EURECOM
Department of Networking and Security
2229, route des Crètes
B.P. 193
06904 Sophia-Antipolis
FRANCE

Research Report RR-10-234

Traffic Classification: Application-based feature selection using logistic regression

Taoufik En-Najjary, Guillaume Urvoy-Keller, Marcin Pietrzyk,
and Jean-Laurent Costeux

Tel : (+33) 4 93 00 81 00
Fax : (+33) 4 93 00 82 00
Email : {ennajjar,urvoy}@eurecom.fr,
{marcin.pietrzyk,jeanlaurent.costeux}@orange-ftgroup.com

[1]EURECOM's research is partially supported by its industrial members: BMW Group Research & Technology - BMW Group Company, Bouygues Téécom, Cisco Systems, France Télécom, Hitachi, SFR, Sharp, ST Microelectronics, Swisscom, Thales.

Abstract

Recently, several statistical techniques using flow features have been proposed to address the problem of traffic classification. These methods achieve in general high recognition rates of the dominant applications and more random results for less popular ones. This stems from the selection process of the flow features, used as inputs of the statistical algorithm, which is biased toward those dominant applications. As a consequence, existing methods are difficult to adapt to the changing needs of network administrators that might want to quickly identify dominant applications like p2p or HTTP based applications or to zoom on specific less popular (in terms of bytes or flows) applications on a given site, which could be HTTP streaming or BitTorrent for instance. We propose a new approach, aimed to address the above mentioned issues, based on logistic regression. Our technique incorporates the following features: i) Automatic selection of distinct, per-application features set that best separates it from the rest of the traffic ii) Real time implementation as it needs only to inspect the first few packets of a flow to classify it, (iii) Low computation cost as logistic regression is implemented by comparing a linear combination of a flow features with a fixed threshold value, (iv) Ability to handle application types that former methods failed to classify. We validate the method using two recent data sets collected on two ADSL platforms of a large ISP.

1 Introduction

Application identification is of major interest for networks operators, especially Internet Service Providers and enterprise network administrators. Motivations behind this need are many fold: (i) enforcement of internal or national rules, e.g., banning p2p traffic from an Intranet, (ii) better understanding of actual and emerging applications (iii) assessment of the impact of those applications on peering agreements and/or the return on investment if some p4p initiative was taken [29] or (iv) possibility to offer additional services based on application, e.g., QoS protection of multimedia transfers.

However mapping flows to applications is not straightforward and has attracted a lot of attention from the research community. Indeed, Internet traffic is the product of a complex multi factor system involving a range of networks, hosts and seemingly uncountable variety of applications. Its complexity is continually increasing as developers keep producing new applications and inventing new usages of the old ones.

Many different methods have been proposed to solve the traffic classification problem. In the early Internet, traffic classification relied on the transport layer identifiers. However, the advent of new protocols like p2p, and the increase of applications tunneled through HTTP make port-based classification significantly misleading. Many studies have confirmed the failure of port-based classification [9]. This triggered the emergence of deep packet inspection (DPI) solutions that identify the application layer protocol by searching for signatures in the payload. The increasing use of encryption and obfuscation of packet content, the need of constant updates of application signatures and governments regulations, might however undermine the ability to inspect packets content.

Recently, several solutions based on statistical classification techniques and per flow features to probabilistically map flows to applications have been proposed [14, 3, 13, 16, 6, 4, 22]. These approaches generally consist of a first phase where flow features are selected based on some intrinsic characteristics like (the lack of) correlation and a second phase where flows are clustered according to the selected features. In general, the overall performance of the proposed statistical classifiers are satisfactory when considering all flows and applications in a given dataset. The latter means that the dominant applications, typically Web transfers and some p2p applications like eDonkey, are well classified but other applications that represent a small fractions of transfers, like streaming, might not be correctly identified by the statistical classifier. The reason behind those varying performance might lay in the feature selection process that tends to pick features that are representative of the dominant applications in the considered dataset. More generally, we identified a number of challenges for traffic classification that current approaches fail to correctly address:

- A feature selection strategy that selects for each specific (family of) application(s) a *distinct* set of features that best discriminates it from the rest of the

traffic.

- The ability to zoom in and out in the traffic as the focus might be on a family of applications like all applications using the HTTP protocol, or on specific applications like HTTP streaming or HTTP Chat.

- Resilience to the problem of data over-fitting observed in cross-site studies [21] whereby the statistical classifier capture so-called local information, like port numbers of p2p applications used by local users, that are detrimental when the classifier is applied on a site different from the one where it was trained.

- A classification method with a low computation cost that is further able to work in real time, i.e., after the observation of the first few packets of a connection.

In this paper, we propose to cast any traffic classification question as a logistic regression problem (Section 3). Using this approach, we develop a method that responds to the above challenges.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 provides formal statements of the problems we address, the background on logistic regression, and the classification process. Section 4 explains how we obtained and processed the data for our validation experiments, and Section 5 provides the results from our experimentation with real traffic. Section 6 summarizes the work and indicates future avenues of research.

2 Related work

Many different methods have been introduced to classify traffic. Traditional approaches relied on port numbers. However, early works [10, 15] quantitatively demonstrated the decrease of accuracy of conventional classification methods based on port numbers. It triggered the emergence of deep packet inspection (DPI) solutions that identify the application based on signatures found in packet payloads or connection patterns [24, 23]. The increasing use of encryption and obfuscation of packet content, the need of constant updates of application signatures and governments regulations, might however undermine the ability to inspect packets content.

