Ambiguity and Concepts in Real Time Online Internet Traffic Classification

Hamza Awad Hamza Ibrahim[#1], Sulaiman Mohd Nor[#2], Ban Mohammed Khammas[#3]
[#] Faculty of Electrical Engineering, Universiti Teknologi Malaysia
UTM Skudai, 81310 Johor, Malaysia.
[1] hamysra76@hotmail.com
[2] sulaiman@fke.utm.my
[3]Banm1978@yahoo.com

Abstract— Internet traffic classification gained significant attention in the last few years. Identifying the Internet applications in the real time is one of the most significant challenges in network traffic classification. Most of the proposed classification methods are limited to offline classification and cannot support online classification. This paper aims to highlight the ambiguity in the definition of online classification. Therefore, some of the previous online classification works are discussed and analyzed. This analysing is to check how far the real time online classification was achieved. The results indicate that most of the previous works consider a real Internet traffic but did not consider a real time online classification. In addition, the paper provides a real time classifier which was proposed and used in [1] [2] [3], to show how to perform a real time online classification.

Keyword- Online Internet Traffic classification, Machine Learning, network traffic classification

I. INTRODUCTION

Internet service providers (ISPs) and network operators are mostly interested in knowing the amount of traffic carried by their networks for the purposes of optimizing network performance and security issues. Therefore, Internet traffic classification is something valuable, particularly for interactive traffic applications such as VoIP and online games.

Simple classification assumes that most applications use well-known port numbers, and the classifier uses this port number to identify the application type. However, most Internet applications use unknown port numbers, or more than one application uses the same port number, which indicates the failure of port base classification [1]. Another classification method is payload based (deep packet inspection), which is individual packet inspection, looking for unique signatures. However, using this technique faces two problems; first, it is difficult to detect non-standard ports by using packet inspection because these packets are encrypted. Second, deep packet inspection touches on users' privacy. In order to solve the problem of past classification methods (base port and payload inspection), machine learning (ML) technique was developed. ML works [2] [3] uses artificial intelligence to classify IP traffic, which provides a powerful solution by extracting the right information from application features [4]. Moreover, some of the ML algorithms are suitable for Internet traffic flow classification at a high speed [5]. Most of the proposed classification methods are limited to offline traffic classification and cannot support online classification [6]. Offline classification is not helpful for online management and control mainly due to the performance reason (Chen et al., 2009).

This paper aims to achieve several gals at the same time. First, addresses the importance as well as the definition of online classification. Second, differentiates between "*real Internet traffic*" and "*online Internet traffic*". Third, the paper also aims to highlight the ambiguity in the definition of online classification that based on previous online classification works.

Online network traffic classification is very valuable because of several reasons such as:

- Online classification is the basis to manage the real network traffic. Thus, in order to manage and control the Internet traffic, there is a real need for online classification.
- The online classification helps to prevent network threats and abnormal behaviors such as denial of service, flooding attack and other such threats.
- Developing of effective online classification algorithms helps to reduce the using of hardware classifier (such as Packet-shaper) which has very high cost.

There is a big difference between "*real Internet traffic*" and "*online Internet traffic*" terms; real traffic can be defined as any real Internet traffic which is captured in any network level, and at the current time this is not live traffic. While online traffic mean the traffic which is currently running in the network (live traffic). By the same manner, there are a big difference between online classification and real traffic classification. Real classification is the identification of the real network traffic which can be called offline classification. This paper defines the online classification as a system which can receive and classify the Internet traffic at the traffic running time. In

other words, online classification means, the decision of which packet belongs to which flow, assuming to be on the traffic speed. Such, like any hardware classifier (PacketShaper, SANGFOR) which installed on the network path to classify with the passage of the traffic.

The main problem that meets the online classification decision is the high speed of Internet traffic. It is difficult to take an online classification decision with huge amounts of Internet traffic. The question is how to divide the Internet traffic into flows, calculate flow patterns, and make classification decisions online with high Internet traffic speed. Most previous literature [6] [7] [8] [9] [10] [11] [12] [13] worked with classifiers using real time traffic, however, only few of them [14] provided a classifier which can make an online decision.

