Peer-to-Peer Netw. Appl. (2017) 10:1182–1203
DOI 10.1007/s12083-016-0471-2

Identifying P2P traffic: A survey

Max Bhatia[1] · **Mritunjay Kumar Rai**[2]

Received: 5 December 2015 / Accepted: 7 June 2016 / Published online: 18 June 2016

Abstract Peer-to-Peer (P2P) traffic is widely used for the purpose of streaming media, file-sharing, instant messaging, games, software etc., which often involves copyrighted data. From the past decade, P2P traffic has been contributing to major portion of Internet traffic which is still rising and hence is consuming a lot of network traffic bandwidth. It also worsens congestion of network traffic significantly and degrades the performance of traditional client–server applications. Popularity of various P2P applications has led Internet Service Providers (ISPs) to face various challenges regarding efficiently and fairly utilizing network resources. The traditional methods of identifying P2P traffic such as port-based and payload-based are proving ineffective due to their significant limitations and can be bypassed. Hence, new approaches based on statistics or behaviour of network traffic needs to be developed and adopted in order to accurately identify existing and new P2P traffic which emerge over the time. This article presents a survey regarding various strategies involved in identifying P2P traffic. Furthermore, conceptual analysis of network traffic measurement and monitoring is also presented.

Keywords Peer-to-Peer · P2P traffic identification · Traffic measurement · Port based classification · Payload based classification · Statistical based classification · Machine learning

✉ Mritunjay Kumar Rai
raimritunjay@gmail.com

Max Bhatia
maxbhatia.cse@gmail.com

[1] Department of Computer Science Engineering, Lovely Professional University, 506/135, New Sukhdev Nagar, Opp. Shiraz Hotel, Hoshiarpur, Punjab 146001, India

[2] Department of Electronics & Communication Engineering, Lovely Professional University, Jalandhar-Delhi, G.T. Road (NH-1), Phagwara, Punjab 144411, India

1 Introduction

In the past, Internet traffic relied on client–server paradigm where client used to request the data and the server provided it leading to network traffic which was asymmetric. With the evolution of Internet and the so called Web 2.0, Internet hosts got the privilege to provide their own multimedia content which could be shared with other peers on Internet. Further, Peer-to-Peer (P2P) traffic started evolving towards the end of the 20th century which incorporated direct distribution of contents between peers on Internet. In such a scenario, peers started acting both as client and server simultaneously; thus downloading the contents which they required from other peers and distributing their contents to other peers on Internet. Due to this, network traffic has become symmetric. From the network management point of view, P2P traffic needs to be identified as it involves traffic flowing in both directions at the same time, thus consuming more bandwidth. In this system, peers share the distribution cost of the service instead of relying on a dedicated server for it. This is actually advantageous for the service providers for distributing the contents, but only at the cost of producing more traffic in the network. In order to search contents with the remote peers, there is increase in number of communications between the peers which has resulted in large number of connections as compared to client–server system where only few connections were formed. Thus, P2P systems produce large amount of traffic as opposed to client–server systems. This poses an issue where network traffic needs to be monitored and controlled so that P2P traffic alone doesn't consume large portion of the available bandwidth. Hence, a

balance needs to be maintained so that other kinds of traffic such as HTTP, FTP, SMTP, etc. also get their fair share of bandwidth. It ensures that Internet Service Provider (ISP) is able to provide Quality of Service for each application by implementing specific policies. Further, conventional devices are unable to control P2P traffic effectively due to which ISPs are facing several other challenges like paying for added traffic requirement, satisfying customers with excellent broadband experience, purchasing costly backbone links and upstream bandwidth.

Internet traffic has been growing rapidly over the past few years [1]. This is attributed to the fact that P2P traffic has grown at such a pace that various types of applications have been emerging over time. Various application protocols such as HTTP, SMTP, etc. no longer dominate Internet traffic which has instead been taken over by P2P traffic to a large extent [2]. P2P file sharing has been a significant trend in recent years. The major content which is shared or distributed through P2P applications are audio, video and games which tend to be large in size [3]. This also includes illegal file sharing. P2P applications nowadays account for more than 60 % of total network traffic [2, 4–6] which consumes major portion of network bandwidth. Azzouna and Guillemin [7] in their study identified that 49 % of traffic was due to P2P applications in Asymmetric Digital Subscriber Line (ADSL) link. A worldwide study conducted by ipoque [8] (in 2007) about Internet traffic showed that P2P file-sharing applications produce more traffic as compared to all other applications taken together. Therefore, identifying the application that produces traffic becomes crucial in order to accomplish the tasks such as implementing billing mechanisms, maintaining Quality of Service for applications, implementing security measures, etc. Now it is a very difficult task as there are umpteen issues associated with it.

Traditional method used to accomplish the task of network traffic classification includes associating port-numbers of transport-layer to the well-known application protocols. But this technique of identifying applications soon became ineffective as various applications started using random port numbers for data transfer. Also, some other applications used masquerading techniques by utilizing well-known port numbers (such as port- number 80 utilized by HTTP) hide their traffic. Karagiannis et al. [9] identified that many P2P applications utilize port number 80 to transfer their data and also found that 30 to 70 % of the traffic generated by P2P applications utilized random port-numbers. Madhukar and Williamson [10] in their study showed that Internet traffic could not be identified correctly by using port-based methods. Due to these issues, another technique based on payload inspection was adopted. Although this technique proved to be of great accuracy, but it also possessed various limitations such as the requirement of large amount of computational resources, privacy issues involved and the inability of this technique to work when payload is encrypted. Hence, another alternative to identify traffic was adopted based on statistical or behavioural methods such as packet length, number of packets sent, number of packets received, etc. which do not possess limitations posed by port-based or payload-based techniques.

The main goal of this survey is to provide comprehensive overview of various traditional techniques as well as the existing ones for classifying P2P traffic. Although there is some research work done regarding survey on internet traffic classification [11, 12], yet this survey explicitly focuses on identifying P2P traffic which is one of the major contributors of internet traffic. It explains about the working of various techniques along with advantages and limitations of each. The remainder of this survey is sectioned as follows. Section 2 describes some related work in traffic classification. In order to have better understanding of traffic identification, Section 3 addresses some important concepts and techniques from the viewpoint of traffic monitoring. Verification about ground truth of traffic is mentioned in Section 4. Section 5 covers various metrics that can be used to evaluate the performance of various techniques. Section 6 covers various P2P classification techniques with published literature which is followed by Conclusion section.

2 Related work

The topic of network traffic identification and hence classification has gained more interest recently in scientific contributions due to various factors associated with it, such as providing network security, quality of service for applications, billing information, among others. As new applications and protocols keep on emerging over time, various studies propose novel techniques to address the challenges posed by them in their identification process.

For identification of P2P traffic, Madhukar and Williamson [10] compared three distinct techniques in terms of efficiency, namely: port-based, payload-based and transport-layer heuristics. In order to provide longitudinal performance study of each technique, they used the sample-data of traffic traces collected over duration of 2-years to evaluate each method. Li et al. [13] compared four different methods of classification in terms of effectiveness and efficiency, namely: port-based, payload-based, C4.5 decision tree and Naive Bayes. The authors collected the traffic traces over duration of several years at two different locations for evaluating the performance on the basis of spatial and temporal perspectives. Nguyen and Armitage [11] provided a survey on traffic classification based on Machine Learning techniques that focused on application-level protocols for identification. The authors also described the issues posed by recent Internet applications in classification process and reasons for developing newer techniques for classification of Internet traffic by highlighting the limitations

of older classification techniques. Callado et al. [12] gave introduction about traffic analysis and described the state-of-art of flow-based traffic analysis using several flow properties of Internet. They also provided the explanation about various research works conducted using distinct traffic classification techniques and theoretically compared the results obtained by them.

This survey focuses mainly on P2P traffic classification and various challenges associated in identifying it. Firstly introduction about traffic measurement from the view-point of traffic classification is given in order to provide better understanding of this topic. Furthermore, various approaches have been compared, analysed and overview regarding various techniques, studies and approaches have been presented for identification of P2P traffic.

3 Network traffic measurement

From the past few decades, various authors have highlighted the role of Internet/network traffic measurement which is crucial to understand the behaviour of computer networks [14–16]. It is not an easy task as it involves many issues and challenges. Paxson in [16] mentioned some of them while performing this task. He also mentioned some approaches for conducting sound Internet measurements. McGregor in [15] also describes several technical challenges in order to conduct quality measurements. The next subsection discusses some important concepts and techniques which should be considered while conducting traffic measurement.

3.1 Measurement of internet traffic

Williamson in [14] categorised the research tools for the purpose of network study as: Online & Offline, LAN & WAN, Hardware & Software, Protocol level, and Active & Passive. The significance of each category depends upon the research purpose. Their brief description for the purpose of traffic classification is given below:

Online and offline Online approach involves analysing traffic while it is currently flowing through the network. Such process requires high computational power and resources in high speed networks but is greatly useful in applications such as in NIDSs and firewalls when instant decisions or actions are required to be made for the packets currently flowing in the network. Whereas Offline approach involves network traces to be collected as an offline file for conducting analysis at a later time when the packets have already crossed the network. This approach is mostly preferred when real-time analysis is not required and it is also useful for the purpose of research and validation, as one can run several approaches on same set of traces which can be compared for results.

LAN and WAN Measurements conducted for traffic classification purpose is preferably done on LAN instead of WAN, since the former involves no loss of information whereas latter one is difficult to get access to.

Hardware and software Dedicated hardware tends to give better solutions in terms of performance which are useful in real-time analysis. For the purpose of traffic measurement, monitoring or capturing, some companies like Endace [17], ipoque [18], Wildpackets [19] and Napatech [20] provide hardware-based solutions. As researchers involved in traffic classification are mostly interested in analyzing IP packets or Ethernet frames in network, hence it is of less significance whether analysis is done using hardware-based or software-based solution.

Protocol level Traffic measurement can be performed at different protocol levels or even multiple protocol levels; but for the purpose of traffic classification, mostly Internet traffic is measured at IP level or Ethernet level by the researchers.

Active and passive Active approach involves injecting actual packets into the network to analyse the behaviour of the traffic. It allows one to control the simulation scenario such as type of traffic flowing in network, its frequency, etc. But its limitation is that it puts extra load on the network bandwidth and can affect the performance of routers or switches. Also, this approach does not truly reflect the actual behaviour of the traffic flowing in the network which may affect the results. On the other hand, Passive approach doesn't need to inject any packets into the network and captures and analyses the actual traffic flowing through the network. Hence, it doesn't affect the performance of bandwidth or any network equipment and measurements made using this approach reflects the actual behaviour or properties of real traffic. But its limitation is that, it produces large amount of data which needs to be processed and analysed in order to obtain useful information.

3.2 Measurements on basis of Per-Flow and Per-Packet

For traffic identification or classification purpose, the researchers mostly focus on IP packets or Ethernet frames. In Per-packet approach, each individual packet travelling in the network is captured for the purpose of analysing the traffic. It can be useful in certain scenarios such as Network Intrusion Detection Systems (using tools like, Snort [21], Bro [22]) where some decisions need to be made on each packet travelling through the network. Also, these packets can be captured and stored for offline analysis by using tools such as Wireshark [23] and Ettercap [24] which have the capability to inspect each individual packet and mine the useful information from all layers of protocol stack.