To address these problems, recent studies have relied on statistical classification techniques to probabilistically map flows and applications [14, 3, 2, 13, 16, 18, 6, 4]. Hereafter, we cite a representative sample of traffic classification research. For a much more complete survey, see the work by Nguyen et al. [25].

Moore et al. in [16] presented an approach based on a naive Bayes classifier to to solve the classification problem of TCP traffic. They used a correlation-based filtering algorithm to select the 10 most relevant flow-behaviour features. The result-

ing accuracy, between 93% and 96%, demonstrated the the discriminative power of a combination of flow features and machine learning algorithms.

Bernaille et al. presented in [3] an approach for early identification of applications using start-of-flow information. The authors used the size and direction of the first 4 data packets and port numbers in each flow as features on which they trained K-means, Gaussian mixture model and spectral clustering respectively. Resulting clusters were used together with labeling heuristics to design classifiers. Their results have shown that information from the first packets of a TCP connection are sufficient to classify applications with an accuracy over 90%. The authors further specialized their work to the identification of encrypted traffic in [2].

Karagiannis et al. [11] studied traffic behavior by analyzing interactions between hosts, protocol usage and per-flow features. Their techniques were able to classify 80%-90% of the traffic with a 95% accuracy. In their recent work [12], they applied those techniques to profile the users activities, and to analyze the dynamics of host behaviours.

Pietrzyk et al. [21] investigated the use of statistical classification algorithms for operational usage. They point out that data over-fitting is a main weakness of statistical classifiers. Indeed, even if a classifier is very accurate on one site, the resulting model cannot be applied directly to other locations. This problem stems from the statistical classifier learning site specific information.

Nechay et al [17] presented tow approaches based on Neyman-Person classification and Learning Satisfiability framework, that allow to set class-specific performance guarantees. Their experiments indicate that these techniques almost achieve the specified constraints, at the expense of a very slight reduction in overall accuracy. However they used a 3G dataset that contains mainly web traffic and mails.

There exists also a lot of works focusing on specific applications. For example, Mellia et al. [4] showed an interesting approach specifically intended to identify Skype traffic by recognizing specific characteristics of Skype. A number of papers have been published focusing on the identification of p2p traffic [10, 9, 5].

In most of the above mentioned studies, the authors collected a large number of flow statistics and selected a subset of them that maximizes the overall accuracy. This raises a number of problems:

- These methods choose the parameters that are the most relevant to the most popular application in the dataset, which is in general Web traffic. It seems more reasonable to choose for each application a subset of features that maximizes its detection accuracy.

- The majority of datasets originate from the academic world. They thus lack the diversity of end users Internet traffic.

- None of these techniques is able to handle qualitative parameters. An example of qualitative parameters is the port numbers. Although it has been

used as quantitative parameter in previous works [13] due to its strong predictive power, it must be used with caution as, e.g., for p2p application, the statistical classifier will learn the port numbers of the local users [21]. Using the port number as a qualitative parameter, e.g. , "Is the port number a well-known port number?" allows to retain the good property of this feature while avoiding its drawbacks, as we will see later.

3 Learning classifier using Logistic regression

The use of logistic regression modeling has exploded during the past decade. From its original use in epidemiological research, the method is now commonly used in many fields including but not nearly limited to biomedical research [27], business and finance [26], criminology [28], health policy [27] and linguistics [20]. Logistic regression is designed for dichotomous variables, i.e., to model the relation between a binary variable (true vs. false) and a set of covariates.

In this work we use logistic regression to classify flows of a given application against the rest of the flows. In the remaining of this section, we introduce the logistic regression model. We show how to estimate its parameters for a given application, how to validate the model and finally, how the model selects the relevant features for the classification of a specific application.

3.1 Problem statements

The problem of traffic classification consists in associating a class to a network flow, given the information or features that can be extracted from this flow. A flow is defined as a sequence of packets with the same source IP address, destination IP address, source port, and destination port. Let X be the n-dimensional random variable corresponding to the flow features. To each flow a vector x consisting of the n the measured features is associated. Each flow is generated by an application y corresponding to a random variable Y that takes values in the set $\{1, 2, \cdots, c+1\}$, where c is the number of applications. This defines $c + 1$ classes; each application defines a class and the $(c + 1)^{th}$ class is the default class that contain flows that cannot be associated with any application. The problem of statistical classification is to associate a given flow x with an application y. Logistic regression is a way of defining the relation between x and y. While using logistic regression, we will consider only one application (we call it A) at a time, i.e. $Y = 1$ if the flow is generated by the application of interest and 0 otherwise.

3.2 Logistic regression model

Consider a flow with the following features vector $x = (x_1, x_2, \cdots, x_n)$. We wish to have a probability of whether this flow is generated by application A or

not. Formally, we can state this as[1]

$$p(Y = 1|X = x) = P(x, \beta), \tag{1}$$

where $p(Y = 1|X = x)$ is the conditional probability that the flow with features $x = (x_1, x_2, \cdots, x_n)$ is generated by the application A and P is a function of x parametrized by the weights vector $\beta = (\beta_0, \beta_1, \cdots, \beta_n)$. Since the function P represents a probability, it must take value between 0 and 1. Within the Logistic regression framework, one assumes a specific function P:

$$P(x, \beta) = \frac{e^{\beta_0 + \sum_{j=1}^{n} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{j=1}^{n} \beta_i x_i}}, \tag{2}$$

From the above equation, we can derive a linear function between the odds of having application A and the features vector x, called the logit model:

$$\log\left(\frac{P(x, \beta)}{1 - P(x, \beta)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n, \tag{3}$$

Unlike the usual linear regression model, there is no random disturbance term in the equation for the logit model. That does not mean that the model is deterministic because there is still room for randomness in the probabilistic relationship between $P(x, \beta)$ and the application A.

