A Survey on Encrypted Traffic Classification

Zigang Cao[1,2], Gang Xiong[2,*], Yong Zhao[2], Zhenzhen Li[2], and Li Guo[2]

[1] Beijing University of Posts and Telecommunications, Beijing, China
[2] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
xionggang@iie.ac.cn

Abstract. With the widespread use of encryption techniques in network applications, encrypted network traffic has recently become a great challenge for network management. Studies on encrypted traffic classification not only help to improve the network service quality, but also assist in enhancing network security. In this paper, we first introduce the basic information of encrypted traffic classification, emphasizing the influences of encryption on current classification methodology. Then, we summarize the challenges and recent advances in encrypted traffic classification research. Finally, the paper is ended with some conclusions.

Keywords: traffic classification, encrypted traffic, statistical classification, fine-grained, behavior based.

1 Introduction

Network traffic classification is the keystone of network management, network planning and network flow model researching. The emergence of new applications using encryption techniques has resulted in a boom in encrypted traffic, bringing new challenges to the traffic classification field. The increase in encrypted traffic results from the following aspects:

- Peer to Peer (P2P) applications, since 2005, have been trying to break the restrictions of Internet Service Providers (ISPs), by using encryption and protocol obfuscation techniques [1-3].
- Internet users are paying more attention to their online security and privacy. Security Socket Layer (SSL), Virtual Private Network (VPN), and Secure Shell (SSH) are widely exploited to ensure network security. Anonymous communications (e.g. Onion Routing [4]) are used to enhance privacy preserving.
- With the fast growth in computing capability of general devices, even a personal computer or mobile devices can easily run complicated encryption and decryption calculation, which provides essential conditions for applications using encryption.

As is known to us, the main purpose of traffic classification is quality of service (QoS). The challenges encrypted traffic brings for traffic classification are in the following aspects, which call for further advances.

* Corresponding author.

L. Batten et al. (Eds.): ATIS 2014, CCIS 490, pp. 73–81, 2014.

- First, it is difficult to achieve accurate and real-time identification of encrypted network applications, such as P2P downloading and online video, to fulfill the QoS requirements.
- Second, enterprise information security is challenged by encrypted channels. Malwares such as botnets, Trojans and advanced persistent threat (APT) [5] are using encrypted techniques to bypass firewall and intrusion detection system (IDS), so they can transmit confidential information to the outside network.
- Third, fine-grained user network behavior management requires accurate classification of encrypted traffic. In most companies and organizations, playing games, watch videos, and P2P downloading are forbidden during working hours. However, some employees try to break the rules by using encrypted tunnels. So it is necessary to know what applications are running inside encrypted tunnels.

Encrypted traffic classification can provide technical support for QoS, network behavior management, as well as the detection and forensics analysis of cybercrime. Some recent survey works in traffic classification are [6,7,8,9]. However, to the best of our knowledge, these survey works are not specially focused on the encrypted traffic problems. Thus, some important and unique challenges in encrypted traffic classification are not discussed, such as the definitions of encrypted traffic, and the different classification requirements. Therefore, it is essential to survey recent works and pinpoint the key problems, which may promote further research in the field.

The rest of this paper is organized as follows. Section 2 introduces the foundation of encrypted traffic, including encrypted traffic categories, classification requirements, classification methodology and challenges. In section 3, recent advances in encrypted traffic classification are reviewed. Finally, Section 4 concludes the paper.

2 Encrypted Traffic Classification Foundation

2.1 Scope and Classification Requirements

In a broad sense, encrypted network traffic should include traffic which has been transformed or generated by an encryption algorithm. While in practice, encrypted traffic mainly refers to the traffic in which the real content to be transferred is encrypted. However, if one uses plaintext HTTP protocol to download an encrypted file, the traffic cannot be taken as encrypted since the protocol itself is not encrypted.

