Identification and Detection of Statistical Characteristics of Encrypted Traffic in Zombie Networks

Ruidong Chen, Kwame Opuni-Boachie Obour Agyekum, Xiaosong Zhang, and Qi Xia

***Abstract*—There is a great significance to encrypted traffic including illegal data regulation, protection of user information and detection of network attacks. Classifying encrypted traffic is critical to effective network analysis and management. With the advent of machine learning techniques, traditional payload-based methods have become powerless and obsolete, in dealing with encrypted traffic. Accurately and efficiently identifying network traffic is very crucial for network management. Machine learning methods, however, are disadvantaged by the creation of overheads in the system. Most traffic encryption methods also focus on single granularities, and hence the full functionality of the network isn't realized. In this paper, we propose a traffic identification method that seeks to combat protocol-independent identification. Our method utilizes an encrypted traffic identification model on the basis of information entropy, which can realize on-line identification without violating user privacy and as higher efficiency analysis and a lower false-alarm rate, and also on multiple granularities. Our experimental results show that the proposed method is able to recognize over 80% of traffic, and achieves an efficient encrypted traffic identification.**

***Index Terms*—Botnet, encrypted traffic identification, information entropy, multiple granularity, zombie networks.**

I. Introduction

Numerous network management tasks such as adaptive Quality of Service (QoS), dynamic access control, encryption, and intrusion detection systems (IDSs) make use of real-time traffic classification methods [1]. Some of these methods include port-based classification [2], payload-based classification [3], and other machine learning techniques [4]. Port-based classification methods cannot deal with applications with dynamic ports while payload-based classification isn't effective when packet payloads are encrypted, as they usually do comparisons. These drawbacks thus introduced the machine learning approach to curb these problems. Several machine learning classification methods have been proposed [5], but these methods still fall short when it comes to their efficiency, real-time capability, and accuracy.

Current traffic identification technology can be broadly divided into two categories; protocol-related identification, and protocol-independent identification. With focus on encrypted traffic identification, and taking Secure Socket Layer (SSL) and Secure Shell (SSH) encryption protocols as examples, they can be classified into the first type of identification without protocol parsing. All other encrypted traffic can be categorized as the second type of protocol identification.

In this paper, we propose a traffic identification method to solve the second type of protocol identification. Because encryption requires a certain amount of computation and time resources, data with higher information and economic value are usually transmitted through an encryption protocol. It is therefore of great significance to implement an effective management of encrypted traffic. However, protocol-dependent encrypted traffic identification is limited in application. There are many kinds of encryption protocols, and designing separate algorithms for each protocol consumes a lot of resources, and from the management point of view, there is no significance in distinguishing these protocols.

On the other hand, due to security considerations, many encryption protocols are not publicly or completely open, making it difficult to identify and manage them. This scheme is however based on information theory coding knowledge, an encrypted traffic identification method on the basis of information entropy, which can realize on-line identification without violating user privacy, can achieve a higher efficiency analysis and also a low false-alarm rate.

The remainder of this paper is structured as follows. Section II addresses related works regarding traffic identification. Section III introduces the system model of our work and also parameters considered in the setup of the experiment. Section IV presents the experimental results, and concluding remarks, and a discussion of future work are made in Section V.

II. Related Works

Mena *et al.* [6] were one of the first researchers to identify an application-specific flow, where they showed the identification of real audio flows by using simple analysis of

Manuscript received March 1, 2018; revised May 10, 2018. This work was supported in part by the applied basic research programs of Sichuan Province under Grant 2015JY0043, in part by the Fundamental Research Funds for the Central Universities under Grant ZYGX2015J154, Grant ZYGX2016J152, and Grant ZYGX2016J170, in part by the programs of International Science and Technology Cooperation and Exchange of Sichuan Province under Grant 2017HH0028, in part by the Key Research and Development Projects of High and New Technology Development and Industrialization of Sichuan Province under Grant 2017GZ0007, in part by the National Key Research and Development Program of China under Grant 2016QY04WW0802, Grant 2016QY04W0800, and Grant 2016QY04WW0803, and in part by the National Engineering Laboratory for Big data application on improving government governance capabilities.

