Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 19(3): 2996–3021.
DOI: 10.3934/mbe.2022138
Received: 30 September 2021
Revised: 17 December 2021
Accepted: 27 December 2021
Published: 17 January 2022

Survey

A Survey of techniques for fine-grained web traffic identification and classification

Xiaolin Gui[1], Yuanlong Cao[1,*], Ilsun You[2,*], Lejun Ji[1], Yong Luo[1] and Zhenzhen Luo[1]

[1] School of software, Jiangxi Normal University, Nanchang 330027, China

[2] Department of Information Security Engineering, Soonchunhyang University, Asan 31538, South Korea

* **Correspondence:** Email: ylcao@jxnu.edu.cn, ilsun@gmail.com.

Abstract: After decades of rapid development, the scale and complexity of modern networks have far exceed our expectations. In many conditions, traditional traffic identification methods cannot meet the demand of modern networks. Recently, fine-grained network traffic identification has been proved to be an effective solution for managing network resources. There is a massive increase in the use of fine-grained network traffic identification in the communications industry. In this article, we propose a comprehensive overview of fine-grained network traffic identification. Then, we conduct a detailed literature review on fine-grained network traffic identification from three perspectives: wired network, mobile network, and malware traffic identification. Finally, we also draw the conclusion on the challenges of fine-grained network traffic identification and future research prospects.

Keywords: web traffic; traffic identification; traffic classification; fine-grained traffic identification

1. Introduction

1.1. Background

Web traffic refers to the network traffic transmitted through hypertext transfer protocol (HTTP) or HTTP over secure socket layer (HTTPS). For the advantages of high flexibility and strong expressiveness, HTTP/HTTPS protocol has become the main protocol for new emerging websites or applications. As a result, web traffic has been the main traffic in the Internet since the mid-1990s [1]. Although the proportion of web traffic was exceeded by P2P traffic at the beginning of this century [2, 3]. However, web traffic surpassed P2P traffic quickly and contributing more than half of Internet traffic [4–6] for the following two reasons. (1) The P2P protocol strictly controlled by network operators due to the excessive consumption of network resources; (2) The rapid rise of rich media web sites represented by Youtube and Flicker leads to web traffic to grow continuous. In addition, most mobile Internet applications built on the web framework for its fast development and cross-platform features, which result in the ratio of web traffic in some mobile networks achieved 90% [7]. With the increase of smart mobile terminals, the growth rate of traffic in mobile internet far exceeds the traditional wire networks [8], which will also promote the proportion of web traffic increasing continuously. As web traffic accounts for a large proportion in various network, dividing web traffic into one category cannot meet the needs of network management. Therefore, it is urgent to identify web traffic in fine-grained with various applications including traffic engineering, billing, service recommendation, network planning, and optimizing. In this article, we provide a comprehensive overview of fine-grained network traffic identification. Then, we conduct a detailed literature review on fine-grained network traffic identification from three perspectives: wired network, mobile network, and malware traffic identification.

In recent years, fine-grained web traffic identification has attracted more and more attention from researchers, and the related results have also emerged continuous [8–19]. This article is a summary of fine-grained web traffic identification in the past few years. We summarize the rules obtained and the challenges faced.

1.2. Related survey

A number of surveys on coarse-grained or fine-grained network traffic have been conducted. Previous study, including the works of T. T. T. Nguyen et al. [20] and A. Callado et al. [21], surveyed the field of network traffic classification in coarse-grained. With the widespread use of HTTP(s) protocol both in the wired and mobile network, there have emerged many research works concerning HTTP(s) traffic [22–33], P. Velan et al. [25] focused on the encrypted traffic classification and analysis, and D. Acarali et al. [26] studied the HTTP-based botnet traffic. These surveys studied network traffic in both coarse-grained and fine-grained. Besides, with the widespread use of encrypted techniques in the network, deep learning (DL) and deep reinforcement learning techniques (DRL) are used to identify network traffic in fine-grained. The works of G. Aceto et al. [32] and A. Shahraki et al. [33] surveyed the applying of DL and DRL techniques in the traffic engineering, respectively. Table 1 is a summary of existing related surveys.

1.3. Research scope and contributions

In this work, we conduct a comprehensive survey of fine-grained web traffic identification. First, we describe the scope of our concerned fine-grained web identification, and why the topic is to be chose. Second, we provide a comprehensive overview of the existing fine-grained web traffic identification is provided. Then, recent research of fine-grained traffic identification are systematically surveyed from three perspectives: wired network, mobile network, malware traffic. Finally, we conclude the existing challenges and future research perspectives for fine-grained web traffic identification. Our key contributions are as follows:

Table 1. Summary of existing related surveys.

Category	Year of Publication	Authors	Topic
Coarse-grained	2008	T. T. T. Nguyen et al. [20]	Coarse-grained internet traffic classification using ML techniques
	2009	A. Callado et al. [21]	Coarse-grained internet traffic classification based on Port/Payload
	2012	A. Dainotti et al. [22]	Coarse-grained Internet traffic identification using hybrid techniques
	2013	M. Finsterbusch et al. [23]	Coarse-grained internet traffic classification based on payload
	2015	D. Naboulsi et al. [24]	Coarse-grained mobile network traffic identification and analysis
Fine/Coarse-grained	2015	P. Velan et al. [25]	Fine-grained encrypted traffic classification and analysis
	2016	D. Acarali et al. [26]	Fine-grained web traffic identification
	2016	W. PAN et al. [27]	Fine-grained internet encrypted traffic identification
	2018	F. Pacheco et al. [28]	Coarse-grained network traffic classification by using ML methods
	2019	S. Rezaei et al. [29]	Coarse-grained mobile network traffic identification and classification
	2019	A. D'Alconzo et al. [30]	Coarse-grained network traffic
Fine-grained	2020	W. M. Shbair et al. [31]	Fine-grained web traffic identification approaches
	2021	G. Aceto et al. [32]	Fine-grained internet traffic identification and classification
	2021	A. Shahraki et al. [33]	Fine-grained internet traffic identification and classification

(i) A comprehensive overview of fine-grained web traffic identification.

(ii) A detailed literature review on wired networks, mobile network and malware traffic.

(iii) For the systematic survey, we conclude the existing challenges and future perspectives for the fine-grained web traffic identification. The conclusion can provide a lot of inspiration for the future researchers.

A comprehensive overview of network traffic identification is provided in section 2. Then, a detailed literature review of recent fine-grained web traffic identification is proposed in section 3 from three aspects: wired network, mobile network, and malware traffic. In section 4, the evaluation criterion for fine-grained identification is presented. Our insight into the challenges and future perspectives of fine-grained web traffic identification are presented in Section 5. In Section 6, a conclusion is drawn. Due to the large number of acronyms involved, Table 2 lists the abbreviations used in this paper.

Table 2. List of abbreviations (alphabetical order).

Abbreviation	Description	Abbreviation	Description
C4.5	An algorithm used to generate a decision tree	OS	Operate System
DL	Deep learning	P2P	Peer to peer
DRL	Deep Reinforcement Learning techniques	QoS	Quality of Service
DPI	Deep packet inspection	QUIC	Quick UDP Internet Connection
HTTP	Hypertext Transfer Protocol	SVM	Support Vector Machine
IP	Internet Protocol	SPDY	A TCP-based session layer protocol developed by Google
IMEI	International Mobile Equipment Identity	URL	Universal Resource Location
ML	Machine Leaning	UA	User Agent

2. Overview of fine-grained web traffic identification

2.1. Introduction of network traffic identification

In the early days of the Internet, there were not many services in the network, and each service was assigned a fixed port number by Internet Assigned Numbers Authority (IANA). Therefore, most of the traffic at this stage can be identified through port-number. With the development and popularization of the Internet, there are more and more services in the network. Emerging internet applications do not necessarily use the service ports recommended by the IANA organization, resulting in the gradual failure of the method of using ports alone for traffic identification [34]. In order to improve the accuracy of traffic identification, deep packet inspection has gradually become popular. Deep packet inspection identifies the flow by detecting the payload characteristics in the packet. Compared with the port number-based methods, the accuracy of identification is greatly improved [35]. With the development of network technology and the increasing attention paid to network security issues, more and more internet content providers use encryption protocols to communication,which resulting in a decrease in the accuracy of traditional port number and payload-based traffic identification methods [36]. At the same time, the research of using machine learning methods for traffic identification and classification has gradually increased [37–43]. Nowadays, deep learning algorithms have achieved important breakthroughs in the fields of image, speech, and text. More and more researchers use deep learning in the field of traffic classification [15, 44].