To the best of our knowledge, encrypted traffic categories studied publicly are as follows: SSH, VPN, SSL, encrypted P2P, encrypted voice over Internet Protocol (VoIP), and encrypted traffic by certain anonymity tools (e.g. Tor [10] and JAP [11]). It should be noted that there may be a cross between two types. For instance, SSL VPN traffic can be grouped into both SSL and VPN. Another example is Skype, which belongs to both encrypted P2P and encrypted VoIP.

As for the classification requirements, encrypted traffic has the natural fine-grained classification. Encrypted traffic often has tunnels inside, which further carry several different applications, so not only encrypted traffic need to be identified, but also the

applications running in the encrypted tunnels need to be classified, too. This is why encrypted traffic classification is much more difficult.

2.2 Classification Methodology

The important premise of classification is that there are features for different applications which can be used to distinguish each other. In our opinion, the essential difference between the classification of encrypted traffic and unencrypted traffic is that the classification methods change as available useful features alter due to the encryption. What the encryption process changes in network communications can be summarized as follows. First, the content inside the IP packets changes from plaintext to ciphertext. Second, the statistics of the payload is changed after encryption, namely randomness or entropy. Third, changes are also in the statistical properties of packet level and flow level, such as packet length, intervals and flow numbers. It is just these changes that greatly challenge the current classification methods.

Port-based method is based on the feature that certain application services use IANA assigned port numbers [12]. This method suffers from the following shortcomings. First, P2P applications use random or dynamic port numbers. Second, common service ports may be used by other services, such as malwares. Third, there are port numbers besides the assigned. Fourth, it is coarse-grained. Finally, port numbers can be hidden by transport layer or IP packet encryption.

Payload based method generally refers to the deep packet inspection DPI [13] technique, which uses static application signatures in the payload to identify protocols. Finamore A et al [14] proposed another payload based method named stochastic packet inspection (SPI) which makes use of statistical properties of payload in packets. DPI is greatly damaged by encryption since the plaintext signatures turn invisible. However, it can be used in coarse classification for certain encrypted traffic such as SSL. As for SPI, it has the fine-grained classification ability in theory since its features are generally specific for application-layer protocols. However, statistical payload properties it relies on will be greatly changed after encryption. It can be useful when the encryption is partial and structured.

Statistical classification mainly refers to the methods based on statistical properties of traffic, in which machine learning is the most common one. The statistics used can be roughly divided into packet level and flow level. The former includes packet length, packet intervals and directions et al, and the latter contains the count and ratio of the upstream and downloading in bytes and packets, the duration of flow, ratio of different types of packets, etc. Though encryption changes the statistics of packet and flow, there are often strong correlations between the unencrypted traffic and original encrypted one. This is the main reason why statistical protocol identification is useful.

Behavior based classification is to analyze the behavioral characteristics of different types of applications from the host perspective, which mainly depends on the connection patterns, so the classification results are usually a series of coarse-grained classes, such as P2P and web. It is not enough due to the following reasons. Firstly, it is coarse-grained. Secondly, it can hardly work in case of transport layer encryption. Thirdly, network environment such as the use of network address translation (NAT)

[15] and asymmetric routing can affect its performance due to incomplete connection information.

To sum up, port based method can be auxiliary. Besides the lack of general fine-grained classification ability, payload based methods have a problem of privacy invasion. Statistical classification and host behavior based classification do not rely on content signatures, and both of them are robust to most application layer encryption. That is why they are the mainstream methods for in the field today.

2.3 Challenges in Encrypted Traffic Classification

Though there are many works, several challenges in this field have not been overcome yet. The main challenges are summarized as follows.

Fine-Grained Classification of Encrypted Traffic Is a Tough Task
It is far from enough to tell the encrypted traffic from the unencrypted, since what the real world need is to distinguish the tunneled application layer protocols to fulfill the network management. So the question is that can encrypted traffic be fine-grained classified in real world by port-based, payload based, statistical properties based, or behavior based method, as well as combinations of different methods.

