# Traffic Classification: Issues and Challenges

Yibo Xue[1], Luoshi Zhang[2], and Dawei Wang[3]
[1] Tsinghua National Lab for Information Sci. & Tech., Beijing, 100084, China
[2] Computer Science & Technology College, Harbin Univ. of Sci. & Tech, Harbin, 150080, China
[3] National Computer Network Emergency Response Technical Team / Coordination Center of China (CNCERT/CC), Beijing, 100029, China
Email: yiboxue@tsinghua.edu.cn; luoshi.zh@gmail.com; stonetools2008@gmail.com

***Abstract*—**Traffic classification has been extensively examined in recent years, as it is widely used in network management, design, security, advertising and research. In the past few years, the traffic classification techniques have been evolved along with the development of Internet protocols and applications, and many approaches have been investigated, proposed and developed. Nowadays, the ever increasing network bandwidth, the constantly sophisticated applications and the growth incentives to confuse classification systems to avoid filtering or blocking are among the reasons that traffic classification remains one of the hot areas in network research. In this paper, we first attempt to present an analysis of the existing traffic classification techniques, and dwell on their issues and challenges, then address some recommendations that can improve the performance of traffic classification systems.

***Index Terms*—**traffic classification, issues, challenges, recommendations

## I. Introduction

Nowadays, Internet has been the necessities of people's life. Besides of the traditional services such as web browsing and mailing, Internet has been providing more new services, including P2P sharing, e-commerce, live video, online game, and so on. Therefore, Internet has been attracting more and more users. For example, according to the 31th statistical report on development of Internet in China published by CNNIC [15], the Internet users of China have reached 5.64 hundred million by the end of 2012. As shown in Fig. 1, the Internet population of China has been increasing very fast in recent years. Currently, almost half Chinese people are using Internet for work or daily life.

The ever increasing users are making Internet change quickly.

- **The ever emerging protocols and applications**. On one hand, the large amount of users need more kinds of service; on the other hand, the IT vendors try their best to develop new applications in order to seize the market share. Therefore, the existing protocols and applications are far more than that of a few years ago.

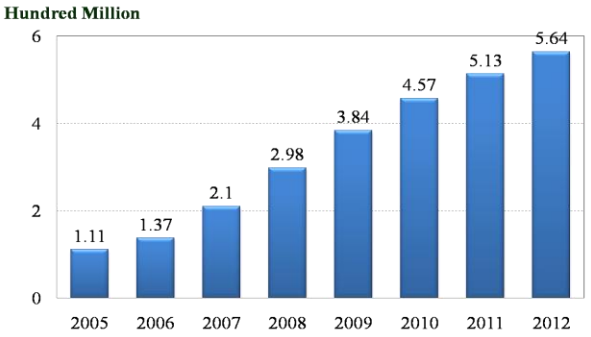



Figure 1. Internet population of China

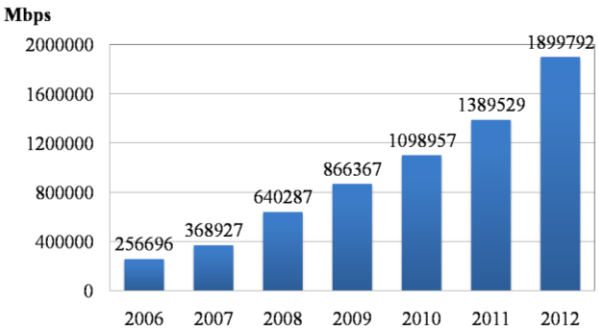



Figure 2. Chinese international bandwidth

- **The ever growing bandwidth**. Due to the ever increasing Internet population as well as the emerging protocols and applications, the Internet bandwidth has to be promoted. The Fig. 2 shows Chinese international bandwidth. From this figure, we can see that Chinese international bandwidth of 2012 has risen by more than seven times than in 2006.

The changes mentioned above also produce more challenges. Firstly, more users, more applications and larger bandwidth means more security threats; secondly, the bandwidth is still not enough, even though most ISPs have been trying to promote bandwidth, because of the bandwidth-occupying applications. Therefore, in order to safeguard Internet as well as maximize its efficiency, we need a powerful tool.

Traffic classification is an automated process which categorizes network traffic into a number of classes according to various parameters (such as protocols) [1]. Traffic classification has great potential use in a wide

Manuscript received March 7, 2013; revised April 17, 2013.
This work was supported by the National Key Technology R&D Program under Grant No.2012BAH46B04, and National High-tech R&D Program Projects under Grant No. 2011AA010601
Corresponding author email: yiboxue@tsinghua.edu.cn.
doi:10.12720/jcm.8.4.240-248

variety of disparate situations, such as network management (Quality of Service, etc.), network design (topological design, etc.) and network security (Intrusion Detection Systems (IDSs), etc.) [2].