To implement any logistic regression model, one needs to choose the $\beta_1, \ldots, \beta_n$ values based on a given training set, i.e., a set of flows for which we know whether they have been generated by A or not. We discuss this issue in the next section.

3.3 Parameter estimation

Assigning the parameters to the logit model boils down to estimating the weights vector β, which is usually done using maximum likelihood estimation.

Consider a training data set of N flows characterized by the features vectors $X = (X_1, X_2, \cdots, X_n)$, where $X_i = (x_1^i, x_2^i, \cdots, x_n^i)$ is the features of flow i, and let the vector $Y = (y_1, y_2, \cdots, y_n)$ be such that $y_i = 1$ if flow i is generated by the application A and $y_i = 0$ otherwise. The likelihood function is given by a standard formula [7]

$$\begin{aligned} P(X, \beta) &= \prod_{j=1}^{N} p(Y = y_j|X_j) \\ &= \prod_{j=1}^{N} (p(Y = 1|X_j)^{y_j} (1 - p(Y = 1|X_j))^{1-y_j} \end{aligned} \tag{4}$$

[1]Please note that, for the sake of clarity, we avoided indexing of many variables with the application A. However we would like to point out the fact that the following procedure is done for each application of interest. In particular, it leads to β vectors that are application dependent.

As the values of p are small, it is common to maximize the log-likelihood $L(X, \beta) = \log P(X, \beta)$ instead [7], to avoid rounding errors,

$$L(X, \beta) = \sum_{j=1}^{N} [y_j log(p(Y = 1|X_j)) + (1 - y_j) log(1 - p(Y = 1|X_j))] \quad (5)$$

By substituting the value of $p(Y = 1|X_j)$ by its value defined in Equation (2) we get the log-likelihood for the logistic regression:

$$L(X, \beta) = \sum_{i=1}^{N} \left[y_i \beta^T X_i - log(1 + e^{\beta^T X_i}) \right] \quad (6)$$

In the logistic regression model, we wish to find β that maximizes Equation (6). Unfortunately, this can not be achieved analytically. In this work, we compute it numerically using the Newton-raphson algorithm [7]. This algorithm requires two main components: the first derivative of the loglikelihood and the Hessien matrix, i.e., the second derivative matrix with respect to β.

From Equation (6) we can derive the first derivative

$$\frac{\partial L(X, \beta)}{\partial \beta} = \sum_{i=1}^{N} X_i (y_i - p(x_i, \beta)) \quad (7)$$

We now derive the Hessien matrix

$$\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T} = - \sum_{i=1}^{N} X_i X_i^T p(x_i, \beta)(1 - p(x_i, \beta)) \quad (8)$$

The pseudo code of Newton-Raphson algorithm is depicted in Algorithm 1. We start with a first guess of β, then we use the first derivative and the Hessien matrix to update β. Using the new β we compute the new loglikelihood. This is repeated until there is no further change of β. The Newton-Raphson algorithm has been shown to converge remarkably quickly [8]. In this work, it takes less than one second to output an estimate of β.

3.4 Overall goodness of fit of the model

The above estimation procedure leads to a set of
$\beta_i, 1 \leq i \leq N$ values optimized according to a given strategy, namely maximum likelihood estimation. It might however be that the set of input features is not an appropriate one, which leads to statistically non significant β_i values. In this section, we introduce a procedure to test the overall significance of the model. Formally, we want to statistically test if the β_i values that have been evaluated are not equal to zero.

Algorithm 1 Newton-Raphson algorithm

initialize β
while $\|\beta_{new} - \beta_{old}\| > thr1$ and abs$(L_{new} - L_{old}) > thr2$) **do**
    Calculate $g = \partial L / \partial \beta$
    Calculate $H = \partial^2 L / \partial \beta^2$
    Set $\beta_{old} = \beta_{new}$
    Calculate $\beta_{new} = \beta_{old} - H^{-1} g$
    Set $L_{old} = L_{new}$
    Calculate L_{new}
end while
Calculate variance matrix $\hat{V}$

In practice, several different measures exist for determining the significance, or goodness of fit, of a logistic regression model. These measures include the G statistic [7], Pearson statistic [7], and Hosmer-Lemeshow statistic [7]. In a theoretical sense, all three measures are equivalent. To be more precise, as the number of rows in the predictor matrix goes to infinity, all three measures converge to the same estimate of model significance.

In this work, we use the G statistic, which is defined as the deviance of the intercept-only model from the whole model:

$$G = -2 \log \frac{\text{likelihood without the variables}}{\text{likelihood with the variables}}. \tag{9}$$

Under the hypothesis of all $\beta_j = 0$, the G statistic follows a chi-squared distribution χ^2_{n-1} with $n-1$ degrees of freedom, where n is the number of parameters in the model [8]. To test the significance of our model, we use classical statistical test [19]. First we decide the *null hypothesis* : $\beta_j = 0 \; for \; j = 1, ...n$. Then we compute the p-value pv:

$$pv = p(\chi^2_{n-1} > G). \tag{10}$$

For a given significance level α, we reject the *null hypothesis* if $\alpha > pv$. Failing to reject the null hypothesis means that the features are not suitable for a good classification of the application of interest. Thus, we have a quick way to judge the quality of our classifier.