II. RELATED WORKS

There are many published articles which include the term "online classification" in the articles title. Unfortunately, most of them have no actual online classification but only real traffic classification. The following paragraphs discuss some of related works that include the words "online classification" in their titles. The study [8] proposed a dynamic method to classify Internet traffic. The method used two levels: overall traffic level and application level. Data mining algorithms are used to continue updated considered datasets. The proposed method has three parts: i) Traffic model; which is: preparing the dataset, selecting the features, and updating the model for the case of new application. ii) Traffic classification; to classify traffic based on the gained features. iii) Change detection; which is run periodically to check if there is a new application. While the paper title includes the words "online traffic classification", but there is no actual online classification. This due to two reasons: first, classifier did not use a classification mechanism which can receive and classify the traffic in the online time. Second, the authors used nine datasets traces which are collected from different places in campus network. These traces were divided between training and testing datasets which indicate that there is no online traffic source.

The authors in [6] claimed that they proposed an approach for online traffic identification. The classification method is based on the observation of the first few packets of a flow. The inter-arrival times and packet size of the individual packets are selected as identification features. The classifier uses Naïve Bayes as ML algorithm which trained in a supervised learning setting. The classifier system includes two phases: offline training and online classification (as the author mentioned). However, the classifier did not consider actual online classification, only real dataset was classified. This because of two reasons: first, when focus on the system structure there is no connection to real traffic source such as a switch or access point or even user NIC. Second, the authors mentioned in their discussion that they have one dataset which divided into two parts; the first part used as training datasets and the second as testing datasets which is not a real online environment.

In order to achieve the requirements of the network activities, a traffic classifier based on Support Vector Machines (SVM) was presented in [10]. The dataset included three traces collected from three different places: 1) Moore_Set which collected from Cambridge University; 2) Handmade_Set which was labelled in the author's laboratory (China); 3) University_Set which collected from Nanjing University of Posts and Telecommunications. Based on statistical features, the classifier used the first ten packets to identify the flow. However, three issues were observed in this work. First, the authors did not consider the capturing of training and testing datasets from the same place, which is an important ML datasets issue. Second, how to classify flow includes a large number of packets based on only ten packets. Third, from online classification point of view; how to name this classifier as online classifier when it used to classify datasets which collected before three years of classification time?

Sun and Chen [13] proposed a method which is suitable to identify the application association with TCP flows. This is based on total data length sent by client (ACK-Len ab) or server (ACK-Len ba) before it received ACK packets. The proposed method was verified by using an ML classifier (C4.5) to classify four types of Internet applications (WWW, FTP, EMAIL and P2P). These applications were collected from three different places as follows: first trace was provided in London UK at 2006, the second trace was used by [14], which was collected from university in France. The last trace was gained from the author work environment (China). In the same manner as other researches, the traces were collected from different network environments as well different times. Because the classifier was trained by datasets collected from three different networks, it's difficult to use this classifier online to identify the traffic of any one of the same networks.

The researchers of [11] proposed a wireless mesh network traffic classification using C4.5. Sub-flow with application behaviour's was applied to solve the problem of how to select represented sub-flow. Based on the statistical features of the first n packets, the classifier clusters the flow to one of the defined applications. The proposed method was used to identify some Internet applications such as http, SMTP, FTP, Kazza etc. Similar to the previous study, the ML datasets were collected from two different networks (campus and residential) which may have different characteristics. As an example, how can it be ensured that the inter-arrival time of http traffic is the same in both networks? In addition, the classifier used to identify datasets which collected before several years of classification time which indicate that there is no actual online classification.

In order to maintain the accuracy of ML classifier the researchers of [15] proposed retraining mechanism. The accuracy of ML classifier is checked from time to time based on the some flows which are labelled by the heuristic training dataset. The mechanism was divided into three stages. In the first stage, flows are extracted from incoming packets by the offline training dataset generator. In the second stage, the accuracy of ML classifier is evaluated against the accuracy generated by the training dataset generator. In the last stage, the online ML classifier is updated in case the current ML accuracy is below a predefined threshold. The used technique is very helpful and it proves ability to enhance ML classifier accuracy. However, from the system structure and paper discussion, there is no online classification only real data are used. This because of there is no discussion about online classification or online algorithm which normally used for online classification.