Due to its wide usage, traffic classification has been extensively researched in recent years, and many techniques have been proposed, including port-based techniques, payload-based techniques [3], flow-based techniques [4], host-based techniques [5] and graph-based techniques [6]. Some of them have been maturely and widely deployed in the current network, and some of them are still under researching. But all the existing techniques still have their critical limitations and issues. In addition, as new coming applications become more and more sophisticated, the existing techniques are facing more and more challenges. In this paper, we attempt to present an analysis of the existing traffic classification techniques, and dwell on the issues and challenges.

The rest of the paper is organized as follows. In section II, we provide a brief summary of existing traffic classification techniques and analyze their issues. Then it is followed by section III with the general challenges of traffic classification. In section IV, we provide some recommendations for solving the current issues and challenges. In section V, we conclude the paper.

## II. Issues of Existing Traffic Classification Techniques

### A. Existing Traffic Classification Techniques

As a fundamental component of various network management and security systems, traffic classification has always been the hot topic, and many techniques have been proposed by a lot of researchers. Due to the fact that most techniques classify traffic according to the protocols or applications, the traffic classification techniques have been evolving with the development of protocols and applications illustrated as Fig. 3.

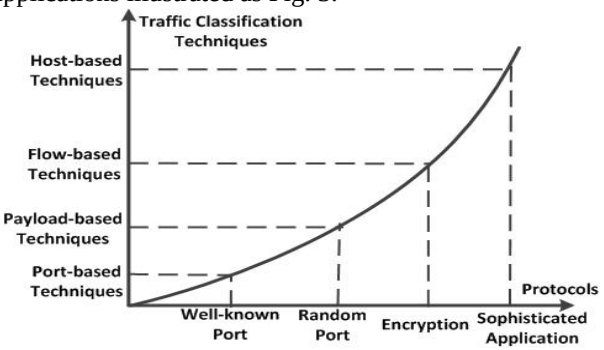



Figure 3. Evolution of protocols and traffic classification techniques.

In the early stage of Internet, protocols were quite simple, and most of them used well-known port numbers assigned by IANA (The Internet Assigned Numbers Authority). Table I shows the most popular well-known port numbers.

To classify these protocols, the port-based techniques can finish the tasks perfectly, as it assumes that most protocols consistently use "well known" TCP or UDP port numbers registered in IANA. However, as more and more protocols and applications used random port numbers (e.g., P2P protocols), the port-based techniques then lost their effectiveness.

TABLE I. Most Popular Well-known Port Numbers

| Port-No. | Service/Protocol |
|---|---|
| 20/21 | FTP - File Transfer Protocol.<br>Transmission of files between the network participants |
| 23 | Telnet – Terminal access to distant system |
| 25 | SMTP – Simple Mail Transfer Protocol<br>Sending of electronic mail |
| 80 | HTTP – Hypertext Transfer Protocol<br>Transmission of HTML-files |
| 443 | HTTPS – HTTP over Secure Socket Layer |

To address this issue, payload-based techniques were proposed. They classify traffic by comparing packet payloads with the known signatures of protocols. Although they can solve the issue of random port, they cannot deal with the encryption protocols, especially the proprietary encryption protocols, because the signatures of these types of protocols can hardly be found.

To address the challenge of (proprietary) encryption protocols, researchers have developed flow-based techniques, which explore flows statistics instead of packet payload patterns. Based on the assumption that flows generated by different protocols have unique statistical characteristics, flow-based techniques can identify the protocol that a flow belongs to, without inspecting packet payloads. However, as new coming applications become more and more sophisticated, many applications tend to integrate more than one protocol at the same time. The new sophisticated applications make flow-based techniques facing more and more challenges. The traditional flow-based techniques can only identify individual protocol but not the whole application.

To classify the traffic generated by these sophisticated applications, researchers proposed host-based [5] techniques. By monitoring all the flows sent or received from the same host, these techniques can classify all the traffic generated by the sophisticated applications. However, these techniques will lose their effectiveness when the monitored host uses more than one application at a time, because they are designed based on a presupposition that one host employs one application at the same time.

Besides of improving the above techniques continuously, the researchers also propose some novel approaches, in which the graph-based [6] techniques are most famous. These techniques employ a new idea, which pays more attention to the relationship between hosts running the same protocols or applications. Based on this idea, the graph-based techniques may have a chance to better deal with the sophisticated applications.

However, these techniques are not mature enough to be deployed in high speed network for online classification.

Although the traffic classification techniques have become more and more complex, all the above traffic classification techniques technically have their own merits and demerits. In this section, we will analyze the issues and challenges that the existing techniques are suffering.

As the port-based techniques are too simple to lose their effectiveness, we don't analyze them in this section.

### B. Challenges of Payload-based Techniques

Payload-based traffic classification techniques classify traffic via comparing packets payloads with some known signatures of target protocols. These techniques usually use two matching methods: string-based matching and regular-expression matching, as shown in Fig. 4. An example of string-based matching is OpenDPI, which supports many protocols such as known P2P protocols, HTTP, and so on [7]. String-based matching can use fast multi-pattern matching algorithms but with limited expressivity [3]. To address this issues, regular expressions are used in the later payload-based approaches. The typical example is L7-filter [8], which can classify many protocols, such as SSH, SSL, and so on. Although payload-based techniques can classify traffic with known signatures very well, they are facing the some challenges.