Problems in Large-Scale Datasets Generation and Labeling Are Still to Be Solved
Datasets used in research generally come from three sources, i.e. public datasets, self-generated ones, and shared ones. Lack of payload information and ground truth is very common in public datasets and shared ones, so some researchers have to rely on port number for labeling [16], causing the benchmark inaccurate. An obvious problem of self-generated data sets is that the amount of data is too small.

As for the labeling, the most common ways are based on port number and DPI, so the accuracy is not trustworthy. Moreover, to meet the fine-grained requirements of encrypted traffic, it is essential to know what is running inside the encrypted tunnels, which makes the problem worse. Therefore, how to obtain accurate benchmark is a great challenge.

Overcoming Countermeasures against Traffic Analysis Is a Long-Term Fight
Since statistical classification is the most commonly used method for encrypted traffic, the adversaries have been developing new countermeasures. A confrontational encryption technique called protocol obfuscation [1] or traffic morph [17] has been developed, which is designed to fight against statistical methods by camouflaging one protocol to look like the target one or normal traffic in statistical properties. Experiment results in related works [17,18] showed that their methods were effective to beat statistical methods. Recently, Dyer K P et al [19] proposed a method called format transforming encryption to mimic any protocol format that can be denoted by a regular expression no matter what the input is. It is a universal framework for bypassing payload based classification.

3 Recent Advances in Encrypted Traffic Classification

First, classification explorations for accurate and fine-grained encrypted traffic classification are introduced in Section 3.1 to 3.3. Finally, advances in traffic analysis countermeasures and solutions are summarized in Section 3.4.

3.1 Accurate Classification Explorations

From *comparison* works in [20, 21], a rough conclusion may be that C4.5 and multi-objective genetic algorithm (MOGA) are two better choices in SSH encrypted traffic classification. Works in [22,23,24] showed that proper combinations of different techniques could overcome already well-performing stand-alone ones.

Exploration of new methods is always a direct path to effective solutions. Bacquet C et al [25] applied genetic programming to encrypt traffic classification. They used an extended MOGA in feature selection and cluster count optimization for K-Means, resulting in that the detection rate got an increase of 2% to 5%, while the FPR did not increased significantly. Xie G et.al [26] used subspace clustering to make the new classifier learn to *identify each application separately* just using its own relevant features instead of distinguishing one application from another using the unified feature sets. The approach showed very high accuracy on five traces from different ISPs, and was adaptable to change.

3.2 Multi-phased Fine-Grained Classification

Since fine-grained classification becomes the general need in real world, it is a common idea to achieve the goal by *multi-phased classification*, in which different tasks are finished respectively. A two-phased method [27] was used to classify SSH tunnel traffic, in which the SSH traffic was identified firstly, and then the statistical attributes such as average packet length et al were used to classify the applications inside the tunnel. Adami D et.al [28] proposed a joint signature-based and statistical approach called Skype-Hunter to detect and classify Skype signaling and data flows in real time. Korczynski M et.al [29] presented a three-phase method for classifying SSL encrypted Skype service TCP flows based on statistical protocol identification, which distinguished voice calls, skypeOut, video conferencing, chat, file upload and download in different phases. In [30], fine-grained classification was done by hierarchical multi-staged classification using multiple different classifiers.

To sum up, a possible general rule for encrypted traffic classification can be a multi-step process. The first phase is to identify certain encrypted traffic, and the left thing to do is fine-grainedly classifying the inner services.

3.3 Behavior Based Fine-Grained Classification

Behavior based methods are one of the common solutions to encrypted traffic classification, which is especially useful for identifying P2P applications. Behaviors here

can be roughly divided into host behavior and application behavior. Host based behaviors are coarse behaviors for a class of similar applications, so the classification results are coarse-grained, such as in [31,32,33]. This type of behavior is robust for encryption, application update, and new protocols, so it may play an important role in real-time coarse-grained classification of the backbone network traffic.