Although, packets flowing through the network are individual data units, but there exist certain relationships between them such as packets generated by same request or response, packets belonging to same application containing data, etc. and hence such hidden information can be mined for by using Per-Flow analysis. A flow is mostly defined as set of packets sharing common characteristics: Source-IP, Destination-IP, Source-Port, Destination-Port and Protocol [25–27]. It is considered as active-flow when time-interval between packets belonging to a particular flow is below certain threshold value, which depends upon the purpose of analysis or study. Claffy et al. [28] identified that threshold value of 64 s is good compromise considering the size of flow and initializing & terminating flows. Also, a flow can be defined as unidirectional if no differentiation is made between packets travelling in each direction and hence considered as single flow [28, 29]; or it can be defined as bidirectional if one considers packets flowing in each direction separately as two independent flows [28]. Unidirectional flows are useful in studies such as measuring network performance and bandwidth management where there is a need to measure differences in traffic in both directions. On the other hand, bidirectional flows are considered useful in scenarios such as analysing TCP sessions and for traffic classification purpose, this approach is more appropriate where traffic flowing between two sides belong to same class and generated by same application. For performing flow-based analysis, there are some tools available such as CoralReef [30] which can perform traffic analysis from network adaptors or from offline packet-traces. Tools such as Cisco Netflow [31] and Internet Protocol Flow Information eXport (IPFIX) [32] can receive the flow information directly from the router and other network elements.

3.3 Traffic data collection and trace reduction

Traffic data collection in network should be done with care in order to protect users' privacy and other data containing sensitive information. Some of the good practices and consideration have been mentioned in [33]. In Passive approach, traffic can be captured by polling of routers to obtain flows data using protocols like IPFIX or the trace files can be made by packet capturing with the help of softwares like tcpdump [34], WinDump (Windows version) [35], or other available tools which are based on libpcap [34] or WinPcap [35] libraries. But, using such techniques results in generation of large trace files, which require more processing power and storage space in case of high speed networks. To handle this issue, trace reduction can be done which reduces amount of data collected by applying packet filtering techniques. One may focus on exclusively capturing traffic belonging to a particular application which can be done using transport-layer port numbers. Alternatively, depending upon the technique used to classify traffic, one may only capture packets that request or establish a connection; or requires only first few packets of a flow for analysis. Trace files can also be reduced: i) by storing the summary of protocol-specific request of each application; ii) by capturing limited amount of packets instead of complete flow packets; iii) by storing only the header information of TCP/IP protocol stack; or iv) by storing just the flow information instead of storing each packet information. Further, packet filtering can also be done using various packet sampling methods where packets are randomly (or pseudo-randomly) chosen for analysis purpose and should be chosen in such a way that they represent the traffic to great extent which one wants to measure. Distinction of each sampling method depends upon study purpose, state of network, traffic characteristics, resource constraints, etc. Jurga and Hulb'oj in [36] and Duffield in [26] elaborated on the subject of packet sampling on traffic measurement.

4 Verification of ground truth of traffic

In early days, traffic identification was an easy task which involved port-based identification by mapping transport layer port numbers with the applications or signature-based identification by matching payload signatures with application protocols. But, as various Internet applications, especially P2P applications evolved, the traditional approaches for traffic identification started becoming ineffective, as applications based on P2P architecture used random or well-known port numbers to hide their traffic. Hence, in order to address various issues involved in traffic identification, several new techniques based on statistical or behavioural methods have been developed and adopted over the time.

In order to test new technique for traffic classification, it is essential to assess the ground truth application information of pre-collected traffic; otherwise it has very limited value [37]. Due to privacy concerns, the packet traces which are available publicly only contain header information which makes it difficult to verify the ground truth regarding the applications. But, this issue can be addressed if the packet traces are labelled for ground truth verification before making the headers publicly available. Another method which can be adopted is to verify ground truth information of the traces manually [38], but it is very slow and only feasible for smaller datasets. One may also assess the ground truth by using port number matching or payload inspection technique [39], but they are have their own limitations since port-based matching is inconsistent as many application use random port numbers, whereas DPI technique is ineffective if traffic is encrypted. Hence, by using such approaches to verify the ground truth of the traffic would produce inconsistent results while testing newer techniques. Due to such issues, researchers mostly collect their own traffic traces to verify the ground truth of the applications and test the accuracy of their techniques; but such approach

gives inconsistent results while comparing various methodologies as their performance is evaluated under different conditions [40]. It is also possible to collect traffic traces from small computer networks which run pre-defined applications in controlled environment but such approach also may not contain properties that reflect human behaviour. Some of the studies also tried to address the ground truth verification subject. Canini et al. [41] presented a framework called GTVS for improving and simplifying the process of ground truth verification of application traffic which makes use of DPI mechanism and multiple heuristic rules. Gringoli et al. [42] proposed a toolset called GT which includes the existence of deamon that is run on each client to return the process information which initiated network connection. A similar client-based approach is also proposed by Szabó et al. in [43].

None of the techniques proposed by various authors is perfect and have their own merits and demerits. Hence, the performance of new classification technique will depend upon the accuracy of the reference classification model which may lose its effectiveness if there arise any change in communication pattern of the applications. Therefore, a proper method should be chosen in order to assess the ground truth by looking the capabilities and limitations of each, as this is one of the factors on which quality of evaluation results depend.

5 Evaluation metrics for performance analysis

All network traffic classification techniques make use of some metrics in order to evaluate the classification results by comparing them with ground truth information of traces. Each individual case falls in one of the following categories:

a) True Positive (TP): It specifies that a case is correctly classified as belonging to a certain class.
b) True Negative (TN): It specifies that a case is correctly classified as not belonging to a certain class.
c) False Positive (FP): It specifies that a case is incorrectly classified as belonging to a certain class.
d) False Negative (FN): It specifies that a case is incorrectly classified as not belonging to a certain class.

A good classifier will minimize FP and FN. In terms of TPs, TNs, FPs and FNs, various metrics can be made for evaluating the performance of classifiers [44, 45], some of which may be equivalent, but most of them measure different classification aspects. Therefore, it is essential to know what is measured by a certain metric. The most commonly used metrics for traffic classification are defined as follows:

$$\text{Accuracy} = (\text{TP} + \text{TN}) \Big/ (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Accuracy measures the capability of classifier to identify positive and negative cases. It measures the overall effectiveness of classification model and hence reflects its predictive power. But, relying only on accuracy to evaluate the classifier is insufficient if imbalanced datasets are used which have large number of positive or negative cases; in which case the importance is given to the more popular class. Therefore, it is desirable to use some more metrics which can evaluate other aspects also. The most popular are: Recall and Precision, which are used together for evaluating classifiers [11] and are defined as follows:

$$\text{Recall} = \text{TP} \Big/ (\text{TP} + \text{FN})$$

$$\text{Precision} = \text{TP} \Big/ (\text{TP} + \text{FP})$$

Recall measures the percentage of overall positive cases present in the dataset that are correctly identified by the classifier. It is also referred to as hit-rate or true positive rate. Precision measures the percentage regarding correctness of the positive cases that are identified by the classifier. It is also referred to as positive-predictive value. Both the precision and recall evaluates the ability to correctly identify positive cases by the classifier; but they also have a limitation. Both cases do not give information about the amount of negative cases correctly classified by the classifier. Therefore, if required, then one can make use of another metric called Specificity [46] which can be used together with Recall for evaluation of positive and negative cases separately (in that case, Recall is usually called Sensitivity [47]) and is defined as follows:

$$\text{Specificity} = \text{TN} \Big/ (\text{FP} + \text{TN})$$

Specificity measures the percentage of cases correctly identified by the classifier as negative. Karagiannis et al. [39] also defined another metric called Completeness, which they used together with Precision to refer to accuracy and is defined as follows:

$$\text{Completeness} = (\text{TP} + \text{FP}) \Big/ (\text{TP} + \text{FN})$$

Completeness measures ratio of cases correctly or incorrectly classified as positive, to the total number of positive cases. Therefore, depending upon the context and purpose of each classifier, a proper metrics should be chosen in order to evaluate it. Table 1 shows the summary of various metrics along with their definition and the aspects they measure.

6 P2P traffic classification techniques

Earlier, traffic identification and hence classification was an easy task. However, as the P2P architecture evolved, it started

Table 1 Various evaluation metrics for performance measurement, where TP → true positive, TN → true negative, FP → false positive, FN → false negative

Metrics	Defined as	Capability/Measures
Accuracy	(TP + TN) / (TP + TN + FP + FN)	Percentage of positive and negative cases correctly identified.
Recall	TP / (TP + FN)	Percentage of overall positive cases correctly identified
Precision	TP / (TP + FP)	Percentage regarding correctness of positive cases identified
Specificity	TN / (FP + TN)	Percentage of negative cases correctly identified
Completeness	(TP + FP) / (TP + FN)	Percentage of positive cases correctly or incorrectly identified among overall positive cases.

using random port numbers or the port numbers assigned to other well-known protocols (such as HTTP), due to which another method based on inspection of payload was adopted to identify the application traffic, but that too had various limitations. So, new approaches employ statistical or behaviour-based methods that overcome various limitations which were present in traditional techniques. The following sections elaborate different types of techniques for traffic classification along with their merits and de-merits.

6.1 Port-based traffic classification

This technique relies on identification of application protocols using TCP or UDP port numbers, since each application is associated with well-defined port numbers which are defined by Internet Assigned Numbers Authority (IANA) [48]. For example, HTTP traffic uses port number 80, DNS traffic uses port number 53 and SMTP uses port number 25. This is a simple technique as it relies on packet headers only to extract to port numbers from it. A classifier placed in the middle of the network analyses for the SYN packets (which are basically TCP packets used for the purpose of 3-way handshake to establish a connection) to know about the server-side of a TCP connection and hence identifies the type of traffic flowing through the network by looking at TCP SYN packet's target port number in IANA's registered list of port numbers [48]. Similarly, UDP traffic can be identified using the port numbers it uses during communication between the hosts, but here connection establishment or its maintenance does not take place. Gomes et al. [49] presented a list of TCP and UDP port numbers utilized by several well-known P2P protocols, which is shown in Table 2.

The main advantage of this technique is that it doesn't involve any calculations and hence is fast to identify network traffic. Also, its implementation is simple which requires addition of port numbers in the database for new applications that have recently emerged. However, with the evolution on Internet, this approach started to become obsolete [10, 50, 51] as some applications such as P2P started using dynamic port numbers and port numbers which may not be registered with IANA (e.g.: Napster and Kazaa) [52]. Also, in order to get through the firewall, many applications masquerade by hiding their traffic behind well-known port numbers such as port number 80, which maps to HTTP traffic. This technique fails if there is encryption at IP layer which obfuscates TCP or UDP port numbers, hence making it impossible to recognize actual port numbers utilized by the applications.

Earlier, some P2P applications utilized port numbers or ranges which were used to identify P2P application protocols. Moore and Papagiannaki [50] identified that byte-accuracy of at most 70 % could be achieved using port-based classification technique. As port-based classification is a traditional technique, so most of its related work is referred in [49].

6.2 Payload-based traffic classification

This technique is usually most accurate and is based on inspecting packet headers and packet payloads. It relies on a database which contains signatures of previously stored application protocols. The packet payload is inspected bit-wise to locate bit-stream that contains the signatures (which are predefined byte sequences) of application protocol. Hence, the traffic can be identified accurately when packet-signatures of network application match with stored-signatures in the database. For example, 'xe3\x38' string is contained in eDonkey P2P traffic, '\GET' string is contained in web traffic and so on. This technique is not only employed for P2P traffic identification [51, 53, 54] but also in scenarios which involve identification of threats such as network intrusion detection [55], malicious data and other traffic anomalies. Such technique is also significant for accounting solutions and charging mechanisms, where accuracy is crucial.