[9] is one of the few papers that proposed actual online classifiers. The proposed online approach is used to classify TCP traffic that based on the first n packets. This approach used information from the first n packets and Bayesian network method to determine the flow belongs to which kind of application. The authors used Correlation-based Feature Selection (CFS) proposed in [16] to select optimum features. The classifier uses TCP dump for data collection on the monitoring computer. The system records the headers of TCP/IP packets with SYN, FIN, or RST flags set. The main advantage of this study among others, that the classifier was connected directly to the main campus router. However, how to classify a flow that include thousands packets based on the first few packets? This question arise here because of, the first packets in many flows can differ from the rest of the packets statistically. In addition, the paper did not provide details of how the online decision was taken.

As defined before the real online classification supposed to be in the same time of the Internet traffic generating. This means, the classifier can achieve online classification only when it is able to identify packet/flow in the time when it passing through network device (NIC, switch, access point). Table 1 analyses some of previous works which are titled by some wards that includes the wards “online classification”.

TABLE 1

Some related works with the wards "*online classification*" in the articles title

Works title	Does the work achieve actual online classification defined by this paper?	The evidence (collected from the paper)
“A Dynamic Online Traffic Classification Methodology based on Data Stream Mining” [8]	no	Equally 200 flows which collected from different traces are mixed and used for online classification
“Identifying online traffic based on property of TCP flow” [9]	no	Different training sets 10%, 30%, 66%, and 100% of trace 1 are used to classify trace 2 as online classification.
“Online Wireless Mesh Network Traffic Classification using Machine Learning” [11]	no	Two traces collected in different times from campus and residential are used for online classification
“Online Internet Traffic Classification Based on Proximal SVM” [10]	no	Three traces collected in 2003 and 2009 are used as online traffic.
“Research of the traffic characteristics for the real time online traffic classification” [13]	no	The online classification used three datasets collected from different networks
“Online Internet Traffic Identification Algorithm Based on Multistage Classifier” [17]	no	One datasets divided into 60% for training and 40% for testing
“Retraining Mechanism for On-Line Peer-to-Peer Traffic Classification” [15]	no	Two datasets was generated one used for training the other used for testing

III. SSPC ONLINE CLASSIFIER

This section provides an example of real online classifier which is able to classify the Internet traffic in real time and before the flow end. The online hybrid Signature Statistical Port Classifier (SSPC) is proposed and used in [18] [19] [20]. SSPC has the ability to achieve an online classification because of the following reasons: first, the existing of online capturing system which is built on the SSPC algorithm to be used at any point of the network such as NIC or any traffic mirror at any network level. The captured algorithm is developed only to capture the needed features and ignore the rest. This can accelerate the online processing and assist to decrease the classification time. Second, SSPC code is easy which can be used by any simple machine (computer), and no

need for high specifications machine. Third, the online computation was decreased as much possible as. This factor is achieved by SSPC system when it considers the statistical rules generated by Rule.PART algorithm (within Weka). This algorithm is more effective than the others because it generates a short number of rules and a high accuracy at the same time. Fourth, SSPC easily can switch between considering full flow or a part of the flow. SSPC achieves an online flow classification before (or shortly after) the end of the considering flow. This flow end is under control either by using of a time function to stop the flow capturing or by the real end of the flow. In addition, after the online classification, SSPC can save the captured traffic for more offline analysis. For more details about SSPC classification system please see the articles mentioned above.

Figure 1 shows the network structure of SSPC online classification environment. This environment includes some monitored user IP addresses. The used Internet applications are run manually through these users (one application per user). Some users were monitored out of more than 700 users located in-campus residential area. Traffic mirror are used to capture the traffic which reflect all traffic passes through the switch. SSPC system only captures and classifies the traffic of the monitored users. Easily and based on the IP addresses, the classification results can be shown. These results include some metrics such as FP, FN, and accuracy. The online classification is done at the same time of the traffic generating and there is no any offline processing.