2.2. Introduction of web traffic identification

The traditional traffic identification method stays at dividing the web traffic into a relatively coarse-grained category without identifying the specific applications carried on it. In the modern network environment, the applications carried by web traffic have more than simple web browsing and include multiple type of applications such as multimedia playback, web mail, and file downloads. Therefore, researchers have begun to try to divide web traffic in a more fine-grained way through a variety of different methods. With reference to the classification method of traditional traffic identification, the related research on web traffic identification can be conclude from two aspects, identification based on statistical features and identification based on packet analysis. Drawing on the traditional idea of traffic identification based on statistical characteristics, some researchers regard HTTP session as a data stream similar to TCP connections, and extract statistical characteristics such as packet length, average arrival interval, and the number of HTTP requests, and use clustering, machine learning algorithms such as C4.5 and SVM to recognize web traffic [38–42]. For the web traffic identification method based on statistical characteristics does not rely on the understanding of the interactive content of web applications, and it is more adaptable to the identification of private protocols or encrypted traffic. However, due to the limitations of machine learning technology, web traffic can only be divided into rough categories, such as web video, web mail, file download, botnet, ordinary web browsing, etc., make it difficult to associate traffic to specific applications. Therefore, the current fine-grained identification of web traffic still mainly relies on methods based on packet analysis. These methods mainly use the host, URL, User Agent, ContentType, and other fields in the HTTP request and response, as well as the content to be transmitted, to identify web traffic [45–48]. Since the header fields of the request and response messages of the HTTP protocol are in readable text form, this provides a useful place for using keyword matching or text pattern matching to associate web traffic with known applications or service providers. To cope with the large number and frequent changes of web applications, some researchers have also proposed several methods for automatically extracting application fingerprints to improve the problem of relying too much on manual extraction of application features based on packet analysis and identification methods [49, 50].

2.3. *Steps and techniques for fine-grained web traffic identification*

The main task of fine-grained identification of web traffic is to determine the identification object and the type of identification according to the requirements. Then, the appropriate identification method is selected according to the identification requirements. The methods of fine-grained web traffic identification can be divided into 6 categories, (1) payload-based methods [8–10]; (2) payload randomness-based methods [12, 14]; (3) methods based on the distribution of data packets [31]; (4) methods based on machine learning [15, 39–44]; (5) methods based on host behavior [13]; (6) hybrid methods combining multiple strategies [48–50]. Figure 1 shows the framework of fine-grained web traffic identification.

The identification object of web traffic refers to the input form of content, including flow-level, packet-level, host-level and session-level web traffic. The corresponding identification object is determined according to the requirement of traffic identification. Among them, the host-level and session-level objects are the most widely used. The session level mainly focuses on the characteristics of the session and the arrival process, such as the large amount of data in response to the video request, and the transmission of multiple sessions for one request. The session-level characteristics include the number of session bytes and the duration of the session.The host-level mainly focuses on the connection mode between hosts, such as all traffic communicating with the host, or all traffic communicating with a certain IP and port of the host. The host-level characteristics include the degree of connection, the number of ports, and so on.The flow level mainly focuses on the characteristics of the flow and the arrival process. The IP flow can be divided into one-way flow and two-way flow according to the transmission direction. The packets of one-way flow come from the same direction; the two-way flow contains packets from two directions, and the connection may not end normally, such as flow timeout. Sometimes a bidirectional flow requires a complete connection between the two hosts from the beginning of the SYN packet to the end of the first FIN packet. Stream-level characteristics include stream duration, number of stream bytes, and so on. The packet level mainly focuses on the characteristics and arrival process of data packets. The packet-level features mainly include packet size distribution and packet arrival time interval distribution.

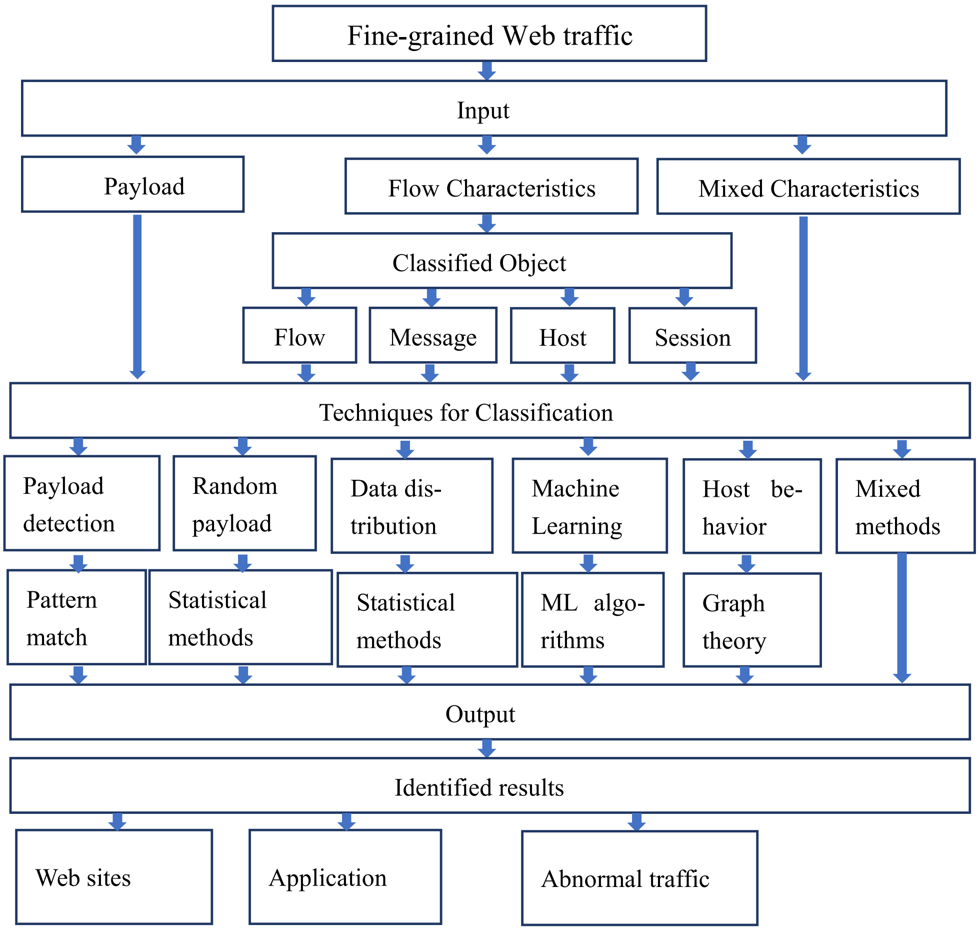

Figure 1. Steps and techniques for fine-grained web traffic identification.

The type of web traffic identification refers to the output form of the identification result. The identification type is determined according to the requirements of the flow identification. The web traffic can be gradually refined from the attributes of applications, websites, protocols, etc., and finally realize application identification, website identification, and abnormal traffic identification. We can describe the five kinds of identification mentioned above in detail as follows. (1) Application identification is to identify the application to which the traffic belongs, such as Google Mail, YouTube, etc. (2) Website identification is to identify the name of the website to which the traffic belongs. (3) Abnormal flow identification is to identify malicious traffic; (4) Encrypted and unencrypted traffic, identify which traffic is encrypted, and the rest is unencrypted. (5) Protocol identification is to identify the encryption protocol used for encrypted traffic, such as SSL, SSH, IPSec.

2.4. Problems of web traffic identification

In recent years, in the research of using the method of packet analysis to identify network traffic, we have found that there are a large number of web traffic in the network that have no clear meaning or semantically ambiguous in the HTTP header. These web traffic cannot use existing technology for identification. To illustrate this problem intuitively, we compare one type of unknown web traffic with the overall traffic. We discover that the web traffic with an unknown IP address in the host field accounted for 20.9% of the total, which is almost the same as the second-ranked business traffic. If we consider the unknown traffic with ambiguous fields such as Host and User Agent, the proportion of web traffic that cannot be processed by the existing recognition technology will be even higher. For these unknown flows, although we cannot identify them by means of message interpretation, we can speculate and identify them based on the relationship between web browsing records. The relationship between web browsing records is closely related to the behavior of users clicking on web pages. Therefore, identifying users' clicks behavior is of great help to the identification of unknown traffic. The latest research on user click recognition is [9, 10, 51]. In reference [9], the StreamStructure recognition method is proposed. This method combines the characteristics of time and file type, divides web browsing records into different blocks, and then determines the area of the first HTTP request of the block is the user click. Literature [26] proposed the ReSurf method. Compared with literature [9], this method has two main innovations. Firstly, it is proposed that the size of HTML documents is usually larger than V bytes. Secondly, after obtaining user access trajectories, backtrack the user's initial URL request in chronological order. The innovations of literature [51] can be concluded as follows: (1) unlike [9, 10] that separates user clicks from numerous URLs, literature [51] directly finds out the URLs clicked by users from referrers; (2) count the number of referrers from all records; Considering the complexity of web pages, a web page usually contains multiple embedded objects, and a URL clicked by a user should appear in the referrer of multiple records; (3) similar to [9], the URL is used The request file extension is to exclude those non-user clicking on the URL, such as: URL file extension is .js, .css, .swf. Fourth, because some web page advertisements are designed with an inline frame, the characteristics of the frame column object are very similar to user clicks, AdBlock library is used to filter non-user clicks on advertisement requests. Although the above studies have achieved satisfactory results in their respective network environments, they all have a common problem: these studies all use IP addresses to distinguish users. However, it is common for multiple users to share an IP address in a fixed network, so the research results have certain limitations. This article will use the data in the mobile Internet to study the user's click behavior. In the mobile Internet, each device has an IMEI number. We can separate the user's traffic according to the IMEI number, and correspond the flow records to the users one-to-one.