The main advantage of this technique is that it performs network traffic identification fairly accurately. However, it also suffers with various limitations. It involves significant amount of complexity and processing load on network equipment which is used to identify network traffic. Such technique is unfeasible in high-speed networks. Hence to resolve this issue, some mechanisms inspect only few packets of each flow which is a compromise between accuracy and efficiency and sometimes in such cases, signatures may not be contained in that part which is captured, which may lead to inaccurate

Table 2 Various P2P protocols utilizing well-known port numbers

Protocols	TCP Ports	UDP Ports
AIM - messages	5190	5190
AIM - video	1024–5000	1024–5000
ARES Galaxy	32285	32285
BitTorrent	6881–6999	
Blubster	41170–41350	41170–41350
Direct Connect	411, 412, 1025–32000	1025–32000
eDonkey	2323, 3306, 4242, 4500, 4501, 4661–4674, 4677, 4678, 4711, 4712, 7778	4665, 4672
FastTrack	1214, 1215, 1331, 1337, 1683, 4329	
Gnutella	6346, 6347	6346, 6347
GoBoogy	5335	5335
HotLine	5500–5503	
ICQ	5190	
iMesh	80, 443, 1863, 4329	
IRC	6665–6669	
Kazaa	1214	1214
MP2P	10240–20480, 22321, 41170	41170
MSN	1863	
MSN - file transfer	6891–6900	
MSN – voice	6901	6901
Napster	5555, 6666, 6677, 6688, 6699–6701, 6257	
PeerEnabler	3531	3531
Qnext	5235–5237	5235–5237
ROMnet	6574	
Scour Exchange	8311	
ShareShare	6399	6388, 6733, 6777
Soribada	7675–7677, 22322	7674, 22321
SoulSeek	2234, 5534	2234, 5534
WASTE	1337	1337
WinMX	6699	6257
XMPP / Jabber	5222, 5269	5222, 5269
Yahoo – messages	5050	
Yahoo – video	5100	
Yahoo – Voice	5000–5001	5000–5010

identification of traffic. The database or the device needs to be kept updated with signatures of newly emerged application protocols or else some new traffic may get unidentified. Furthermore, it is difficult to maintain signatures with high hit and low false-positive ratio. For example, payloads of both Gnutella and HTTP traffic contain '\GET' string and hence arises ambiguity. The major drawback of this technique is that identification of network traffic becomes almost impossible if traffic is encrypted or if traffic contains proprietary protocols. Direct analysis of packet payload may also breach the privacy policies of some organisations or violate relevant privacy legislation.

Song and Zhou [56] proposed file-aware P2P traffic classification mechanism based on DPI technique to identify a file and its associated flows; which consists of two strategies based on: i) per-file bandwidth consumption, and ii) number of per-file concurrent active flows. This approach maintained 6-tuple (source-ip, destination-ip, source-port, destination-port, protocol and file-id) file-level information in flow table. In order to reduce computational overhead involved in

traditional DPI technique, pattern matching (involving only simple pattern-sets) occurred at beginning of payload and depth of inspection involved only dozen of bytes. Authors evaluated their approach on dataset collected from campus network, where majority of P2P applications include: BitTorrent, eDonkey and Gnutella; and their ground truth was verified using GTVS. The proposed approach achieved 100 % accuracy and completeness ranging from 88 to 93 %. As payload-based classification is traditional technique, so most of its related work is referred in [49].

6.3 Classification of traffic in the dark

As various limitations exists in the port-based and payload-based techniques, therefore new approaches have been developed and adopted which do not rely on port number and inspection of payload to identify the traffic. Such approach is often called classification in the Dark [39, 57] which classifies the traffic using generic properties of packets [38] such as packet size, total bytes sent, ports, etc. or by observing behavioural or statistical patterns of the flows. The main advantage of this technique is that it is able to classify the traffic without inspecting payload or relying on port numbers. However, it is not as accurate as payload-based technique but recent studies have achieved good accuracy in classifying the traffic. Also, this approach is applicable to any unknown application since methods based on it classify the traffic in a particular class instead of identifying specific applications. Various methods which fall under this approach are discussed as follows.

a) **Statistical or behavioural signatures:** Such method rely on packet or flow level properties of traffic such as packet size, totals bytes sent or received, flow duration, flow size, packet inter-arrival time, TCP or UDP ports used, etc.; which can be used individually or collectively for calculation of statistical measures such as average, variance and probability density function. In order to classify the traffic, such method requires prior learning phase to build a reference model.

Freire et al. in [58] and [59] proposed a technique to identify VoIP calls hidden in Web traffic by analysing several properties of network data, which are: size of Web request and response, number of per-page requests, inter-arrival time between requests and retrieval time of page. They evaluated their approach on VoIP data of Google-Talk and Skype which was collected from ISP and university links and achieved recall rates of about 90 % for VOIP calls and 100 % for VoIP calls hidden in Web traffic. Gomes et al. [60] analysed several P2P and non-P2P applications to identify their behaviour pattern and found that there is high heterogeneity in P2P packet sizes when compared to that of non-P2P traffic. Heterogeneity degree was represented using entropy and its value was calculated for a sliding window containing fixed number of packets. It was found that P2P traffic related to VoIP services returned high entropy values while regular client–server traffic returned consistently smaller values. Sun and Chen [61] proposed a novel technique suitable based on C4.5 decision tree for identifying application associated with a TCP flow, using two characteristics: the ACK-Len ab and ACK-Len ba; which are the data volume first sent by communicating parties continuously. Using this approach, authors classified four different types of applications: www, ftp, e-mail and P2P; where P2P traffic was identified by analysing that both parties involved in communication send considerable volumes of data to each other, thus reflecting P2P behaviour. Three dataset were used, where first was taken from Moore [62], second from the working environment (called Set1) and third was extracted from Set1 by using characteristic mentioned in ref. [63]. The proposed approach can be used for online traffic classification as it only depends on data's total length of first few packets on the flow which greatly save storage space and classified P2P traffic with accuracy, recall and precision rates ranging from 97.648 to 99.694 %, 30 to 80 % and 65 to 93 %, respectively. He et al. [64] proposed fine-grained host-based P2P traffic classification by simply counting special flows (i.e. clustering flows). This approach locates all P2P hosts within monitored network and identifies the types of P2P application running. It builds application profiles of each P2P application by using the flow information that describes its most significant network activity pattern and is learned from traffic traces generated by corresponding P2P application. The performance is evaluated on traffic datasets consisting of P2P applications namely BitComet, BitTorrent, eMule, Vagaa and Thunder. The ground truth verification is done by manually investigating each host running P2P application. The experimental results achieved average true positive and false positive rate of 97.22 and 2.78 % respectively. The proposed approach does not use complicated statistical features of traffic or machine learning algorithms and can readily include new P2P applications in classification scope. It is also able to classify encrypted traffic in real-time**.** Yang et al. [65] proposed a method to identify P2P live streaming based on union features by analyzing its behavioural characteristics. The datasets consisted of mixture of traffic from BitTorrent and Thunder which are file sharing applications and traffic from PPTV, PPStream, QQlive and UUSE which are on-demand and live streaming applications. The experimental results achieved 95 % accuracy in identifying P2P live streaming traffic. Qin et al. [66] developed a framework named CUFTI (Core Users Finding and Traffic Identification) for identifying and managing P2P traffic of core users (i.e. long-lived peers). They studied peer's lifetime in PPlive system and identified core users from the overlay. The model utilized payload length and direction of first few control packets of different P2P applications (PPlive,

BitTorrent and Thunder) as statistical features that were extracted using the longest common subsequence (LCS) and performed flow identification. The experimental results achieved false positive and false negative rates of 3.49 and 8.47 %, respectively in identifying PPlive traffic. Further the model can be employed for real-time identification of traffic. Zhang et al. [67] proposed component based method to detect P2P traffic utilizing UDP for communication. In graph theory, component is defined as connected sub-graphs from a disjoint graph. The approach uses graph-level statistics to detect P2P traffic (utilizing UDP) and does not use packet level information. The dataset consisted of records taken from netflow version 5 and exported from university campus network borderlink.

b) **Heuristic-based methods:** This method classifies the traffic by observing the behavioural patterns of traffic using pre-defined set of heuristics such as hosts acting both as client and server, number of connections made by host, number of distinct addresses or ports a host is connected to, hosts using both TCP and UDP for communication, etc. The set of heuristics are analysed sequentially and the packets or flows are classified as belonging to a particular class depending upon the results obtained. There are some studies that make use of heuristics to identify P2P traffic.

Per'enyi et al. [68] proposed a technique for identification of P2P traffic that is based on set of six heuristics: usage of UDP and TCP simultaneously; well-known P2P port numbers; number of consecutive connections existing between two peers; several flows having same flow identities; flow-duration greater than 10 min or flow-size greater than 1 MB; and an IP address using same port number more than 5 times in measurement period. A small labelled traffic traces were used for validation of this approach, which achieved recall rate of 99.14 % for P2P traffic and 97.19 % for non-P2P traffic. John and Tafvelin [69] redefined the combination of heuristics used in [68] and [54] and proposed the heuristics: usage of UDP and TCP simultaneously; well-known port numbers of P2P protocols; the port numbers that are used very often; relationship between number of ports and IP addresses; flow-duration greater than 10 min or flow-size greater than 1 MB. They collected the traffic traces from university link and achieved recall rate of 98 %. Hong [70] proposed a novel method to identify P2P traffic utilizing UDP protocol and revealed & validated three unique characteristics that will not appear together in TCP or UDP traffic produced by non-P2P applications, which are: i) almost all UDP traffic of local host transfers by fixed port number; ii) nearly all remote peers use single port number for communication with local host, and iii) size of UDP packets produced by P2P applications is relatively fixed. These characteristics were examined by collecting 100 blocks of P2P traffic (consisting of BitSpirit, Emule and other P2P applications), each ranging from 100 M bytes to 200 M bytes and evaluation of this approach achieved an accuracy ranging from 98.4 to 99.6 %. Reddy and Hota [71] proposed a new set of heuristics to identify P2P host based on its connection patterns and they do not require any payload signatures. The datasets used was realistic in nature ad consisted of applications namely Http, FTP, Dropbox, SMTP, eMule, Frostwire, Skype, uTorrent and Vuze. The authors verified their approach in real time and only 0.2 % of P2P traffic remained unclassified. As their approach consisted of minimal heuristics, it can be used for real-time identification; but it can only identify broad P2P applications rather than different P2P applications. Bashir et al. [72] proposed an approach based on heuristics to identify BitTorrent activities using Netflow records by observing 3 major segments of traffic: a) traffic from peers contacted via DHT, b) TCP traffic from peers contacted via trackers and c) UDP traffic from peers contacted via trackers. The approach was tested on 5 real life datasets having mixture of applications consisting of BitTorrent, p2p radio streaming application, Skype, SopCast and PPStream. The experimental results achieved the byte accuracy ranging from 91.3 to 95.4 % in identifying BitTorrent activity.

c) **Machine Learning methods:** Machine learning techniques based on supervised or un-supervised methods have been adopted in various studies such as clustering [73], Bayesian estimators or networks [74] and decision trees [75]; which work on set of traffic characteristics by correlating them using probability functions and hence classify the packets or flows as belonging to particular class.