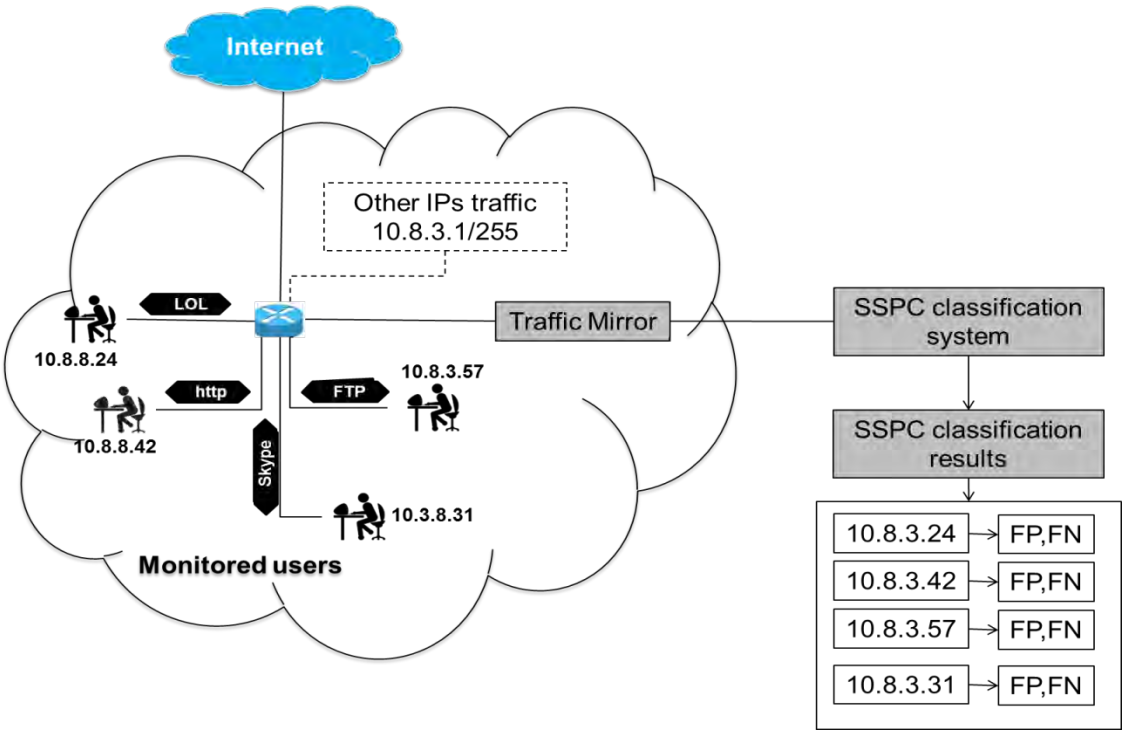

Fig 1 SSPC online classification environment

IV. Experiments and Discussion

This section provides experimental work which performed by SSPC system to identify Internet applications. Real time Internet traffic was collected from UTM campus network in order to perform the classification in two stages offline and online. All the side factors in offline and online classification are the same. This means traffic features, algorithm rules, and application classes are the same in offline and online classification. The only difference is the used datasets. The classification includes four types of Internet application classes (WWW, FTP, Skype, online game (LOL)).Table 2 shows the number of flows of the four classes which are used for the offline classification (decisions). These flows are generated manually through the monitored clients (IPs). Using monitored IPs ensures that the training and testing datasets were collected from the same network without the need for standard (labelled) datasets.

TABLE 2

Used flows for offline decisions

Class	Applications	Number of flows
WWW	http, https	1857
FTP	FTP-data, FTP-control	304
Skype	Skype	2044
Online games (LOL)	LOL	52

In this experiment, the statistical classifier in SSPC uses only interarrival time and packet length as traffic features to identify the Internet applications. For each of these two features, some statistical factors are calculated as shown in table 3. In addition, the signature classifier checks the signatures in the following fields: DNS query and response, http host, http referrer, and http user.