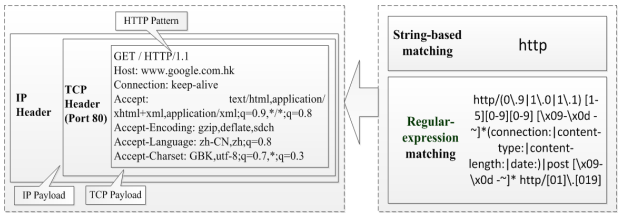



Figure 4. String-based and regular expression

The Fig. 5 shows the challenges and issues of payload-based techniques. Generally, the main challenges resulting in issues are the ever emerging protocols (especially encryption protocols), and the ever growing bandwidth. We list these issues as follows.

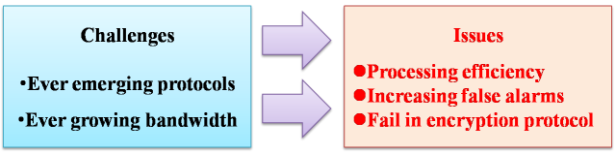



Figure 5. String-based and regular expression.

- **Processing efficiency**. As shown in Fig.2, the era of Terabit bandwidth has arrived. Currently, the traffic classification systems usually have to process Gigabits even Terabits per second, which is a fairly challenging task for payload-based techniques. All the payload-based techniques need to compare every packet with payload of a flow with the signature set until the protocol of the flow is determined. As the packets in high speed network arrive quickly and indefinitely, the payload-based classification systems need higher processing capability. Furthermore, as more and more new protocols appear, the size of pattern or regular-expression set becomes larger and larger, so the traffic classification systems need to compare packet payload with more signatures or regular expressions. Formally, we can use time complexity to demonstrate this issue. Assume that systems need to process m packets, and the size of patterns or regular expressions is n. Then in the worst case, the time complexity of payload-based techniques is $O(mn)$. As the network speed increases or the size of patterns and regular expressions grows, the processing efficiency of payload-based techniques will drop quickly.
- **Increasing false alarms**. Theoretically, the false alarm rate of payload-based techniques can be calculated by the following equation.

$$\text{false alarm rate} = \frac{1}{2^{\text{Bits of String Pattern}}}. \quad (1)$$

From the above equation, we can see that the false alarm rate of payload-based techniques is fixed. However, as the bandwidth continuously increases, the network packets arrive quickly. Although the false alarm rate is fixed, the false alarms will increase in high speed network, because more packets have more opportunities to strike false alarms. For example, there is a protocol which has a payload signature with 2 bytes. According to equation (1), its theoretical false arm rate is 1/65536. Under the low speed network, such as access networks, there would be little false alarms. But in the backbone network, this signature may strike too many false alarms. Therefore, to classify protocols using payload-based techniques in high-speed network, one should choose longer payload signatures.

- **Proprietary encryption protocol**. As people have paid more attention to the individual privacy, the encryption protocols play more and more important role in the current Internet. Besides some public encryption protocols, such as SSL, more and more applications start to use proprietary protocol for communication and encryption, such as Skype. As the specifications of these proprietary encryption protocols remain private, the signatures of packet payload can hardly be found. Thus, the payload-based techniques lose their effectiveness when classifying such kinds of encryption protocols.

The payload-based techniques play main role in the traffic classification, because of their accurate classification capability and easy deployment. However, as the protocols and applications are continuing to evolve, these techniques inevitably lose their effectiveness to the new coming protocols and applications. The flow-based techniques are proposed in this situation.

### C. Challenges of Flow-based Techniques

Different from the payload-based techniques, flow-based techniques use flow statistical features instead of packet payload patterns to classify traffic. Most flow-based techniques use Machine Learning (ML) techniques as their classification algorithms. In 2004 McGregor et al. [9] published one of the earliest methods that applied ML in traffic classification using Expectation Maximization algorithms. Then many other approaches are proposed based on ML, such as Bayesian methods [10] and SVM Support Vector Machine [11]. Fig. 6 shows the general process of flow-based techniques. Most flow-based techniques contain two main stages: training stage and classifying stage. At the training stage, the flow-based techniques use features extracted from the training dataset to train classification model. The obtained model is employed to classify target protocols or applications at the classifying stage.

