Encrypted Internet Traffic Classification Method based on Host Behavior

[1,*]Chengjie GU, [1]Shunyi ZHANG, [2]Xiaozhen XUE
[1]*Institute of Information Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*
[2]*Department of Computer Science, Texas Tech University, Texas 79415, USA*
jackiee.gu@gmail.com
doi:10.4156/jdcta.vol5. issue3.16

Abstract

Accurate network traffic classification plays important roles in many areas such as traffic engineering, QoS and intrusion detection etc. Encrypted Peer-to-Peer (P2P) applications have dramatically grown in popularity over the past few years, and now constitute a significant share of the total traffic in many networks. To solve the drawback of the previous classification scheme for encrypted network traffic, we propose an encrypted Internet traffic classification method based on host behavior. Experiment results illustrate this method can classify encrypted traffic using source-destination IP pairs and connection characteristics with high accuracy and faster computational time.

Keywords*: Encrypted Traffic, Host Behavior, Traffic Classification, Profile*

1. Introduction

Internet traffic classification is to associate the observed traffic with a specific application, and the classification results are used for profiling network usage and controlling the traffic under institutional policies etc [1]. The demand for bandwidth management methods that optimize network performance and provide QoS guarantees has increased substantially in recent years [2]. Therefore, network traffic classification plays important roles in many areas such as traffic engineering, QoS, and intrusion detection etc.

Internet traffic classification could be easily implemented by reading port numbers in the early Internet [3]. This method is no longer valid because of the inaccuracy and incompleteness of its classification results. Several payload-based analysis techniques have been proposed to inspect the packets payload searching for specific signatures [4]. Although this solution does can achieve high classification accuracy, it can't work with encrypted traffic or newly P2P applications [5]. At the same time, traffic classification method based on flow statistics shows effective performance in this field. Substantial attention has been invested in data mining techniques and machine learning algorithms using flow features for traffic classification.

While traffic classification methods based on flow statistics offer various degrees of successes, there are two challenges in identifying encrypted applications using flow properties. First, different flows in the same application may have different flow statistics, since encrypted applications are usually complicated. For example, in the encrypted P2P file-sharing applications, some flows are used to get peer information, other flows are to negotiate between peers, and other flows are involved in the actual file transfer. These various kinds of flows have different statistical properties and even different transport protocols. Second, some flows in the given application do not have obvious and specific flow statistics. If we only look at per-flow statistics, these flows are very similar to some flows in other applications.

To address the above-mentioned problems, we take two aspects to improve the accuracy and speed of this method for network traffic classification. 1) Observing statistics of individual flows, we build IP flow profile for a given application, which describe the communication patterns of this application. 2) We use source-destination IP pairs and connection characteristics to classify the traffic with high accuracy and faster computational time.

The remainder of this paper is structured as follows. Related work is represented in Section 2. Section 3 discusses the host behavior frame of traffic classification in detail. Section 4 illustrates how

to construct encrypted traffic classification architecture based on host behavior. Section 5 presents the experimental results and analysis. The conclusion and potential future work are listed in Section 6.

2. Related Work

The field of Internet traffic classification has received continuous interest. In the early Internet, traffic classification relied on the use of transport layer port numbers, typically registered with IANA to represent a well-known application [6]. But this approach is often inaccurate, due to the dramatic increase in network applications using random ports or in tunneling through HTTP.

In order to deal with the disadvantages of the above method, payload-based classification method is proposed to inspect the packet payload. Payload-based classifiers rely on the application specific signatures in the payloads, but they need highly computational costs [7]. After early work showed the value of payload signatures in traffic classification, others have proposed automated ways to identify such signatures, while they evaluated the automated schemes only on conventional applications such as FTP, SMTP, HTTP, HTTPS, DNS, and SSH, not on newer applications such as P2P, Games, and Streaming [8-10].

