

Using host profiling to refine statistical application identification

Mohamad Jaber, Roberto G. Cascella and Chadi Barakat
INRIA - France

Email: {mohamad.jaber, roberto.cascella, chadi.barakat}@inria.fr

Abstract—The identification of Internet traffic applications is very important for ISPs and network administrators to protect their resources from unwanted traffic and prioritize some major applications. Statistical methods are preferred to port-based ones and deep packet inspection since they don't rely on the port number and they also work for encrypted traffic. These methods combine the statistical analysis of the application packet flow parameters, such as packet size and inter-packet time, with machine learning techniques. Other approaches rely on the way the hosts communicate and their traffic patterns to identify applications.

In this paper, we propose a new online method for traffic classification that combines the statistical and host-based approaches in order to construct a robust and precise method for early Internet traffic identification. We use the packet size as the main feature for the classification and we benefit from the traffic profile of the host (i.e. which application and how much) to decide in favor of this or that application. This latter profile is updated online based on the result of the classification of previous flows originated by or addressed to the same host. We evaluate our method on real traces using several applications. The results show that leveraging the traffic pattern of the host ameliorates the performance of statistical methods. They also prove the capacity of our solution to derive profiles for the traffic of Internet hosts and to identify the services they provide.

I. INTRODUCTION

The identification of Internet traffic applications is very important for ISPs and network administrators to protect their resources from unwanted traffic and prioritize some major applications. On the one hand, this allows to treat flows in a different way based on their quality of service requirements. On the other hand, it can serve for security reasons by blocking or looking closely at those users who run non legacy applications.

The identification of Internet traffic becomes more and more complex. Historically the recognition was done by using the port number [1]. Yet, some applications use dynamic non-standard port numbers; this is typically the case of telephony over IP. Other applications hide themselves using standard ports stolen from other applications. These ports are usually given by the end host and thus they can be easily changed.

Current techniques of "Deep Packet Inspection" (DPI) [2], [3] make it possible to go further in the identification of the applications but they require a complete and costly exploration of the payload of the packets. This induces an important load and requires updates with the appearance of new applications. Furthermore, when packets are encrypted, the recognition is not possible.

The statistical techniques [4]–[8] seem to be today an interesting alternative. They allow to recognize and to classify the applications according to their statistical signatures. These signatures can be volumes (number of bytes) per connection, connection durations, rates, inter-packet delays, packet sizes, and direction.

Most of the techniques that use statistical features require a machine learning phase to perform the classification of connections (or flows) into applications. In [5], McGregor et al. show the utility of using clustering algorithms for the identification of the traffic. They propose to use an unsupervised machine learning, called auto class, and the following statistical criteria: packet size, inter-arrival time, byte count, and connection duration. In [4], Moore et al. use a Naive Bayesian classifier for TCP traffic, and they try to find the best set of statistical criteria. In [6], Bernaille et al. test three clustering algorithms (K-Means, Gaussian mixture model, and the Spectral clustering) and the input features to assign flows to applications are the size and the direction of the first four packets jointly used. In [7], Crotti et al. classify Internet traffic by using the packet size and inter-arrival time. In our previous work [8], we develop a method to iteratively classify Internet traffic while using the size and the direction of the packets.

The common feature of statistical methods is that they classify every flow independently of each other using the pattern of its packets (size, time, and direction). Indeed, they don't use any information about the traffic pattern of the originating host or the type of services that run on the destined server. The same thing applies to peer-to-peer communications. We believe that the classification of previous flows sharing the same IP address either as source and/or destination is important to refine the classification of future flows and hence Internet applications in general. For instance, a host browsing the web is more prone to open several consecutive HTTP connections. A machine hosting a POP3 mail server is very likely to receive POP3 flows. In general, hosts have profiles for their flows either because of the behavior of users or the services run on them, and these profiles can help in the identification of flows in which they are implied. In this paper, we propose to build the traffic profile of hosts, based on the result of the classification of previous flows, and then to use this information to refine the classification of subsequent flows. On one hand these profiles help in flow classification and on the other hand they point to the behavior of the users behind them and on the network services they deploy.

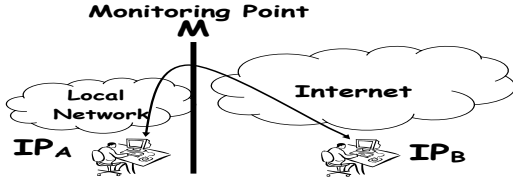


Fig. 1: The system.

Our solution differs from the previous works that consider the role of the host [9], [10] by the fact that it only relies on the information that a monitor collects passively from the packet flows. In [9], Trestian et al. characterize the role and type of traffic of an end-point by collecting publicly available information on the web based on the IP address of the host. BLINC [10] is a solution for Internet traffic identification that considers the role of the host. It focuses on the source and destination of the flows to determine the host behavior, which is studied across three levels: social to account for the host popularity and communities of hosts (groups of communicating hosts), functional to identify the functional role of a host (offered services, used services), and protocol patterns of the host. The main difference, which is also the strength, of our approach consists of only considering the flows sent and received by the monitored hosts and in crossing the information between flows of the same host so as to build profiles and reach better classification. We construct and leverage the profiles of the communicating hosts simultaneously and on the fly without requiring the traffic monitor to maintain a detailed history of their interaction.