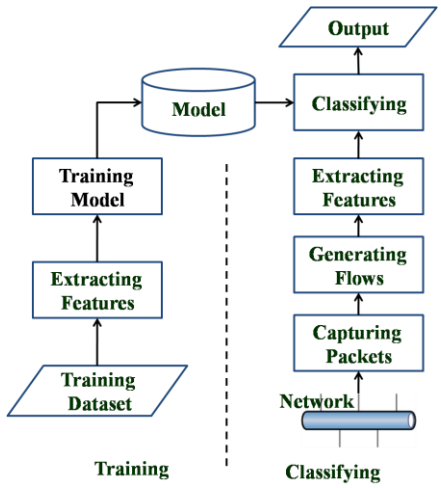



Figure 6. Process of flow-based techniques

TABLE II. FLOW STATISTICAL FEATURES

| No. | Description of Flow Statistical Feature |
|---|---|
| 1 | The bytes of payload of first *n* packets |
| 2 | The inter-arrival time of first *n* packets |
| 3 | Mean of bytes in first *n* packets |
| 4 | Mean of inter-arrival time in first *n* packets |
| 5 | Standard deviation of bytes in first *n* packets |
| 6 | Standard deviation of inter-arrival time in first *n* packets |
| 7 | The count of packets from client to server and server to client |
| 8 | For TCP flows, the count in first *n* packets that set some flags, such as PSH, RST and URG |
| 9 | The count in first *n* packets with at last a byte of data payload from client to server and server to client. |
| 10 | The count in first *n* packets without data payload from client to server and server to client. |

As these techniques do not need to inspect packet payload, they can classify traffic without invading user privacy. The issues of the proprietary encryption protocols can also be solved using flow-based techniques. Now the flow-based techniques is the hot spot in traffic classification research area, the proposed approaches can be used to classify the traffic generated by many protocols, such as HTTP, SSL and so on. However, the flow-based techniques are far away from being usable due to the following reasons.

- **Selection of flow statistical features**. In [12], Moore et al. listed more than 200 flow statistical features for flow-based techniques. Table II shows some common flow statistics, these features have already been employed by many systems to classify traffic. As the goal of flow-based techniques is to classify the target protocols or applications as early as possible, the most features are extracted from the first *n* packets of a flow. Obviously, we can not simply use all those 200 features to classify traffic. The reasons are: (1) many ML models are sensitive to the dimensionality of feature vector, which is also named as curse of dimensionality. If all the proposed 200 flow features are used to train ML classification models, the dimensionality will be too high to classify traffic effectively. (2) Due to the different protocol specifications, different protocols may have different measures in different flow statistical features. Some features maybe become meaningless, even make classification results worse. Thus, the first challenge of flow-based techniques is how to select the proper flow statistical features. (3) The flow statistical features contain both continuous and discrete ones. The feature of the length of first n packets in a flow is the continuous one. Meanwhile, the Flag of TCP flow, such as SYN, RST and FIN, belong to discrete feature. However, some ML models cannot deal with these two types of feature at the same time. In summary, a key problem for flow-based techniques is how to extract features from traffic, in other word, what flow statistical features should be selected.
- **Uniqueness proof of flow statistical features**. In recent years, many flow-based methods have been proposed. All the proposed methods were claimed that they can obtain perfect classification results. We also realized some flow-based approaches, including SVM and Bayesian, and deployed them in the backbone network to classify SSL/TLS traffic. According to testing in our lab, these two models can obtain more than 95% precision. However, when we deployed them in the backbone network, in which 1 million flows are needed to classify within a second in one server, the precision of the two models reduce to less than 5%. Through data analysis, we found that in the backbone network, there are too many non-SSL/TLS flows that have the similar statistical features as SSL/TLS flow. In other words, the selected flow statistical features for classifying SSL/TLS traffic are not unique. In fact, the same issue also bothers the payload-based techniques. Therefore, How to prove

the uniqueness of the selected statistical features is an another big issue for the flow-based techniques.