3.5 Selection of relevant features

At first sight, it might seem that a model (set of input features and β_i values) is good if it fits the observed data very well, i.e. it can accurately classify the flows in the training set. By including a sufficiently large number of features in our model, we can, in theory, make the fit as close as we wish. However, simplicity, represented by the minimum number of parameters, is a desirable feature of any model. We thus would like to include as little features as possible to perform the classification. Reducing the number of input features can be done through

the formulation and testing of a statistical hypothesis to determine whether the corresponding variables in the model are "significantly" related to the outcome variable Y. In other words, for each feature j, we test the hypothesis that the corresponding weight β_j is equal to zero. If we can't reject this hypothesis, this means that this parameter is not relevant to classify this application and, thus, can be removed from the model [8].

In this work, we use the Wald test [8] that tests, individually, for each β_j the null hypothesis that $\hat{\beta}_j = 0$. The Wald statistic $W(j)$ is obtained by comparing the maximum likelihood estimate of each parameter $\hat{\beta}_j$ to an estimate of its standard deviation $\hat{V}(\hat{\beta}_j)$.

$$W(j) = \frac{\hat{\beta}_j}{\hat{V}(\hat{\beta}_j)} \tag{11}$$

The standard deviation $\hat{V}(\hat{\beta}_j)$ of β_j is given by the j^{th} diagonal element of the variance matrix given by Equation (12) [7], that is computed as the last iteration of the Newton-Raphson algorithm (Alg. 1).

$$\hat{V} = \left\{ -\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T} \right\}^{-1} \tag{12}$$

Under the *null hypothesis* that $\beta_j = 0$, $W(j)$ follows a standard student *t-distribution* with $n - 1$ degree of freedom t_{n-1}.

For a given significance level α, for each β_j we compute the p-value $pv_j = p(t_{n-1} > W(j))$, and we reject the hypothesis of $\beta_j = 0$ if $\alpha > pv_j$. Otherwise, if we fail to reject the hypothesis of $\beta_j = 0$, we exclude the corresponding feature from our model. By doing so, we keep a minimum number of features relevant to the application under study.

3.6 Classification process

Logistic regression falls into supervised type of machine learning [25], thus it consists of two main steps:

3.6.1 Training

Algorithm 2 describes the learning process for a given application A. First, we estimate β using Newton-raphson algorithm using all the features. Then we test the hypothesis $\beta = 0$ using the G-statistic introduced in Section 3.4. If the test is rejected, we proceed by selecting the relevant parameters using the *student* test as explained in section 3.5. Then, we estimate the new β using only the set of selected features.

A crucial aspect of using logistic regression is the choice of an α (see section 3.4) level to judge the model and the importance of features. Bendel et al [1] have

Set	Date	Start	Dur	Size [GB]	Flows [M]	TCP [%]	TCP Bytes [%]	Local users	Distant IPs
MS-I	2008-02-04	14:45	1h	26	0.99	63	90.0	1380	73.4 K
R-III	2008-02-04	14:45	1h	36	1,3	54	91.9	2100	295 K

Table 1: Traces summary

shown that the choice of α smaller than 0.01 is too stringent, often excluding important variables from the model. In this work, we use $\alpha = 0.01$, and we will show in section 5.4 that it enables to reduce the number of features for each application without decreasing the classification scores.

Algorithm 2 Parameters estimation and features selection

1: Estimate β
2: Test of model significance,
3: If the hypothesis $\beta = 0$ is rejected
4: Select the relevant features for the application
5: Estimate the new β

3.6.2 Classification

A given feature vector $x = (x_1, \cdots, x_p)$, is classified as generated by application A if $P(x, \beta)$ is larger than a threshold th. The usual choice in statistic is $th = 0.5$ [8, 7]. By using Equation (3), this boils down to deciding the new flow x is generated by the application A if $\sum_{i=1}^{n} x_1 \beta_i > 0$.

The choice of $th = 0.5$ is very conservative, as the logistic regression has a strong discrimination power. For example, Figure 2 shows the cumulative distribution functions of the probability $p(y = p2p|x)$ for P2P and non-P2P flows in one of the trace used in Section 5. A choice of th corresponds to a vertical line at value th on the x-axis. Figure 2 shows that the classification in p2p/non-p2p is almost unaffected by the exact th value. Indeed, more than 80% of non-p2p flows have a probability to be p2p flow less than 0.01, and more than 90% of p2p flows have a probability of being a p2p larger than 0.95. This is even more pronouced in the case of HTTP flows (Figure 1) where 99% of non-HTTP flows have a probability of being HTTP flows less than 0.005, and more than 90% of HTTP flows have a probability larger than 0.99. These figures show clearly that the choice of a larger threshold would change only slightly the classification results.

4 Experiment setting

In this section, we present our dataset, how we establish the reference point (ground truth) that is used as benchmark for our statistical classifier, the definition of our traffic classes and the traffic breakdown.

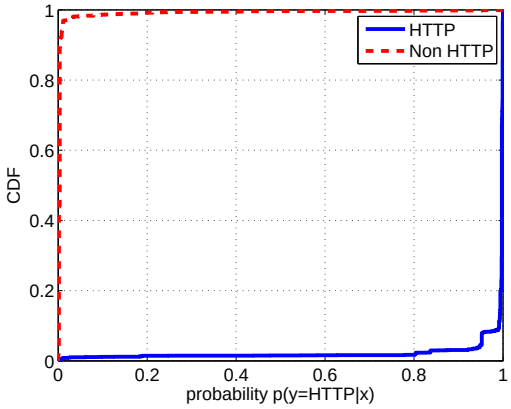

Figure 1: CDFs of the probability of being a HTTP flows for HTTP flows and Non HTTP flows. Training and test data are from R-III trace (see table 3)