2.5. Introduction of network traffic datasets

There are more and more public available network traffic datasets with network security drawing more and more attention. Table 3 lists some common datasets in the field of network traffic classification and identification. UNIBS is the group of telecommunication networks from University of

Brescia. The traces from this group were mainly collected on the edge router of campus network of the University of Brescia. CIC is the group of Canadian Institute for Cybersecurity. The datasets from this group are mainly foucus on intrusion detection. UMass is a trace repository which maintained by the Laboratory for Advanced System Software, and everyone can contribute to the repository. CAIDA conducts network research and builds research infrastructure to support large-scale data collection, curation, and data distribution to the scientific research. WIDE is a traffic data repository maintained by the MAWI Working Group of the WIDE Project.

Table 3. Publicly available network traffic datasets.

Datasets	Description
UNIBS [52]	The traces were collected on the edge router of the campus network of the University of Brescia
CIC [53]	Canadian Institute for cybersecurity datasets
UMass [54]	The UMass Trace Repository is maintained by the Laboratory for Advanced System Software
CAIDA [55]	Containing various anonymized Internet Traces
WIDE [56]	This is a traffic data repository maintained by the MAWI Working Group of the WIDE Project

3. Review of recent fine-grained web traffic identification

3.1. Fine-grained web traffic identification in wired network

Wired network internet connectivity is a mature service in many countries. According to the different areas of wired network, it can be divided into residential broadband network, school network and work network. All these kinds of network have rich access which encourage users to closely involve network into their lives-from checking the weather or breaking news to shopping and banking or to communicating with family and friends in many aspects. However, the nature of these network differs from each other in use. For example, users of residential broadband connections will often have more entertainment needs than users of work environment, and school broadband connections will often have more study needs than other environment, and work broadband connections may have strict acceptable policies that may regulate their access at work, such as prohibitions against accessing certain web sites or employing certain applications. The identification of web traffic in wired networks is of great significance to network operators' management of the network. At present, there are mainly four types of methods for web traffic identification in wired networks: (1) Based on pattern matching; (2) Based on statistics; (3) Based on ML or DL; (4) Based on graph theory.

3.1.1. Pattern matching based methods in wired network

Internet content providers usually use 80, 8080, and 443 as the access ports for websites. In addition, the HTTP header has a unique fingerprint feature, which can be used as a basis for traffic identification. By using pattern matching method, Literature [57] is the first paper to study residential broadband network. The authors in this paper describe observations from monitoring the network activity for residential DSL customers in an urban area and reveals a number of surprising results, such as HTTP-not peer-to-peer-traffic dominates by a significant margin, more often than not the home user's immediate ISP connectivity contributes more to the round-trip times the user experiences than the WAN portion of the path, etc.

3.1.2. Statistics based methods in wired network

In the early stage of the research on fine-grained identification of web traffic, researchers usually described web traffic with the statistical characteristics of complete flow. Since these statistical features are based on the description of the complete flow, they can only be applied in offline recognition. Therefore, in recent years, the extraction of early features of traffic has become the focus of research. In a real application scenario, it is meaningful only if the identifier extracting its characteristics in the early stages of traffic occurrence. Bernaille et al. [58] pointed out that the first few data packets have a decisive significance in identifying the type of flow. Generally speaking, the first few data packets of the flow are communication the negotiation process between the two parties, and this negotiation process is completely determined by the application itself, which means that the first few data packets have the most obvious application-specific characteristics. This discovery provides a technical basis for online real-time identification of web traffic. Then, they continued to apply semi-supervised algorithms to the early recognition of Internet application traffic [59] and the early recognition of encrypted traffic [60] to do in-depth research. Este et al. [61] studied the early simple characteristics of traffic and found that the early characteristics of traffic contained rich information about application behavior characteristics. They applied mutual information and other methods to analyze the RTT, packet size, packet arrival time (IAT), and packet direction of several data packets in the early stages of the flow. The analysis results show that packet size is the most effective feature of early traffic. The research results provide experimental basis for the early feature extraction and recognition of traffic. Huang et al. also studied the early behavior characteristics of Internet applications, and conducted effective identification experiments based on these characteristics [62]. They further studied the conversation and negotiation behavior characteristics of different applications in the early stages of traffic. And based on these early characteristics, the ML model is used for early traffic recognition, and the ideal recognition effect is achieved [63]. Nguyen et al. [64] extended the concept of early recognition. They extracted statistical features from a small sequence of packets at any time, and applied C4.5 decision trees and naive bayes classifiers to online games and IP voice traffic and obtained a high recognition rate. He Gaofeng et al. [65] proposed an identification method based on TLS fingerprints and message length distribution, and successfully applied this identification method to online identification of Tor anonymous traffic. Chen Liang et al. [66] extracted features from NetFlow records and applied these features to realize high-speed flow identification. Dong Shi et al. [67] put forward an efficient flow identification model by studying the behavioral characteristics of traffic in ports, message lengths, and flow record preference. These research works are of great significance to the early and rapid identification of web traffic.

3.1.3. ML or DL based methods in wired network

In recent years, the flow identification method based on ML or DL has attracted more and more researchers' attention. These methods extract a series of independent statistical features of the payload from the network traffic, such as the number of packets, the amount of bytes carried by the packets, the duration of the flow, the average interval between packet arrivals. Then, researcher should uses ML's method to train a recognition model to perform the next step of traffic recognition. In these methods, network traffic is characterized by a series of traffic statistics features. Researcher obtains a recognition model by training some known application traffic data, which can then be used to identify unknown network traffic. From the perspective of data mining, ML can be divided into two class: supervised learning and unsupervised learning, which correspond to classification and clustering techniques respectively. For supervised learning, we need to provide a known data set for the target problem firstly, called the training data set. The function of the data set is to train a classification model, such as deep neural networks (DNN), support vector machine (SVM) and decision tree etc. The training process is general an iterative process. The parameters of the theoretical model are continuously adjusted through random optimization or analytical methods to make it as close as possible to the real situation of the training data set. After the model is trained, it can be used to identify unknown samples. This process is called testing or actual classification.

The feature description and extraction of traffic samples are the basic problems that need to be solved when using ML methods for traffic identification. At present, researchers use the statistical characteristics of the flow to formally describe the description of the flow sample. The effectiveness of this method is based on the following two assumptions: (1) The traffic of different applications has certain statistical characteristics at the network level, such as the duration of the flow, the idle time of the flow, the average between packets interval time, packet length; (2) The traffic characteristics of each application are unique, so it can be used to distinguish different network applications. In the mid-1990s, Paxson [69] used the statistical characteristics of streams to identify a series of TCP network applications.Then, Dewes et al. [69] analyzed the Internet chat system through a series of statistical characteristics including the duration of the stream, the average interval between groups, and the size of the group. A large number of subsequent studies [70,71] have shown that statistical characteristics were quite effective in the network traffic identification. Theoretically, although the statistical characteristics of streams can also be confused by disguise, it is very difficult in practice compared to technologies such as payload encryption. Literature [18] is a recent paper studying network traffic classification by using of deep convolutional recurrent autoencoder neural networks. The author find that the traffic classifier obtained by stacking the autoencoder with a fully-connected neural network, achieves up to a 28% improvement in average accuracy over state-of-the-art machine learning-based approaches.This is a huge improvement in the field of traffic classification.

3.1.4. Graph theory based methods in wired network

Traffic dispersion graph (TDG) is a common used method to represent network traffic. Each node in the graph is an IP address, and each edge represents a specific interaction between two nodes. In the early days when TDG was proposed, TDG was mainly used to solve network security problems, such as intrusion detection [72] and worm propagation [73, 74]. In reference [74], TDG is applied to the backbone network to study the interaction within the network. Its purpose is to automatically group and analyze network applications using information about the degree and port distribution of network applications. Besides, TDG could have a wider range of functions and could be directly applied to traffic classification.

3.2. Fine-grained web traffic classification in mobile network

In recent years,the rapid growth of smartphone users has led to the vigorous growth in the traffic volume of mobile networks. According to the prediction from Cisco, mobile data traffic will grow at compound annual growth rate of 46% from 2017 to 2022 [75]. Similar to wired networks, there are also various types of network traffic in mobile networks, including web traffic, P2P traffic, and network traffic based on other proprietary protocols. But research shows that web traffic is still the mainstream [12]. In some mobile networks,the ratio of web traffic even exceeds 90% [76]. In addition, new apps on mobile networks generally use HTTP to provide services to the public, further boosting the proportion of web traffic in mobile networks. Therefore, using traditional methods such as ports-based and payloads-based methods to identify web traffic can only identify web traffic in coarse-grained. In the condition of the web traffic accounts for up to 90% of total traffic, identifying web traffic as a type is disadvantage for network operators' management. Therefore, fine-grained web traffic identification is meaningful for operators to perform network management, including: traffic engineering, billing, service recommendation, network planning and optimization. There are mainly four types of methods for web traffic identification in mobile networks: (1) Based on pattern matching; (2) Based on statistics; (3) Based on ML or DL.