Machine learning technique which is a powerful tool in data separation in many disciplines aims to classify data based on either a priori knowledge or statistical information extracted from raw dataset. The branch that appears to solve the limitations of the network traffic classification methods is flow statistics analysis based on machine learning (ML) [11-18]. Nguyen et al. [11] provided context and motivation for the application of ML techniques to IP traffic classification, and reviewed some significant works. Machine learning algorithms are generally divided into supervised learning and unsupervised learning. Unsupervised learning essentially clusters flows with similar characteristics together [12-14]. Supervised learning requires training data to be labeled in advance and produces a model that fits the training data [15-18].

3. Host Behavior

3.1. Host behavior description for application

The basic insight exploited by our method is that interactions between network hosts display diverse patterns across the various application types. We model each application by capturing its interactions through empirically derived signatures. We visually capture these signatures using graph that reflect the most common behavior for a particular application. A sample of application graph is presented in Figure 1. Each graph captures the relationship between the use of source and destination ports, the relative cardinality of the sets of unique destination ports and IPs as well as the magnitude of these sets.

In our view, each graph for network traffic application has four factors corresponding to the 4-tuple (source IP, destination IP, source port and destination port). Figure 1(a) presents a mail server supporting IMAP, POP, and SMTP, while also acting as a DNS server. Figure 1(b) displays a typical P2P application where a host scans the address space to identify vulnerability at a particular destination port. In such cases, the source host may or may not use different source ports, but such attacks can be identified by the large number of flows destined to a given destination port. In some cases, hosts offering services on certain ports exhibit similar behavior. For instance, P2P, VoIP, and games all result in the similar type of graph which is illustrated in Figure 1(b) and Figure 1(c). Figure 1(d) presents the host behavior pattern of WWW application. A single source IP communicates with multiple destinations using the same source port on several different destination ports. In such cases, we need further analysis to distinguish between applications. Applications such as ftp, streaming or mail present more complicated graphs have more than one service ports, or have both source and destination service ports. Lastly, graphs become even more complex when services are offered through multiple application and/or transport protocols.

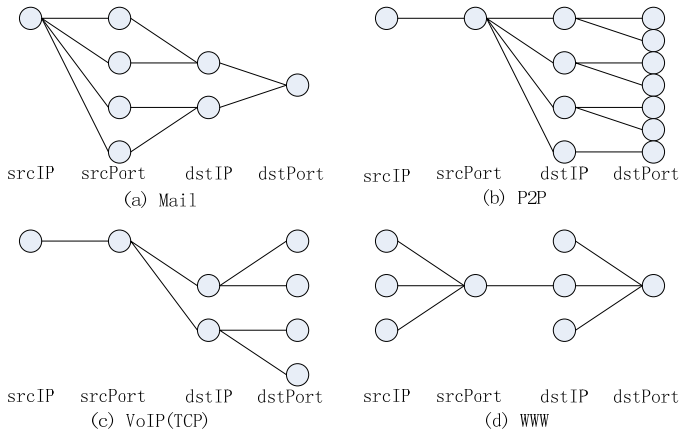

Figure 1. Host behavior for different network traffic

3.2. IP flow profile based on host behavior

We now describe our proposed IP flow profile. Our heuristics are based on observing connection patterns of source and destination IPs. While some of these patterns are not unique to hosts, examining the flow history of IPs can help eliminate false positives and reveal distinctive features. We employ two main heuristics that examine the behavior of two different types of pairs of flow keys. The first examines source-destination IP pairs that use both TCP and UDP to transfer data (TCP/UDP heuristic). The second is based on how host peers connect to each other by studying connection characteristics of {IP, port} pairs.