Our contribution in this work can be summarized as follows. First, we define the host profile and we determine the host-based probability that a flow is of a given application in both the incoming and outgoing direction. We develop a new method that relies on the result of the classification of flows from the same host to determine the profiles of hosts and to use these profiles later as an initial guess before the classification of future flows. The main idea is to combine the statistical properties of a flow with the traffic profile of the end-points to better associate flows to applications. The host profiles are updated after each classification using an exponential weighted moving average filter to absorb any transient behavior; the way the profile accounts for past classified flows depends on some discounting parameter, which can be decided by the network administrator. Once described, we use two real traces to test our method and to show how to characterize the traffic pattern of each host in the traces.

The rest of the paper is organized as follows. Section II introduces and discusses the host profiling. Section III explains our classification method. Section IV and Section V describe the traces and the evaluation results, respectively. Section VI concludes the paper.

II. HOST TRAFFIC PROFILE

In this section we discuss how we determine the traffic profile of a host and what are the benefits of using this

information to refine the classification of Internet applications. The methodology herein described is general and it can be integrated to any classification method transparently. In our model and without loss of generality we consider a monitoring point at the edge of the network, located in the ISP network, as shown in Fig. 1. The monitor passively captures the flows between any two users; a flow consists of the packets with the same 5-tuple (IP source and destination, port source and destination, IP protocol). For each flow, we consider the two end points: a host located inside the ISP network (IP_A in Fig. 1) and a destination host downstream the monitoring point (IP_B in Fig. 1). We don't assign any specific role to the hosts, IP_A and IP_B , which can act indifferently as client or server during a session. The monitor inspects the packets of each flow and extracts statistical information, such as packet size, inter-packet time, direction of the packet, etc. This information is used to create the signature of the flow and to assign the flow to the application that matches the signature. We will discuss in Section III the definition of this signature and classification procedure.

The traffic profile of a host is defined based on type of previous flows. This requires that the monitor collects statistical information about a flow, classifies the flow, and stores the result of the classification to track the activity of a host. In a real setting, we assume that the monitor logs only the traffic for the hosts inside the ISP network, or those of interest, and might store information about some IP addresses that runs dedicated services. The traffic profile, so computed, gives an indication of the preferred applications that run at the host.

The novelty of our approach consists of using this traffic pattern to predict future flows that involve the same host. In this section, we first discuss how a monitor computes the probability that a flow of packets between two hosts is of a certain application solely using the traffic patterns of these hosts. Then we discuss how the monitor computes and updates the host profile.

A. Host based probability of a flow

Let F denote a function that associates a packet flow between a source S and destination D to an application $A(i)$, with $1 \leq i \leq N_A$ and N_A the number of monitored applications. Let $P(F_S = A_S|S)$ (or $P(F_D = A_D|D)$) be the probability that, given the host traffic profile, the flow is of an application A_S for the source (or A_D for the destination). Then, the probability $P(F = A(i))$ that the flow is of application $A(i)$ is computed as follows:

$$\begin{aligned}
 P(F = A(i)) &= P(F_S = A_S \cap F_D = A_D | A_S = A_D) \\
 &= \frac{P(F_S = A(i)|S) * P(F_D = A(i)|D)}{\sum_{j=1}^{N_A} P(F = A(j)|S \cap D)} \quad (1) \\
 &= \frac{P(F_S = A(i)|S) * P(F_D = A(i)|D)}{\sum_{j=1}^{N_A} P(F_S = A(j)|S) * P(F_D = A(j)|D)}
 \end{aligned}$$

Equation (1) means that we compute the probability by considering the cases when the prediction for each host is in accordance by considering the traffic profiles of S and D

separately. Equation (1) also holds when the monitor only records the traffic profile of one of the two hosts. In fact, if we assume a uniform probability for the other host, e.g., $P(F_D = A_D|D) = \frac{1}{N_A}$, then, equation (1) simplifies to $P(F = A(i)) = P(F_S = A(i)|S)$.

B. Host profile definition and update

We now discuss how the monitor computes the host profile and updates this information. Each host can be source or destination for different flows and it depends on whether it does send the first packet of the flow. The monitor captures the flows and decides about a flow by using the source or destination profile of the host. This results in two traffic profiles for the same host. In the rest of the section, we discuss a generic host and the computation of the source profile for this host; the destination profile is defined in the same way.

Let S denote the generic source host of a flow and F_S the function that maps the flow to an application. The monitor computes the host profile by using previous classified flows, in this case when the host is the source of the first packet. The profile $P(\mathcal{A}|S)$ is thus defined as the prior distribution for the flows in the domain \mathcal{A} , which defines the applications $A(i), 1 \leq i \leq N_A$. If the monitor has not any information about previous traffic of a host, then, the monitor considers a uniform prior distribution. The prior distribution is updated after each classification of a new collected flow.