Mohammadi et al. [2] proposed a hybrid approach using genetic algorithm neural networks to classify P2P traffic. Genetic algorithm was used in calculating minimum classification error (MCE) matrix which is then used to map features of dataset into new space where they can easily be separated into different classes. The mapped dataset is fed into classifier named neural networks. Three different indexes namely mutual information, Dunn and SD were measured to compare proposed methodology against standard MCE-based and normal (i.e. no feature mapping) approaches. The experimental results showed that proposed mapping technique reduces overlap among classes and gives improved classification accuracy of 96 %. Schmidt and Soysal [76] proposed a technique involving Bayesian network to identify P2P traffic by using the parameters: well-known port numbers, IP packets-per-flow distribution, packet-size distribution, octets-per-flow distribution and flow-time distribution. They collected the traffic from academic network to evaluate the performance of classifier in their technique as well as in signature-based

method and showcased the results of false positive ranging from 22 to 28 % and false negative ranging from 16 to 26 %. Cao et al. [77] proposed a technique using Classification And Regression Tree (CART) for real-time identification of application protocols at both flow-level and host-level. They collected the traffic traces of HTTP, SMTP & FTP from enterprise network by port number filtering method and traces of BitTorrent were collected actively at home environment in controlled manner to assess the ground truth. By evaluating this technique, the classification results obtained showed false positive rate ranging from 0.05 to 12.7 % and false negative rates ranging from 0 to 17.9 %. Raahemi et al. [47] proposed a technique using set of network level packet attributes to identify P2P traffic by using Concept-adapting Very Fast Decision Tree (CVFDT). In order to evaluate the performance of their technique, they used labelled datasets and achieved the accuracy ranging from 79.50 to 98.65 % and specificity ranging from 82.96 to 95.89 %. Angevine and Zincir-Heywood [78] classified TCP and UDP flows of Skype using C4.5 decision tree and AdaBoost algorithms. They collected the labelled traffic traces from university network and achieved recall rate ranging from 94 to 99 % with their technique. Wang et al. [79] identified traffic of multiple P2P protocols using classifier based on decision tree called Random Forest. They captured the traffic traces from academic and residential networks and evaluated their technique using manually labelled dataset to achieve accuracy ranging from 89.38 to 99.98 % and precision ranging from 32.69 to 100 %. Dainotti et al. [80] proposed a classification technique based on hidden Markov models and using parameters: packet size & inter-packet time. They carried out classification on real-traffic traces of HTTP, SMTP, eDonkey, P2P-TV, MSN messenger, PPlive & two multi-player games; whose traces were verified manually as well as using DPI technique, to achieve recall rates ranging from 90.23 to 100 %. Valenti et al. [81] adopted a mechanism based on Support Vector Machine (SVM) and number of packets exchanged between peers during short interval of time; to identify P2P-TV applications. They tested their approach on traffic captured in larger test-bed to achieve recall rates ranging from 91.3 to 99.6 %. Liu et al. [82] proposed a mechanism by utilizing supervised ML algorithm and ratio of amount of downloaded and uploaded traffic in each minute as an identification pattern. They classified P2P applications of Maze, PPlive, BitTorrent, eDonkey and thunder and achieved accuracy ranging from 78.5 to 99.8 %. Raahemi et al. [83] identified P2P traffic using the neural network: Fuzzy Predictive Adaptive Resonance Theory; which was built by utilizing IP headers data. This approach utilized labelled datasets to achieve the classification accuracy ranging from 78 to 92 %. Hu et al. in [84, 85] proposed a novel approach to identify the various applications by building behavioural profiles using association rule mining. They extracted flow statistics by selecting five flow tuples and correlated them using Apriori algorithm. The authors collected the traffic traces from on-campus network, which were verified manually as well as using DPI technique and tested this mechanism on BitTorrent and PPlive to achieve the recall rates ranging from 90 to 98 %.

Liu and Sun [86] proposed a new approach called P2PTIAL that doesn't require fully labeled samples-set for P2P traffic identification by active learning which consists of two parts: Support Vector Machine (SVM) and uncertainty selection policy. SVM acts as learner which repeats learning process on both labelled & unlabelled sample; whereas uncertainty selection (which is based on distance) selects unlabelled sample to be labelled by oracle (e.g., a human annotator). Further, to improve its effectiveness, authors employed support vector data description (SVDD) technique to filter unlabelled samples having little contribution in active learning to reduce storage space & save computation cost; and used unlabeled sample's pre-labeled information to avoid imbalanced learning. They utilized Moore-dataset [38, 87], which includes traffic from applications: P2P, www, bulk, database, interactive, mail, services, attack, games & multimedia and evaluated their technique on both un-balanced & balanced learning to achieve the accuracy rate ranging from 79.65 to 86.86 % and 93.00 to 93.07 %, respectively. Jiang and Tao [88] proposed P2P traffic identification model based on SVM that can work on encrypted traffic and selected 3 characteristics: i) change of mean square value of packet size, ii) average flow duration, and iii) ratio of IP address and port numbers. The performance achieved in terms of precision, false-positive and false-negative rates range from 96.55 to 97.89 %; 2 to 2.8 % and 2.45 to 5.29 %, respectively. Gong et al. [89] proposed improved SVM incremental learning algorithm for P2P traffic identification which is able to save storage space and increase identification accuracy (87.89 %), when its performance is compared with standard SVM incremental learning algorithm (having 80.35 % accuracy) and SVM-based retraining algorithm (having 78.90 % accuracy) for increased number of test samples. Deng et al. [90] proposed the ensemble learning model which integrates Random Forests and feature weighted Naive Bayes for P2P traffic identification. Network traces considered for evaluation consisted of both P2P traffic (BaiDuYingYin, BaoFengYingYin, PPS, PPlive, QQlive, XunLeiKanKan and Thunder) and non-P2P traffic (Web, Youku and Souhu) and achieved accuracy of 92.47 %; which overall performs better when compared to simple machine learning methods. Jie et al. [91] proposed a novel and fine-grained P2P traffic classification approach that relied on count of most frequent and steady flows generated by corresponding P2P applications called Clustering Flows. This approach exploited only basic properties of flows (protocol, packets size and number) to perform the classification using SVM algorithm and doesn't require any other complicated traffic statistical or behavioural features. The experiment

performed on traffic traces of P2P applications include BitTorrent, eMule, PPTV & Cbox and achieved true positive rate ranging from 95.4 to 98.63 % and false positive rate of 0.01 %. Bozdogan et al. [92] evaluated the performance of machine learning algorithms for classification of P2P applications, which include BitCommet, uTorrent and BitTorrent. Four supervised algorithms (C4.5, Ripper, SVM and Naive Bayes) and one un-supervised algorithm (K-means) were evaluated using the metrics: detection rate, false positive rate, f-measure and correctly classification rate. The experimental results showed that Ripper algorithm performs better in identifying P2P network traffic.

d) **Methods involving combined approaches:** There also exist some studies which combine different classification approaches to identify network traffic, which are discussed below.

Karagiannis et al. [54] adopted cross-validation mechanism to identify traffic from FastTrack, eDonkey, Gnutella, BitTorrent, Direct-Connect, MP2P & Ares; by using port-numbers, payload signatures and behavioural patterns. In addition to using payload-signatures for particular applications, the non-payload based method used two heuristics to identify flows belonging to P2P applications, which are: (i) identification of source & destination IP pairs that use both TCP and UDP; and (ii) identification of number of distinct IP addresses connected to destination IP is equal to number of distinct ports used for making connections. The behavioural approach achieved the recall rates ranging from 90 to 99 %. Also, they compared the results of payload-based approach with behavioural approach to find the false positive rates ranging from 8 to 12 % of overall P2P traffic. Dedinski et al. [93] adopted an approach for identification of P2P traffic that made use of active crawlers for collecting information of peers of a certain application to infer the topology of the overlay network. In addition, for analysing behavioural patterns, the authors used wavelet analysis technique on traffic to analyse network-level properties: per-packet or inter-packet arrival times. The performance of this architecture evaluated on traffic belonging to eDonkey and FTP. Adami et al. [94] proposed a real-time mechanism using payload-based method & statistical method to identify different Skype clients in the network, which have the communication of: file transfer, direct calls, calls to phone service and calls using relay nodes. They collected the traffic traces from a university network and ADSL link of a small network. The performance of this mechanism (which was conducted both online and offline) was tested for both TCP & UDP with other five classifiers, to achieve false positive rates ranging from 0 to 0.01 % and false negative ranging from 0.06 to 0.64 %, in terms of bytes and flows.

Yan et al. [95] proposed a novel technique for P2P identification based on host heuristics & flow statistics. In order to find out if host is participating in P2P application, authors first matched its behaviour with pre-defined heuristic rules:- IP-popularity ratio, port-pair difference, ephemeral-port ratio, failed-connection ratio; and secondly refined the identification by comparing statistical features of each flow with flow features:- Flow-bytes & flow-duration, and byte-ratio of forward & backward direction. The traffic traces were collected at edge router of the campus network and consists of Web (http and https), Mail (pop3, pop3s, imap, imaps) and P2P (bittorrent, edonkey, skype) traffic; and accuracy rate achieved by this technique in terms of flows and bytes were 93.9 and 96.3 %, respectively. Ye and Cho [96] proposed two-step hybrid P2P traffic classification approach by combining packet-level and flow-level classifier. First step (which is packet-level classification) is the combination of signature-based and heuristic-based technique; where the packets if not classified with former approach, are checked with the latter one for classification. The second step (which is flow-level classification) is based on combination of statistical & pattern-heuristics approach; which is applied on the traffic that remains unclassified in first step. The authors used REPTree algorithm with statistical approach after comparing six ML algorithms for their performance and then applied pattern heuristics (set of rules) to rectify faulty results caused by the former approach. Four datasets were used for evaluation of this technique; where the first two were taken from University of Brescia and Ericsson Research in Hungary other two in controlled environment inside the Dankook University that were labelled with actual application types. The proposed scheme showed low overhead & high scalability and was able to achieve the accuracy rates of 98.19 & 99.82 % in terms of flows and bytes. The authors in [97] used similar hybrid approach to classify and distinguish between P2P botnet traffic from P2P traffic. The botnet traffic of Storm, Waledac, Conficker, C&C and Zeus were mixed to create three datasets. The proposed approach provides low overhead and achieved flow and byte accuracy of 97.10 and 97.06 % respectively using real datasets. Wang et al. [98] proposed a novel Application Behavior Characterization technique for P2P identification. It extracts behavioural features (number of external IP addresses, number of flows, number of packets and number of bytes) from set of flows belonging to certain applications and classifies P2P traffic using machine learning algorithm: C4.5 decision tree. The datasets used involved TCP and UDP flows belonging to Skype, Thunder, PPTV and non-P2P applications. The experimental results achieved for PPTV, Skype and Thunder include precision values of 93.66, 91.01 and 90.96 % and recall values of 92.82, 86.69 and 95.73 %, respectively. Yang et al. [99] proposed a cocktail approach consisting of three sub-methods for identifying BitTorrent traffic. First sub-method uses signature-based approach to identify un-encrypted BitTorrent traffic. Second sub-method uses message-based approach to perform identification of

Table 3 Summary of traffic classification studies involving different approaches, including P2P traffic involved and their performance in terms of accuracy (A), precision (P), recall (R), completeness (C), sensitivity (SN), specificity (SP), false-positive (FP), false-negative (FN) or true-positive (TP)