R. Chen is with Mr. Ray Co. Ltd., Chengdu, China and Youe Data Co. Ltd., Beijing, China (e-mail: crdchen@163.com).

K. O.-B. Obour Agyekum is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China (e-mail: obour539@yahoo.com).

X. Zhang and Q. Xia are with the Center for Cyber Security, University of Electronic Science and Technology of China, Chengdu, China (e-mail: johnsonzxs@uestc.edu.cn, xiaqi@uestc.edu.cn).

packet length and inter-arrival time. Traffic pattern similarity between different application layer protocols was exploited by McGregor *et al.* in [7]. Their work considered grouping observed flows into hierarchical clusters, by making use of machine learning techniques. QoS classes were implanted by Roughan *et al* [8] in which they used nearest neighbor and linear discriminate analysis. Moore and Zuev applied a supervised Naïve Bayes estimator to classify application protocols and further improved the accuracy of refined variants [4]. They utilized manually-classified data corresponding to the actual category of flows as the training data set. Following these above algorithms, a lot of machine learning models were applied to traffic classification, such as simple K-Means, Nearest Neighbor, Decision Tree, and Bayesian Network [9]. Authors in [10]-[12] all discussed a method for identifying the use of port or payload information, with each achieving respectable higher accuracies and false positive rates.

Some authors have also focused on the classification of encrypted traffic [13]-[17]. Bernaille and Teixeira in [13] proposed a method based on the encryption connection of the size of the first few packets. They achieved an accuracy of more than 85%. [14] made use of a more recent hybrid method that tries to identify SSL/TLS encrypted application layer protocols with a combination of a signature-based and a flow-based statistical analysis scheme. Bissias *et al.* presented a traffic analysis against encrypted HTTP streams to identify the source of the traffic by analyzing distributions of packet sized and inter-arrival times of requests made to web services [15]. Lee *et al.* and Levillain *et al.* evaluated the practices of SSL/TLS servers by investigating server replies [16], [17]. They studied the details of encryption parameters and protocol features, and their extensions.

Most of these traffic identification methods are based on single granularity and also have the disadvantage of a high false-alarm rate. Our work, therefore, proposes a traffic identification scheme based on multi-granularity feature extraction. Compared with traditional single-granularity traffic identification schemes, this scheme can extract the features from multiple granularities, and greatly reduce the false-alarm rate of the traditional methods.

III. System Model

The various components of this work, and the terminologies utilized in this work are elaborated below.

A. Encrypted Traffic Identification Based on Statistical Test Information Entropy

Information entropy can be defined as a measure of the amount of uncertainty in the information contained in a message. The more orderly a system is, the lower the entropy of information. The entropy is therefore the measure of the degree of systemization. The main work flow of this method is given in Fig. 1.

Assuming that the payload is N bytes, the alphabet has a total of $m = 256$ letters, and the number of occurrences of a byte is n, then the frequency of occurrence is given as $f_i = n_i / N$. The entropy H for a data packet w is given by

$$H(w) = -\sum_{i=0}^{m-1} f_i \log_2(f_i) \tag{1}$$

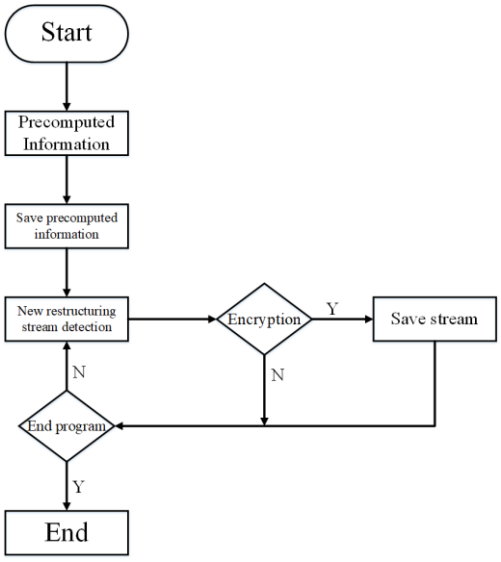

Fig. 1. Statistical test flow chart.