3.2.1. Pattern matching based methods in mobile network

There are three research directions of fine-grained HTTP traffic classification in mobile network by using pattern matching: (1) classify HTTP traffic into different applications (such as web browsing, E-mail and Stream) [8]; (2) associate HTTP traffic with a specific website [13, 77, 78]; (3) describe and model HTTP traffic in the dimensions of operating system and device [79]. The first paper on fine-grained HTTP traffic is [8]. This paper divides HTTP traffic into 14 categories in accordance with different application activities. Then, it brings several works to analyze HTTP traffic from different perspectives.The authors in [45] analyze the usage of HTTP-based applications on residential broadband Internet and find that the HTTP traffic dominates the whole downstream traffic. On the basis of the traffic similarity, the author proposed a classification scheme in [46], which can classify various traffic types in a single application. Reference [79] propose a detailed measurement study on the HTTP traffic characteristics of cellular network from the perspective of operating systems as well as device-types. These measurement study will helpful for network operators managing their network resources.

3.2.2. Statistics based methods in mobile network

Statistics based methods were widely used in web traffic classification. The authors in [77] studied the websites in the cellular data network and obtained nine different traffic distributions during the day. Reference [80] describes the mobile Internet traffic generated by multiple operating systems. However, this work only analyzes the traffic dynamics and application usage in one day, so it is impossible to find the characteristics in the billing cycle, which is very essential for billing and network planning. In [81], the authors describe and model Internet traffic dynamics from two aspects: device type and application. For different user markets: ordinary and commercial consumers, this approach is limited to coarse-grained description of these two types of smart phone devices. Reference [82] focuses on understanding how, where and when applications are used compared with traditional web services. However, the data sets used were collected in 2010, some conclusions may change now.

3.2.3. ML or DL based methods in mobile network

In 2016, Taylor et al. [83] proposed classification based on burst data streams, considering the two directions of data stream transmission (source and destination address swapping), respectively, count the packet size sequence of the stream, and calculate the average value for each sequence. There are 18 statistical features such as, minimum, maximum, quantile, etc. Finally, the support vector regression algorithm and random forest algorithm are used to achieve a classification accuracy of 99%. In 2019, Shen et al. [84] proposed a decentralized application recognition method, which proposed to use the kernel function for feature fusion based on the statistical characteristics of the two-way data stream, and then further feature screening, and finally achieved a classification accuracy of 92%. The main disadvantage of traffic classification methods based on machine learning is that they require expert experience to extract and filter features. Therefore, these methods are time-consuming and expensive, and are prone to human error. As a result, researchers gradually set their sights on deep learning that can learn features independently.

Traffic classification methods based on deep learning are divided into two categories: based on the original byte characteristics of the data packet and based on the sequence characteristics of the data packet in the flow. The method refers to classification based on the original byte characteristics of the data packet the input of the classifier is the original byte content of the data packet. The method based on the characteristics of the data packet sequence in the stream means that the input of the classifier is the data packet size in the stream, the packet time interval sequence and other characteristics. The DeepPacket proposed by Lotfollahi et al. [85] is a representative of the deep learning method based on the original byte characteristics of the data packet. It proposes to use each data packet as an input sample, and does not require expert experience to extract features, only the original bytes of the data packet As features, the classification model is a one-dimensional convolutional neural network (1DCNN) and a sparse automatic encoder (SAE), and finally achieved a classification accuracy of 98%. Wang et al. [3] proposed to use the first 784 bytes of each data stream (one-way stream/two-way stream) as the model input, based on one-dimensional convolutional neural network (1DCNN) and two-dimensional convolutional neural network (2DCNN), respectively Experiments with two models have shown that 1DCNN has a better effect and can reach an accuracy rate of more than 90%. Li et al. [86] introduced recurrent neural network (RNN) into network traffic classification, and designed a new neural network-byte segment neural network (BSNN). BSNN directly inputs data packets as a model. The experimental results show that in the process of classifying 5 protocols, the average F1-score of BSNN is about 95.82%. Xie et al. [87] proposed a flow classification method SAM based on a self-attention mechanism, using the original bytes of each packet header as the model input. This method achieved 98.62% and 98.62% in protocol recognition and application recognition, respectively. 98.93% F1-score average value. The FS-Net proposed by Liu et al. [88] is a representative of the deep learning method based on the characteristics of the packet sequence in the stream. The timing feature uses the packet size sequence in the stream. Based on this, an automatic encoder (auto-encoder) encoder) reconstruction mechanism, this reconstruction mechanism enables the model to learn the features that are most conducive to classification and the most representative of this data stream, and the final classification accuracy rate is as high as 99%. Lopez-Martin et al. [89] proposed to form a 20×6 matrix based on the port number, packet load length, packet interval time, window size and other attributes of the first 20 data packets in the data stream, and input it to the convolutional neural network (CNN) and The combined model of long and short-term memory recurrent neural network (LSTM) can achieve a final accuracy rate of over 96%. Shapira et al. [90] proposed to convert the data stream into pictures according to the packet size and packet arrival time of the one-way data stream, and then classify them through the CNN model. The final classification accuracy rate can reach 99.7%.

3.3. Malware web traffic identification

Similar to personal computers, the widespread use of mobile devices has aroused the interest of malware developers. Among the many mobile devices, smartphones are ideal targets for attackers because: (1) they are ubiquitous, that is, the number of potential targets is large; (2) they have sensitive information about the owner, such as identity, contact people, GPS location; (3) they have network capabilities, and they usually connect to the Internet. We define malware detection as trying to understand whether it is malicious by analyzing the network traffic generated by a mobile application. In mobile networks, the detection of malicious traffic usually detects apps.

In June 2004, the first known smartphone malware appeared in public view. It exists in Symbian OS [94], named cabir, and propagates through Bluetooth. In fact, between 2004 and 2007, more than 95% of malware came from Symbian OS [92]. Since then, there have been more and more Android and IOS devices. Therefore, malware against these operating systems is also emerging. In July 2008, the first autobiographical worm for iPhone was detected, named Ikee [93]. The worm only uses the installed SSH server and the default root password to attack the jail-broken iPhone, but the threat to users is low. Then, in 2009, a variant named Ikee.B was found. This is the first botnet with obvious malicious attacks. In 2010, the first malware against Android was discovered, named FakePlayer [94]. In fact, between 2012 and 2014, more than 90% of the malware detected were targeted at the Android platform [95]. As of March 31, 2015, nearly 4000 mobile families and variants have been identified for mobile devices [96]. Between 2012 and 2014, the number of malwares per quarter was about 200 [95].

Protecting intelligent devices from malware attacks is also a hot topic in research [97–99]. There are two main analysis directions for identifying malware: static analysis and dynamic analysis. Dynamic analysis refers to the technology of executing the sample software and verifying the behavior of the sample in practice, while static analysis will verify the software based on its source code rather than actually executing the sample. In fact, static analysis can only detect malware with unavailable signatures, which is invalid for polymorphic and deformed code. Literature [100] points out that the commonly used static analysis only detects 20.2% to 79.6% of malware. Dynamic analysis is very promising. It can use multiple behavioral characteristics to analyze samples. For example, Trojan horse always needs to call multiple system processes. Therefore, literature [101, 102] proposed a method to detect Trojan horse through system behavior analysis. In addition, traffic characteristics are very useful for identifying malware spreading through the network. For instance, according to the traffic characteristics, a model is constructed in literature [103] to identify fast traffic botnet attacks. Literature [104] uses traffic characteristics to detect a new class of active worms.

XcodeGhost's servers and clients communicate with each others via Internet. Therefore, there were a lot of XcodeGhost-related traffic in our collected data. This unique vantage makes our work distinguish with the works [105, 106]. We can gain some XcodeGhost features by analyzing the collected data. And it may be helpful to study XcodeGhost or to identify other malware like XcodeGhost.

Literature [107] introduced a malware detection application in the Android environment. This application can monitor multiple aspects of the device,such as memory, network, power and extract different characteristics, some of which are related to network traffic such as the number of packets received.

Then, training is performed based on the collected traffic statistical feature sample data to obtain a classifier. And use the obtained classifier to check whether the installed application is malicious. The article used 40 benign and 4 malicious Android applications to evaluate the model and achieved good results.

Besides, with the development of network theory and technology, some new ideas have emerged that can tolerate network attacks [108] or resist malicious attacks on network terminals from the scratch of network design [109, 110]. For example, the authors in [108] proposed the use of multiple paths to transmit data to avoid network attacks, and reference [109] proposed an smart collaborative balance scheme to dynamically adjust network functions. This scheme can effectively resist malicious attack from terminals.

4. Evaluation criteria for web traffic classification

At present, the evaluation of traffic identification and classification are mainly uses accuracy-related indicators. This indicator is relatively single. To meet the ever-increasing flow analysis requirements, on the basis of accuracy-related evaluation indicators, comprehensive indicators of compatibility, robustness, integrity, and directionality are introduced. The following is a detailed introduction to the evaluation indicators of network traffic identification and classification.

1) Accuracy

Accuracy reflects the ability of traffic identification technology to identify network applications. Assuming that N is the number of traffic samples, m is the number of application types, and n_{ij} represents the actual number of samples of type i applications marked as type j. True Positive (TP) represents the number of correctly labeled samples among the samples of the actual type i, $TP_i = n_{ii}$. False Positive (FP) represents the number of samples incorrectly identified as typeiamong samples whose actual type is not i, $FP_i = \sum n_{ji}$.

By using of all parameters mentioned above, confusion matrix is a more clear way to describe classification. It can tell us how the classification model is confused when it makes predictions. The confusion matrix includes TN, FP, FN and TP. When the classification problem is two classifications, the content of the confusion matrix is shown in Table 4.