Technique	Method	Ref.	Studies	Performance%	P2P traffic involved
Port-based approach	inspecting port numbers	[50]	Moore and Papagiannaki [2005]	A: 70	P2P
		[106]	Saroiu et al. [2002]	–	Akamai content delivery network, Kazaa, Gnutella
		[107]	Leibowitz et al. [2002]	–	Kazaa, Grokster, Morpheus
		[108]	Gerber et al. [2003]	–	P2P
Payload-based approach	inspecting packet payload	[53]	Sen et al. [2004]	FP: 0, FN: 0.0–9.9	Gnutella, eDonkey, Kazaa, Direct Connect, BitTorrent
		[50]	Moore and Papagiannaki [2005]	R: 99.99	Kazaa, BitTorrent, Gnutella
		[51]	Karagiannis et al. [2004]	–	FastTrack, eDonkey2000, WinMX, BitTorrent, Gnutella, Soulseek, Direct Connect, MP2P
		[109]	Spognardi et al. [2005]	–	OpenNap, WPN, FastTrack
		[110]	Bin et al. [2007]	–	eDonkey
		[111]	Dewes et al. [2003]	R: 91.7, P: 93.13	Chat-traffic
		[112]	Guo and Qiu [2008]	FP: 0–11, FN: 0.33–10.50	BitTorent
		[113]	Cascarano et al. [2010]	–	eDonkey, BitTorent, Skype, PPlive, Tvants, Sopcast
		[114]	Carvalho et al. [2009]	–	BitTorrent
		[115]	Carvalho et al. [2009]	–	LiveStation, TVUplayer (P2PTV traffic)
		[116]	Freire et al. [2009]	–	eDonkey
		[117]	Park et al. [2008]	A: 97.39, FP: 0.39–10.40, FN: 0	LimeWire, Fileguri, BitTorrent
		[56]	Song and Zhou [2013]	A: 100, C: 88–93	BitTorrent, eDonkey, Gnutella
Classification in the Dark	Statistical, behavioural, heuristic, machine-learning and combined methods	[58, 59]	Freire et al. [2008]	R: 90–100, FP: 2–5	Skype, Google-Talk
		[60]	Gomes et al. [2008]	–	P2P
		[61]	Sun and Chen [2011]	A: 97.648–99.694, R: 30–80, P: 65–93	Xunlei, PPTV, BTComet, etc.
		[68]	Per'enyi et al. [2006]	R: 97.19–99.14, FP: 0.3; FN: 0.8	Direct Connect, Gnutella, BitTorrent, eDonkey, Napster, File navigator, WinMX
		[69]	John and Tafvelin [2008]	R: 98	P2P
		[70]	Hong [2011]	A: 98.4–99.6	BitSpirit, Emule, etc.
		[76]	Schmidt and Soysal [2006]	A: 62, FP: 22–28, FN: 16–26	P2P
		[77]	Cao et al. [2008]	FP: 0.05–12.7, FN: 0–17.9	BitTorrent
		[47]	Raahemi et al. [2008]	A: 79.50–98.65, SN: 82.96–	P2P

Table 3 (continued)

Technique	Method	Ref.	Studies	Performance%	P2P traffic involved
				95.89, SP: 67.96–99.72	
		[78]	Angevine and Zincir-Heywood [2008]	R: 94–99, FP: 1–26	Skype
		[79]	Wang et al. [2008]	A: 89.38–99.98, P: 32.69–100.00, FP: 0.00–12.61	P2P
		[80]	Dainotti et al. [2008]	R: 90.23–100.00	Gaming, eDonkey, PPlive, MSN
		[81]	Valenti et al. [2009]	R: 91.3–99.6, FP: 0.3–8.7	P2P-TV
		[82]	Liu et al. [2007]	A: 78.5–99.8	Maze, BitTorrent, PPlive, eDonkey, Thunder
		[83]	Raahemi et al. [2008]	A: 78–92, SN: 68–90, SP: 85–96	P2P
		[84]	Hu et al. [2008]	R: 90–98, FP: 0.2–5.0	BitTorrent, PPlive
		[85]	Hu et al. [2009]	R: 90–98, FP: 0.2–5.0	BitTorrent, PPlive
		[2]	Mohammadi et al. [2011]	A: 96	P2P
		[86]	Liu and Sun [2014]	A: 79.65–93.07	P2P, Attack, Games, Multimedia
		[54]	Karagiannis et al. [2004]	R: 90–99, FP: 8–12	FastTrack, eDonkey, Gnutella, BitTorrent, Direct-Connect, MP2P, Ares
		[93]	Dedinski et al. [2005]	–	eDonkey
		[94]	Adami et al. [2009]	FP: 0.00–0.01, FN: 0.06–27.46	Skype
		[95]	Yan et al. [2013]	A: 93.9–96.3	BitTorrent, eDonkey, Skype
		[96]	Ye and Cho [2014]	A: 98.19–99.82	P2P
		[88]	Jiang and Tao [2013]	P: 96.55–97.89, FP: 2–2.8, FN: 2.45–5.29	P2P
		[89]	Gong et al. [2014]	A: 87.89	P2P
		[98]	Wang et al. [2014]	P: 90.96–93.66, R: 86.69–95.73	Skype, Thunder, PPTV
		[90]	Deng et al. [2014]	A: 92.47	BaiDuYingYin, BaoFengYingYin, PPS, PPlive, QQlive, XunLeiKanKan, Thunder
		[91]	Jie et al. [2013]	TP: 95.40–98.63, FP: 0.01	BitTorrent, eMule, PPTV, Cbox
		[100]	Korczynski and Duda [2014]	TP: 98.6, FP: 0.1	Gadu-Gadu, Skype
		[101]	Alshammari and Zincir [2015]	TP: 80.3–99.6, FP: 0.7–3.8	Skype
		[102]	Kumano et al. [2014]	A: 79.3–92.5	P2P
		[103]	Wang et al. [2015]		Chat

Table 3 (continued)

Technique	Method	Ref.	Studies	Performance%	P2P traffic involved
				TP: 96.4–99.1, FP: 0–3.6	
		[104]	Du and Zhang [2013]	TP: 92.64–99.76	BitTorrent, BitSpirit, eMule
		[71]	Reddy and Hota [2015]	A: 99.16–99.99	eMule, Frostwire, Sype, uTorrent, Vuze
		[97]	Ye and Cho [2015]	A: 97.06–99.10	Storm, Waledac, Conficker, C&C, Zeus
		[64]	He et al. [2015]	TP: 97.22, FP: 2.78	BitComet, BitTorrent, eMule, Vagaa, Thunder
		[92]	Bozdogan et al. [2015]	–	BitComet, BitTorrent, uTorrent
		[72]	Bashir et al. [2013]	A: 91.3–95.4	BitTorrent
		[99]	Yang et al. [2012]	FP: 1.31–2.47 P: 98.26–99.03 R: 85–98	BitTorrent
		[105]	Datta et al. [2015]	R: 99.98–100	Google-hangout
		[65]	Yang et al. [2013]	A: 95	PPTV, PPStream, QQlive, UUSE
		[66]	Qin et al. [2015]	FP: 3.49 FN: 8.47	PPlive
		[67]	Zhang et al. [2014]	–	P2P

encrypted BitTorrent traffic. Here, after resembling the bidirectional flows into message streams, if the direction and length of first three messages satisfy certain criteria of message stream encryption (a protocol used to obfuscate traffic), then it classifies the flow as encrypted BitTorrent traffic. Third sub-method uses signalling-based approach to perform pre-identification of BitTorrent traffic. Here, prediction of BitTorrent flows takes place using first packet with SYN flag only. The authors evaluated their approach by using modified Vuze clients which not only generated real BitTorrent traffic but also labelled the traffic in benchmark traces by themselves. The experimental results achieved false positive, precision and recall rates ranging from 1.31 to 2.47 %, 98.26 to 99.03 % and 85 to 98 %, respectively. This approach has the ability for real-time identification with low overhead.

6.4 Classification of encrypted traffic

Nowadays, due to widespread use of encrypted communication to protect personal information and/or to conceal exchanged information; identification accuracy is dropping. For example, encryption is used in P2P file sharing, VoIP and ISPs offering virtual private networks for communication. These factors reflect that encryption in going to increase and it makes harder for network administrators to identify applications, since the traffic and its characteristics gets changed when it is encrypted. Hence, most identification methods classify encrypted traffic as either unknown traffic or wrongly infer encrypted traffic as belonging to same application, even though different encrypted applications are mixed in traffic. Hence, most of the existing methods can be expected to become less effective. There exist some studies that make use of P2P traffic classification techniques (discussed in previous section) for addressing this issue, which are discussed below.

The Korczynski and Duda [100] proposed stochastic fingerprints based on first-order homogeneous Markov chains to identify encrypted traffic flows of various applications. They studied twelve representative applications (which includes Skype), whose parameters were identified by observing training application traces. Their technique achieved good accuracy as fingerprint parameters of applications differ considerably. The issue with this technique is that, as application fingerprints change over time; they need to be updated periodically. For P2P application (Skype), the experimental results achieved true positive rate of 98.6 % and false positive rate of 0.1 %. The Alshammari and Zincir [101] proposed a novel technique to identify VoIP encrypted traffic that is based on machine learning which generated robust signatures. They used statistical calculation on network flows to extract feature set without the use of information regarding payload or port numbers & IP addresses of source and destination. Three different sampling techniques (uniform random sampling, stratified sampling, continuous data stream) were studied on three

Table 4 A summary of cited papers that mentions references (Ref), publication year (Year), authors (Studies) and classification technique used/applicable which includes: port (Port), payload (Payl), statistical/behavioural signature (Stat/Beha), machine learning (Mach), heuristic (Heu), machine learning algorithm (Algorithm), Real-time classification (Real) and encryption (Encryption)

Ref	Studies	Technique					Algorithm	Real	Encryption
		Port	Payl	Stat/ Beha	Mach	Heu			
[106]	Saroiu et al. [2002]	✓						✓	Yes
[107]	Leibowitz et al. [2002]	✓						✓	Yes
[108]	Gerber et al. [2003]	✓						✓	Yes
[50]	Moore and Papagiannaki [2005]	✓						✓	Yes
[111]	Dewes et al. [2003]		✓						No
[53]	Sen et al. [2004]		✓						No
[51]	Karagiannis et al. [2004]		✓						No
[50]	Moore and Papagiannaki [2005]		✓						No
[109]	Spognardi et al. [2005]		✓						No
[110]	Bin et al. [2007]		✓						No
[112]	Guo and Qiu [2008]		✓						No
[117]	Park et al. [2008]		✓						Yes
[114]	Carvalho et al. [2009]		✓						Yes
[115]	Carvalho et al. [2009]		✓						Yes
[116]	Freire et al. [2009]		✓						Yes
[113]	Cascarano et al. [2010]		✓						No
[56]	Song and Zhou [2013]		✓						No
[58, 59]	Freire et al. [2008]			✓				✓	Yes
[60]	Gomes et al. [2008]			✓				✓	Yes
[65]	Yang et al. [2013]			✓					No
[67]	Zhang et al. [2014]			✓					Yes
[64]	He et al. [2015]			✓				✓	Yes
[66]	Qin et al. [2015]			✓				✓	Yes
[76]	Schmidt and Soysal [2006]				✓		Bayesian network		Yes
[82]	Liu et al. [2007]				✓		–	✓	Yes
[47]	Raahemi et al. [2008]				✓		Concept-adapting Very Fast Decision Tree		No
[78]	Angevine and Zincir-Heywood [2008]				✓		C4.5, AdaBoost		Yes
[79]	Wang et al. [2008]				✓		Random Forest		No
[80]	Dainotti et al. [2008]				✓		Hidden Markov Model		Yes
[83]	Raahemi et al. [2008]				✓		Neural network (Fuzzy ARTMAP)	✓	Yes
[81]	Valenti et al. [2009]				✓		SVM		Yes
[2]	Mohammadi et al. [2011]				✓		Neural network		No
[88]	Jiang and Tao [2013]				✓		SVM		Yes