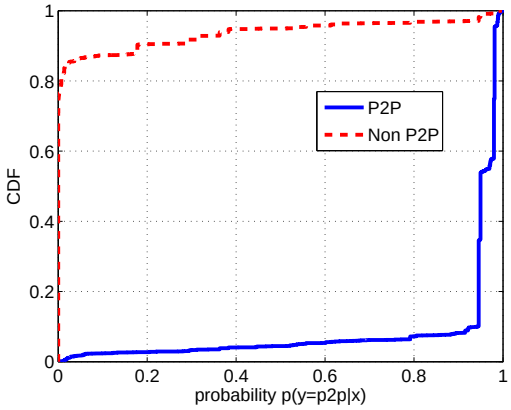

Figure 2: CDFs of the probability of being a P2P flows for P2P flows and Non HTTP flows. Training and test datasets are from R-III trace (see table 3)

4.1 Datasets

Our dataset consists of two recent packet traces collected at two *different* ADSL Points of Presence (PoPs) in France from the same ISP. Both traces were collected at *the same time* using passive probes located behind a Broadband Access Server (BAS), which routes traffic to and from the digital subscriber line access multiplexers (DSLAM) to the Internet. Captures were performed without any sampling or loss. Traces contains one hour of full bidirectional traffic, with similar number of active users. More details are provided in table 1.

4.2 Application breakdown

In order to benchmark the performance of any classification method, a dataset with pre-labeled classes of traffic is needed. We term such a dataset our reference point. Establishing a correct reference point is fundamental when evaluating traffic classification mechanisms to provide trust-worthy results. As a human-labeled dataset is almost impossible to have, we rely on DPI tools. In [22], We have compared an internal tool of Orange, that we term Orange_DPI_Tool or ODT for short, to Tstat [24], whose latest version features DPI functions. ODT and Tstat v2 offer similar performance and outperform signature based tools used in the literature [13, 6]. More details about the reference point issue can be found in [22].

To label applications in our dataset, we rely on ODT. ODT is constantly under development and is in use on several PoPs of Orange in France. It can detect several types of applications, including encrypted ones.

Traffic classes used in this work are summarized in Table 2. Breakdown of traffic is presented in Tables 3 and 4. Traffic proportions are very different in both locations even though both traces were collected in the same country and at the same time. Web and eDonkey are the dominant classes in terms of flows while in terms of bytes, these are Web, eDonkey and HTTP streaming, the latter reflecting the popularity of streaming service providers like YouTube. While HTTP traffic is broken into many classes, it is important to note that the most important ones for HTTP applications in our datasets are Web browsing, HTTP-streaming and HTTP chat. We will term those three categories as HTTP in Section 5, neglecting the minority of HTTP flows in the mail and games classes.

4.3 Flow Definition

We restrict our attention to TCP flows as they carry the vast majority of the bytes in both traces. We are still left with the issue of defining the set of flows to be analyzed. Restriction is imposed by the classification method itself as we are using as features information derived from the first 4 data packets. We de facto exclude all flows with less than 4 data packets as well as the ones for which we did not observe the initial three way handshake. This typically leaves around 70% of volume for the analysis. Details about the impact of the flow definition on the amount of data excluded for each application class can be found in [21].

5 Evaluation

5.1 Flow Features

Most studies on traffic classification rely on statistics computed once all the packets of a flow have been observed, e.g., duration, number of packets, mean packet size, or inter-arrival time [25]. This clearly prevents any online classification. In contrast, we evaluate the feasibility of application identification in the early

Class	Application/protocol
WEB	HTTP and HTTPs browsing
HTTP-STR	HTTP Streaming
EDONKEY	eDonkey, eMule obfuscated
BITTORRENT	Bittorrent
GNUTELLA	Gnutella
CHAT	MSN, IRC, Jabber
	Yahoo Msn, HTTP Chat
MAIL	SMTP, POP3, IMAP, IMAPs
	POP3s, HTTP Mail
FTP	Ftp-data, Ftp control
GAMES	NFS3, Blizzard Battlenet, Quake II/III
	Counter Strike, HTTP Games
STREAMING	Ms. Media Server, Real Player
	iTunes, Quick Time
OTHERS	NBS, Ms-ds, Epmap, Attacks
UNKNOWN	-

Table 2: Application classes

stage of a connection. A few works have tackled this challenge. In particular, [3] and [14] showed that statistical features extracted from the first k packets of each connection, where k is typically in the range of 4 to 5 packets, lead to a good overall classification performance. We however uncovered in [21] some weaknesses of those approaches related to the ability to detect some key applications like HTTP streaming, which is gaining in popularity and a data overfitting issue when one wants to apply a classifier on a trace collected on a location different from the one it was trained on. The latter situation could typically be the one of an ISP that trains the classifier on its major PoP, where DPI tools are available, before deploying it on its other PoPs. We will show in this section that logistic regression is able to overcome those weaknesses.

The choice of flow level features turns out to be a major task in traffic classification. As explained before, state of the art approaches often rely on a preliminary feature selection phase, e.g. the correlation based filter technique in [13, 16]. Such method outputs a single set of features which is *the same* for all applications. In contrast, logistic regression picks for each application of interest *distinct features* that best separates it from the rest of the traffic.

As we want to evaluate the ability of logistic regression to perform traffic classification on the fly, we selected an initial set of features that can be computed by the observation of the beginning of the flow: size and direction of the first 4 data packets, presence of push flags and port numbers. Out of this set, logistic regression picks the most relevant features for each application. Size and direction of the

first data packets have been shown to lead to good classification results in [3]. We enrich this set with a push flag indicator that indicates whether a data packet has its PUSH flag set or not.