TCP/UDP heuristic identifies source-destination IP pairs that use both TCP and UDP transport protocols. To identify P2P hosts we can thus look for pairs of source-destination hosts that use both transport protocols TCP and UDP. Six out of nine analyzed P2P protocols use both TCP and UDP as layer-4 transport protocols. These protocols include Xunlei, PPlive, PPStream, BitTorrent, Kougou and eDonkey. Generally, control traffic, queries and query-replies use UDP, and actual data transfers use TCP. While concurrent usage of both TCP and UDP is definitely typical for the aforementioned P2P protocols, it is also used for other application layer protocols such as DNS or streaming media. To determine non-P2P applications in our traces that use both transport protocols, we examined all source-destination host pairs for which both TCP and UDP flows exist. In summary, if a source-destination IP pair concurrently uses both TCP and UDP as transport protocols, we consider this flow as P2P flow or Game.

{IP, port} heuristic is based on monitoring connection patterns of {IP, port} pairs. Since the lawsuit against Napster, the prevalence of centralized P2P networks has decreased dramatically, and distributed or hybrid P2P networks have emerged. To connect to these distributed networks, each P2P client maintains a starting host cache. This pool of hosts facilitates the initial connection of the new peer to the existing P2P network. While in first-generation P2P networks the listening port was well-defined and specific to each network, simplifying P2P traffic classification, newer versions of all P2P clients allow the user to configure a random port number. The super peer must propagate this information, mainly the {IP, port} pair of the new host A, to the rest of the network. This {IP, port} pair is essentially the new host's ID, which other peers need to use to connect to it. In summary, when a P2P host initiates either a TCP or a UDP connection to peer, the destination port will also be the advertised listening port of host, and the source port will be an ephemeral random port chosen by the client.

4. Encrypted Internet Traffic Classification Method based on Host Behavior

In this section, we propose encrypted Internet traffic classification method based on host behavior. Internet traffic classification schemes operate on the notion of network flows. A flow is defined to be as a series of packet exchanges between two hosts, identifiable by the 5-tuple (source address, source

port, destination address, destination port, transport protocol), with flow termination determined by an assumed timeout or by distinct flow termination semantics. For each flow, network monitors can record statistics such as source-destination IP pairs and connection characteristics.

As illustrated in Figure 2, network capture function is responsible for collection of all the needed information. According to flow information from flow behavior preprocessing module, flow behavior is computed in terms of each selected feature and stored them to the corresponding database when reaching some milestone. Then, classifier model which classifies the traffic according to host behavior is to implement network traffic classification. Eventually, the classification output would be applied to network activities such as network surveillance and QoS.

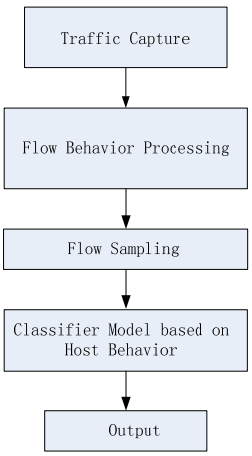

Figure 2. Architecture of encrypted Internet traffic classification based on host behavior

5. Experimental Results and Analysis

5.1. Empirical traces

This subsection describes the empirical traces in our work. Moore_Set was collected from the experiment of Pro. Moore from Cambridge University. Univetsity_Set was collected from Nanjing University of Posts and Telecommunications. To simplify the presentation, we group the applications by category. For example, the P2P category includes all identified P2P traffic from protocols including PPlive, PPstream, BitTorrent, Gnutella, Xunlei and KaZaA.

Moore_Set trace consists of bidirectional network traffic of some biological research institute during 0 to 24 o'clock on Aug 20th, 2003. We extract 10 subsets with an average sampling time of 1680s to form our dataset, which contains 377526 samples of network flow. These samples are divided into 10 types. The application names of each type and the quality as well as the respective proportion of each network flow are shown in Table 1.

Table 1. Statistics of Moore_Set

Type of flow	Application names	Num of flow	Percent(%)
WWW	http, https	328091	86.91
MAIL	Imap, pop3, smtp	28567	7.567
BULK	ftp	11539	3.056
DATABASE	oracle, mysql	2648	0.701
SERVER	ident,ntp,x11,dns	2099	0.556
P2P	kazaa,bittorrent	2094	0.555
ATTACK	worm, virus	1793	0.475
MEDIA	real, media player	1152	0.305
INT	telnet,ssh,rlogin	110	0.029
GAME	half-life	8	0.002
Total	26 applications	377526	100

Each network flow sample of Moore Set is derived from a complete bidirectional TCP flow and contains 248 attributes, among which the first and second attributes are port numbers of source and destination respectively.