The length is given by N, and all possible average information entropies are calculated as

$$H_N(u) = \frac{1}{m^N} \sum_{n_0+\ldots+n_{M-1}=N} \left[\binom{N}{n_0,\ldots,n_{M-1}} \times -\sum_{i=0}^{m-1} f_i \log_2(f_i) \right] \tag{2}$$

where u is evenly distributed and is given by

$$\binom{N}{n_0,\ldots,n_{M-1}} = \frac{N!}{n_0!\ldots n_{M-1}!} \tag{3}$$

Equation (2) is the entropy size under the average distribution of the word combination of all possible cases. In other words, the closer the sample's calculated maximum likelihood estimate to the value found in (2), the more likely it is that the sample under test is evenly distributed. Similarly, the closer it is, the more likely it is to encrypt data or compress data.

According to Monte Carlo method, we can get all the average information entropy $H_N(p)$ and variance σ of length N, if we only compare $H(w)$, and find out whether the interval $H_N(p) \pm 3\sigma$ can determine if the data packet is encrypted. Similarly, for a stream composed of multiple data packets, we determine a few packets in a stream for indication as to whether the stream is encrypted or not. The specific process is as follows:

First, given N, we need to figure out the value of (2). In the network, traffic is usually used to express a byte 0xXX, the bytes of space 0x00 to 0xFF out of a total of 256 possibilities; so $m = 256$. For example, a network traffic can be intercepted as follows: 0x55 0x89 0xe5 0x83 0xec 0x58 0x83 0xe4 0xf0 0xb8 0x00 0x00 0x00 0x00 0x29 0xc4 0xc7 0x45 0xf4 0x00 0x00 0x00 0x00 0x83 0xec 0x04 0xff 0x35 0x60 0x99 0x04 0x08.

The maximum likelihood entropy estimate for this traffic, H_1, can be calculated using Eq. (1). Here, $H_1 = 3.97641$, the value of $H_N(u)$ is calculated to be 4.87816, and then the standard deviation σ of $H_N(u)$ is calculated. Finally, it is checked whether H_1 is between $H_N(u) - 3\sigma$ and $H_N(u) + 3\sigma$. If yes, then it is categorized as an encrypted traffic.

B. Encrypted Traffic Identification Based on Statistical Test Information Entropy

The traditional method for traffic identification has the disadvantage of high false alarm rate. There are many traditional network traffic identification methods, such as port number-based, fingerprint-based, machine-based and host-based network identification algorithms. However, the traditional method is only limited to one granularity traffic feature extraction, such as single, from the flow granularity or package granularity feature extraction. This therefore ignores the hidden features at other granularities, for example, the similarity between the network data streams. Characteristics such as the direct similarity of data packets, etc., are often very useful for network traffic identification. This key technology proposes a rigid network traffic identification method based on multi-granularity feature extraction.

The purpose of the method based on traffic identification is to find out which host in the target network uses the deadlock network software. Therefore, this method first separates all aggregated traffic according to the monitored host. Then, using the collected traffic of the host, the traffic is extracted from multiple granularities. For example, feature field analysis is performed from the packet granularity, analyzed from a streaming point of view, or feature extraction is performed from the host's network behavior. And then an update is provided to the host corresponding to the host's profile eigenvector. Finally, the existing supervisory learning algorithm is used to establish a recognition model to identify these types of deadlocked network traffic so as to reach the purpose of finding a deadlocked host in a predetermined network segment.

The basic idea of this method is to save all filtered packets of a host as a Host profile (in essence, a packet capture (pcap) file), and then extract the features of the Host profile from different perspectives. This key technology intends to extract the following features from the Host profile:

- **Package level**: port number, keywords and so on;
- **Stream Level**: The average entropy of the stream, the average length of the stream, the average size of the stream, etc.
- **Host level**: broken network probability (initiate sensitive DNS query), host network behavior characteristics, host data packet length distribution;
- **Other levels**: similarity between streams, or similarity between rooms.