TABLE 3

Statistical features

Max of inter-arrival time
Min of inter-arrival time
Mean of inter-arrival time
Variance of inter-arrival time
Standard deviation of inter-arrival time
Max of packet length
Min of packet length
Mean of packet length
Variance of packet length
Standard deviation of packet length

Before going into online decision experiments, offline works were performed to evaluate the methodology. In parallel, each of the partial classifier i.e. port, signature and statistical was used over each class dataset (Table 2) and the result of each case was recorded. Second, in the same manner, SSPC algorithm was used over each class dataset separately. Figure 2 and Table 4 show the individual classifiers accuracy and SSPC accuracy. SSPC shows a higher accuracy compared to the other partial classifiers with WWW and Skype classes. For FTP and LOL, SSPC equals to statistical classifier accuracy which proves that SSPC cannot be less than the statistical classifier

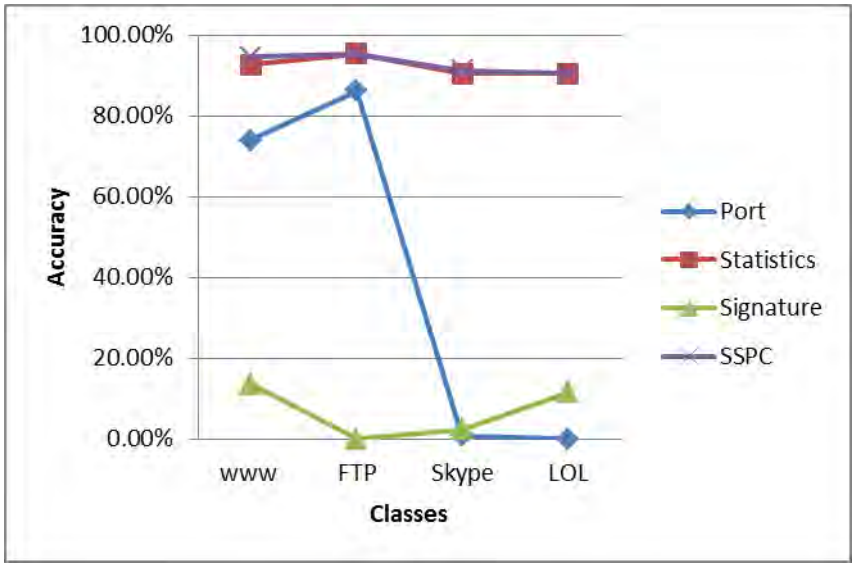

Fig 2: Offline classifiers accuracy

TABLE 4

Offline classification accuracy

	Signature	**Port**	**Statistics**	**SSPC**
WWW	13.62%	73.94%	92.62%	94.56%
FTP	0.00%	86.18%	95.39%	95.39%
Skype	2.35%	0.78%	90.36%	91.24%
LOL	11.54%	0.00%	90.38%	90.38%

In the online decision, the same offline applications were used through running of two different experiments. All the online stages discussed in section were applied. Similar as in the offline experiments, the applications were run through the monitored clients. The testing dataset generated were totally different from the training dataset. As an example in some clients, we ran only WWW applications and then checked in parallel (at the same time) what the decision of each classifier was and what is SSPC decision. It is important to note that this experiment did not consider any FTP signatures or LOL port number which justify why the signature and port classifiers score 0% with FTP and LOL respectively.

TABLE 5

Number of flows for online decisions

	Experiment 1	Experiment 2
www	548	823
Skype	217	44
LOL	695	145
FTP	668	526

Table 5 shows the number of flows generated by each of the online experiments. These online flows traffic is generated manually through the volunteer users. Table 6 and figure 3 illustrate the results of online decisions. The WWW classification accuracies are 90.15% and 88.34%, which are higher when compared with the other three classifiers. Skype signatures and port numbers were found only when considering the traffic of user's login (experiment 1). This results show that SSPC is the higher accuracy between the other classifiers. However, when the login traffic (experiment 2) does not consider, no signatures or well port numbers was found, which makes SSPC accuracy equal to statistics classifier. In the case of LOL online game, SSPC accuracy is higher than others, but it still low (68.78 & 87.59%). This is because of two reasons: first, the low amount of datasets which considered in the offline training stage; second, no LOL signatures are used in this stage. In FTP data, SSPC accuracy is acceptable but it less than port classifier. This because of two reasons, first: there are no FTP signature was added to support SSPC decision. Second, the used FTP traffic was generated by some monitored clients, which used real FTP port numbers (20 & 21). The port accuracy was expected to be low in case some clients used static IPs or VPN network. The last column in TABLE shows the average of the classification time (in seconds) for each flow. As an example, classification of single WWW flow in experiment 1 by SSPC was taken 0.01 seconds after end of flow capturing.