Application behavior based methods relying on periodic application operations, communication mode inside the certain network, et al can be useful for fine-grained classification. Schatzmann D et.al [34] exploited host and protocol correlations, as well as periodic behavior features to detect encrypted webmail out of HTTPS traffic by Netflow data. In [35], the count of packets and bytes exchanged among peers during small time-windows were relied on for fine-grained classification of P2P-TV traffic. Xiong G et.al [36] proposed a real-time detection method for encrypted P2P traffic based on host behavior association. Based on some priori knowledge, P2P connections were identified by the communication mode between peers, peer and server, and so on. However, priori knowledge is needed and DNS traffic has to be inspected for correlation analysis. In [37], the behavior of an SSL-encrypted application, i.e. the possible SSL message type sequences, was modeled by a first-order homogeneous Markov chain and used in fine-grained application classification.

Though it seems reasonable that application behavior based approach is useful for fine-grained classification, the fact may be that only a small set of encrypted application traffic can be classified by the method. The performance of application behavior based method for encrypted traffic is still to be explored.

3.4 Countermeasures of Statistical Traffic Analysis

The countermeasures for traffic analysis of encrypted traffic are mostly against statistical classification. As is mentioned above, Wright C et.al [17] proposed a method to morph one class of traffic to look like another class in packet size distribution, which used convex optimization techniques to modify the packets real-time. SkypeMorph [18] was able to disguise the traffic from Tor clients to the bridges of Tor network as Skype video traffic, both in packet size and packet intervals, as well as the behaviors. Meanwhile, the official site of Tor provided its encryption proxy called OBFSProxy [38], claiming that the Tor traffic can be obfuscated as traffic of another protocol, such as regular HTTP, with multi-protocol supported. Similar statistical packet features masking work can be seen in [39, 40].

On the contrary, Dyer K et al [41] provided a comprehensive analysis of traffic analysis countermeasures in HTTP traffic over encrypted tunnels and showed that nine known countermeasures were vulnerable to simple attacks which exploited coarse features of traffic. In [42], It was clearly pointed out that the protocol mimicry in [18, 39] et al were not good enough, and the partial imitation could be easily identified by several passive and active methods.

In a word, protocol imitation is widely used by anonymity tools, malwares (such as P2P botnet) and attackers today to resist traffic classification techniques. It can be predicted that more anti-classification techniques will appear in future, and current classification methods much evolve to confront the coming challenges.

4 Summary and Conclusions

Encrypted traffic classification is one of the most challenging problems in traffic classification field. In this paper, we exhibit the landscape of encrypted traffic classification comprehensively. Firstly, the necessity of encrypted traffic classification is introduced. Then, the basis of encrypted traffic is summarized, followed by the classification methodology and challenges. After that, recent advances and the focuses are reviewed. We believe our work can benefit researchers in the field. In our opinion, main challenges in the field are:

- How to build large-scale datasets of encrypted traffic with accurate fine-grained ground truth is a huge burden.
- With a large number of applications using encryption techniques, how to fine-grainedly classify the traffic of a wide variety of different encryption applications is another challenge.
- How to tackle the countermeasures against statistical classification and payload-based identification is a difficult issue in future.

From the current practice, no single method is good enough. A possible practical solution to fine-grained classification may be a multiple layer classification framework composed of two or more methods. Moreover, better performance can be achieved if data related constraints are considered as in [43]. A conclusion in encrypted traffic classification is that fine-grained classification is both an essential need and a promising direction. Besides, we believe that encrypted traffic classification will benefit network forensics in future.

Acknowledgements. This work is supported by the National Science and Technology Support Program under Grant No. 2012BAH46B02 and 2012BAH45B01; the National High Technology Research and Development Program (863 Program) of China under Grant No. 2011AA010703; and the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDA06030200.