C. Experiment Set-Up

A botnet detection and analysis system is developed for the evaluation of this system. The evaluation is based on "GB / T 17544-1998 information technology, software packages, quality requirements and testing", "GB / T 16260-2006 information technology, software product evaluation, quality characteristics and its user Guide." Reference is also given to the "Zombie Network testing and analysis system development and application user manual."

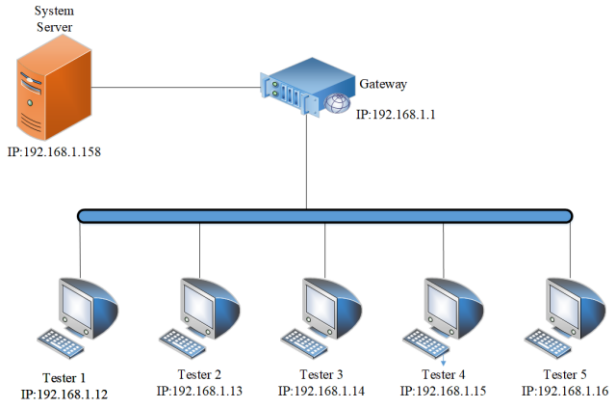

Fig. 2. Network topology.

The evaluation is based on the product's functionality, reliability, ease of use, and maintainability. "Botnet detection and analysis system" is mainly written in C language, with some plug-ins using Java, and python. The system is set-up at the LAN gateway. By monitoring the collected traffic of the gateway, the Client software data packets are analyzed and real-time warnings are quickly provided on the system integrated web pages. The tested Client software and browser plug-ins include IRC, Eggdrop, vpn, Zeus, and Mirai. Table I gives a description of the test environment.

Due to the fact that the entire control system needs to process access requests and traffic data from different users, there is the requirement to provide a system of higher processing capability. The following is the configuration of the hardware:

Server: Dell PowerEdge R730XD
Operating System Version: Cent OS 7 server
2 Intel® Xeon® E5-2630 v3 processors
8 16GB RDIMM memory chips
8 1.8TB 10K RPM SAS 6Gbps 512e 2.5-inch hot-swappable hard drives; making up a RAID 0 disk array
2 Intel Ethernet I350 QP 1Gb Network Cards.

Deployment of the server needs to run botnet traffic marking subsystem. The system is therefore a 64-bit Windows 7 operating system. The main hardware configuration is as follows:

A 40G hard drive
A virtual machine running on one machine

Network connection in Network Address Translation (NAT) mode.

The topology of the network is shown in Fig. 2.

IV. Results and Discussion

This section analyses and discusses all the various tests that were performed pertaining to this system. As mentioned earlier, four (4) major tests were taken into consideration. They are functionality test, where we tested for both the system and the web services' user management, the reliability

test, system usability test, and finally, the maintainability test. Tests results for the various domains are given in Tables II-V, with either the test result showing a “Pass” or “Fail” or a "Y" and "N" representing "Yes" and "No", respectively.

TABLE I: EXPERIMENTAL SETUP PARAMETERS

Device Name	Operating System	32/64 bit	Client Software	LAN IP	Processor
Server	Centos 7 server	64	None	192.168.1.158	E5-2630 v3
Tester 1	Windows 7	64	Mirai	192.168.1.12	Intel® Core™ i3-4170
Tester 2	Windows 8	64	IRC	192.168.1.13	Intel® Core™ i3-4170
Tester 3	Windows 10	32	Zeus	192.168.1.14	Intel® Core™ i3-4170
Tester 4	Windows 7	64	VPN	192.168.1.15	Intel® Core™ i3-4170
Tester 5	Windows 7	64	Zeus	192.168.1.16	Intel® Core™ i3-4170

A. *Functionality Test*

The evaluation on the system part was based on the ability to detect the various plug-ins, in a LAN environment. The system server, after successful detection of Client Host and Client Software, would send signals to the website. Upon the numerous tests conducted, the desired results were obtained. Fig. 3a, b and c show the various results obtained for the VPN configurations. Fig. 4 shows the Eggdrop set-up.