The update works as follow. Let $P_{(n-1)}(A(i)|S)$ be the prior probability for application $A(i)$ computed from the past $(n-1)$ flows that the monitor affects to the application $A(i)$ with probability $P(F_S = A(i)|S)$ for each application, then the posterior probability for each application is computed as follows:

$$P_{(n)}(A(i)|S) = \lambda * P_{(n-1)}(A(i)|S) + (1 - \lambda) * P(F_S(n) = A(i)|S) \quad (2)$$

$P(F_S(n) = A(i)|S)$ is the result of the classification of flow n and $\lambda, 0 \leq \lambda \leq 1$, represents the discounting factor for past classifications. When λ is close to 0, the profile is computed by associating a higher weight to the most recent flows. When λ is close to 1 the profile is calculated over a longer period, which means that the profile is determined in equal measure by all previous classified flows. When $\lambda = 1$ the profile corresponds to the initial prior distribution, which in our case assigns a uniform probability to all applications. The best choice of λ depends on the traffic pattern of the host and on the performance of the classifier. We will discuss more about λ in Section V. The amount of information that the

TABLE I: Example of a traffic profile of a host

Applications:	FTP	HTTP	POP3	SMTP	SSH
Source:	0.02	0.76	0	0.2	0.02
Destination:	0.22	0	0.1	0.23	0.45

monitor maintains to update the profile of the host is limited to the two prior distributions. Table I shows an example of the source and destination profiles of a host.

III. METHOD DESCRIPTION

Our purpose for the classification of Internet traffic is to detect online which flow belongs to which application. We use a statistical and iterative method that computes the probability that packets are generated by an application. We have defined and used this method to classify Internet traffic based on the size of the packets in [8]. The method allows an iterative classification of the flows for each packet size independently and uses more packet sizes for the identification of an application until the classifier reaches a predefined threshold. Each flow corresponds to a sequence of N packets Pkt_k , where k indicates the position of the packet in the flow independently of its direction.

In this section we first propose an overview of our method and then we detail how the method uses the host profile to refine the classification. The method consists of three main phases which are detailed in the following sections: the model building phase, the classification phase, and the application probability or labeling phase.

A. Model building and classification phase

We use K-Means as supervised machine learning algorithm to partition the input in a predefined number of clusters. Given the number of clusters N_C , K-Means assigns each input feature to a cluster so as to minimize the Euclidian distance of each input from the centroid of the cluster.

Pkt_k denotes the packet size, i.e., the observations, and for each packet size we train separately K-Means to obtain different set of classes. The input feature corresponds to the value of the size of the packet associated with a sign that represents the direction of the packet. A positive sign corresponds to a packet from the source to the destination. In the learning algorithm, every class is affected by all applications with different probabilities proportional to the number of flows from each application present in the class. Hence, each class defines the probability that the elements within this class are generated by the applications.

The model building phase consists of constructing these sets of classes (clusters) by using a training data set, described in Section IV. This learning phase is used to compute $P(C(j)|A(i))$, i.e., the per-class probability, knowing the application $A(i)$. The probability is computed for all clusters $C(j)$, where $1 \leq j \leq N_C$ and N_C is the number of clusters. We build a separate model, i.e., set of classes, for every packet size noted by Pkt_k and we use these classes for the classification phase.

The classification consists of using the classes defined in the learning phase to test and assign the Internet flows to a class. The test is performed by computing the Euclidian distance between the input feature from Pkt_k and the centroid of each class determined for the k -th packet size. We affect the point to the closest class. The test is repeated for all the packet sizes of a flow iteratively until we reach a predefined threshold. The classification result consists in the probability that the Pkt_k

TABLE II: Traces Description

Source and Date	Application	training	testing
Brescia University April 2006 [7]	HTTP	8000	17,263
	SMTP	8000	19,835
	POP3	8000	19,935
Brescia University Fall 2009 [11]	HTTP	500	30422
	HTTPS	500	3608
	EDONKEY	500	3702
	BITTORENT	500	3608

identifies an application and it is given as input to a labeling function described in the following section.

B. Application probability or labeling phase

In the labeling phase we assign a flow to an application knowing the result of the classification and the probability computed from the profiles of the source and destination, as discussed in Section II. We combine iteratively the results of the classification for each single packet size and we calculate the probability ($P(A(i))$) that a flow belongs to an application $A(i)$ given the prediction from the host profiles and the classification results of the first N packet sizes (i.e., class $C(j(1))$ for the first packet size, class $C(j(2))$ for the second packet size and so on).

$$\begin{aligned}
 P(A(i)) &= P(A(i)|Result) \cap P(F = A(i)) \\
 &= \frac{P(F = A(i)) * \prod_{k=1}^N P(C(j(k))|A(i))}{\sum_{i=1}^{N_A} [P(F = A(i)) * \prod_{k=1}^N P(C(j(k))|A(i))]} \quad (3)
 \end{aligned}$$

$P(F = A(i))$ is the probability that a flow between a source and a destination comes from application $A(i)$ based on their traffic profiles. $P(C(j(k))|A(i))$ is the probability that Pkt_k of a flow belongs to the class $C(i)$ knowing the application $A(i)$. N_A is the total number of applications. We call this probability the assignment probability. It combines the result of the classification, obtained with the K-Means clustering method, and the pattern of the hosts, which gives an indication of the next type of application flow. The assignment probability is computed when the monitor captures each packet of the same flow. Thus, we do not require to wait for a given number of packets to start the classification procedure. We stop this iterative process when the highest assignment probability is above a predetermined threshold or the maximum allowed number of tests is reached. This way the threshold is seen as a way to leave the classification phase earlier when we are sure about the flow. The monitor updates the profiles of the hosts, both the source and the destination, when the classification ends and it has assigned a flow to a given application. The update of the profiles is done as described in Section II.