Table 4. Confusion matrix.

Prediction Real	0	1
0	TN	FP
1	FN	TP

According to the above analysis and Table 4, the precision is defined as follows.

$$P = \frac{TP}{TP_i + FP_i} \tag{4.1}$$

False negative (FN) represents the number of samples whose actual type is i that are misidentified as other types. $FN_i = \sum n_{ij}$ True negative (TN) represents the number of samples marked as non-i among the samples whose actual type is non-i, $TN_i = n_{jj}$. The recall rate is defined as follow.

$$R = \frac{TP_i}{TP_i + FN_i} \tag{4.2}$$

Similarly, true negative rate (TNR) represents the ratio of negative outcomes that are actually predicted to be negative. This metric is also called specificity and is defined as follows.

$$TNR = \frac{TN_i}{FP_i + TN_i} \tag{4.3}$$

In Mathematics, the Geometric Mean is the average value or mean which signifies the central tendency of the set of numbers by finding the product of their values. In the field of network traffic classification, we use it to balance Sensitivity and Specificity at the same time. The definition of g-mean is shown in Eq (4.4).

$$g - mean = \sqrt{R * TNR} \tag{4.4}$$

The precision rate and recall rate reflect the recognition effect of the recognition method on each individual protocol category. Especially when the sample categories are unevenly distributed, recall and precision can accurately know the classification of each category. The accuracy rate reflects the overall recognition performance of the recognition method. A good algorithm should have a high accuracy rate, precision rate, and recall rate at the same time. The accuracy is defined as follow.

$$Acc = \frac{\sum_{i=1}^{m} (TP_i + TN_i)}{\sum_{i=1}^{m} (TP_i + TN_i + FP_i + FN_i)} \tag{4.5}$$

F-Measure is an evaluation index obtained by comprehensive precision and recall. The higher the F-Measure, the better the classification performance of the algorithm in each type.

$$F - Measure = \frac{2PR}{P + R} \tag{4.6}$$

Besides, top-*k* accuracy is an important evaluation index used to evaluate the classification accuracy of the k categories with the most number.

2) Completeness

The completeness reflects the recognition coverage of the recognition method. Completeness refers to the ratio of the sample identified as *i* to the sample of the actual type i, which is equivalent to the ratio of the precision rate to the recall rate, and the value range may exceed 1. Completeness is defined as follow.

$$completeness = \frac{R}{P} \tag{4.7}$$

3) Unrecognized rate

The unrecognized rate reflects the ability of the traffic identification tool to identify unknown traffic types. Unrecognized rate refers to the ratio of traffic that does not belong to a known traffic type to the total traffic. F_{total} represents the total number of bytes or streams of traffic, and F_{known} represents the number of bytes or streams of identified traffic.

$$unrecognized = \frac{F_{total} - F_{known}}{F_{total}} \tag{4.8}$$

4) Robustness

Robustness reflects the ability of traffic identification tools to maintain high identification performance for a long time. Robustness refers to the ability of the traffic recognition technology to maintain a high recognition rate for a long time. acc_k represents the accuracy rate of period k, acc_0 represents the initial accuracy.

$$robustness = \sqrt{\frac{\sum_{k=1}^{r} (acc_0 - acc_k)}{r}} \tag{4.9}$$

5) Compatibility

Compatibility reflects the ability of traffic identification tools to be used in different network environments. Compatibility indicates the ability of traffic identification technology to be used in different network environments. acc_j represents the accuracy in the network environment j, $\overline{acc}$ represents the average accuracy in all environments.

$$compatibility = \sqrt{\frac{\sum_{j=1}^{m} \left(acc_j - \overline{acc}\right)}{m}} \tag{4.10}$$

6) Evaluation index

In addition, there are still some problems in the quantification of some evaluation indicators, such as real-time, directional, and computational complexity. The real-time performance reflects the ability of the traffic identification method to identify network applications online and quickly. We can identify an application in time by use of the characteristics of some data packets rather than waiting for the end of the entire flow.

The directionality reflects the ability of the flow identification method to identify different flow transmission directions. IP flow can be divided into unidirectional flow and bidirectional flow. Unidirectional flow can be divided into upstream and downstream according to the transmission direction. If the first data packet is packet loss, it is impossible to judge the upstream and downstream directions. Directionality can be embodied in unidirectional flow (upstream, downstream) or bidirectional flow.

The computational complexity reflects the overhead required by the traffic identification method to accurately identify network applications. Complex identification features consume a lot of storage space and computing power, which seriously affects the traffic analysis of the backbone network. Computational complexity can be embodied in time and space complexity.

5. Discussion

In summary, there are still many problems to be solved in the field of traffic classification. In the future, scientific research can be carried out from the following aspects.

1) Fine-grained unknown web traffic identification

We discover that the web traffic with an unknown IP address in the host field accounted for 20.9% of the total traffic in the backbone. Identifying these unknown traffic is still a big challenge.

2) Fine-grained identification of encrypted traffic

With the increasing demand for fine-grained identification of traffic, it is far from enough to identify whether the traffic is encrypted. In the actual scenario of network management,netwok operators need to identify the applications or services under the encryption protocol or tunneling protocol. To achieve the goal of fine-grained recognition, multi-stage progressive fine-grained recognition and hybrid methods are better solutions. Each stage completes different identification tasks, or combines different algorithms to identify different applications.

3) Application recognition under SSL protocol

To ensure the security of communication, there are increasingly network applications using the SSL protocol. The SSL protocol is widely used in web browsing, watching videos, social networks, etc., so that the application based on the SSL protocol has become increasingly complex. The SSL protocol is impeccable in protecting user data and privacy. At the same time, the protocol also pushes the difficulty of traffic identification to a new level. How to identify network applications under the SSL protocol has become a challenge for current network management.

4) Encrypted video content information recognition

As video services become increasingly widely used and the proportion of video traffic continues to increase, network operators and video service providers need to know the current quality of video experience services to improve video QoS. As the most commonly used video website, YouTube uses encryption technology for more than 90% of traffic, and increasingly video websites use encryption technology. In the scenario of encryption, it is difficult to obtain the parameters related to the quality of the video experience service, such as the playback bit rate..Therefore, how to identify the bit rate and the encrypted video is of great significance for evaluating and improving QoS.

5) Accurate marking of encrypted traffic data sets

In recent years, some new algorithms and techniques with good classification performance have been proposed. However, these algorithms and techniques cannot be compared with each other for the collected network traffic is always different, most public data sets have no payload information and marking information, and even the payload of encrypted traffic is difficult to mark with DPI tools. Therefore, some researchers have to use common port numbers to add filtering rules for marking, which leads to inaccurate benchmarks. In addition, to meet the requirements of fine-grained identification of encrypted traffic, the key is to mark different applications running under the encryption protocol, making it more difficult to mark. The self-generated data set mainly adopts the method of monitoring the host kernel or the DPI method to obtain the labels. Although the self-generated data set is relatively easy to obtain the label information, the self-generated data set of each will cause the problem of incomparability between different algorithms. Therefore, it is urgent to build some labeled data sets for various traffic classification.

6) Traffic masquerading

The identification method based on flow characteristics is the most widely used approach for encrypted flow identification. Therefore, the corresponding flow pattern disguising techniques, such as flow filling, flow standardization, and flow masking, are constantly being studied. Wright [111] proposed a convex optimization method for real-time modification of data packets, disguising the packet size distribution of one traffic as the packet size distribution of another traffic, and the transformed traffic can effectively avoid traffic classification such as VoIP and web recognition. In the future, traffic masquerading technology will integrate multiple methods such as traffic filling, traffic standardization, and traffic masking to deal with traffic analysis, and the diversity and adaptive capabilities of traffic masquerading will be greatly enhanced. In addition, anonymous communication, tunneling technology, and proxy technology are all different manifestations of traffic masquerading. Anonymous communication prevents tracking by hiding the identity information and the communication relationship. Tunneling technology uses L2TP and SSTP. Data packets are re-encapsulated by other protocols, and the data compression proxy technology changes the flow statistics characteristics to save traffic. Therefore, it is necessary to improve the current identification methods to cope with the upcoming challenges.

7) New protocols and changes in traffic distribution

Due to the improvement and optimization of application protocols and the continuous development of new versions to hinder traffic identification, the protocol signatures and behavior characteristics are changed accordingly. Therefore, the original identification methods need to be updated periodically. As the public's demands for network security and network performance increase, new encryption protocols such as SPDY, HTTP/2.0, and QUIC are constantly being introduced to solve the bottleneck of TCP and UDP-based protocols, and achieve low latency, high reliability and security network communication. In the near future, HTTP/2.0 and QUIC protocols will be widely used, and how to identify the applications carried under the protocol faces new challenges.

In addition, the methods based on DL have been widely used in encrypted traffic classsification and have made a great progess [90,112,113]. To solve the expensive of the general DL model, The authors in [114] proposed a Incremental Learning techniques to add new classes to models without a full retraining,This techniques can save a lot of calculations as well as automatically adjust the model with the input of data. With the rise of the Internet of Things, the identification and classification of traffic in the Internet of Things is also an important research orientation in the future [115,116]. Additionally, another emerging trend in ML/Dl-based traffic classifiers is explainable AI. All these new emerging techniques will helpful to overcome the challenge of new protocols and changes in traffic distribution.