- **Selection of classification models**. There are thousands ML models available, and over hundreds are published each year. Currently, many ML models have been improved for traffic classification, such SVM, Decision Tree, Bayesian and so on. Theoretically, all the ML models classify traffic by dividing space using divisional plane. As all the models have the same theoretical foundation, there will be no best classification model but the more suitable one. There are two factors that affect the performance of ML models, including the goal of traffic classification and the selection of flow statistical features. For example, someone wants to classify traffic generated by multiple protocols, C4.5 would be better than SVM. Otherwise, if the target traffic only occupies small part of the background traffic, SVM would be the more suitable model. Therefore, how to select ML models according to the traffic characteristics is a key point to improve the classification performance.
- **Classifying all traffic according to protocols**. As it is assumed that every protocol has its unique payload signatures, payload-based techniques can classify the traffic according to protocols in background traffic. For flow-based techniques, however, this goal can hardly be achieved, because of the following reasons: (1) the selected flow statistical features of different protocols may be similar with each other. In this case, the classification models cannot classify these protocols very well. (2) Most ML models are binary classification, which means one model can only identify one protocol. If we want to classify all the protocol traffic, many ML models are needed. In a word, with the constantly increasing of Internet protocols, it is becoming more and more difficult for flow-based techniques to classify all protocols.
- **Training set**. For the flow-based techniques, the data is network traffic, meanwhile a training set is a traffic set, which is used to train ML models. In order to provide more training information, the training set usually contains traffic generated by various protocols. However, at the training stage, the target protocol can be easily ignored, if its traffic only occupy a small proportion in the training set. In order to train a model for classifying the target protocol, one can increase its traffic artificially. However, the artificial training set cannot describe the real network environment, and tend to cause the bias of models. So building the training set is a tradeoff process. The training set should be built based on the application context and network environment. Moreover, the trained models are still not accuracy enough sometimes, no matter how we adjust the training set. Generally, there are two main choices: cleverer ML models or more data. Although it seems that designing a new algorithm and retraining a cleverer model is a better way, the pragmatically retraining models using more data provides a quickest path to success. In summary, it can be said that the training set is the most important foundation of flow-based techniques. The training set must be built using the real network traffic, and consider the classification context at the same time.
- **Self-adaptation of ML models**. This issue contains two parts: Be adaptive to the change of target protocols and be adaptive to different background traffic. Currently, the protocols and applications tend to update quickly, because of many factors, such as avoiding classification, security or service updating. To classify these protocols and applications, the current flow-based techniques need to retrain ML models as soon as the update is spotted. It is no doubt to increase workload. On the other hand, different from payload-based techniques, the performance of flow-based techniques is strongly related with background traffic, because the training set certainly contains background traffic. Theoretically, using different training set will train different ML models. The model trained by lab traffic may not suit for the classification in backbone network. One feasible solution is to retrain ML models as long as the network environment changes. By doing so, however, may create more work. Especially, the training complexity of some ML models, such as SVM, is very high. Therefore, improving the self-adaptive capability of ML models is the better way to both improve classification performance and reduce workloads.
- **Sophisticated applications**. As new coming applications become more and more sophisticated, the flow-based techniques are facing more and more challenges. First, many applications use both TCP and UDP as their transport layer protocols at the same time. Because most existing flow-based techniques focus on either TCP or UDP flows, they cannot deal with multiple transport protocols simultaneously. Second, many applications employ multiple flows to obtain services from several servers or peers. Existing flow-based techniques do not explicitly explore such information. Currently, there emerge amount of sophisticated applications, including instant message application such as Skype, QQ, and MSN, download management application such Thunder, eMule, and so on. How to classify their traffic completely and accurately has become to a big challenge for flow-based techniques.

The flow-based techniques have provide a promising solution for traffic classification. However, they are suffering from some issues and challenges in practice. As applications become more and more sophisticated, such challenges become more significant.

### D. Challenges of Host-based Techniques

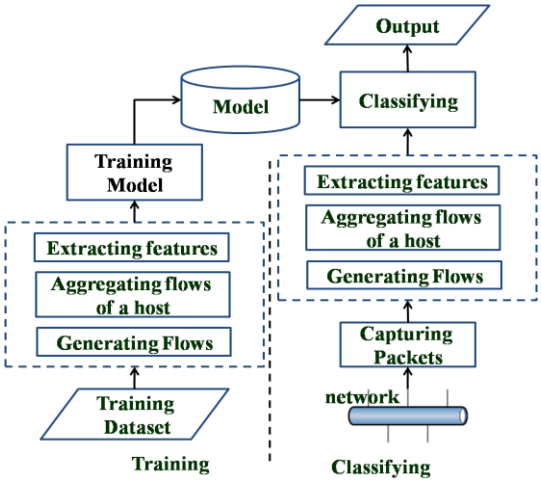



Figure 7. Process of host-based techniques

To classify the traffic generated by the sophisticated applications, researchers have proposed host-based techniques [5]. Fig. 7 illustrates the model of host-based techniques. In fact, these techniques also employ statistical features as their input, while the major difference from flow-based techniques is that the host-based techniques employ the features abstracted from the multi-flows. To obtain the features from multi-flows, the first step is to integrate flows. It is obvious that integrating flows based on a host is most feasible way. This is why we call these techniques as host-based.

BLINC is the most typical host-based technique [5]. This method shifts the focus from classifying individual flows to associating a host with applications, and then classifying their flows accordingly. BLINC follows a different philosophy from the traditional methods. It attempts to capture the inherent behavior of a host at three different levels: (a) social level, (b) network level and (c) the application level. The host-based techniques can classify all the traffic generated by an application, but they are suffering from the following issues.