IV. TRACE DESCRIPTION

In our analysis we use two real traces, see Table II for details. The two traces have been collected at the edge gateway of the Brescia University's campus network. The first trace, noted trace I [7], was collected during April 2006 and the second trace, noted trace II [11], was collected on three consecutive working days during fall 2009. Every trace consists of two sets,

a training set and a testing set, and the type of application associated with each flow is determined with a deep packet inspection method.

In the learning phase we use the training set, which consists of an equal number of flows per application to ensure that there is no bias in our learning phase. During this phase we do not compute the host profile and the application flows are used only to construct the classes in K-Means. The host profile is only computed during the testing phase. We initially consider a uniform prior distribution for the source and destination profiles of an unknown host. Then, we update the profiles once flows of this host are collected. The testing set is used to evaluate how well our iterative method behaves in identifying the application.

V. EXPERIMENTAL RESULTS

In this section we present the evaluation results of our method when the traffic profile of the hosts is used to refine the classification. We consider the case of a monitor that maintains the profile of the hosts that are located inside the ISP domain, since it is interested to understand what is the usage of the network by the ISP customers. In a real setting, the monitor might also decide to maintain the information about popular Internet servers. Indeed, it can use these profiles to compute the probability that a flow is originated by an application, as we discuss in Section II. We use the traces described in Section IV and we profile the hosts with the same IP prefix, i.e., those inside the Brescia campus. We have counted an average of 10 flows per IP address outside the Brescia campus and we have decided to not show the results since there is not a sufficient number of flows per IP to compute the profile. Thus, we only compute the client profile for these machines. The flows are all TCP connection and the hosts within the campus initiate the connection.

The metrics used for the evaluation are:

- *False Positive (FP) rate* is the percentage of flows of other applications classified as belonging to an application I .
- *True Positive (TP) rate* is the percentage of flows of application I correctly classified.
- *Precision* is the ratio of flows that are correctly assigned to an application, $TP/(TP + FP)$. The overall precision is the weighted average over all applications given the number of flows per application.

We run the test for all the available packet sizes to test its significance as a feature for identifying applications. We set the number of clusters equal to 400 for K-Means. We have tested the supervised machine learning algorithm with different number of clusters and 400 gives the best results as it allows to group the features in small clusters and account for possible noise in the observations.

A. Packet size distribution

We initially analyze the first 10 packet sizes of the testing flows of trace I to understand how well the packet size characterizes an application. Fig. 2 and Fig. 3 show the distributions for each packet number, where the number indicates the order

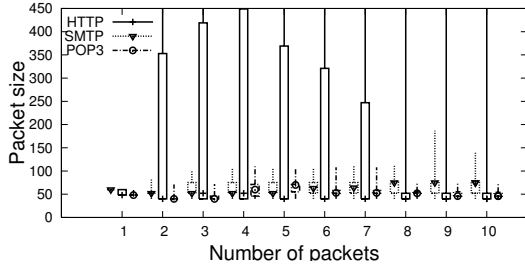


Fig. 2: Packet size distribution: sent by hosts inside the campus

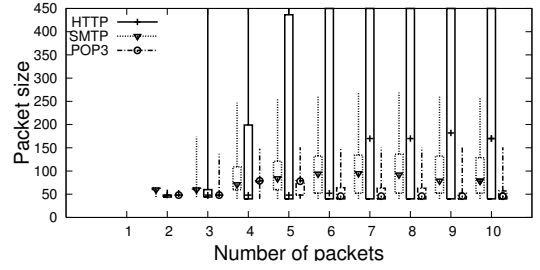


Fig. 3: Packet size distribution: got by hosts inside the campus

in a flow; we limit the plot to 450 bytes. The results from the packets sent and received by the hosts inside the Brescia campus are shown in Fig. 2 and Fig. 3 respectively. For each packet number, the plots show the median, and the bars indicate the quartiles, and the 2nd and 98th percentiles. From the distribution we first notice that all the testing flows are initiated by the hosts with the same IP prefix because there are no first packets of a flow in Fig. 3.

Fig. 2 shows that the first packet for the three applications has the same size in almost all the flows. In particular, the SMTP packet is larger than the others. POP3 and HTTP have the same value for the median but the latter spans more values for the size of the packet, shown by the 75th percentile. The second packet has a similar distribution for the three applications. In Fig. 3, the second packet has a size equal to the median and POP3 and HTTP have a similar distribution. By analyzing the first packets we can conclude that it is possible to distinguish an SMTP application from a POP3 or HTTP flow. From the 6th packet, the three applications have different distributions for the packet size. This means that it is possible to differentiate one application from the others both for packets sent or received by the hosts. We can conclude from these two figures that using the packet size can be a very good parameter to differentiate between Internet applications. For the lack of space we will not show the distributions related to Trace II.

B. Classification results

In this section we discuss the performance of the classification method when the host profile is used to refine the probability that a flow is of a given application type.

1) *Precision*: Fig. 4 and 5 plot the total precision of trace I and trace II respectively versus the number of packets used for the classification. Our method classifies a flow at each packet iteratively, as we discuss in Section III. The different lines in the plot correspond to the precision of the classifier when different values of the discounting factor λ are used. The value of λ determines the weight assigned to the last classification results. When $\lambda = 0.1$, the most recent classification results determine the profile of the host. When $\lambda = 0.9$, the host profile is computed over a longer period. The value of $\lambda = 1$ means that a uniform probability is associated to each application, thus, the host profile is not used, as we have discussed in Section II-B.