Table 4 (continued)

Ref	Studies	Technique					Algorithm	Real	Encryption
		Port	Payl	Stat/ Beha	Mach	Heu			
[104]	Du and Zhang [2013]				✓		K-means	✓	Yes
[86]	Liu and Sun [2014]				✓		SVM		No
[89]	Gong et al. [2014]				✓		SVM		No
[90]	Deng et al. [2014]				✓		Random Forest, Naive Bayes		Yes
[100]	Korczynski and Duda [2014]				✓		Markov Chains		Yes
[102]	Kumano et al. [2014]				✓		C4.5, SVM	✓	Yes
[103]	Wang et al. [2014]				✓		Hidden Markov model		Yes
[101]	Alshammari and Zincir-Heywood [2015]				✓		C5.0, AdaBoost, Genetic		Yes
[92]	Bozdogan et al. [2015]				✓		C4.5, Ripper, SVM, Naive Bayes		No
[68]	Per'enyi et al. [2006]					✓			Yes
[69]	John and Tafvelin [2008]					✓		✓	Yes
[70]	Hong [2011]					✓			Yes
[72]	Bashir et al. [2013]					✓		✓	Yes
[71]	Reddy and Hota [2015]					✓		✓	Yes
[54]	Karagiannis et al. [2004]	✓	✓			✓			Yes
[94]	Adami et al. [2009]		✓	✓				✓	Yes
[99]	Yang et al. [2012]		✓	✓				✓	Yes
[77]	Cao et al. [2008]			✓	✓		Classification & Regression Tree	✓	Yes
[84]	Hu et al. [2008]			✓	✓		Apriori, Association Rule mining		Yes
[85]	Hu et al. [2009]			✓	✓		Apriori, Association Rule mining		Yes
[61]	Sun and Chen [2011]			✓	✓		C4.5	✓	Yes
[91]	Jie et al. [2013]			✓	✓		SVM	✓	Yes
[98]	Wang et al. [2014]			✓	✓		C4.5	✓	Yes
[105]	Datta et al. [2015]			✓	✓		Naive Bayes, AdaBoost, J48		Yes
[95]	Yan et al. [2013]			✓		✓			Yes
[96]	Ye and Cho [2014]		✓	✓	✓	✓	REPTree	✓	Yes
[97]	Ye and Cho [2015]		✓	✓	✓	✓	REPTree	✓	Yes

machine learning algorithms (C5.0, AdaBoost, Genetic programming) that were trained on various training datasets; where uniform random sampling was found to be most appropriate for enhancing automatic generation of robust signatures. Experimental results showed that C5.0 performs much better than GP and AdaBoost algorithms in classifying multiple VoIP applications and classified Skype traffic with detection rate ranging from 80.3 to 99.6 % and false positive rate ranging from 0.7 to 3.8 %. But, for other network applications, this technique needs to be explored for its accuracy. The Kumano et al. [102] focused on identifying encrypted traffic in real-time by reducing no. of packets needed to obtain traffic features and maintaining high accuracy. They used two types of encryption (IPSec and PPTV) and employed two machine learning algorithms (C4.5 and SVM) for classifying type of encryption and identification of application. Their work shows how accuracy degrades by reducing no. of packets and also proposed a procedure to identify sufficient no. of packets for each traffic feature. They compared overall accuracy by varying no. of features and packets; which ranged from 79.3 to 92.5 %. The number of packets can further be reduced for some features by eliminating initialization packets but detailed exploration and estimation is required to be done. The Wang et al. [103] proposed a novel approach based on Hidden Markov Model for identifying network activities of encrypted traffic. In their technique, time series and statistical characteristics of packets are considered for analysis. Four time series sequences during the interaction of four activities (session request, data transfer, response to session request, and response to data transfer) are analysed for distinction; due to which packet inter-arrival time is considered as feature element. Similarly for statistical characteristics, due to distinction in packet sequences of four activities; packet length and packet inter-arrival time are selected as feature elements. To verify the effectiveness of the approach, TeamViewer (which allows encrypted communication between hosts) is used. The datasets utilized includes audio, video, transfer and chat traffic types. Experimental results achieved true positive rate ranging from 96.4 to 99.1 % and maximum false positive rate of 3.6 %. However, unsupervised learning methods of modelling and further analysis of complex activities needs to considered further. Du and Zhang [104] identified P2P traffic by utilizing k-means algorithm that monitors flow information of TCP connections and calculates distance. Their approach focused on three TCP file-sharing P2P applications namely BitTorrent, BitSpirit and eMule. Experimental results achieved average true positive rate of 92.64, 96.22 and 99.76 % for BitTorrent, BitSpirit and eMule, respectively. The algorithm proposed by authors is simple, feasible, low overhead of time and can be used for real-time detection of traffic. Datta et al. [105] proposed a novel technique using application behaviour based feature extraction to detect Google-hangout traffic by taking it as a case study. Three machine algorithms were used namely Naive Bayes, J48 decision tree and AdaBoost to classify traffic. The datasets consisted of traffic traces of google-hangout, gmail and google-plus, since these google services share common behaviour between them. The classification results had the recall values of 100 % with J48 and AdaBoost separately and 99.98 % with Naive Bayes.

Table 3 provides the summary of different P2P classification approaches along with the methodologies adopted by various studies. For each study, the performance evaluation is also mentioned; which makes use of the metrics: accuracy, precision, recall, completeness, sensitivity, specificity, false-positive, false-negative or true-positive (TP). Additionally, P2P traffic involved in a study is also mentioned to give an idea of the kind of traffic on which the corresponding performance is achieved. The comparison between various methods in the Table 3 cannot be done, as evaluations were made by authors using distinct metrics and under different conditions.

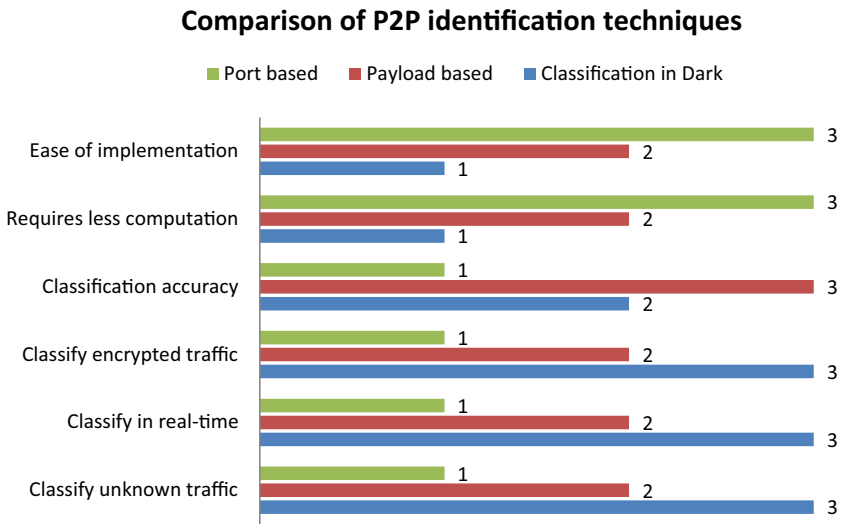

Fig. 1 Comparison of P2P identification techniques based on their performance by considering various factors

Hence, it only provides an overview of various methods used for classifying P2P traffic which are presented in this literature.

Table 4 presents the summary of various studies conducted to classify P2P traffic along with their references and publication year. The technique/approach adopted by various studies for classifying P2P traffic is categorised based on: port, payload, statistical (or behavioural signatures), machine learning and heuristic. If a study uses machine learning approach for classifying P2P traffic, then corresponding algorithms used in it have also been specified. In addition, two columns are added to describe the ability of a method to be applied to encrypted traffic and for real-time classification. Although the studies based on port numbers did not address the issues of encryption and real-time classification, they still have the ability to identify the traffic. It is because TCP and UDP port numbers are not usually encrypted and traffic can be quickly categorized online by matching their port numbers with the stored database of applications.

By considering various studies discussed in previous sections and advantages as well as limitations of various identification techniques (i.e. port-based, payload-based and classification in dark), Fig. 1 compares them by considering their implementation, resource requirements and performance in classifying traffic. Hence, the comparison factors include: ease of implementation, requiring less computation, classification accuracy, classification of encrypted traffic, classification in real-time and classification of unknown traffic. Each technique is given a value on a particular factor ranging from 1 to 3, where value 3 represents comparatively highest performing technique and value 1 represents comparatively lowest performing technique. Port-based technique has highest value while considering the factors of ease of implementation and less computation requirement. This technique has the ability to classify encrypted traffic and real-time classification, but it has lowest value in all remaining factors (i.e. classification of encrypted traffic, classification in real-time, classification accuracy and classify unknown traffic) since current generation P2P applications masquerade or utilizes random port-numbers due to which it will not give accurate results. Payload-based classification has highest performance when classification accuracy is of prime importance. Due to this fact it is widely used for ground truth verification of traffic which is discussed in section 4; but comparatively it doesn't perform well on other remaining factors. Classification in Dark has highest performance while considering encrypted traffic classification, real-time classification and unknown traffic classification.

7 Conclusion

Major portion of Internet is composed of P2P traffic which consumes a lot of network bandwidth. With the evolution of P2P applications and services and more hosts keep on joining/adopting them; it poses various challenges for network administrators or ISPs to address or manage the network issues concerned with billing, security, fault diagnosis, quality of service, among others. Hence, it is necessary for network administrator or ISPs to accurately and efficiently identify the kind of traffic flowing through their network. Traditionally, port-based mechanism was used for traffic identification, but has lost its utility as applications started masquerading or using random port numbers. Due to such limitations, payload-based mechanism was adopted which has very high accuracy, but also suffers from various limitations or issues such as traffic encryption, privacy, etc. Therefore, newer approaches based on Classification in Dark have been adopted to identify network traffic which overcomes various limitations of previous approaches.

This paper presents a survey on P2P traffic identification approaches and analyses some of the methodologies & achievements of each approach. Nowadays, due to widespread use of encryption for communication by most applications, the existing approaches lose effectiveness and make harder for network administrators or ISPs to accurately classify network traffic, since the traffic as well as its characteristics gets changed resulting in reduced accuracy. Real-time traffic classification also has great importance. So, future work needs to focus on identifying encrypted P2P traffic efficiently in real-time that can also work in high-speed networks. Research should be focused on developing technique that can identify traffic from individual P2P applications (i.e. fine-grained classification) instead of just identifying P2P traffic (i.e. course-grained classification) so that ISPs or network administrators can manage traffic in better way. Also, a new generic technique should be developed that can identify not only existing P2P applications, but any new P2P application which emerges in the future. This requires detailed knowledge of already existing techniques and their loopholes.