We thus end up having a mix of quantitative and qualitative features. While logistic regression can handle both types of parameters, it is recommended to transform quantitative parameters into qualitative ones [8]. We proceeded as follows:

- **Size of data packets**: we classify each data packet as small or not small packet. We used a fixed threshold, derived from empirical distributions of packet sizes, of 200 bytes for all applications and all traces.
- **Port numbers**: the quantization technique used depends on the application of interest. For applications using the HTTP protocol, we assign the port variable to 1 if the source or destination port number belongs to the set $80, 8080, 443$ and 0 otherwise. For P2P applications, we assign the port variable to 1 if both the source and destination ports are above 1024. Note that other quantization strategies are possible. For instance, for p2p applications, one could have used legacy port numbers of considered p2p applications. It turned out however that the quantization technique we use, which makes no use of such a priori information, offers satisfactory results.

Table 3: Traffic breakdown RIII. For flows ≥ 4 data packets

	Flows		Size	
	Number	%	MB	%
WEB	160802	49.16	5519.56	24.61
HTTP-STR	4282	1.31	2654.14	11.84
EDONKEY	119057	36.40	8295.35	36.99
BITTORRENT	8789	2.69	1529.69	6.83
GNUTELLA	4718	1.44	1093.83	4.89
CHAT	4365	1.33	46.66	0.22
MAIL	4206	1.29	244.47	1.10
STREAMING	679	0.21	451.09	2.02
FTP	437	0.13	156.06	0.71
GAMES	182	0.06	3.87	0.02
OTHERS	835	0.26	12.54	0.07
UNKNOWN	18501	5.66	2248.00	10.03

5.2 Performance metrics

We present results in terms of True Positives (TPs) and True Negatives (TNs) ratios. These notions are defined with respect to a specific class. Let us consider such a specific class, say the HTTP streaming class. TPs are the fraction of HTTP

Table 4: Traffic breakdown MSI. For flows ≥ 4 data packets

	Flows		Size	
	Number	%	MB	%
WEB	319009	78.91	8368.85	52.41
HTTP-STR	6901	1.71	2777.43	18.72
EDONKEY	23212	5.75	1106.59	9.06
BITTORRENT	2313	0.57	649.81	4.15
GNUTELLA	223	0.06	104.19	0.66
CHAT	7539	1.87	86.87	0.55
MAIL	18406	4.56	856.33	5.46
STREAMING	207	0.05	372.43	2.39
FTP	1129	0.28	470.52	3
GAMES	183	0.05	1.68	0.02
OTHERS	8803	2.19	196.23	1.25
UNKNOWN	13535	3.36	275.96	1.76

streaming flows that are labeled as such by the statistical classifier, i.e., logistic regression. TNs are the fraction of flows not labeled as HTTP streaming by our DPI tool that are also not labeled as HTTP streaming by logistic regression. For an ideal classifier, TPs and TNs should be equal to 100%.

5.3 Overall Performance

For both traces we have, the logistic regression achieve overall TPs and TNs ratios over 98% and 97% respectively (when training set and testing set are from the same trace). These results are similar to the results obtained by most statistical classifier, see [25]. This is because dominant applications like web or edonkey are well classified in all cases.

In the next sections, we will focus on two sets of applications: (i) applications that use the HTTP protocol like Web browsing, HTTP streaming (e.g., YouTube) and HTTP chat and (ii) p2p applications. In each case, we will evaluate the ability of logistic regression to either detect the whole family, e.g. all HTTP applications or specific members like HTTP streaming. Please note that in each experiment, including cross site case, we use only 5% of flows for training and the remaining for testing.

5.4 HTTP driven applications

In this section we focus on the HTTP applications found in our datasets, namely: Web browsing, HTTP streaming and HTTP chat.

In Table 5, we present on the right column ('before selection'), TPs and TNs ratios for all HTTP applications taken together and each type of HTTP application in isolation for MS-I trace. We observe very high TPs and TNs for the 'All HTTP'

and 'Browsing' cases and high values for 'HTTP streaming'. The latter result is a noticeable one as, to the best of our knowledge, no statistical classification technique has been able so far to isolate HTTP streaming traffic only – see for instance [21] where the features selected in [3] and [14] are used on dataset MS-1 and lead to poor TNs results.

The left column of Table 5 shows results of logistic regression where only features corresponding to statistically significant β values are considered. We do observe no significant changes before and after the selection procedure. This reveals that logistic regression indeed gives no significance to the parameters that have no discriminative power for the considered applications or set of applications. Thus, we can safely the non relevant features without accuracy degradation which reduces the computational cost of the classification.

The list of selected features is presented in Table 7. We observe that the set of features kept for HTTP applications is (almost) the intersection of the ones kept for each individual HTTP applications. Indeed, logistic regression selects, for a given application, the features that maximize the difference between the flows of this application and the rest of the flows in the datasets. When focusing on HTTP streaming, it might thus use most of the specific features used for detecting all HTTP applications and add a few additional ones (e.g., the push variable for the third data packet here) to further differentiate those flows from other flows. Conversely, when logistic regression has to handle all HTTP applications, it keeps only features that allow to distinguish those flows from the non HTTP flows in the dataset, thus getting rid of features that might be important to specifically detect HTTP chat or HTTP streaming for instance.