The results show that the precision of the classifier improves considerably when the profile of the hosts is used to decide in favor of this or that application.

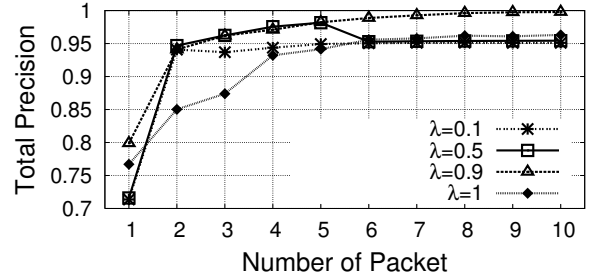


Fig. 4: Total precision versus the number of packets (Trace I)

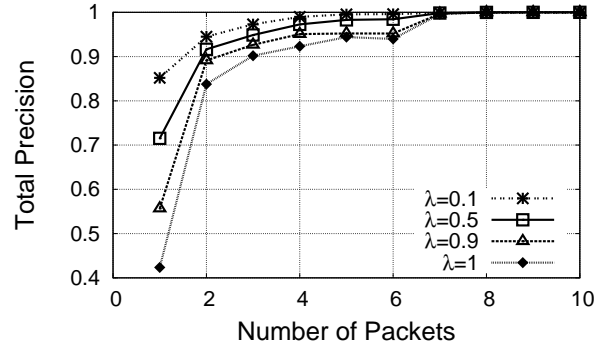


Fig. 5: Total precision versus the number of packets (Trace II)

For Trace I, we can observe in Fig. 4 that a value of $\lambda = 0.9$ gives the best performance for the classifier. We obtain a precision of 94% already after two packets, 97% after four packets and 99.9% when 10 packets are used for the classification. When $\lambda = 0.1$ the classifier predicts with less accuracy the applications. With this value of λ the classifier is more sensitive to recent flows. Thus, it is more prone to a wrong classification when the host has a uniform traffic behavior over all applications. For this trace we have a big number of flows that belong to two different applications generated uniformly by the same host. Thus, the method classifies the applications with less precision for small values of λ . For Trace II and for all the selections of λ , we have better performance compared to the classification without host profile information ($\lambda = 1$). We can observe in Fig. 5 that a small value of λ increases the precision already after the first packet. However, large values of λ require more packets to classify correctly the flows. The precision for all values of λ converges to 99.99% after 7 packets. We can conclude from these results that the profile of the host gives an early characterization of a flow because of the traffic pattern of the host. For instance, we can consider that a host that is browsing the web is more

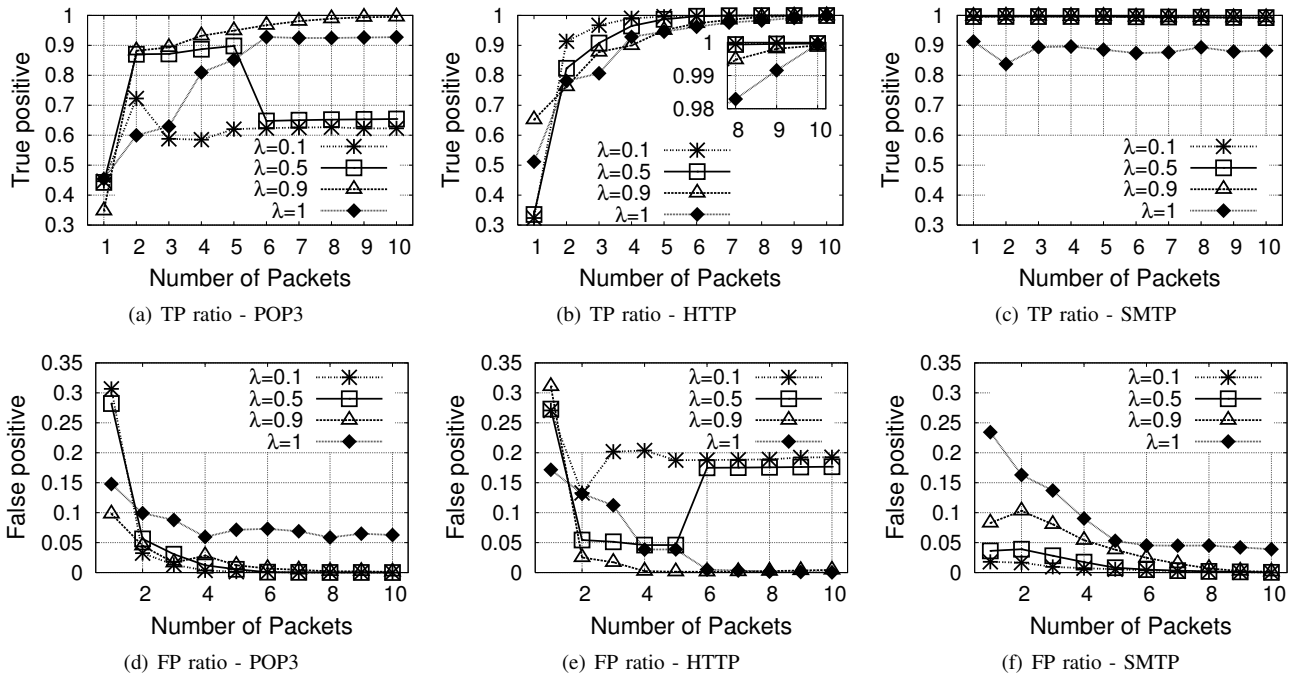


Fig. 6: Classification results for Trace I

prone to have a sequence of HTTP connections.