References

1. Hurley J, Garcia-Palacios E, Sezer S (2009) Classification of P2P and HTTP using specific protocol characteristics. Lecture notes in The Internet of the Future, Proceedings of 15th Open European Summer School and IFIP TC6.6 Workshop, EUNICE 2009, Barcelona, Spain, vol. 5733, pp 31–40
2. Mohammadi M, Raahemi B, Akbari A, Moeinzadeh H, Nasershari B (2011) Genetic-based minimum classification error mapping for accurate identifying Peer-to-Peer applications in the internet traffic. Expert Syst Appl: Int J 38(6):6417–6423
3. Sen S, Wang J (2004) Analyzing peer-to-peer traffic across large networks. IEEE/ACM Trans Networking 12(2):219–232
4. Dai L, Yang J, Lin L (2010) A comprehensive system for P2P classification. In: 2nd IEEE international conference on network infrastructure and digital content, pp 561–563

5. Chu H, Yi H, Zhang H (2011) A new P2P traffic identification methodology based on flow statistics. In: 3rd IEEE international conference on communication software and networks (ICCSN 2011), pp 277–281
6. Keralapura R, Nucci A, Chuah C-N (2010) A novel self-learning architecture for p2p traffic classification in high speed networks. Comput Netw 54(7):1055–1068
7. Azzouna NB, Guillemin F (2003) Analysis of ADSL traffic on an IP backbone link. In: IEEE Global Telecommunication Conference (GLOBECOM'03), vol. 7, pp. 3742–3746
8. Schulze H, Mochalski K (2007) Internet study 2007. Tech. report, ipoque
9. Karagiannis T, Broido A, Brownlee N, Claffy KC, Faloutsos M (2004) File-sharing in the Internet: a characterization of P2P traffic in the backbone. Tech. report
10. Madhukar A, Williamson C (2006) A longitudinal study of P2P traffic classification. In: Proceedings of 14th IEEE international symposium on modeling, analysis, and simulation of computer and telecommunication systems, Washington, DC, USA, pp. 179–188
11. Nguyen TTT, Armitage G (2009) A survey of techniques for internet traffic classification using machine learning. In: IEEE Communications surveys and tutorials, vol. 10, no.4
12. Callado A, Kamienski C, Szabo G, Gero B, Kelner J, Fernandes S, Sadok D (2009) A survey on internet traffic identification. IEEE Commun Surv Tutorials 11(3):37–52
13. Li W, Canini M, Moore AW, Bolla R (2009) Efficient application identification and the temporal and spatial stability of classification schema. Comput Netw 53(6):790–809
14. Williamson C (2001) Internet traffic measurement. IEEE Internet Comput 5(3):70–74
15. McGregor T (2002) Quality in measurement: beyond the deployment barrier. In: Proceedings of the symposium on applications and the internet workshops (SAINT), IEEE Computer Society, pp 66–73
16. Paxson V (2004) Strategies for sound Internet measurement. In: Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC 2004), NY, USA, pp 263–271
17. Enterprise network monitoring tools – network security system – application performance monitoring. http://www.endace.com
18. IPOQUE (2015) Bandwidth management with deep packet inspection. http://www.ipoque.com
19. WildPackets: Network analyzer, voip monitoring, protocol analysis. http://www.wildpackets.com
20. Intelligent real-time network analysis. http://www.napatech.com
21. SNORT. http://www.snort.org
22. Bro intrusion detection system. http://bro-ids.org
23. Wireshark, go deep. http://www.wireshark.org
24. ETTERCAP. http://ettercap.sourceforge.net
25. Claffy KC, McCreary S (1999) Internet measurement and data analysis: passive and active measurement. In: American Statistical Association
26. Duffield NG (2004) Sampling for passive internet measurement: a review. Stat Sci 19(3):472–498
27. Duffield N, LUND C, Thorup M (2005) Estimating flow distributions from sampled flow statistics. IEEE/ACM Trans Netw 13(5): 933–946
28. Claffy KC, Braun H-W, Polyzos GC (1995) A parameterizable methodology for Internet traffic flow profiling. IEEE J Sel Areas Commun 13(8):1481–1494
29. Apisdorf J, Claffy KC, Thompson K, Wilder R (1996) OC3MON: flexible, affordable, high performance statistics collection. In: Proceedings of the 10th USENIX conference on systems administration (LISA 1996), USENIX Association, Berkeley, CA, USA, pp. 97–112
30. Moore D, Keys K, Koga R, Lagache E, Claffy KC (2001) The CoralReef software suite as a tool for system and network administrators. In: Proceedings of the 15th USENIX conference on system administration (LISA 2001), USENIX Association, Berkeley, CA, USA, pp 133–144
31. CISCO NETFLOW. http://www.cisco.com/web/go/netflow
32. IETF (2008) Specification of the IP flow information export (IPFIX) protocol for the exchange of IP traffic flow information. In: RFC 5101
33. Allman M, Paxson V (2007) Issues and etiquette concerning use of shared measurement data. In: Proceedings of the 7th ACM SIGCOMM conference on Internet Measurement (IMC 2007), ACM, New York, NY, USA, pp 135–140
34. TCPDUMP/LIBPCAP public repository. http://www.tcpdump.org
35. WINDUMP. tcpdump forWindows usingWinPcap. http://www.winpcap.org/windump
36. Jurga RE, Hulbój MM (2007) Packet sampling for network monitoring. Technical report, CERN - HP Procurve openlab project
37. Sperotto A, Sadre R, Vliet F, Pras A (2009) A labeled data set for flow-based intrusion detection. In: Proceedings of the 9th IEEE international workshop on ip operations and management(IPOM 2009) (Venice, Italy, Oct.), LNCS Series, Springer-Verlag, Berlin Heidelberg, vol. 5843, pp 39–50
38. Zuev D, Moore AW (2005) Traffic classification using a statistical approach. In: Proceedings of the passive and active measurement conference (PAM 2005), LNCS Series, Springer-Verlag, Berlin Heidelberg, vol. 3431, pp 321–324
39. Karagiannis T, Papagiannaki K, Faloutsos M (2005) BLINC: multilevel traffic classification in the dark. In: Proceedings of the ACM SIGCOMM conferenece. on applications, technologies, architectures, and protocols for computer communications, ACM, New York, NY, USA, vol. 35, no. 4, pp 229–240
40. Salgarelli L, Gringoli F, Karagiannis T (2007) Comparing traffic classifiers. ACM SIGCOMM Comput Commun Rev 37(3):65–68
41. Canini M, Li W, Moore AW, Bolla R (2009) GTVS: boosting the collection of application traffic ground truth. In: Proceedings of the 1st international workshop on traffic monitoring and analysis (TMA'09) (Aachen, Germany), Springer Verlag, Heidelberg, Germany, pp 54–63
42. Gringoli F, Salgarelli L, Dusi M, Cascarano N, Risso F, Claffy KC (2009) GT: picking up the truth from the ground for Internet traffic. ACM SIGCOMM Comput Commun Rev 39(5):13–18
43. Szabó G, Orincsay D, Malomsoky S, Szabó I (2008) On the validation of traffic classification algorithms. In: Proceedings of the passive and active measurement conference (PAM 2008) (Cleveland, OH, USA), LNCS Series, Springer-Verlag, Berlin Heidelberg, vol. 4979, pp 72–81
44. Makhoul J, Kubala F, Schwartz R, Weischedel R (1999) Performance measures for information extraction. In: Proceedings of the DARPA Broadcast News Workshop (Herndon, VA, USA), pp 249–252
45. Olson DL, Delen D (2008) Advanced data mining techniques, 1st edition, Springer
46. Wang Y (2008) Statistical techniques for network security: modern statistically-based intrusion detection and protection. In: Premier Reference Source, Information Science Reference
47. Raahemi B, Zhong W, Liu J (2008) Peer-to-peer traffic identification by mining IP layer data streams using concept-adapting very fast decision tree. In: Proceedings of the 20th IEEE international conference on tools with artificial intelligence (ICTAI'08), IEEE, vol. 1, pp. 525–532
48. Internet Assigned Numbers Authority (IANA), http://www.iana.org/assignments/port-numbers
49. Gomes JV, Inacio PRM, Pereira M, Freire MM, Monteiro PP (2013) Detection and classification of peer-to-peer traffic: A

survey. In: ACM Computing Surveys (CSUR), NY, USA, vol. 45, no. 3
50. Moore AW, Papagiannaki K (2005) Toward the accurate identification of network applications. In: Proceedings of the 6th international conference on Passive and Active Network Measurement (PAM 2005), pp 41–54
51. Karagiannis T, Broido A, Brownlee N, Claffy KC, Faloutsos M (2004) Is P2P dying or just hiding?. In: Proceedings of the IEEE global telecommunications conference (GLOBECOM'04), vol. 3, pp. 1532–1538
52. Roughan M, Sen S, Spatscheck O, Duffield N (2004) Class-of-service mapping for QoS: A statistical signature-based approach to IP traffic classification. In: Proceedings of the 4th ACM SIGCOMM conference on Internet Measurement (IMC 2004), ACM, New York, NY, USA, pp 135–148
53. Sen S, Spatscheck O, Wang D (2004) Accurate, scalable in network identification of P2P traffic using application signatures. In: Proceedings of the 13th international conference on World Wide Web (WWW 2004), ACM, New York, NY, USA, pp 512–521
54. Karagiannis T, Broido A, Faloutsos M, Claffy KC (2004) Transport layer identification of P2P traffic. In: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, (IMC 2004), pp 121–134
55. Wang K, Stolfo SJ (2004) Anomalous payload-based network intrusion detection. In: Lecture Notes in Computer Science, Springer, Berlin, vol. 3224, pp 203–222
56. Song T, Zhou Z (2013) File aware P2P traffic classification: an aid to network management. J Peer-to-Peer Netw Appl (Springer) 6(3):325–339
57. Turkett WH, Karode AV, Fulp EW (2008) In-the-dark network traffic classification using support vector machines. In: Proceedings of the 20th National Conference on innovative applications of artificial intelligence (IAAI 2008), AAAI Press, pp 1745–1750
58. Freire EP, Ziviani A, Salles RM (2008) Detecting Skype flows in web traffic. In: Proceedings of the IEEE network operations and management symposium (NOMS 2008), IEEE, pp. 89–96
59. Freire EP, Ziviani A, Salles RM (2008) Detecting VoIP calls hidden in web traffic. IEEE Trans Netw Serv Manag 5(4):204–214
60. Gomes JVP, Inácio PRM, Freire MM, Pereira M, Monteiro PP (2008) Analysis of peer-to-peer traffic using a behavioural method based on entropy. In: Proceedings of the 27th IEEE International Performance Computing and Communications Conference (IPCCC 2008), IEEE Computer Society Press, Austin, Texas, pp. 201–208
61. Sun M, Chen J (2011) Research of the traffic characteristics for the real time online traffic classification. J China Univ Posts Telecommun (Elsevier) 18(3):92–98
62. Moore AW, Zuev D (2006) Discriminators for use in flow-based classification. In: Proceedings of the 20th BCS HCI Group Conference (HCI'06), London, UK, Sep 11–15
63. Bernaille L, Teixeira R, Akodkenou I, Soule A, Salamatian K (2006) Traffic classification on the fly. ACM SIGCOMM Comput Commun Rev 36(2):23–26
64. He J, Yang Y, Qiao Y, Deng W (2015) Fine-grained P2P traffic classification by simply counting flows. Front Inf Technol Electron Eng 16(5):391–403
65. Yang Kai, Wang B, Zhang Z (2013) A method of identifying P2P live streaming based on union features. In 4th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, pp. 426–429
66. Qin T, Wang L, Zhao D, Zhu M (2015) CUFTI: methods for core users finding and traffic identification in P2P systems. J Peer-to-Peer Netw Appl (Springer) 9(2):424–435
67. Zhang Q, Ma Y, Zhang P, Wang J, Li X (2004) Netflow Based P2P detection in UDP traffic. In IEEE 5th International Conference on Intelligent Control and Information Processing (ICICIP), Dalian, pp. 250–254
68. Perényi M, Dang TD, Gefferth A, Molnár S (2006) Identification and analysis of peer-to-peer traffic. J Commun 1(7):36–46
69. John W, Tafvelin S (2008) Heuristics to classify Internet backbone traffic based on connection patterns. In: Proceedings of the international conference on information networking (ICOIN 2008), IEEE, pp. 1–5
70. Hong W (2011) A novel method for P2P traffic identification. In: Procedia Engineering (Elsevier), vol. 23, pp. 204–209
71. Reddy JM, Hota C (2015) Heuristic-based real-time P2P traffic identification. In IEEE international conference on emerging information technology and engineering solutions (EITES), Pune, pp. 38–43
72. Bashir A, Huang C, Nandy B, Seddigh N (2013) Classifying P2P activity in netflow records: a case study on BitTorrent. In IEEE International Conference on Communications (ICC), Budapest, pp. 3018–3023
73. Mcgregor A, Hall M, Lorier P, Brunskill J (2004) Flow clustering using machine learning techniques. In: Proceedings of the passive and active measurement workshop (PAM 2004) (Antibes Juanles-Pins, France). LNCS Series, Springer-Verlag, Berlin Heidelberg, vol. 3015, pp 205–214
74. Moore AW, Zuev D (2005) Internet traffic classification using bayesian analysis techniques. ACM SIGMETRICS Perform Eval Rev 33(1):50–60
75. Branch PA, Heyde A, Armitage GJ (2009) Rapid identification of Skype traffic flows. In: Proceedings of the 18th international workshop on network and operating system support for digital audio and video (NOSSDAV'09), ACM, NY, USA, pp 91–96
76. Schmidt SEG, Soysal M (2006) An intrusion detection based approach for the scalable detection of P2P traffic in the national academic backbone network. In: Proceedings of the International Symposium on Computer Networks (ISCN 2006), IEEE, pp. 128–133
77. Cao J, Chen A, Widjaja I, Zhou N (2008) Online identification of applications using statistical behavior analysis. In: Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM 2008), IEEE, pp. 1–6
78. Angevine D, Zincir-heywood AN (2008) A preliminary investigation of Skype traffic classification using a minimalist feature set. In: Proceedings of the 3rd International Conference on Availability, Reliability and Security (ARES 08), IEEE Computer Society Press, pp. 1075–1079
79. Wang Y-H, Gau V, Bosaw T, Hwang J-N, Lippman A, Liebennan D, Wu I-C (2008) Generalization performance analysis of flow-based peer-to-peer traffic identification. In: Proceedings of the IEEE Workshop on Machine Learning for Signal Processing (MLSP 2008), IEEE, pp 267–272
80. Dainotti A, de Donato W, Pescape A, Rossi PS (2008) Classification of network traffic via packet-level hidden markov models. In: Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM 2008), IEEE, pp. 1–5
81. Valenti S, Rossi D, Meo M, Mellia M, Bermolen P (2009) Accurate, fine-grained classification of P2P-TV applications by simply counting packets. In: Proceedings of the 1st international workshop on traffic monitoring and analysis (TMA'09), LNCS Series, vol. 5537. Springer-Verlag, Berlin, Heidelberg, pp. 84–92
82. Liu H, Feng W, Huang Y, Li X (2007) A peer-to-peer traffic identification method using machine learning. In: Proceedings of the international conference on networking, architecture, and storage (NAS 2007), IEEE, pp. 155–160
83. Raahemi B, Kouznetsov A, Hayajneh A, Rabinovitch P (2008) Classification of peer-to-peer traffic using incremental neural networks (fuzzy ARTMAP). In: Proceedings of the Canadian