Table 5: The percentage of true positives (TP) and true negatives (TN) of HTTP flows using all the features (before selection) and only the features selected by the algorithm

	after selection		before selection	
	TP	TN	TP	TN
All HTTP	99%	99%	98%	99%
Web	98%	97%	98%	97%
HTTP streaming	83%	84%	83%	84%
HTTP Chat	94%	98%	94%	98%

5.5 P2P Application

In this section, we focus on the p2p applications observed in our datasets. As for the HTTP case, we obtain similar results for all applications for our two traces. In Table 6, we present results for the MS-I trace. Logistic regression achieves very good performance in all cases. The lowest value is the TPs ratio of Gnutella. However, even in this case, we limit the risk of misclassifying a flow as Gnutella

as the TNs ratio is very high. The only risk is to miss a small fraction of actual Gnutella transfers.

Table 6: The percentage of true positives (TP) and true negatives (TN) of P2P flows using all the features (before selection) and only the features selected by the algorithm

	After selection		Before selection	
	TP	TN	TP	TN
All P2P	96%	95%	96%	95%
eDonkey	97%	96%	97%	95%
BitTorrent	88%	96%	88%%	96%
Gnutella	83%	98%	83%	98%

Table 7: The set of selected features for each application

	1st packet			2nd packet			3rd packet			4th packet			port number
	direction	push	size	direction	push	size	direction	push	size	direction	push	size	
All HTTP			X	X	X		X			X	X		X
Web		X	X	X	X		X			X			X
HTTP streaming			X		X		X	X		X	X	X	X
All P2P	X	X	X	X	X	X	X	X	X	X		X	X
eDonkey	X		X	X		X	X	X	X	X	X	X	X
BitTorrent	X			X	X	X		X	X	X	X	X	X
Gnutella			X	X	X	X					X		X

5.6 Cross-site Evaluation

We performed a cross-site evaluation where, for each case (application or set of application), we train the classifier, using the selected features given in table 7, on one trace, e.g., MS-I and apply it on the other trace, e.g., R-III. Such a validation is important for practical usage of any classifier as it verifies whether the statistical model we build is representative of application and does not incorporate site dependent data.

We present the corresponding results in Table 8. We observe good performance in all cases. [2]. While this result was to be expected in the case of HTTP applications, it constitutes a major achievement in the case of p2p applications as it was demonstrated in [21] that a data overfitting issue could occur with p2p applications. The latter stems from the fact that the classifier learns ports used by p2p applications of local users, which then fool the classifier when the set of local users is changed. We attribute the good performance observed here with logistic regression to the quantization technique used for the port number that gets rid of specific

[2]The only exception is Gnutella when training is done on MS-I and testing on R-III. This is because we have only 223 Gnutella flows in trace MS-I in contrast to 4718 in the other trace and with only 223 flows, we apparently miss part of the diversity of this class. Note that when training and testing is done in the other direction, the TP ratio reaches 84%, as now we have a higher diversity in the training set

port values but simply check if the two ports correspond to well-known ports or not.

To further investigate this hypothesis, we applied again logistic regression for each trace and for the cross site test using the initial port number rather than its quantized version. As expected, we observed slightly worse performance on a trace basis and significant performance decrease in the cross site case. A sticking example is the one of Gnutella whose TPs ratio decreases from 83% to 70% on R-III trace when no discretization is applied and from 84% to 42% when the logistic regression algorithm is trained on R-III and applied to MS-I.

As a conclusion, the ability of logistic regression to handle qualitative and not only quantitative values as well as per application feature selection enables us to minimize the risk of data over-fitting in cross site studies that were observed in previous work.

Table 8: The percentage of true positives (TP) and true negatives (TN) in cross case : training on 5% of flows from MS-I dataset (resp. R-III) and test on all the flows from R-III (resp. MS-I) dataset.

	R-III to MS-I		MS-I to R-III	
	TP	TN	TP	TN
All HTTP	98%	99%	99%	99%
Web browsing	95%	91%	98.5%	96%
HTTP Streaming	80%	81%	90%	82%
HTTP Chat	75%	98%	75%	98%
All P2P	94%	91%	90%	95%
eDonkey	97%	95%	94%	96%
BitTorrent	89%	96%	88%	96%
Gnutella	84%	98%	22%	99.7%

6 Conclusion and Future Work

In this paper, we have proposed a novel on-line classification algorithm based on the logistic regression model. It is a flexible classification framework that overcomes important weaknesses of state of the art methods proposed so far. We have validated the performance of the proposed methods using ADSL traffic traces obtained from a French ISP. This method incorporates the following new features:

- It automatically selects the best possible subset of distinct features relevant to each (family of) application(s).
- It can be used for application based or protocol based classification. For instance, it can classify all P2P file-sharing at once, or focus on one of them only, e.g., eDonkey.

- It can handle both quantitative and qualitative features, while current approaches are able to handle quantitative features only. This is important as some features might be more useful when considered as qualitative rather than quantitative information. For instance, the port numbers are more useful when considered as qualitative indicators.

- Due to its ability to handle qualitative and not only quantitative features, it can be made resilient to the data over-fitting problem encountered in cross-site studies: it can be trained on data collected on one location and used for traffic data from other sites. This turns out to be a very useful feature for companies or ISPs managing several sites.

- It has a constant and low computational cost as logistic regression boils down to comparing a linear combination of the flow features with a fixed threshold to take its classification decision.

- It can work in real-time as it needs to consider features extracted from the first four data packets of a transfer only to take an accurate classification decision.

We consider a number of future extensions to this work. So far we considered TCP traffic only, however with the growing trend of UDP traffic, we would like to generalize the method to handle UDP traffic as well. We also intend to address issue of temporal stability of the classifier,i.e., determining what is a correct retraining strategy.

References

[1] RB Bendel and AA. Afifi. Comparison of stopping rules in forward regression. *Journal of the American Statistical Association*, pages 46–53, 1972.