2) *True Positive*: Fig. 6(a) shows the True Positive (TP) ratio for the POP3 application as a function of the number of packets used for the classification. We can see that for $\lambda = 0.9$ the TP ratio increases when we use more packets for the classification, until it reaches a value of 0.99 with ten packets. In comparison, when the host information is not taken into account, the TP ratio reaches a maximum value of 0.91 with 10 packets. The TP ratio drops when $\lambda = 0.5$ and 6 packets are used for the classification. Similar performance can be observed for $\lambda = 0.1$ at the second packet. This means that the classifier fails to identify correctly POP3 flows as they are assigned to other applications. We will confirm this behavior when we analyze the False Positive ratio to understand what is the output application of our classification method. The fact that this happens only for these small values of λ means that we have some hosts who generate flows belonging to different applications uniformly.

Fig. 6(b) and Fig. 6(c) plot the True Positive (TP) ratio for HTTP and SMTP as a function of the number of packets respectively. Fig. 6(b) shows that the TP ratio increases for all the values of λ , even when we do not use any host information. The classifier has better performance with $\lambda = 0.1$, which means that there are consecutive HTTP flows in general. From Fig. 6(c) it is interesting to notice that the TP ratio is 1 for any number of packets of SMTP classified traffic when we use the profile of the host to refine the classification. The fact that the result is not sensitive to the value of λ means that the SMTP traffic is predominant in some hosts. We have also already observed in Fig. 2 and Fig. 3 that the size of the SMTP traffic has a different distribution than the one of other applications, which eases the classification of SMTP traffic. Besides, when

packets of other applications have a size which is recognized as a SMTP signature then the flow is also classified as SMTP. This explanation can only be confirmed from the analysis of the false positive ratio.

In Fig. 7(a-d) we plot the True Positive ratio for different applications (HTTP, HTTPS, EDONKEY and BITTORENT) of Trace II. We can clearly observe that for all the values of λ we have better performance compared to the classification without host profile information ($\lambda = 1$). The True Positive ratio keeps increasing when more packets are used for the classification. For all the applications we can observe that we have better performance when we use a small value for λ , which means that we don't have a lot of changes in the traffic pattern of these hosts. For HTTP and HTTPS applications we obtain a good precision with ($\lambda = 0.1$) around 80% after the first packet, 90% after the second and around 99.99% after seven packets for HTTP and 93% after the first packet and converge to 99.99% after the seventh packet for HTTPS. For the peer to peer applications we have a very high precision. As for EDONKEY, we converge to 99.99% after three packets for all the values of λ , and for $\lambda = 0.1$ we obtain 100% already at the first packet. As for BITTORENT, we get a precision around 99.99% from the first packet and for all values of λ even when we don't use the host information, because the distribution of the size of the first packet is very different compared to the one of other applications.

3) *False Positive*: Finally we discuss the results of the False Positive ratio of Trace I and Trace II to confirm our previous hypotheses. For Trace I we can immediately notice in Fig. 6(d) that the percentage of misclassified flows of other applications, assigned to POP3, drops significantly after 4 packets. Thus, the True Positive ratio of 6(a) indicates the correctly classified