References

1. eMule-Project.net - Protocol Obfuscation, http://www.emule-project.net/home/perl/help.cgi?l=1&rm=show_topic&topic_id=848
2. BitTorrent protocol encryption-Wikipedia, http://en.wikipedia.org/wiki/BitTorrent_protocol_encryption
3. Help for Skype: Does Skype use encryption, https://support.skype.com/en/faq/FA31/does-skype-use-encryption?frompage=search&q=encryption
4. Goldschlag, D., Reed, M., Syverson, P.: Onion routing. Communications of the ACM 42(2), 39–41 (1999)
5. Tankard, C.: Advanced Persistent threats and how to monitor and deter them. Network Security 2011(8), 16–19 (2011)

6. Valenti, S., Rossi, D., Dainotti, A., Pescapè, A., Finamore, A., Mellia, M.: Reviewing traffic classification. In: Biersack, E., Callegari, C., Matijasevic, M., et al. (eds.) Data Traffic Monitoring and Analysis. LNCS, vol. 7754, pp. 123–147. Springer, Heidelberg (2013)
7. Dainotti, A., Pescape, A., Claffy, K.: Issues and future directions in traffic classification. IEEE Network 26(1), 35–40 (2012)
8. Hu, B., Shen, Y.: Machine learning based network traffic classification: A Survey. Journal of Information and Computational Science 9(11), 3161–3170 (2012)
9. Nguyen, T., Armitage, G.: A survey of techniques for internet traffic classification using machine learning. IEEE Communications Surveys and Tutorials 10(4), 56–76 (2008)
10. Tor Project, https://www.torproject.org/
11. JAP – ANONYMITY & PRIVACY, http://anon.inf.tu-dresden.de/index_en.html
12. Service Name and Transport Protocol Port Number Registry, http://www.iana.org/assignments/service-names-port-numbers
13. Dubrawsky, I.: Firewall evolution - deep packet inspection. Infocus (July 2003), http://www.symantec.com/connect/articles/firewall-evolution-deep-packet-inspection
14. Finamore, A., Mellia, M., Meo, M., et al.: Kiss: Stochastic packet inspection. In: The First International Workshop on Traffic Monitoring and Analysis, pp. 117–125 (2009)
15. Tsirtsis, G.: Network address translation-protocol translation (NAT-PT). RFC 2766, IETF (2000)
16. Alshammari, R., Zincir-Heywood, A.N.: A flow based approach for SSH traffic detection. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 296–301 (2007)
17. Wright, C., Coulls, S., Monrose, F.: Traffic morphing: an efficient defense against statistical traffic analysis. In: The 14th Annual Network and Distributed Systems Symposium (2009)
18. Mohajeri, M.H., Li, B., Derakhshani, M., et al.: Skypemorph: protocol obfuscation for tor bridges. In: 2012 ACM Conference on Computer and Communications Security, pp. 97–108 (2012)
19. Dyer, K.P., Coull, S.E., Ristenpart, T., et al.: Protocol misidentification made easy with format-transforming encryption. In: 2013 ACM SIGSAC Conference on Computer & Communications Security, pp. 61–72 (2013)
20. Alshammari, R., Zincir-Heywood, A.: Machine learning based encrypted traffic classification: identifying SSH and skype. In: the 2009 IEEE Symposium on Computation Intelligence in Security and Defense Applications, pp. 1–8 (2009)
21. Bacquet, C., Gumus, K., Tizer, D., Zincir-Heywood, A., Heywood, M.: A comparison of unsupervised learning techniques for encrypted traffic identification. Journal of Information Assurance and Security 5, 464–472 (2010)
22. Bar - Yanai, R., Langberg, M., Peleg, D., Roditty, L.: Realtime classification for encrypted traffic. In: Festa, P., et al. (eds.) SEA 2010. LNCS, vol. 6049, pp. 373–385. Springer, Heidelberg (2010)
23. Dainotti, A., Pescapé, A., Sansone, C.: Early classification of network traffic through multi-classification. In: Domingo-Pascual, J., Shavitt, Y., Uhlig, S. (eds.) TMA 2011. LNCS, vol. 6613, pp. 122–135. Springer, Heidelberg (2011)
24. Jaber, M., Cascella, R.G., Barakat, C.: Using host profiling to refine statistical application identification. In: The 2012 IEEE INFOCOM, pp. 2746–2750 (2012)
25. Bacquet, C., Zincir-Heywood, A., Heywood, M.: Genetic optimization and hierarchical clustering applied to encrypted traffic identification. In: IEEE Symposium on Computational Intelligence on Cyber Security, pp. 194–201 (2011)