6. Conclusions

We present an up-to-date survey on fine-grained web traffic identification in this paper. A comprehensive overview of fine-grained web traffic identification is presented firstly. Then, we introduce the recent research work of fine-grained web traffic identification from three aspects: wired network, mobile network, and malware traffic identification. Finally, we conclude the challenges and future perspectives on the basis of our systematic survey. The detailed literature review and in-depth investigations may inspire more endeavour to further improve fine-grained web traffic identification.

Acknowledgement

This work was supported in part by the Natural Science Foundation of Jiangxi Province under Grant No. 20192ACBL21031, and in part by the Soonchunhyang University Research Fund.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. F. Hernández-Campos, K. Jeffay, F. D. Smith, Tracking the evolution of web traffic: 1995–2003, in *11th IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer Telecommunications Systems*, (2003), 16–25. https://doi.org/10.1109/MASCOT.2003.1240638
2. H. Schulze, K. Mochalski, Internet study 2008/2009, *Ipoque Rep.*, **37** (2009), 351–362.
3. T. Zimmermann, J. Rüth, B. Wolters, O. Hohlfeld, How HTTP/2 pushes the web: An empirical study of HTTP/2 server push, in *2017 IFIP Networking Conference (IFIP Networking) and Workshops*, (2017), 1–9. https://doi.org/10.23919/IFIPNetworking.2017.8264830
4. O. Hohlfeld, J. Rüth, K. Wolsing, T. Zimmermann, Characterizing a meta-CDN, in *International Conference on Passive and Active Network Measurement*, (2018), 114–128. https://doi.org/10.1007/978-3-319-76481-8_9
5. F. Lichtblau, F. Streibelt, T. Krüger, P. Richter, A. Feldmann, Detection, classification, and analysis of inter-domain traffic with spoofed source IP addresses, in *Proceedings of the 2017 Internet Measurement Conference*, (2017), 86–99. https://doi.org/10.1145/3131365.3131367
6. A. Al-Najjar, S. Teed, J. Indulska, M. Portmann, Flow-based load balancing of web traffic using OpenFlow, in *2017 27th International Telecommunication Networks and Applications Conference (ITNAC)*, (2017), 1–6. https://doi.org/10.1109/ATNAC.2017.8215411
7. Cisco, *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021*, 2017. Available from: https://www.ramonmillan.com.
8. W. Li, A. W. Moore, M. Canini, Classifying HTTP traffic in the new age, *ACM SIGCOMM*, **8** (2008), 17–22.
9. J. Liu, C. Fang, N. Ansari, Request dependency graph: A model for web usage mining in large-scale web of things, *IEEE Internet Things J.*, **3** (2016), 598–608. https://doi.org/10.1109/JIOT.2015.2452964
10. L. Vassio, I. Drago, M. Mellia, Detecting user actions from HTTP traces: toward an automatic approach, in *2016 International Wireless Communications and Mobile Computing Conference (IWCMC)*, (2016), 50–55. https://doi.org/10.1109/IWCMC.2016.7577032
11. G. Scavo, Z. B. Houidi, S. Traverso, R. Teixeira, M. Mellia, WeBrowse: mining HTTP logs online for network-based content recommendation, preprint, arXiv:1602.06678.
12. P. Fiadino, A. Bar, P. Casas, HTTPTag: a flexible on-line HTTP classification system for operational 3G networks, in *2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, (2013), 71–72. https://doi.org/10.1109/INFCOMW.2013.6970744
13. X. Gui, J. Liu, Q. Lv, C. Dong, Z. Lei, Probabilistic top-k query: model and application on web traffic analysis, *China Commun.*, **13** (2016), 123–137. https://doi.org/10.1109/CC.2016.7513208
14. J. Sun, L. She, H. Chen, W. Zhong, C. Chang, Z. Chen, et al., Automatically identifying apps in mobile traffic, *Concurrency Comput. Pract. Exper.*, **28** (2016), 3927–3941. https://doi.org/10.1002/cpe.3703
15. G. Aceto, D. Ciuonzo, A. Montieri, A. Pescapé, Mobile encrypted traffic classification using deep learning: experimental evaluation, lessons learned, and challenges, *IEEE Trans. Network Serv. Manage.*, **16** (2019), 445–458. https://doi.org/10.1109/TNSM.2019.2899085

16. P. Białczak, W. Mazurczyk, Characterizing anomalies in malware-generated HTTP traffic, *Secur. Commun. Networks*, **2020** (2020). https://doi.org/10.1155/2020/8848863

17. J. Li, H. Zhang, Z. Wei, The weighted word2vec paragraph vectors for anomaly detection over HTTP traffic, *IEEE Access*, **8** (2020), 141787–141798. https://doi.org/10.1109/ACCESS.2020.3013849

18. G. D'Angelo, F. Palmieri, Network traffic classification using deep convolutional recurrent autoencoder neural networks for spatial–temporal features extraction, *J. Network Comput. Appl.*, **173** (2021), 102890. https://doi.org/10.1016/j.jnca.2020.102890

19. S. Dong, Y. Xia, T. Peng, Traffic identification model based on generative adversarial deep convolutional network, *Ann. Telecommun.*, (2021), 1–15. https://doi.org/10.1007/s12243-021-00876-6

20. T. T. Nguyen, G. Armitage, A survey of techniques for internet traffic classification using machine learning, *IEEE Commun. Surv. Tutorials*, **10** (2008), 56–76. https://doi.org/10.1109/SURV.2008.080406

21. A. Callado, C. Kamienski, G. Szabó, B. P. Gero, J. Kelner, S. Fernandes, et al., A survey on internet traffic identification, *IEEE Commun. Surv. Tutorials*, **11** (2009), 37–52. https://doi.org/10.1109/SURV.2009.090304

22. A. Dainotti, A. Pescape, K. C. Claffy, Issues and future directions in traffic classification, *IEEE Network*, **26** (2012), 35–40. https://doi.org/10.1109/MNET.2012.6135854

23. M. Finsterbusch, C. Richter, E. Rocha, J. Muller, K. Hanssgen, A survey of payload-based traffic classification approaches, *IEEE Commun. Surv. Tutorials*, **16** (2013), 1135–1156. https://doi.org/10.1109/SURV.2013.100613.00161

24. D. Naboulsi, M. Fiore, S. Ribot, R. Stanica, Large-scale mobile traffic analysis: a survey, *IEEE Commun. Surv. Tutorials*, **18** (2015), 124–161. https://doi.org/10.1109/COMST.2015.2491361

25. P. Velan, M. Cermak, P. Celeda, M. Drasar, A survey of methods for encrypted traffic classification and analysis, *Int. J. Network Manage.*, **25** (2015), 355–374. https://doi.org/10.1002/nem.1901

26. D. Acarali, M. Rajarajan, N. Komninos, I. Herwono, Survey of approaches and features for the identification of HTTP-based botnet traffic, *J. Network Comput. Appl.*, **76** (2016), 1–15. https://doi.org/10.1016/j.jnca.2016.10.007

27. W. Pan, G. Cheng, X. Guo, S. Huang, Review and perspective on encrypted traffic identification research, *J. Commun.*, **37** (2016), 154–167. https://doi.org/10.11959/j.issn.1000-436x.2016187

28. F. Pacheco, E. Exposito, M. Gineste, C. Baudoin, J. Aguilar, Towards the deployment of machine learning solutions in network traffic classification: A systematic survey, *IEEE Commun. Surv. Tutorials*, **21** (2018), 1988–2014. https://doi.org/10.1109/COMST.2018.2883147

29. S. Rezaei, X. Liu, Deep learning for encrypted traffic classification: an overview, *IEEE Commun. Mag.*, **57** (2019), 76–81. https://doi.org/10.1109/MCOM.2019.1800819

30. A. D'Alconzo, I. Drago, A. Morichetta, M. Mellia, P. Casas, A survey on big data for network traffic monitoring and analysis, *IEEE Trans. Network Serv. Manage.*, **16** (2019), 800–813. https://doi.org/10.1109/TNSM.2019.2933358

31. W. M. Shbair, T. Cholez, J. François, I. Chrisment, A survey of HTTPS traffic and services identification approaches, preprint, arXiv:2008.08339.

32. G. Aceto, D. Ciuonzo, A. Montieri, A. Pescape, Toward effective mobile encrypted traffic classification through deep learning, *Neurocomputing*, **409** (2020), 306–315. https://doi.org/10.1016/j.neucom.2020.05.036

33. A. Shahraki, M. Abbasi, A. Taherkordi, A. D. Jurcut,. Active learning for network traffic classification: a technical study, preprint, arXiv:2106.06933.

34. S. Dong, R. Li, Traffic identification method based on multiple probabilistic neural network model, *Neural Comput. Appl.*, **31** (2019), 473–487. https://doi.org/10.1007/s00521-017-3081-x

35. H. Tang, Z. Li, Design and implementation of a DPI-Based P2P traffic control system, *Inf. Secur. Commun. Privacy*, **6** (2007).