To facilitate our work, we collect traces in all academic units and laboratories on the campus from the Internet gateway of Nanjing University of Posts and Telecommunications. Univetsity_Set was collected over a span of six months from April 10, 2009 to October 10, 2009. Table 2 summarizes the applications found in the 40 1-hour Campus traces. On the campus network, HTTP and DATABASE traffic contribute a significant portion of the total flows. P2P application also accounts for a considerable portion, approximately 42.14%.

Table 2. Statistics of University _Set

Type of flow	Num of flow	Percent(%)
WWW	1151678	37.05
MAIL	141004	4.54
BULK	2947	0.09
DATABASE	447578	14.39
SERVER	613	0.02
P2P	1309896	42.14
MEDIA	1217	0.04
GAME	31068	1.01
UNKNOWN	22617	0.72
Total	3108618	100

5.2. Evaluation metrics

To measure the performance of our proposed method, we use three metrics: *accuracy, precision and recall*. In this paper, TP, FP, and FN are the numbers of true positives, false positives, and false negatives, respectively. True Positives is the number of correctly classified flows, False Positives is the number of flows falsely ascribed to a given application, and False Negatives is the number of flows from a given application that are falsely labeled as another application.

Accuracy is the ratio of the sum of all True Positives to the sum of all the True Positives and False Positives for all classes. We apply this metric to measure the accuracy of a classifier on the whole trace set. The latter two metrics are to evaluate the quality of classification results for each application class.

$$accuracy = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \times 100\% \quad (1)$$

Precision of an algorithm is the ratio of True Positives over the sum of True Positives and False Positives or the percentage of flows that are properly attributed to a given application by this algorithm.

$$precision = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

Recall is the ratio of True Positives over the sum of True Positives and False Negatives or the percentage of flows in an application class that are correctly identified.

$$recall = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

5.3. Comparing performance among different technology

We compare classification accuracy of encrypted traffic classification method based on host behavior with the other classification approaches solely based on port-based approach and payload-based approach and machine learning-based using SVM. For our experiments, we classify network traffic using flows from Moore_Set.

Table 3. Classification accuracy among different traffic classification method

Type of flow	Port-based		Payload-based		ML-based		Proposed method	
	Recall	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***	***Recall***	***Precision***
WWW	74.62	76.24	80.66	81.22	89.02	92.44	95.74	97.64
MAIL	73.73	72.91	79.81	77.18	88.17	88.32	94.89	93.92
BULK	66.27	65.42	71.98	69.88	80.34	81.10	87.03	86.71
DATABASE	71.76	69.84	74.06	73.45	82.42	84.62	89.14	90.22
SERVER	71.89	68.88	76.72	74.35	85.08	85.57	91.89	91.92
P2P	64.32	61.54	72.83	68.75	81.19	79.97	87.91	85.54
ATTACK	69.78	62.02	68.81	66.06	77.17	77.28	83.89	82.88
MEDIA	74.93	66.32	77.61	70.49	85.97	81.71	92.69	87.31
INT	66.48	61.18	70.17	66.95	78.53	78.17	85.25	83.79
GAME	63.07	58.56	71.21	66.11	79.57	77.32	86.29	82.92

As illustrated in Table 3, we can see that the classification precision which uses port-based methods to classify network traffic is 76.24%, 61.54% for WWW and P2P application respectively. At the same time, the classification precision which extracts payload signatures to identify specific protocol by payload-based approach is 81.22%, 68.75% for WWW and P2P application respectively. In addition, in order to classify network traffic, we choose the machine learning method using C4.5 which has 92.44%, 79.97% for WWW and P2P application respectively. Finally, we construct our encrypted traffic classification method based on host behavior to classify Internet traffic using Moore_Set, we get 97.64% and 85.54% for WWW and P2P application respectively. Compared with the other classification approaches, this method proposed in the paper can achieve higher overall accuracy.