TABLE 6

Online classification accuracy

	Port	Statistics	Signature	SSPC	Flow/s
Experiment 1					
WWW	30.47%	68.07%	7.12%	90.15%	0.01
Skype	6.45%	79.72%	5.99%	82.95%	0.06
LOL	0.00%	68.49%	20.86%	68.78%	0.04
FTP	98.80%	96.86%	0.00%	96.86%	0.14
Experiment 2					
WWW	46.66%	56.01%	12.03%	88.34%	0.01
Skype	0%	95.45%	0%	95.45%	0.07
LOL	0.00%	81.38%	15.86%	87.59%	0.02
FTP	96.77%	94.49%	0.00%	94.49%	0.17

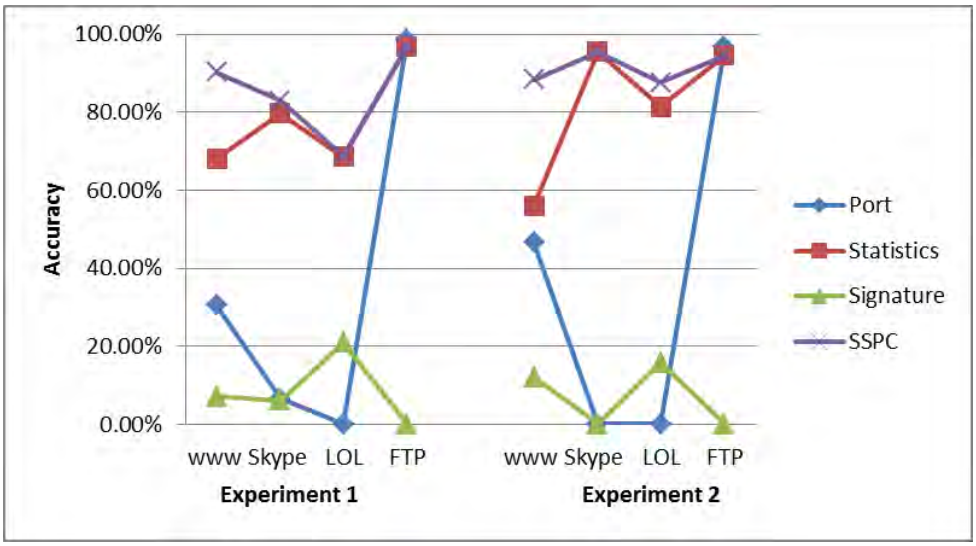

Fig 3: Online classification accuracy.

V. Conclusion

Real time online Internet traffic classification is one of the most important challenges in network traffic classification. This paper highlighted the ambiguity which covers the real time online classification. The paper studied some of related works which proposed online classification methods. When all of these works have a title which includes the words “online classification”, however, most of them did not consider a real time online classification. Therefore, the term “online classification” in most previous cases did not used for a real time online classification. In addition, signature statistical port classifier (SSPC) is considered as an online classifier which applied a real time online classification. Real time datasets (more than 7900 flows) were captured in the campus environment, which includes WWW (http, https) and non-WWW (FTP data, FTP control, online gaming, and Skype) traffic. The SSPC was tested in two stages, offline and online. The results of both stages show that the SSPC has higher accuracy among the three partial classifiers. In the online classification, the classification decision was made during the traffic generating and before the flow end. Thus, SSPC is achieved a real time online traffic classification since it identified the Internet traffic in real time and without any compromise in delay.

REFERENCES

[1] Nguyen, T.T.T., Armitage, G.: A Survey of Techniques for Internet Traffic Classification using Machine Learning. Ieee Commun Surv Tut 10(4), 56-76 (2008). doi:Doi 10.1109/Surv.2008.080406

[2] Jesudasan, R.N., Branch, P., But, J.: Generic Attributes for Skype Identification Using Machine Learning. Technical Report 100820A (2010).