The user management test required that a user could login, make changes to the password, log out, and make other functional tests. Results proved to be affirmative. The sample update feature test requested an upload of the latest Client software to the server for independent modeling. Upon successful modeling, the latest machine learning model was successfully generated. The Botnet traffic identification page test on the other hand needed the system project on the server to run properly, and there were no negative results obtained.

The Mirai link mapping page test was set up to display information in the Mirai Link Mapping Data Sheet. Results showed that new nodes on the page were displayed. The confidential IP management page test was setup for the botnet traffic identification page to flag a signal if an IP address was detected in the Client. The traffic mirror feature page test was to enable downloads to the local Client traffic, and it was positive with the ability to download the traffic saved by the host of the client where the user is logged in. Table II gives a summary of the functionality test results.

B. *Reliability Test*

The system operability check was to realize the maximum duration with which the system could work. Tests results could run in the test environment for a long time, and the network traffic was found to be 13MB/s. The system server detection accuracy required that each botnet software flow could be detected. The system was mounted to verify its ability to support Gigabit network, and the response was favorable. The ability of the system to identify encrypted traffic for real-time filtering was also measured. It was found that within a limited test period, the system was able to identify each Client host traffic generated.

TABLE II: FUNCTIONALITY TEST RESULTS

Test Content	Test Parameter	Pass/Fail
System	Mirai Traffic Test plug-in	Pass
	Zeus flow plug-in	Pass
	IRC-Botnet flow test plug-in	Pass
	VPN traffic detection plug-in	Pass
	Eggdrop flow plug-in	Pass
Web-service	User management test	Pass
	Sample update feature test	Pass
	Botnet traffic identification plug-in	Pass
	Mirai link mapping page test	Pass
	Confidential IP management page test	Pass
	Traffic mirror feature page test	Pass

(a)

(b)

(c)

Fig. 3. (a) VPN setup, (b) VPN working environment and (c) Google functionality.

The system was also able to collect and mark not less than five host traffic training sets, a network sample flow of not less than 5G, and a background flow of not less than 100G. With reference to the completion of topology mapping of not less than 100 Mirai nodes, the system was able to do that. Lastly, within a test period, the system was able to detect and identify the botnet flow generated by each client host in an offline environment. The accuracy rate was measured to be 82%. Figs. 5 and 6 respectively show the pcap module and Mirai nodes topology mapping. A summary of the test results is given in Table III.

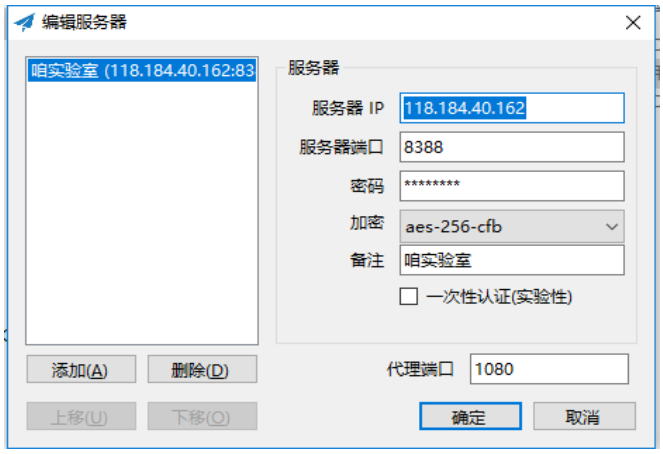

Fig. 4. Eggdrop working environment.