We calculate computational performance such as computational time, memory and accuracy through the experiments among different approaches using University_Set. We can find that the encrypted traffic classification method based on host behavior is able to greatly improve the accuracy, while only minimally impacting computational time and memory. The computational performance for different approaches is demonstrated in Table 4.

Table 4. Performance among different traffic classification method

Performance	Port-based	Payload-based	ML-based	Proposed in this paper
Time(s)	34.986	652.901	1104.652	238.671
Memory(M)	9.451	56.082	108.347	34.091
Accuracy(%)	68.59	76.45	86.03	91.62

5.4. Classifying encrypted Skype from P2P traffic

To address the above-mentioned encrypted problems by exploring host behavior pattern traffic, we mainly focus on P2P traffic classification which is the most challenging problem in Internet traffic classification. Skype is an encrypted P2P VoIP network which has the similar characteristic with the Xunlei, PPlive, PPStream, BitTorrent, Kougou and eDonkey. The purpose of this subsection is to verify whether the proposed method is robust or flexible enough to detect encrypted Skype traffic from P2P traffic. We choose the P2P applications from 29 different applications identified in the University_Set to test our classification method. The main applications in our experiments include Xunlei, PPlive, PPStream, BitTorrent, Kougou, eDonkey and unkonwn P2P. Table 5 shows statistics of P2P flow from University_Set.

Table 5. Statistics of P2P flow from University_Set

Application	Num of flow	Percent(%)
Skype	10897	0.83
Xunlei	447263	34.14
PPlive	288086	21.99
PPStream	109884	8.39
BitTorrent	186794	14.26
Kougou	99636	7.61
eDonkey	67398	5.15
Unkonwn P2P	99938	7.63
Total	**1309896**	**100**

With the method proposed in this paper, the results in Figure 3 show that Skype precision and recall in University_Set are 94.51% and 83.63% respectively, which indicates that Skype can be effectively identified. Specifically, the experimental results show that precision and recall is 94.81% and 83.26% for Kougou application respectively. At the same time, the average identification accuracy of unknown P2P traffic is 86.28%, which indicates that the methods can classify unknown P2P traffic with considerable accuracy. Thus the methods are robust enough to classify Skype traffic based on the fact that the P2P applications share the similar characteristics of peer behavior, connection feature and transportation statistics.

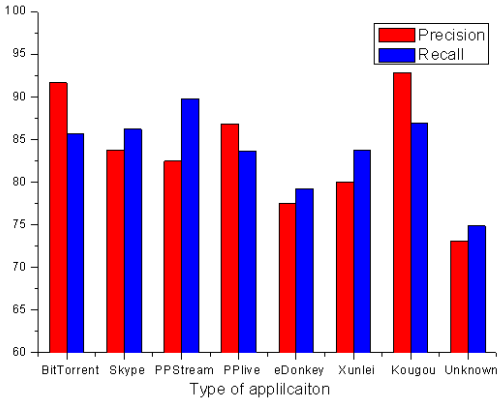

Figure 3. Identification accuracy of P2P traffic

6. Conclusions

As many newly-emerged encrypted applications use dynamic port numbers and masquerading techniques, it causes the most challenging problem in network traffic classification. One of the challenging issues for existing classification methods is that they can't classify encrypted traffic. In this paper, we propose an encrypted traffic classification method based on host behavior, which is implemented by collecting source-destination IP pairs and connection characteristics. Experiment results illustrate encrypted traffic identification methodology can meet the key criteria, such as low complexity, high accuracy and robustness to provide QoS guarantees according to all kinds of Internet application.