36. M. Soysal, E. G. Schmidt, Machine learning algorithms for accurate flow-based network traffic classification: evaluation and comparison, *Perform. Eval.*, **67** (2010), 451–467. https://doi.org/10.1016/j.peva.2010.01.001

37. S. Dong, Multi class SVM algorithm with active learning for network traffic classification, *Expert Syst. Appl.*, **176** (2021), 114885. https://doi.org/10.1016/j.eswa.2021.114885

38. F. Haddadi, A. N. Zincir-Heywood, Benchmarking the effect of flow exporters and protocol filters on botnet traffic classification, *IEEE Syst. J.*, **10** (2016), 1390–1401. https://doi.org/10.1109/JSYST.2014.2364743

39. T. Bakhshi, B. Ghita, On internet traffic classification: A two-phased machine learning approach, *J. Comput. Networks Commun.*, **2016** (2016). https://doi.org/10.1155/2016/2048302

40. S. Dong, X. Zhang, D. Zhou, Auto adaptive identification algorithm based on network traffic flow, *Int. J. Comput. Commun. Control*, **9** (2014), 672–685. http://dx.doi.org/10.1145/1080091.1080119

41. Y. Dong, J. Zhao, J. Jin, Novel feature selection and classification of Internet video traffic based on a hierarchical scheme, *Comput. Networks*, **119** (2017), 102–111. https://doi.org/10.1016/j.comnet.2017.03.019

42. S. Dong, W. Liu, D. Zhou, Y. Qi, NSVM: A new SVM algorithm based on traffic flow metric, *J. Internet Technol.*, **16** (2015), 1005–1014.

43. R. Dubin, A. Dvir, O. Pele, O. Hadar, I know what you saw last minute—encrypted http adaptive video streaming title classification, *IEEE Trans. Inf. Forensics Secur.*, **12** (2017), 3039–3049. https://doi.org/10.1109/TIFS.2017.2730819

44. H. D. Trinh, A. F. Gambin, L. Giupponi, M. Rossi, P. Dini, Mobile traffic classification through physical control channel fingerprinting: a deep learning approach, *IEEE Trans. Network Serv. Manage.*, **2020** (2020). https://doi.org/10.1109/TNSM.2020.3028197

45. M. Xie, J. Fu, Y. Wang, G. Peng, Monitoring and blocking methods of HTTP traffic injection in mobile web browser, *J. Wuhan Univ.*, **63** (2017), 385–396.

46. G. Rizothanasis, N. Carlsson, A. Mahanti, Identifying user actions from HTTP (S) traffic, in *2016 IEEE 41st Conference on Local Computer Networks (LCN)*, (2016), 555–558. https://doi.org/10.1109/LCN.2016.91

47. J. Manzoor, I. Drago, R. Sadre, How HTTP/2 is changing web traffic and how to detect it, in *2017 Network Traffic Measurement and Analysis Conference (TMA)*, (2017), 1–9. https://doi.org/10.23919/TMA.2017.8002899

48. J. Muehlstein, Y. Zion, M. Bahumi, I. Kirshenboim, R. Dubin, A. Dvir, et al., Analyzing HTTPS encrypted traffic to identify user's operating system, browser and application, in *2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, (2017), 1–6. https://doi.org/10.1109/CCNC.2017.8013420

49. T. Petsas, A. Papadogiannakis, M. Polychronakis, E. P. Markatos, T. Karagiannis, Measurement, modeling, and analysis of the mobile app ecosystem, *ACM Trans. Model. Perform. Eval. Comput. Syst.*, **2** (2017), 7. https://doi.org/10.1145/2993419

50. M. Rapoport, P. Suter, E. Wittern, O. Lhotak, J. Dolby, Who you gonna call? Analyzing web requests in Android applications, in *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, (2017), 80–90. https://doi.org/10.1109/MSR.2017.11

51. Z. B. Houidi, G. Scavo, S. Ghamri-Doudane, A. Finamore, S. Traverso, M. Mellia, Gold mining in a river of internet content traffic, in *International Workshop on Traffic Monitoring and Analysis*, Springer, (2014), 91–103. https://doi.org/10.1007/978-3-642-54999-1_8

52. *UNIBS*, 2011. Available from: http://netweb.ing.unibs.it/ ntw/tools/traces/.

53. *CIC*, 2021. Available from: https://www.unb.ca/cic/datasets/.

54. *UMass*, 2021. Available from: http://skuld.cs.umass.edu/traces/network/README-webident2.

55. *CAIDA*, 2021. Available from: https://catalog.caida.org/search?query=types=dataset.

56. *WIDE*, 2021. Available from: http://mawi.wide.ad.jp/mawi/.

57. G. Maier, A. Feldmann, V. Paxson, A. Mark, On dominant characteristics of residential broadband internet traffic, in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, (2009), 90–102. https://doi.org/10.1145/1644893.1644904

58. L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, K. Salamatian, Traffic classification on the fly, *ACM SIGCOMM Comput. Commun. Rev.*, **36** (2006), 23–26. https://doi.org/10.1145/1129582.1129589

59. L. Bernaille, R. Teixeira, K. Salamatian, Early application identification, in *Proceedings of the 2006 ACM CoNEXT Conference*, (2006), 1–12. https://doi.org/10.1145/1368436.1368445

60. L. Bernaille, R. Teixeira, Early recognition of encrypted applications, in *International Conference on Passive and Active Network Measurement*, (2007), 165–175. https://doi.org/10.1007/978-3-540-71617-4_17

61. A. Este, F. Gringoli, L. Salgarelli, On the stability of the information carried by traffic flow features at the packet level, *ACM SIGCOMM Comput. Commun. Rev.*, **39** (2009), 13–18. https://doi.org/10.1145/1568613.1568616

62. N. Huang, G. Jai, H. Chao, Early identifying application traffic with application characteristics, in *2008 IEEE International Conference on Communications*, (2008), 5788–5792. https://doi.org/10.1109/ICC.2008.1083

63. N. Huang, G. Jai, H. Chao, Y. Tzang, H. Chang, Application traffic classification at the early stage by characterizing application rounds, *Inf. Sci.*, **232** (2013), 130–142. https://doi.org/10.1016/j.ins.2012.12.039

64. T. T. Nguyen, G. Armitage, P. Branch, S. Zander, Timely and continuous machine-learning-based classification for interactive IP traffic, *IEEE/ACM Trans. Networking*, **20** (2012), 1880–1894. https://doi.org/10.1109/TNET.2012.2187305

65. G. He, M. Yang, J. Luo, L. Zhang, Online identification of tor anonymous communication traffic, *J. Commun.*, **24** (2013), 540–556.

66. L. Chen, J. Gong, Fast application-level traffic classification using NetFlow records, *J. Commun.*, **33** (2012), 145–152. https://doi.org/1000-436X(2012)01-0145-08

67. S. Dong, W. Ding, Traffic classification model based on fusion of multiple classifiers with flow preference, *J. Commun.*, **34** (2013), 143–152. https://doi.org/10.3969/j.issn.1000-436x.2013.10.017

68. V. Paxson, Empirically derived analytic models of wide-area TCP connections, *IEEE/ACM Trans. Networking*, **2** (1994), 316–336. https://doi.org/10.1109/90.330413

69. C. Dewes, A. Wichmann, A. Feldmann, An analysis of Internet chat systems, in *Proceedings of the 3rd ACM SIGCOMM Conference on Internet Measurement*, (2003), 51–64. https://doi.org/10.1145/948205.948214

70. T. Lang, G. Armitage, P. Branch, H. Choo, A synthetic traffic model for half-life, in *Aust. Telecommun. Networks Appl. Conference*, **2003** (2003), 1–5.

71. T. Lang, P. Branch, G. Armitage, A synthetic traffic model for Quake3, in *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology*, (2004), 233–238. https://doi.org/10.1145/1067343.1067373

72. S. Cheung, R. Crawford, M. Dilger, J. Frank, J. Hoagland, K. Levitt, et al. The design of GrIDS: A graph-based intrusion detection system, in *Technical Report CSE-99-2, UC Davis Computer Science Department*, (1999).

73. M. Iliofotou, P. Pappu, M. Faloutsos, M. Mitzenmacher, S. Singh, G. Varghese, Network monitoring using traffic dispersion graphs (tdgs), in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, (2007), 315–320. https://doi.org/10.1145/1298306.1298349

74. M. Iliofotou, H. Kim, M. Faloutsos, M. Mitzenmacher, P. Pappu, G. Varghese, Graption: a graph-based P2P traffic classification framework for the internet backbone, *Comput. Networks*, **55** (2011), 1909–1920. https://doi.org/10.1016/j.comnet.2011.01.020

75. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update,2017–2022 White Paper, 2019. Available from: https://branden.biz/wp-content/uploads/2019/05/white-paper-c11-738429.pdf.

76. A. Gember, A. Anand, A. Akella, A comparative study of handheld and non-handheld traffic in campus wi-fi networks, in *International Conference on Passive and Active Network Measurement*, (2011), 173–183. https://doi.org/10.1007/978-3-642-19260-9_18

77. J. Liu, T. Li, G. Chen, Y. Hua, Z. Lei, Mining and modelling the dynamic patterns of service providers in cellular data network based on big data analysis, *China Commun.*, **10** (2013), 25–26. https://doi.org/10.1109/CC.2013.6723876

78. S. Dong, D. Zhou, W. Ding, Traffic classification model based on integration of multiple classifiers, *J. Comput. Inf. Syst.*, **8** (2012), 10429–10437.