- **Noise**. Most host-based techniques have a strong presupposition, that is, the monitored hosts only use one application at a moment. In reality, however, this situation may never happen. Most users tend to use many applications at the same time. In this case, the host-based techniques may classify other application traffic into the goal application. This presupposition indeed simplifies the research of host-based techniques, but it will be a big challenge to its deployment.
- **Not suitable for backbone network**. In order to obtain features from multi-flows, the host-based techniques must maintain the host information of either inside or outside of gateway, even both. As deployed in backbone network, there will be hundreds of millions of hosts that are needed to be monitored. In addition, each host in backbone network will generate plenty of traffic. The host-based classification systems will run out of computing and storage resource, and break down ultimately.
- **Identifying specific application sub-types**. The host-based techniques are able to identify the type of an application, but cannot classify their sub-types. For example, they can identify P2P flows, but it is difficult to determine which P2P application (e.g. eMule or BitTorrent) generates these flows. Classifying all traffic generated by an application may be the only first step, and how to identify their sub-types is a more important goal to achieve.
- **NAT**. As the host-based techniques classify traffic based on the knowledge of the traffic sent or received by a host, these techniques lost their effectiveness as long as monitored host hides behind the NAT.

In summary, the host-based techniques is a potential solution for classifying the sophisticated applications. But they are far way in practical use, and need more researching.

### E. Other Techniques

To analyze the behavior characteristics of network traffic, researchers proposed graph-based techniques. The classical approach is named as Traffic Activity Graphs (TAGs) [6]. This technique uses graph theory to model the communication behavior of hosts engaging in certain types of communications and their collective behavior, and then employ a novel statistical traffic graph decomposition technique to classify the traffic belongs to the sophisticated applications. The graph-based techniques are the new idea for dealing with the challenges of sophisticated applications. However, these techniques are not mature enough to be deployed in high speed network for online classification.

## III. General Challenges

In section II, we analyze the issues and challenges of the existing traffic classification techniques. In this section, we would like to analyze some general challenges from the perspective of traffic classification.

### A. Large-scale Traffic—Real-time Big Data

Currently, one of the hottest topic in academic and industry is Big Data, which is a collection of data sets so large and complex that becoming difficult to process using on-hand database management tools or traditional data processing applications[13]. The traffic of backbone network is a kind of Big Data, because the current traffic is hard to be captured, stored and analyzed. Moreover, due to the fact that most systems process traffic online, the large-scale traffic is accurately a kind of real time Big Data. For a typical campus network, its bandwidth has already reached Gigabits per second. Meanwhile, in the national backbone network, the bandwidth already reaches Terabits per second. For example, the Chinese international bandwidth has reached to 1800 Gigabits per second in 2012. Fig. 8 illustrates the forecast of global data center IP traffic growth [14], it is really a huge data.

Classifying traffic of Gigabits per second is still a fairly challenging, while handling Terabits per second is a tough task. In order to classify traffic under such high speed network, we cannot simply depend on improving the processing performance of the existing classification techniques. Fortunately, the maturing parallel hardware architectures and the distributed processing techniques bring new opportunities to further improve the performance of traffic classification systems.

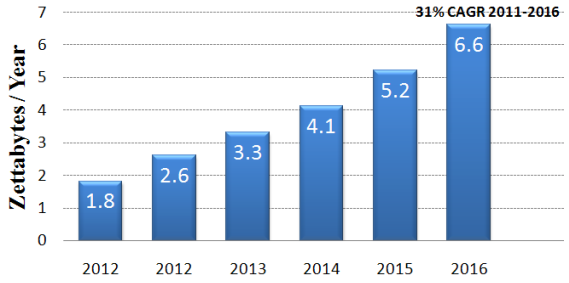



Figure 8. Forecast of global data center IP traffic growth[14]

### B. New Coming Protocols

As network technologies continue to be mature and evolve, more and more sophisticated protocols and applications are developed and deployed in the current network. In the early stage of Internet, some simple protocols, such as HTTP, SMTP and POP3 occupy more than 80% of all background traffic. As these traditional protocols are relatively simple, the traffic classification systems can easily identify and classify them. However, with the appearance of new protocols, especially P2P protocols, the traffic classification becomes more and more difficult. Firstly, almost all the new protocols encrypt their traffic for the purpose of privacy, which makes signature-finding more and more difficult; secondly, many new coming protocols employ distributed processing mode, which makes it difficult to classify all the protocol traffic completely and accurately. Although the host-based techniques promise to classify their traffic without inspect packet payloads, they are far away from practice. As the traffic of new coming protocols has exceeded traditional protocols, classifying their traffic has become a big challenge.