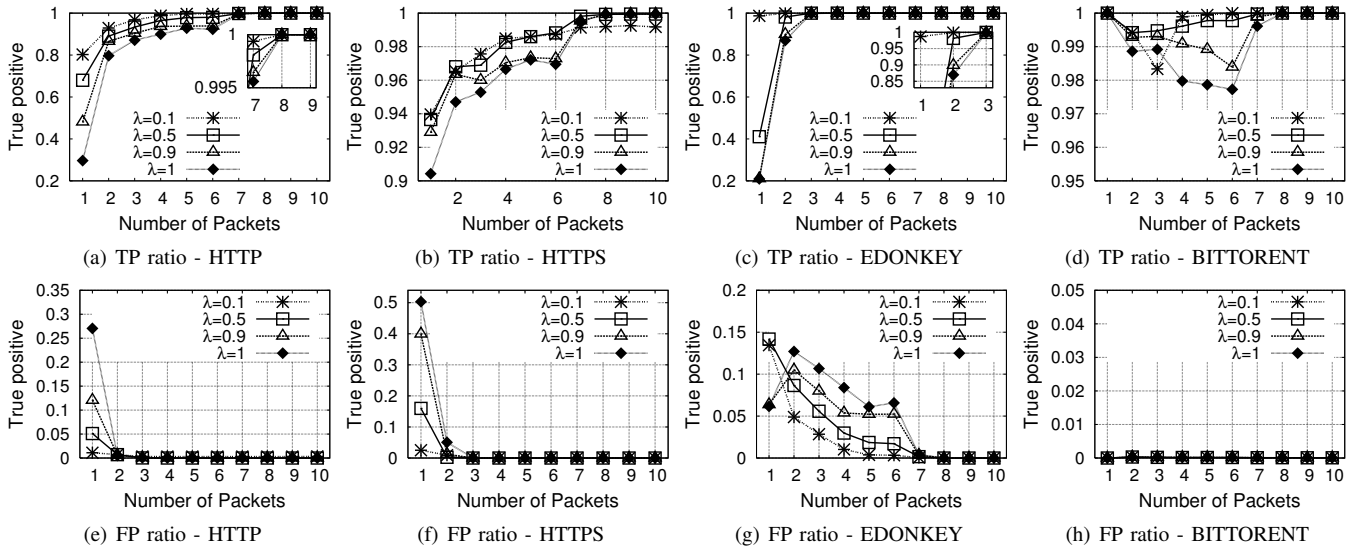


Fig. 7: Classification results for Trace II (They are showed with different scales)

POP3 flows. Fig. 6(e) also confirms that most of the POP3 flows that have not been detected are indeed classified as HTTP traffic. This is shown for values of $\lambda = 0.1$ and $\lambda = 0.5$. In Fig. 6(d) and Fig. 6(b), we also observe that when $\lambda = 0.9$ the False Positive ratio drops to 0 after four packets. Fig. 6(f) shows that the classification for most of the SMTP traffic is indeed correct when the classification of recent flows weighs more. When $\lambda = 0.9$, the classifier labels other flows as SMTP, which means that some hosts have SMTP flows that interleave the ones of other applications. In Fig. 7(e-h) we plot the False Positive ratio for the different applications (HTTP, HTTPS, EDONKEY, BITTORENT) of Trace II. We can observe that for all values of λ we get better performance in comparison with the classification without host profile information. For HTTP the False Positive ratio reaches to 0% for all values of λ only after two packets. We can notice the same behavior for HTTPS application. For BITTORENT, the False Positive ratio is always around 0% from the first packet and for all λ values, this observation can be explained by the fact that the size of the first packet for the BITTORENT application is different from the other applications. Finally for the EDONKEY application the False Positive ratio is around 12% for the first packets and still decreasing until it reaches 0% after 7 packets. The next section discusses the impact of λ for the classification. Then we have analyzed the profile of one host to understand which services the host runs and determine its traffic pattern.

C. Importance of the discounting factor λ

The discounting factor λ determines how previous flows are considered for the classification of a new one. We recall that for λ close to 1 the host profile is computed over a longer period and previous flows have similar weights. The opposite case is when λ is close to 0. Finally $\lambda = 1$ means that the classification does not account for the profile of the host. A precision above 0.9 for all values of λ shows that our iterative

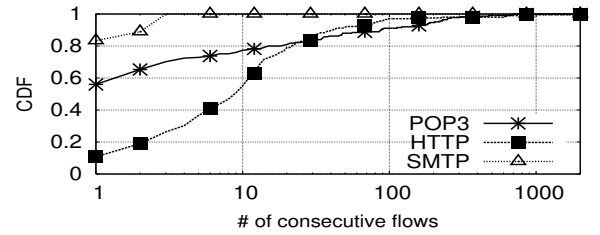


Fig. 8: IP1: number of consecutive application flows

method is effective in classifying the applications. However, we can see that the precision is a function of λ . In particular we notice that when few packets are used for the classification, the profile of the user helps for an early detection of the Internet traffic. In this case, the machine learning algorithm does not have much information about a flow and it associates to the applications comparable probabilities. Thus, the profile of the host has higher influence on the final classification.

However, a correct λ is important to improve the classification even if more information is available. From results of Trace I, we can notice that high values perform better and from results of Trace II we can notice that small values of λ perform better. So the network administrator should choose the best λ for its network based on the applications and type of users. The host profile prediction of the application is more accurate because we associate a probability to the application of the next flow based on the fraction of the past traffic due to this application. In the next section, we characterize this traffic and how the flows are distributed.

D. Traffic pattern of a host

In the Trace I, described in Table II, we have identified different types of hosts. In particular, there are some hosts within the Brescia campus that are dedicated to single services and other hosts that use all the three applications. In this section, we analyze in details the traffic of the latter type of

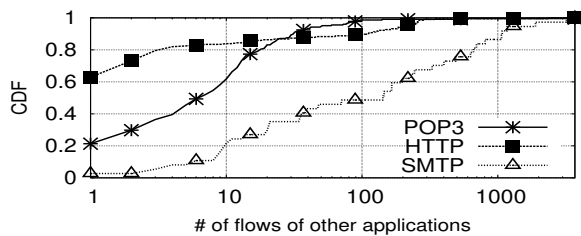


Fig. 9: IP1: number of flows of other applications

host and we determine its profile. This will shed light on the importance of λ for the identification of the Internet traffic. The total number of flows generated by this IP, noted $IP1$, is 14,151 flows, which is subdivided as follows: 7,101 HTTP flows, 7,014 POP3 flows, and only 35 SMTP flows.

Fig. 8 plots the cumulative distribution (CDF) of consecutive flows of the same type of application in semi-log scale. Since there are only 35 SMTP flows, most of these flows are isolated so that there are only sequences of 1 to 3 consecutive flows (burst). Almost 60% of the POP3 traffic consists of a single flow. The rest of the POP3 traffic consists of single larger bursts and the largest one consists of almost 700 flows. The number of small HTTP consecutive flows is 10% of the total number of HTTP bursts and 80% of the burst consists of less than 100 flows. This corresponds to a typical browsing activity of a user, who surfs from one web site to another by following the hyperlinks.