[3] Alshammari, R., Zincir-Heywood, A.N.: An investigation on the identification of VoIP traffic: Case study on Gtalk and Skype. In: Network and Service Management (CNSM), 2010 International Conference on, 25-29 Oct. 2010 2010, pp. 310-313

[4] Yu, J., Lee, H., Im, Y., Kim, M.S., Park, D.: Real-time Classification of Internet Application Traffic using a Hierarchical Multi-class SVM. Ksii T Internet Inf 4(5), 859-876 (2010). doi:DOI 10.3837/tiis.2010.10.009

[5] Soysal, M., Schmidt, E.G.: Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison. Perform Evaluation 67(6), 451-467 (2010). doi:DOI 10.1016/j.peva.2010.01.001

[6] Gu, R., Wang, H., Ji, Y.: Early traffic identification using Bayesian networks. In: 2010, pp. 564-568

[7] Xu, C., Tang, H., Zhao, G.F.: Traflow: Design and complementation of a real time Traffic Measurement System in High-Speed Networks. 2008 Ifip International Conference on Network and Parallel Computing, Proceedings, 341-344 (2008). doi:Doi 10.1109/Npc.2008.10

[8] Xu, T., Qiong, S., Xiaohong, H., Yan, M.: A Dynamic Online Traffic Classification Methodology Based on Data Stream Mining. In: Computer Science and Information Engineering, 2009 WRI World Congress on, March 31 2009-April 2 2009 2009, pp. 298-302

[9] Hong, M.-h., Gu, R.-t., Wang, H.-x., Sun, Y.-m., Ji, Y.-f.: Identifying online traffic based on property of TCP flow. The Journal of China Universities of Posts and Telecommunications 16(3), 84-88 (2009). doi:http://dx.doi.org/10.1016/S1005-8885(08)60231-9

[10] Gu, C., Zhang, S., Huang, H.: Online internet traffic classification based on proximal SVM. Journal of Computational Information Systems 7(6), 2078-2086 (2011).

[11] Gu, C., Zhang, S., Xue, X., Huang, H.: Online wireless mesh network traffic classification using machine learning. Journal of Computational Information Systems 7(5), 1524-1532 (2011).

[12] Nguyen, T.T.T., Armitage, G., Branch, P., Zander, S.: Timely and Continuous Machine-Learning-Based Classification for Interactive IP Traffic. Networking, IEEE/ACM Transactions on PP(99), 1-1 (2012). doi:10.1109/tnet.2012.2187305

[13] Sun, M.F., Chen, J.T.: Research of the traffic characteristics for the real time online traffic classification. Journal of China Universities of Posts and Telecommunications 18(3), 92-98 (2011).

[14] Bernaille, L., Teixeira, R., Akodkenou, I., Soule, A., Salamatian, K.: Traffic classification on the fly. Comput Commun Rev 36(2), 23-26 (2006).

[15] Zarei, R., Monemi, A., Marsono, M.N.: Retraining Mechanism for On-Line Peer-to-Peer Traffic Classification. Adv Intell Syst 182, 373-382 (2013).

[16] Nguyen, T.T.T., Armitage, G.: Training on multiple sub-flows to optimise the use of Machine Learning classifiers in real-world IP networks. Conf Local Comput Ne, 369-376 (2006).

[17] Min, D., Xingshu, C., Jun, T.: Online Internet traffic identification algorithm based on multistage classifier. Communications, China 10(2), 89-97 (2013). doi:10.1109/cc.2013.6472861

[18] Hamza Ibrahim, H.A., Mohd Nor, S., Mohamed Abdelaziz, I.I., Alfaki Abdalla, A.A.: SSPC algorithm based on three different methods for online Skype traffic classification. Journal of Theoretical and Applied Information Technology 53(3), 422-429 (2013).

[19] Ibrahim, H.A.H., Mohd Nor, S., Ahmed, A.: Internet Traffic Classification Algorithm Based on Hybrid Classifiers to Identify Online Games Traffic. Jurnal Teknologi 64(3) (2013).

[20] Ibrahim, H.A.H., Nor, S.M., Ismail, I., Abdalla, A., Abdalla, A.A.A., Jamil, H.A.: Enhance the accuracy of Machine Learning Internet Traffic Classifier by Applying Datasets Validation Issues and Using a Hybrid Classifier. JCIT 8(15), 44-62 (2013).