### C. Protocols VS. Applications

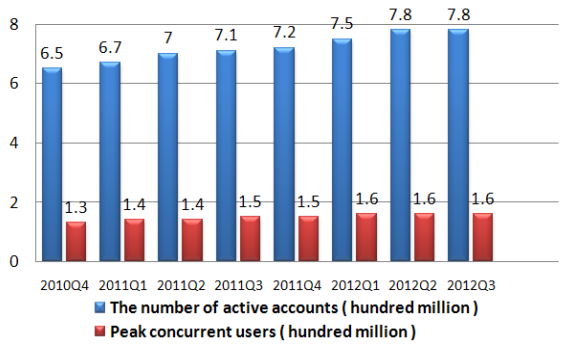



Figure 9. Number of users of QQ in China

Currently, Internet tries to provide various services conveniently for users. Meanwhile, the users want to get more services from a small quantity of applications. This requirement stimulates the development of sophisticated applications. Take QQ as an example, which is a popular Instant Messaging tool in China. According to the 31th statistical report on development of Internet in China published by CNNIC [15], the number of QQ registered users have reached to 7.8 hundred million by the end of 2012. The Fig. 9 illustrates the number of QQ users in recent years. Besides its proprietary, it also integrates VoIP, Video Chat, File transmission, advertisement based on HTTP, information push, Email, game and so on. Now most new coming applications belong to this kind of type. Classifying the traffic of these sophisticated applications is very difficult. Firstly, as the sophisticated applications are used to employing plenty of application layer protocols at a time, it is difficult to classify all their traffic from the background. Secondly, the traditional techniques tend to classify traffic according to the protocols but applications. For example, the existing techniques will classify advertisement of QQ to HTTP, but QQ. Sometimes, we do need to classify traffic according to the applications.

### D. Confusion and Imitation

Technically, the processing of traffic classification can be seen as information countermeasure. In order to avoid being filtered, identified and classified, many protocols and applications employ confusing and imitating techniques. The confusing techniques can fail the classification systems via hiding the signatures of packet payload. Meanwhile, the imitating techniques can lead to a high false alarm rate via imitating other protocols. Besides the confusion and imitation, there are other methods to avoid classification, such as random filling. As such techniques are widely used, the accuracy of traffic classification would be much reduced.

### E. UDP Flows

According to statistics, the UDP flows have occupied 40% of the background traffic on average. In peak, this proportion will raise to 80%. Different from TCP flows, which provide reliable, ordered delivery of a stream, UDP flows have no state and no direction. Classifying UDP flows is fairly challenging work. Firstly, UDP flows provide less information than TCP flows; Secondly, there are a lot of short UDP flows, which have less than 10 packets; Thirdly, the network attacks tend to use UDP as their transport layer protocol. Improving the current techniques to classify UDP flows is an urgent task.

## IV. Recommendations

Researches on traffic classification have produced a lot of novel approaches, but there are still a lot of issues and challenges yet to be solved. In this paper, we provide a comprehensive analysis of traffic classification

techniques, and point out some issues and challenges. As the current network bandwidth becomes larger and larger as well as the protocols and applications become more and more sophisticated, the traffic classification become to a more and more tough task. Here, we outline some recommendations to improve the performance of traffic classification systems as follows:

### A. Parallel Classification

As analyzed before, the current traffic classification systems usually have to handle Gigabits or Terabits per second. Only using one processing unit cannot deal with such huge network traffic. A promising solution for this challenge is to employ multi-core processors or distributed processing architecture. There already exist some approaches that can deploy traffic classification on multi-core processors or distributed processing architecture, such as OCTEON CN5860 and TILE Pro 64. However, as the traffic and classification rules remain increasingly growing, how to balance the traffic to every processing unit as well as reasonable arrange rule set should be carefully considered and designed.

### B. Hierarchical Classification

Besides parallel processing, the hierarchical processing is another way to improve the performance of traffic classification systems. Although the background traffic contains plenty of protocols, there are still a large proportion of protocols that can be easily identified and classified. Firstly designing a hierarchical traffic classification architecture to classify a big part of traffic which can easily be identified, then identifying the little part of the sophisticated applications. This can significantly increase the effectiveness and overall performance of classification system.

The combination of parallel and hierarchical processing can obtain a better result. Fig. 10 demonstrates a feasible solution.

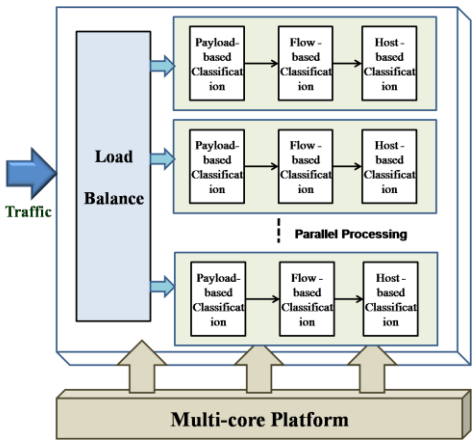



Figure 10. Parallel and hierarchical classification

### C. Finding Payload Signatures Automatically

Traditionally, finding payload signatures of a protocol basically relies on analyzing protocol RFC or reversing its application software. However, as more and more new protocols keep private and use software protection techniques, the traditional signatures finding approaches lose their effectiveness gradually. Based on the assumption that every protocol has its payload signatures, we can design an automatic signatures extracting and deriving approach.

### D. Traffic Analysis Approaches for Flow-based Techniques

The payload signatures are determined values, because they are found via analyzing protocol RFC or reversing protocol applications. But the statistical features are the empirical values which are extracted from protocol communication behaviors. Although theoretically different protocols have their own flow statistical features, we still cannot believe their effectiveness, because the current approaches cannot prove the uniqueness of flow statistical features. Therefore, we need a strong traffic analysis approach. On one hand, it should prove the uniqueness of flow statistical features; on the other hand, it can help to choose more suitable statistical features.