26. Xie, G., Iliofotou, M., Keralapura, R., et al.: SubFlow: towards practical flow-level traffic classification. In: IEEE INFOCOM, pp. 2541–2545 (2012)
27. Hirvonen, M., Sailio, M.: Two-phased method for identifying ssh encrypted application flows. In: The 7th International Conference on Wireless Communications and Mobile Computing (IWCMC), pp. 1033–1038 (2011)
28. Adami, D., Callegari, C., Giordano, S., et al.: Skype-Hunter: A real-time system for the detection and classification of Skype traffic. International Journal of Communication Systems 25(3), 386–403 (2012)
29. Korczynski, M., Duda, A.: Classifying service flows in the encrypted skype traffic. In: 2012 IEEE International Conference on Communications (ICC), pp. 1064–1068 (2012)
30. Grimaudo, L., Mellia, M., Baralis, E.: Hierarchical learning for fine grained internet traffic classification. In: The 8th International Wireless Communications and Mobile Computing Conference (IWCMC), pp. 463–468 (2012)
31. Karagiannis, T., Papagiannaki, K., Faloutsos, M.: BLINC: multilevel traffic classification in the dark. ACM SIGCOMM Computer Communication Review 35(4), 229–240 (2005)
32. Li, B., Ma, M., Jin, Z.: A VoIP traffic identification scheme based on host and flow behavior analysis. Journal of Network and Systems Management 19(1), 111–129 (2011)
33. Hurley, J., Garcia-Palacios, E., Sezer, S.: Host-based P2P flow identification and use in real-time. ACM Transactions on the Web (TWEB) 5(2), 7 (2011)
34. Schatzmann, D., Mühlbauer, W., Spyropoulos, T., et al.: Digging into HTTPS: flow-based classification of webmail traffic. In: 10th ACM SIGCOMM Conference on Internet Measurement, pp. 322–327 (2010)
35. Bermolen, P., Mellia, M., Meo, M., et al.: Abacus: Accurate behavioral classification of P2P-TV traffic. Computer Networks 55(6), 1394–1411 (2011)
36. Xiong, G., Huang, W., Zhao, Y., Song, M., Li, Z., Guo, L.: Real-time detection of encrypted thunder traffic based on trustworthy behavior association. In: Yuan, Y., Wu, X., Lu, Y. (eds.) ISCTCS 2012. CCIS, vol. 320, pp. 132–139. Springer, Heidelberg (2013)
37. Korczynski, M., Duda, A.: Markov chain fingerprinting to classify encrypted traffic. In: 2014 IEEE INFOCOM, pp. 781–789 (2014)
38. Tor Project: obfsproxy, https://www.torproject.org/projects/obfsproxy.html
39. Weinberg, Z., Wang, J., Yegneswaran, V., et al.: StegoTorus: a camouflage proxy for the Tor anonymity system. In: The 2012 ACM Conference on Computer and Communications Security, pp. 109–120 (2012)
40. Iacovazzi, A., Baiocchi, A.: From ideality to practicability in statistical packet features masking. In: The 8th International Wireless Communications and Mobile Computing Conference (IWCMC), pp. 456–462 (2012)
41. Dyer, K., Coull, S., Ristenpart, T., Shrimpton, T.: Peek-a-boo, i still see you: why efficient traffic analysis countermeasures fail. In: The 2012 IEEE Symposium on Security and Privacy, pp. 332–346 (2012)
42. Houmansadr, A., Brubaker, C., Shmatikov, V.: The parrot is dead: observing unobservable network communications. In: 2013 IEEE Symposium on Security and Privacy (SP), pp. 65–79 (2013)
43. Wang, Y., Xiang, Y., Zhang, J., et al.: Internet traffic clustering with constraints. In: The 8th International Wireless Communications and Mobile Computing Conference (IWCMC), pp. 619–624 (2012)