0.9600	0.9290	0.9443

The overall accuracy is 94.52%

TABLE VI. THE PERFORMANCE OF TOR/HTTPS IDENTIFICATION WITH NO INDEPENDENT ASSUMPTION BASED ON BAYESIAN MODEL (PER FLOW)

Protocol	Training Data Set	Test Data Set	Precision (P %)	Recall (R %)	F-score (F %)
HTTPS	1200	155	0.9313	0.9613	0.9460
TOR	1500	155	0.9470	0.9226	0.9346

The overall accuracy is 93.55%

As shown in Table V, TOR and HTTPS flows are well classified based on Naïve Bayes model and a few of feature templates shown in Table I. More features may help further improve the performance, but we found that the features used in our experiment are effective. Improving the scale of training data sets is more useful than the utilization of different types of features.

Moreover, we further relax the assumption of features independence, because the features are dependent in theory. Based on the distribution computation with the covariance matrix, the results with dependent assumption are shown in Table VI. The performance is not good as the results of Naïve Bayes model. It is confirmed that the independent assumption has little effects in the performance of Bayesian model.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a hybrid method for identifying encrypted application traffic, combining a signature-based method and a statistical analysis method. We extract the signatures of SSL/TLS by the analysis of protocol payloads. We use a matching method to classify the SSL/TLS traffic. The signature-based method achieves the accuracy of classification over 99%.

Two Bayesian models and six types of features are developed to further associate SSL/TLS flows to applications. We use HTTPS and TOR flows to validate the performance of our models. The F-score of identification is 94.52%. Through the comparison of two Bayesian models, the independent assumption of Bayesian model has little effects in classifying performance.

We are pursuing this work in several directions. The first is obtaining more classified data and expanding our method to identify more encrypted protocols. The second is analyzing the payloads of protocols and searching effective features. The third is applying our method to actual systems in the network.

ACKNOWLEDGMENT

We thank the anonymous referees for their useful and illuminating comments. We thank Changxing Liu, Zhifeng Chen and Luoshi Zhang for their assistance in gathering the traces.

REFERENCES

[1] V. Paxson, “Bro: A system for detecting network intruders in real-time,” Computer Networks, vol. 31(23-24), pp. 2435–2463, 1999.

[2] T. Karagiannis, A. Broido, N. Brownlee, and K. Claffy, “Is P2P dying or just hiding?” in Proceedings of Globecom 2004, Dallas, Texas, USA, November/December 2004.

[3] S. Sen, O. Spatscheck, and D. Wang, “Accurate, scalable in network identification of P2P traffic using application signatures,” in WWW2004, New York, NY, USA, May 2004.

[4] A. Moore and D. Zuev, “Internet traffic classification using Bayesian analysis techniques,” in Proc. of ACM SIGMETRICS 2005, Banff, Alberta, Canada, June 2005.

[5] M. Dusi, A. Este, F. Gringoli, and L. Salgarelli, “Using GMM and SVM-based Techniques for the Classification of SSH-Encrypted Traffic,” The 2009 IEEE International Conference on Communications, Dresden, June 14-18, 2009.

[6] T. Nguyen, G. Armitage. "A survey of techniques for Internet traffic classification using machine laerning," IEEE Communications Surveys and Tutorials, 2008.

[7] P. Haffner, S. Sen, O. Spatscheck, and D. Wang, “ACAS: Automated Construction of Application Signatures,” in SIGCOMM’05 Workshops, Philadelphia, USA, August 22-26, 2005.

[8] A. McGregor, M. Hall, P. Lorier, and J. Brunskill, “Flow clustering using machine learning techniques,” in Proc. of PAM2004, Antibes Juan-les-Pins, France, April 2004.

[9] N. Williams, S. Zander, and G. Armitage, “A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification,” SIGCOMM Computer Communication Review, vol. 36, no.5, pp.5–16, 2006.

[10] T. Dierks, E. Rescorla, "The Transport Layer Security (TLS) Protocol, version 1.2", IETF RFC, 2008.

[11] J. Duda, P. Hart and D. Stork, "Pattern Classification (Second Edition)", John wiley&sons Inc, 2001.