79. X. Gui, J. Liu, C. Li, Q. Lv, Z. Lei, Fine-grained analysis of cellular smartphone usage characteristics based on massive network traffic, *J. China Univ. Posts Telecommun.*, **23** (2016), 70–75. https://doi.org/10.1016/S1005-8885(16)60035-3

80. Y. Li, J. Yang, N. Ansari, Cellular smartphone traffic and user behavior analysis, in *2014 IEEE International Conference on Communications (ICC)*, (2014), 1326–1331. https://doi.org/10.1109/ICC.2014.6883505

81. M. Z. Shafiq, L. Ji, A. X. Liu, J. Wang, Characterizing and modeling internet traffic dynamics of cellular devices, in *Proceedings of the ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems*, (2011), 305–316. https://doi.org/10.1145/2007116.2007148

82. Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, S. Venkataraman, Identifying diverse usage behaviors of smartphone apps, in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, (2011), 329–344. https://doi.org/10.1145/2068816.2068847

83. F. T. Vincent, R. Spolaor, M. Conti, I. Martinovic, Appscanner: automatic fingerprinting of smartphone apps from encrypted network traffic, in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, (2016), 439–454. https://doi.org/10.1109/EuroSP.2016.40

84. M. Shen, J. Zhang, L. Zhu, K. Xu, X. Du, Y. Liu, Encrypted traffic classification of decentralized applications on ethereum using feature fusion, in *2019 IEEE/ACM 27th International Symposium on Quality of Service (IWQoS)*, (2019), 1–10. https://doi.org/10.1145/3326285.3329053

85. G. Aceto, D. Ciuonzo, A. Montieri, A. Pescape, MIMETIC: mobile encrypted traffic classification using multimodal deep learning, *Comput. Networks*, **165** (2019), 106944. https://doi.org/10.1016/j.comnet.2019.106944

86. G. Aceto, D. Ciuonzo, A. Montieri, A. Pescape, Multi-classification approaches for classifying mobile app traffic, *J. Network Comput. Appl.*, **103** (2018), 131–145. https://doi.org/10.1016/j.jnca.2017.11.007

87. G. Xie, Q. Li, Y. Jiang, D. Tao, G. Shen, R. Li, et al., SAM: self-attention based deep learning method for online traffic classification, in *Proceedings of the Workshop on Network Meets AI&ML*, (2020), 14–20. https://doi.org/10.1145/3405671.3405811

88. C. Liu, L. He, G. Xiong, Z. Cao, Z. Li, Fs-net: a flow sequence network for encrypted traffic classification, in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, (2019), 1171–1179. https://doi.org/10.1109/INFOCOM.2019.8737507

89. M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, J. Lloret, Network traffic classifier with convolutional and recurrent neural networks for internet of things, *IEEE Access*, **5** (2017), 18042–18050. https://doi.org/10.1109/ACCESS.2017.2747560

90. T. Shapira, Y. Shavitt, Flowpic: encrypted internet traffic classification is as easy as image recognition, in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, (2019), 680–687. https://doi.org/10.1109/INFCOMW.2019.8845315

91. F-SECURE, *Threat Description Bluetooth-Worm: SymbOS/Cabir*, 2021. Available from: https://www.f-secure.com/v-descs/cabir.shtml.

92. F-SECURE, *Mobile Threat Report Q4 2011*, 2021. Available from: https://www.f-secure.com/documents/996508/1030743/mobile_threat_report_q4_2011.pdf.

93. F-SECURE, *Threat Description Bluetooth-Worm: SymbOS/Cabir*, 2021. Available from: https://www.f-secure.com/v-descs/worm_iphoneos_ikee.shtml.

94. GDATASECURITYLAB, *FakePlayer*, 2021. Available from: https://www.gdata.at/securitylabs/mobile/mobile-malware/.

95. F-SECURE, *Mobile Threat Report 2012–2014*, 2021. Available from: https://www.f-secure.com/en/web/labs_global/whitepapers.

96. APPTHORITY, *Mobile Threat Report*, 2021. Available from: http://info.appthority.com/hubfs/website-LEARN-content/Appthority-Mobile-Threat-Report-Q12015.pdf.

97. S. C. Peng, A survey on malware containment models in smartphones, *Appl. Mech. Mater.*, **263** (2013), 3005–3011. https://doi.org/10.4028/www.scientific.net/AMM.263-266.3005

98. S. PENG, S. Yu, A. Yang, Smartphone malware and its propagation modeling: a survey, *Commun. Surv. Tutorials*, **16** (2014), 925–941. https://doi.org/10.1109/SURV.2013.070813.00214

99. G. Suarez-Tangil, J. E. Tapiador, P. Peris-Lopez, A. Ribagorda, Evolution, detection and analysis of malware for smart devices, *Commun. Surv. Tutorials*, **16** (2014), 961–987. https://doi.org/10.1109/SURV.2013.101613.00077

100. Y. Zhou, X. Jiang, Dissecting android malware: characterization and evolution, in *2012 IEEE Symposium on Security and Privacy*, (2012), 95–109. https://doi.org10.1109/SP.2012.16

101. Y. Liu, L. Zhang, J. Liang, S. Qu, Z. Ni, Detecting trojan horses based on system behavior using machine learning method, in *2010 International Conference on Machine Learning and Cybernetics*, (2010), 855–860. https://doi.org/10.1109/ICMLC.2010.5580591

102. V. K. Gudipati, A. Vetwal, V. Kumar, A. Adeniyi, A. Abuzneid, Detection of trojan horses by the analysis of system behavior and data packets, in *2015 Long Island Systems, Applications and Technology*, (2015), 1–4. https://doi.org/10.1109/LISAT.2015.7160176

103. J. Nazario, T. Holz, As the net churns: fast-flux botnet observations, in *2008 3rd International Conference on Malicious and Unwanted Software (MALWARE)*, (2008), 24–31. https://doi.org/10.1109/MALWARE.2008.4690854

104. W. Yu, X. Wang, P. Calyam, D. Xuan, W. Zhao, Modeling and detection of camouflaging worm, *IEEE Trans. Dependable Secure Comput.*, **8** (2011), 377–390. https://doi.org/10.1109/TDSC.2010.13

105. NSFOCUS Information Technology Co. Ltd., *XcodeGhost automatically Checking*, 2015. Available from: https://cloud.nsfocus.com/#/krosa/views/initcdr/secalertindex.

106. PANGU JAILBREAK, *Statistical Report for XcodeGhost Virus*, 2015. Available from: http://x.pangu.io/.

107. A. Shabtai, U. Kanonov, Y. Elovici, C. Glezer, Y. Weiss, "Andromaly": a behavioral malware detection framework for android devices, *J. Intell. Inf. Syst.*, **38** (2012), 161–190. https://doi.org/10.1007/s10844-010-0148-x

108. Y. Cao, R. Ji, L. Ji, X. Shao, G. Lei, H. Wang, MPTCP-*me*Learning: a multi-expert learning-based MPTCP extension to enhance multipathing robustness against network attacks, *IEICE Trans. Inf. Syst.*, **E104-D** (2021). https://doi.org/10.1587/transinf.2021NGP0009

109. F. Song, L. Li, I. You, H. Zhang, Enabling heterogeneous deterministic networks with smart collaborative theory, *IEEE Network*, **35** (2021), 64–71. https://doi.org/10.1109/MNET.011.2000613

110. F. Song, Z. Ai, H. Zhang, I. You, S. Li, Smart collaborative balancing for dependable network components in cyber-physical systems, *IEEE Trans. Ind. Inf.*, **17** (2021), 6916–6924. https://doi.org/10.1109/TII.2020.3029766

111. C. J. Wright, *Towards Real Time Characterization of Grain Growth from the Melt*, Columbia University, 2020.

112. G. Aceto, D. Ciuonzo, A. Montieri, A. Pescape, DISTILLER: encrypted traffic classification via multimodal multitask deep learning, *J. Network Comput. Appl.*, **183** (2021), 102985. https://doi.org/10.1016/j.jnca.2021.102985

113. Z. Bu, B. Zhou, P. Cheng, K. Zhang, Z. Ling, Encrypted network traffic classification using deep and parallel network-in-network models, *IEEE Access*, **8** (2020), 132950–132959. https://doi.org/10.1109/ACCESS.2020.3010637

114. G. Bovenzi, L. Yang, A. Finamore, A first look at class Incremental Learning in Deep Learning Mobile Traffic Classification, preprint, arXiv:2107.04464.

115. F. Song, M. Zhu, Y. Zhou, I. You, H. Zhang, Smart collaborative tracking for ubiquitous power IoT in edge-cloud interplay domain, *IEEE Internet Things J.*, **7** (2020), 6046–6055. https://doi.org/10.1109/JIOT.2019.2958097

116. F. Song, Z. Ai, Y. Zhou, I. You, R. Choo, H. Zhang, Smart collaborative automation for receive buffer control in multipath industrial networks, *IEEE Trans. Ind. Inf.*, **16** (2020), 1385–1394. https://doi.org/10.1109/TII.2019.2950109

AIMS Press