conference on electrical and computer engineering (CCECE 2008), IEEE, pp. 719–724
84. Hu Y, Chiu D-M, Lui JCS (2008) Application identification based on network behavioral profiles. In: Proceedings of the 16th International Workshop on Quality of Service (IWQoS 2008), IEEE, pp. 219–228
85. Hu Y, Chiu D-M, Lui JCS (2009) Profiling and identification of P2P traffic. Comput Netw 53(6):849–863
86. Liu S-M, Sun Z-X (2014) Active learning for P2P traffic identification. In: Journal of Peer-to-Peer Networking and Applications (Springer)
87. Moore AW, Zuev D (2005) Internet traffic classification using Bayesian analysis techniques. In: Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, ACM, NY, USA, pp. 50–60
88. Jiang D, Tao L (2013) P2P traffic identification research based on the SVM. In: 22nd Wireless and Optical Communication Conference (WOCC, 2013), IEEE, Chongqing, China, pp. 683–686
89. Gong J, Wang W, Wang P, Sun Z (2014) P2P traffic identification method based on an improvement incremental SVM learning algorithm. In: Proceeding of international symposium on wireless personal multimedia communications (WPMC 2014), IEEE, Sydney, NSW, pp. 174–179
90. Deng S, Luo J, Liu Y, Wang X, Yang J (2014) Ensemble learning model for P2P traffic identification. In: 11th international conference on fuzzy systems and knowledge discovery (FSKD 2014), IEEE, Xiamen, pp. 436–440
91. Jie H, Yuexiang Y, Yong Q, Chuan T (2013) Accurate classification of P2P traffic by clustering flows. Commun China IEEE 10(11):42–51
92. Bozdogan C, Gokcen Y, Zincir I (2015) A preliminary investigation on the identification of peer to peer network applications. In Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation, ACM, NY, USA, pp. 883–888
93. Dedinski I, Meer HD, Han L, Mathy L, Pezaros DP, Sventek JS, Xiaoying Z (2005) Cross-layer peer-to-peer traffic identification and optimization based on active networking. In: Proceedings of the 7th annual international working conference on active and programmable networks (IWAN 2005) (Sophia Antipolis, France, Nov.). Springer-Verlag, Berlin Heidelberg, pp. 13–27
94. Adami D, Callegari C, Giordano S, Pagano M, Pepe T (2009) A real-time algorithm for Skype traffic detection and classification. In: Proceedings of the 9th international conference on next generation wired/wireless networking (NEW2AN'09) (St. Petersburg, Russia, Sept.), LNCS Series, vol. 5764. Springer-Verlag, Berlin Heidelberg, pp. 168–179
95. Yan J, Wu Z, Luo H, Zhang S (2013) P2P traffic identification based on host and flow behaviour characteristics. Cybern Inf Technol 13(3):64–76
96. Ye W, Cho K (2014) Hybrid P2P traffic classification with heuristic rules and machine learning. J Soft Comput, Springer, Berlin Heidelberg 18(9):1815–1827
97. Ye W, Cho K (2015) P2P and P2P botnet traffic classification in two stages. In Journal of Soft Computing, Springer Berlin Heidelberg, pp. 1–12
98. Wang D, Zhang L, Yuan Z, Xue Y, Dong Y (2014) Characterizing application behaviors for classifying P2P traffic. In: International Conference on Computing, Networking and Communications (ICNC 2014), IEEE, Honolulu, HI, pp. 21–25
99. Yang Z, Li L, Ji Q, Zhu Y (2012) Cocktail method for BitTorrent traffic identification in real time. J Comput 7(1):85–95
100. Korczynski M, Duda A (2014) Markov chain fingerprinting to classify encrypted traffic. In: Proceedings of 2014 IEEE, INFOCOM, Toronto, pp. 781–789
101. Alshammari R, Zincir-Heywood AN (2015) Identification of VoIP encrypted traffic using a machine learning approach. J King Saud Univ Comput Inf Sci NY, USA 27(1):77–92
102. Kumano Y, Ata S, Nakamura N, Nakahira Y, Oka I (2014) Towards real-time processing for application identification of encrypted traffic. In: International conference on computing, networking and communications (ICNC), Honolulu, pp. 136–140
103. Wang X, Yang Y, He J (2014) Identifying P2P network activities on encrypted traffic. In: 13th IEEE international conference on trust, security and privacy in computing and communications (TrustCom), Beijing, pp. 893–899
104. Du Y, Zhang R (2013) Design of a method for encrypted P2P traffic identification using K-means algorithm. J Telecommun Syst (Springer) 53(1):163–168
105. Datta J, Kataria N, Hubballi N (2015) Network traffic classification in encrypted environment: a case study of google hangout. In 21st IEEE National Conference on Communications (NCC), Mumbai, pp. 1–6
106. Saroiu S, Gummadi KP, Dunn RJ, Gribble SD, Levy HM (2002) An analysis of Internet content delivery systems. In: Proceedings of the 5th symposium on operating systems design and implementation (OSDI'02), ACM, New York, NY, USA, vol. 36, pp 315–327
107. Leibowitz N, Bergman A, Ben-shaul R, Shavit A (2002) Are file swapping networks cacheable? Characterizing P2P traffic. In: Proceedings of the 7th international workshop on web content caching and distribution (WCW)
108. Gerber A, Houle J, Nguyen H, Roughan M, Sen S (2003) P2P, the gorilla in the cable. In: Proceedings of the national cable & telecommunications association (NCTA), pp 8–11
109. Spognardi A, Lucarelli A, Pietro RD (2005) A methodology for P2P file-sharing traffic detection. In: Proceedings of the 2nd international workshop on hot topics in peer-to-peer systems (HOT-P2P'05), IEEE Computer Society, Washington, DC, USA, pp. 52–61
110. Bin L, Zhi-Tang L, Hao T (2007) A methodology for P2P traffic measurement using application signature work-in-progress. In: Proceedings of the 2nd international conference on scalable information systems (InfoScale'07), ICST, Brussels, Belgium
111. Dewes C, Wichmann A, Feldmann A (2003) An analysis of Internet chat systems. In: Proceedings of the ACM SIGCOMM internet measurement conference (IMC 2003), ACM, New York, NY, USA, pp. 51–64
112. Guo Z, Qiu Z (2008) Identification peer-to-peer traffic for high speed networks using packet sampling and application signatures. In: Proceedings of the 9th international conference on signal processing (ICSP 2008), IEEE, pp. 2013–2019
113. Cascarano N, Ciminiera L, Risso F (2010) Improving cost and accuracy of DPI traffic classifiers. In: Proceedings of the 2010 ACM symposium on applied computing (SAC 2010), ACM, New York, NY, USA, pp. 641–646
114. Carvalho DA, Pereira M, Freire MM (2009) Towards the detection of encrypted BitTorrent traffic through deep packet inspection. In: Proceedings of the international conference on security technology (SecTech 2009), communications in computer and information science series, Springer-Verlag, Berlin Heidelberg, vol. 58, pp. 265–272
115. Carvalho DA, Pereira M, Freire MM (2009) Detection of peer-to-peer TV traffic through deep packet inspection. In: Acta da 9a Conference sobre Redes de Computadores (Oeiras, Portugal, Oct.). INESC-ID and Instituto Superior Técnico, 6
116. Freire MM, Carvalho DA, Pereira M (2009) Detection of encrypted traffic in eDonkey network through application signatures. In: Proceedings of the 1st international conference on advances in P2P systems (AP2PS 2009), IEEE Computer Society Press, Los Alamitos, CA, USA, pp.174–179
117. Park B.-C, Won YJ, Kim M.-S, Hong JW (2008) Towards automated application signature generation for traffic identification. In: Proceedings of the IEEE/IFIP Network Operations and Management Symposium (NOMS 2008), IEEE, pp. 160–167

Max Bhatia has received his Master of Engineering degree in CSE from Panjab University, Chandigarh, India. He is working as an Assistant Professor in Lovely Professional University. His research interest includes network traffic engineering and network security.

Mritunjay Kumar Rai received his Doctorate degree from ABV-Indian Institute of Information Technology and Management, Gwalior, India, after the completion the Master of Engineering degree in Digital system from Motilal Nehru National Institute of Technology, Allahabad, India. Presently he is working as an Associate Professor in Lovely Professional University, Phagwara, India. He has more than 10 years experience in teaching and research. His research interest includes wireless networks, computer networks and Network Security. He has published more than 35 research articles in reputed International Conferences and International Journals.