[2] Laurent Bernaille, Renata Teixeira, Universit Pierre, and Marie Curie Lipcnrs. Early recognition of encrypted applications. In *In Passive and Active Measurement conference (PAM 07)*, 2007.

[3] Laurent Bernaille, Renata Teixeira, and Kave Salamatian. Early application identification. In *CoNEXT '06: Proceedings of the 2006 ACM CoNEXT conference*, pages 1–12, 2006.

[4] Dario Bonfiglio, Marco Mellia, Michela Meo, Dario Rossi, and Paolo Tofanelli. Revealing skype traffic: when randomness plays with you. *SIGCOMM Comput. Commun. Rev.*, 37(4):37–48, 2007.

[5] F. Constantinou and P. Mavrommatis. Identifying known and unknown peer-to-peer traffic. In *Network Computing and Applications, 2006. NCA 2006. Fifth IEEE International Symposium on*, pages 93–102, July 2006.

[6] Jeffrey Erman, Martin Arlitt, and Anirban Mahanti. Traffic classification using clustering algorithms. In *MineNet '06: Proceedings of the 2006 SIGCOMM workshop on Mining network data*, 2006.

[7] James W. Hardin and Joseph W. Hilbe. *Generalized Linear Models and Extensions, 2nd Edition*. StataCorp LP, 2007.

[8] David W. Hosmer and Stanley Leshow. *Applied Logistic Regression*. New York ; Chichester ; Brisbane : J. Wiley and Sons, cop., 2001.

[9] T. Karagiannis, A. Broido, N. Brownlee, K.C. Claffy, and M. Faloutsos. Is p2p dying or just hiding? [p2p traffic measurement]. In *Global Telecommunications Conference, 2004. GLOBECOM '04. IEEE*, pages 1532–1538 Vol.3, Nov.-3 Dec. 2004.

[10] Thomas Karagiannis, Andre Broido, Michalis Faloutsos, and Kc claffy. Transport layer identification of p2p traffic. In *IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 121–134, 2004.

[11] Thomas Karagiannis, Konstantina Papagiannaki, and Michalis Faloutsos. Blinc: multilevel traffic classification in the dark. *SIGCOMM Comput. Commun. Rev.*, 35(4), 2005.

[12] Thomas Karagiannis, Konstantina Papagiannaki, Nina Taft, and Michalis Faloutsos. Profiling the end host. In *In Passive and Active Measurement conference (PAM 07)*, April 2007.

[13] Hyunchul Kim, KC Claffy, Marina Fomenkov, Dhiman Barman, Michalis Faloutsos, and KiYoung Lee. Internet traffic classification demystified: myths, caveats, and the best practices. In *CONEXT '08: Proceedings of the 2008 ACM CoNEXT Conference*, pages 1–12, 2008.

[14] Wei Li, Marco Canini, Andrew W. Moore, and Raffaele Bolla. Efficient application identification and the temporal and spatial stability of classification schema.

[15] Andrew W. Moore and Konstantina Papagiannaki. Toward the accurate identification of network applications. In *In Passive and Active Measurement conference (PAM 05)*, pages 41–54, 2005.

[16] Andrew W. Moore and Denis Zuev. Internet traffic classification using bayesian analysis techniques. In *SIGMETRICS '05: Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 50–60, 2005.

[17] D. Nechay, Y. Pointurier, and M. Coates. Controlling false alarm/discovery rates in online internet traffic flow classification. In *INFOCOM 2009, IEEE*, April 2009.

[18] T.T.T. Nguyen and G. Armitage. Training on multiple sub-flows to optimise the use of machine learning classifiers in real-world ip networks. pages 369–376, Nov. 2006.

[19] Thomas W. O'Gorman. *Applied adaptive statistical methods; Tests of Significance and Confidence Intervals*. Society for Industrial Mathematics, 1987.

[20] John C. Paolillo. *Variable Rule Analysis: Using Logistic Regression in Linguistic Models of Variation*. Chicago University Press, 2002.

[21] Marcin Pietrzyk, Jean-Laurent Costeux, Taoufik En-Najjary, and Guillaume Urvoy-Keller. Challenging statistical classification for operational usage : the adsl case. In *IMC '09: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, 2009.

[22] Marcin Pietrzyk, Guillaume Urvoy-Keller, and Jean-Laurent Costeux. Revealing the unknown adsl traffic using statistical methods. In *COST 2009 : Springer : Lecture Notes in Computer Science, Vol 5537, 2009.*, May 2009.

[23] Bro Intrusion Detection System. http://bro-ids.org.

[24] Tstat. http://tstat.tlc.polito.it/.

[25] G Armitage TTT Nguyen. A survey of techniques for internet traffic classification using machine learning. *Communications Surveys and Tutorials, IEEE*, 10(4):56–76, 2008.

[26] Jon Tucker and Dr Jon Tucker. Neural networks versus logistic regression in financial modelling: A methodological comparison. In *in Proceedings of the 1996 World First Online Workshop on Soft Computing (WSC1*, 1996.

[27] Eric Vittinghoff, Davis V. Glidden, and Stephen C. Shiboski. *Regression methods in biostatistics : linear, logistic, survival, and repeated measures models*. Springer New York, 2005.

[28] Jeffery T. Walker and Sean Maddan. *Statistics in criminology and criminal justice, Third Edition*. Sudbury, Mass., USA, Jones and Bartlett Publishers, 2009.

[29] Haiyong Xie, Y. Richard Yang, Arvind Krishnamurthy, Yanbin Grace Liu, and Abraham Silberschatz. P4p: provider portal for applications. *SIGCOMM Comput. Commun. Rev.*, 38(4), 2008.