The fact that the medium size burst of HTTP traffic are interleaved with POP3 is evident when we use a small value of λ . In particular, since there are more consecutive HTTP flows than POP3, the host based classification with a small value of λ is more sensitive to the presence of a new burst. Moreover, the POP3 and HTTP packet size has similar distribution for certain packet numbers, as discussed in Section V-A, thus the host profile has more impact in the classification decision.

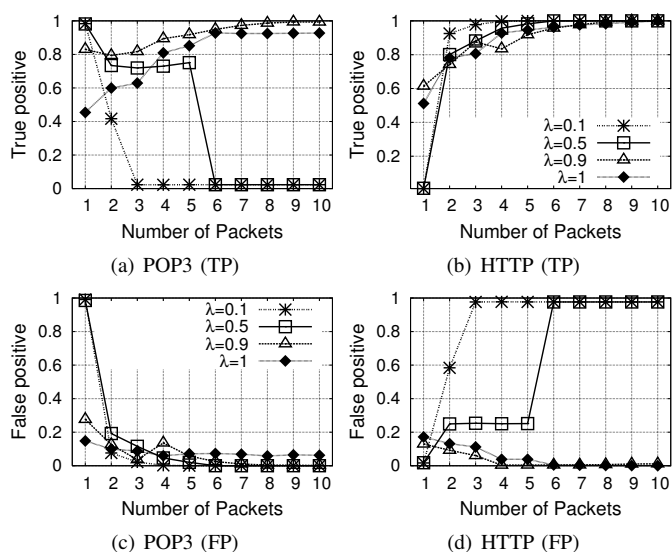


Fig. 10: Classification results for the analyzed host IP1

To conclude the characterization of the profile, Fig. 9

plots the cumulative distribution (CDF) of flows of other applications that separate two flows of the same application. There are more than 60% of HTTP bursts separated by only 1 flow of another application and 80% of HTTP bursts separated by less than 10 consecutive flows. As for POP3, there are 20% and 60% of burst respectively. The presence of small burst of other applications between two HTTP flows justifies the high percentage of false positives (see Fig. 6(e)). In general this is more pronounced with small values of λ . In Fig. 10 we show the classification results for $IP1$. The results confirm the analysis of Trace I. The impact of the profile of the host is also evident.

VI. CONCLUSION

In this paper we present our new method for Internet traffic identification that combines the statistical and host-based approaches. The statistical parameters that we use are the size and direction of the first N packets. The novelty of our approach consists in leveraging the host profile to refine the classification. First we define the profile of the host and how it is updated. Then we show how the profiles of the source and destination hosts are used to assign a prediction probability to the new flow.

We evaluate our solution on two real traces and we profile the hosts with the same IP prefix. We test our method for different values of the discounting factor λ and discuss the optimal choice based on the traffic pattern of the host. The results show a great improvement for the classification of applications when the host profile is used. In particular, the classifier reaches a precision of 0.99. Finally, we characterize the host profile and show the distribution of the flows, i.e., the traffic pattern of a representative host.

REFERENCES

- [1] IANA, "Internet assigned numbers authority," <http://www.iana.org/>.
- [2] A. Moore and K. Papagiannaki, "Toward the accurate identification of network applications," in *Proceedings of the 6th Passive and Active Measurement Workshop (PAM 2005)*, October 2005, pp. pages 41–54.
- [3] S. Sen, O. Spatscheck, and D. Wang, "Accurate, scalable in-network identification of p2p traffic using application signatures," in *WWW 2004 Conference*, Philadelphia, USA, May 2004.
- [4] A. Moore and D. Zuev, "Internet traffic classification using bayesian analysis techniques," in *ACM Sigmetrics*, 2005.
- [5] A. McGregor, M. Hall, P. Lorier, and J. Brunskill, "Flow clustering using machine learning techniques," in *PAM*, 2004, pp. 205–214.
- [6] L. Bernaille, R. Teixeira, and K. Salamatian, "Early application identification," in *The 2nd ADETTI/ISCTE CoNEXT Conference*, Lisboa, Portugal, December 2006.
- [7] M. Crotti, M. Dusi, F. Gringoli, and L. Salgarelli, "Traffic classification through simple statistical fingerprinting," in *ACM-Sigcomm Computer Communication Review*, vol. 37, January 2007, pp. 5–16.
- [8] M. Jaber and C. Barakat, "Enhancing application identification by means of sequential testing," in *NETWORKING '09: Proceedings of the 8th International IFIP-TC 6 Networking Conference*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 287–300.
- [9] I. Trestian, S. Ranjan, A. Kuzmanovi, and A. Nucci, "Unconstrained endpoint profiling (googling the internet)," in *ACM SIGCOMM '08*. Seattle, WA, USA: ACM, 2008, pp. 279–290. [Online]. Available: <http://doi.acm.org/10.1145/1402958.1402991>
- [10] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "Blinc: Multilevel traffic classification in the dark," in *SIGCOMM '05*, New York, USA, August 2005.
- [11] T. II, "Brescia university," <http://www.ing.unibs.it/ntw/tools/traces/>, 2009.