### E. Developing Suitable Cleverer Learning Algorithms

Since machine learning theory is created, it has been a hot research topic. So far, there are still hundreds more new algorithms are proposed every year. These algorithms are very carefully and creatively designed, and may help to improve the performance of traffic classification systems. But there is no a mature practical one which can be deployed on backbone network yet. So we should improve or develop cleverer machine learning algorithm to be self-adaptive to not only changes of deployment but also emerging of new sophisticated applications or their updates.

## V. Conclusion

Traffic classification is a traditional and important research area. In this paper, we first state the evolution of traffic classification along with the development of protocols. Then the issues and challenges of existing traffic classification techniques are examined in details. We also present some general challenges from the perspective of traffic classification, and finally outline some recommendations.

## Acknowledgment

This work was supported by the National Key Technology R&D Program under Grant No.2012BAH46B04, and National High-tech R&D Program Projects under Grant No. 2011AA010601.

## References

[1] IETF RFC 2475, "An architecture for differentiated services," Section 2.3.1 IETF Definition of Classifier.

[2] A. Sperotto, G. Schffrath, and R. Sadre, *et al.*, "An overview of IP flow-based intrusion detection," *IEEE Communications Surveys and Tutorials*, vol. 12, pp. 1-14, 2010.

[3] R. Smith, C. Estan, and S. Jha *et al.*, "Deflating the big bang: fast and scalable deep packet inspection with extended finite automata," in *Proc. ACM SIGCOMM 2008*, Seattle, USA, August 2008, pp. 207-218.

[4] T. Nguyen and G. Armitage. "A survey of techniques for Internet traffic classification using machine learning," *IEEE Communications Surveys and Tutorials*, vol. 10, no. 4, pp. 56-76, 2008.

[5] T. Karagiannis, K. Papagiannaki, and M. Faloutsos. "BLINC: multilevel traffic classification in the dark," in *Proc. SIGCIMM*, Kyoto, Japan, August 2007, pp. 229-240.

[6] Y. Jin, E. Sharafuddin, and Z. L. Zhang. "Unveiling core network wide communication patterns through application traffic activity graph decomposition," in *Proc. SIGMETRICS*, New York, NY, USA, 2009, pp. 49-60.

[7] OpenDPI Intetration Manual. (September 4, 2009). [Online]. Available: http://opendpi.googlecode.com/files/OpenDPI-Manual.pdf

[8] L7-filter. [Online]. Available: http://l7-filter.sourceforge.net/.

[9] A. Mcgregor, M. Hall, and P.0 Lorier *et al.*, "Flow clustering using machine learning techniques," in *Proc. 5th International Workshop PAM*, Antibes, France, April, 2004, pp. 205-214.

[10] T. Auld, A. W. Moore, and S. F. Gull. "Bayesian neutal networks for Internet traffic classification," *IEEE Transaction on Neural Networks*, no. 1, pp. 223-239, January 2007.

[11] B. H. Yang, G. S. Hou, and L. Y. Ruan, *et al*., "SMILER: Towards practical online traffic classification," in *Proc. ACM/IEEE Seventh Symposium on Architectures for Networking and Communications Systems*, Washington, DC, USA, 2011, pp. 178-188.

[12] A. W. Moore and D. Zuev. "Discriminators for use in flow-based classification," *Technique Report of Queen Mary University*, 2005.

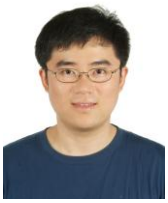
[13] T. White, *Hadoop: The Definitive Guide*, O'Reilly Press, 2012.

[14] Cisco Global Cloud Index: Forecast and Methodology. (2011-2016). [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ ns705/ns1175/Cloud_Index_White_Paper.pdf

[15] CNNIC: Statistics Report of Development of China Internet Network. (Jan 2013). [Online]. Available: http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201301/P020130122600399530412.pdf

**Yibo Xue**, is the corresponding author of this paper. He is IEEE/ACM member and CCF senior member. He received his B.S. degree and M.S. degree in Computer Science from Harbin Institute of Technology in 1989 and 1992, respectively, Ph.D degree in Institute of Computer Technology from Chinese Academy of Science in 1995. He is currently a professor in the Research Institute of Information Technology (RIIT) at Tsinghua University. His main research interests are in the areas of computer architecture and network security.

**LuoShi Zhang**, born in 1983, Ph. D. candidate. His research interests include network security, protocol identification and traffic management.



**Dawei Wang** received his B.S. degree in Computer Science from Nanjing University of Post and Telecommunication in 2005, and got his M.S and Ph.D. degree in Computer Science from Harbin University of Science and Technology in 2007 and 2010, respectively. Now he is working at National Computer Network Emergency Response Technical Team/Coordination Center of China. His research interests include intrusion detection, traffic management, statistical pattern recognition and artificial immune system.