TABLE III: SYSTEM RELIABILITY TEST

Test Parameter	Yes/No
System operability	Y
System server detection accuracy	Y
Support for Gigabit network	Y
Encrypted traffic identification module to complete real-time filtering	Y
Complete the collection and marking of 5 host training sets in Gigabit network	Y
Complete topology mapping of not less than 100 Mirai nodes	Y
Offline analysis mode client recognition accuracy	Y

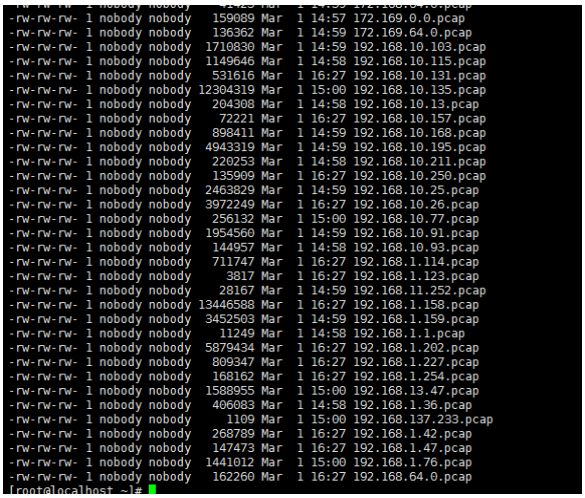

Fig. 5. Pcap module.

C. Usability Test

The first test performed under this domain was to detect the simple use of the web service. It was found that the web interface is very easy to use. It is fully functional, and gives clear tips as well to new users. When the overall interface test was performed, we realized that the web interface was in line with standards and norms; it maintained consistency and also described the correct function. The interface element test was done to inspect the consistency to the norms of the web interface window, icons, mouse, and texts. All the tests gave positive results. Table IV summarizes the test results.

Fig. 6. Topology mapping of Mirai nodes.

TABLE IV: SYSTEM USABILITY TEST RESULTS

Test Parameter	Yes/No
Functional usability test	Y
Overall interface test	Y
Interface element test	Y

TABLE V: SYSTEM MAINTAINABILITY TEST RESULTS

Test Parameter	Yes/No
Ease of analysis	Y
Ease of interruption	Y

TABLE VI: SUMMARY OF TEST RESULTS

Test Content	Evaluated items	Pass	Fail
Function Test	11	11	0
Reliability Test	7	7	0
Usability Test	3	3	0
Maintainability Test	2	2	0
Total	23	23	0
Percentage		100%	0

D. Maintainability Test

The ease of analysis test was conducted to indicate signals for any abnormalities on the system. Results showed that the system could provide tips and alarm information, and a user-friendly analysis of the cause of the malfunction. Tests were also carried out as to whether the system could be interrupted at any time. Results indicated that the system could be interrupted at any time, thereby providing flexibility. A summary of the test results is shown in Table V.

V. CONCLUSION

In this paper, an encrypted traffic identification module based on statistical test information entropy is developed. The rationale behind this approach is the ability to realize on-line identification without any violation to user privacy. The system uses multi-granularity feature extraction, which has the benefit of greatly reducing the false alarm rate compared

to traditional single-granularity, for the traffic identification scheme. Our experimental results proved that exceptional results can be obtained with this method. The detection accuracy for this multi-granularity method was above 80%. In a future work, more tests will be carried out aside the ones in this work, and also increase the detection accuracy to a higher percentage. We also hope to explore other granularity methods. Table VI gives a summary of all the events.

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 *Proc. of Globecom 2004*, Dallas, Texas, USA, 2004.

[3] S. Sen, O. Spatscheck, and D. Wang, "Accurate, scalable in network identification of P2P traffic using application signatures," *WWW2004*, 2004.

[4] A. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," in *Proc. of ACM SIGMETRICS 2005*, 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," in *Prof. 2009 IEEE International Conference on Communications*, 2009.

[6] A. Men and J. Heidemann, "An empirical study of real audio traffic," in *Proc. of the IEEE Infocom*, pp. 101-110, 2000.

[7] A. McGregor, "Flow clustering using machine learning techniques," *PAM 2004, Antibes Juan-les-Pins*, April 19-20, 2004.

[8] M. Roughan, "Class-of-service mapping for QoS: A statistical signature-based approach to IP traffic classification," *IMC '04, Taormina*, 2004.