7. Acknowledgement

This work is supported by National High-Tech Research and Development Plan (863) of China (No.2006AA01Z232, No.2009AA01Z212, No.2009AA01Z202), Natural Science Foundation of China (No.61003237), Natural Science Foundation of Jiangsu Province (No.BK2007603), High-Tech Research Plan of Jiangsu Province (No.BG2007045).

8. References

[1] Antonio Martin, Carlos Leon, Felix Biscarri, "Intelligent Integrated Management for Telecommunication Networks ", International Journal of Advancements in Computing Technology, vol. 2, no. 2, pp. 158-171, 2010.

[2] Fazal Wahab Karam, Terje Jensen, "A Survey on QoS in Next Generation Networks", Advances in Information Sciences and Service Sciences, vol. 2, no. 4, pp. 91-102, 2010.

[3] Qindong Sun, Qian Wang, Jie Ren , "Modeling and Analysis of the Proactive Worm in Unstructured Peer-to-Peer Network ", Journal of Convergence Information Technology, vol. 5, no. 5, pp. 111-117, 2010.

[4] Haffner P, Sen S, Spatscheck O, Wang D. "ACAS: Automated Construction of Application Signatures". In Proceedings of SIGCOMM'05 Workshops, pp. 197-202, 2005.

[5] Moore A W, Papagiannaki K. "Toward the accurate identification of network applications". In Proceedings of Passive and Active Measurement Workshop, pp. 41-54, 2005.

[6] Constantinou F, Mavrommantis P. "Identifying known and unknown peer-to-peer traffic". In Proceedings of IEE NCA'06 Conference, pp. 93-102, 2006.

[7] T. Karagiannis, A. Broido, N. Brownlee, K. Claffy, M. Faloutsos. "Is P2P dying or just hiding". In Proceedings of IEEE Globecom, pp. 1532-1538, 2004.

[8] T. Karagiannis, K. Papagiannaki, M. Faloutsos. "BLINC: multilevel traffic classification in the dark". In Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications, pp. 229-240, 2005.

[9] S. Sen, O. Spatscheck, D. Wang. "Accurate, scalable in-network identification of p2p traffic using application signatures". In Proceedings of the 13th international conference on World Wide Web, pp. 512-521, 2004.

[10] B.Marco, Mellia, Antonio Pescape, Luca Salgarelli. "Traffic classification and its applications to modern networks". Computer Networks, vol. 53, no. 6, pp. 759-76, 2009.

[11] T. Nguyen, G. Armitage. "A Survey of Techniques for Internet Traffic Classification using Machine Learning". IEEE Communications Surveys and Tutorials, vol. 11, no.3, pp. 37-52, 2008.

[12] Soysal, Murat, Schmidt, Ece Guran. "Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison". Performance Evaluation, vol. 67, no. 6, pp. 451-467, 2010.

[13] J. Erman, M. Arlitt, A. Mahanti. "Traffic classification using clustering algorithms". In Proceedings of SIGCOMM workshop on Mining network data, pp. 281-286, 2006.

[14] Bernaille L, Teuxeira R, Akodkenous I, Soule A, Slamatian K. "Traffic classification on the fly". In Proceedings of ACM SIGCOMM, pp. 23-26, 2006.

[15] A. W. Moore, D. Zuev. "Internet traffic classification using bayesian analysis techniques". In Proceedings of international conference on measurement and modeling of computer systems, pp. 50-60, 2005.

[16] N. Williams, S. Zander, G. Armitage. "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification". ACM SIGCOMM Computer Communication Review, vol.30, no. 5, pp. 5-16, 2006.

[17] T. Auld, A. W. Moore, and S. F. Gull. "Bayesian neural networks for internet traffic classification". IEEE Transaction on Neural Network,vol.18, no. 1, pp. 223-239, 2007.

[18] Yongli Ma, Zongjue Qian, Guochu Shou, Yihong, Hu. "Study of information network traffic identification based on C4.5 algorithm". In Proceedings of 2008 International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1-5, 2008.