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

[10] R. Alshammari and A. N. Zincir-Heywood, "Machine learning based encrypted traffic classification: Identifying ssh and skype," in *Proc. of the Second IEEE International Conference on Computational Intelligence for Security and Defense Applications*, pp. 289-296, 2009,

[11] C. Bacquet, A. N. Zincir-Heywood, and M. I. Heywood, "An investigation of multi-objective genetic algorithms for encrypted traffic identification," in *Proc. International Workshop on Computational Intelligence in Security for Information Systems*, 2009.

[12] J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson, "Offline/realtime traffic classification using semi-supervised learning," *Perform. Eva.* vol. 64, no. 9-12, pp. 1194-1213, 2007.

[13] L. Bernaille and R. Teixeira, "Early recognition of encrypted applications," in *Proc. of the PAM Conference*, vol. 4427, pp. 165-175, 2007.

[14] G. L. Sun, Y. Xue, Y. Dong, D. Wang, and C. Li, "An novel hybrid method for effectively classifying encrypted traffic," in *Proc. of IEEE*, GLOBECOM, 2010, pp. 1-5.

[15] G. D. Bissias, M. Liberatore, D. Jensen, and B. N. Levine, "Privacy vulnerabilities in encrypted HTTP streams," in *Proc. of the 5th Int. Conference on Privacy Enhancing Technologies*, 2006, pp. 1-11.

[16] H. K. Lee, T. Malkin, and E. Nahum, "Cryptographic strength of SSL/TLS servers: Current and recent practices," in *Proc. of ACM IMC*, 2007, pp. 83-92.

[17] O. Levillain, A. Ébalard, B. Morin, and H. Debar, "One year of SSL internet measurement," *in Proc. of ACM ACSAC*, 2012, pp. 11-20.

Ruidong Chen received his M.Sc degree in computer science from UESTC, China. He is currently pursuing his Ph.D in the same field.

He is a software security engineer at Cyber Network Security Center, Chengdu, China, and also works with Youedata as a data security business manager. He also works with Mr. Ray Corporation as the Chief Technical Officer, on developing a data protection infrastructure.

Mr. Chen has his research focus in the area of software security and privacy, vulnerability discovery and program hardening in mobile and web platforms. He has published several papers in the area of security.

Kwame Opuni-Boachie Obour Agyekum received his B.Sc degree in telecommunications engineering from Kwame Nkrumah University of Science and Technology, Ghana in 2014, and an M.Eng. in communication and information engineering in UESTC, China, in 2017.

He is a researcher in RGlobal and is currently pursuing his Ph.D in computer science and technology in UESTC. His current research interests include Blockchain technologies, big data security and privacy.

Xiaosong Zhang received his B.Sc degree in dynamics engineering from Shanghai Jiaotong University, Shanghai, in 1990, and the M.S. and Ph.D degrees in computer science from the University of Electronic and Technology of China (UESTC), Chengdu, in 2011.

He has worked on numerous projects in both research and development roles. He is currently an associate director with the National Engineering Laboratory of Big Data application to improving the Government governance capacity in China. He is also a professor in computer science with UESTC. He is also a Professor in computer science with UESTC.

Prof. Zhang is the dean with the center for cyber security of UESTC. He has coauthored a number of research papers on computer network security. His current research involves software reliability, software vulnerability discovering, software test case generation, and reverse engineering.

Qi Xia received her bachelors, masters and doctorate degrees in computer science and engineering from the University of Electronic and Technology of China (UESTC) in 2002, 2006, and 2010, respectively. She was a visiting scholar with the University of Pennsylvania, Philadelphia, PA, USA, from 2013 to 2014. She is an associate professor at the University of Electronic Science and Technology of China.

She is currently the deputy director of the big data security institute and the vice director of the center for cyber security at UESTC, China. She serves as chair and head for several key projects in the Peoples Republic of China. She is the PI of the national key research and development program of China in cyber security.

Dr. Xia has won several awards including the national scientific and technological progress award in which she placed second in 2012. She has published numerous papers in fields of cloud computing, network and information security, cyber, data and information systems security among others. Her current research includes networking, security and blockchain.