Strategies for Automatic Labelling of Web Traffic Traces

Luis Miguel Torres, Eduardo Magaña, Mikel Izal and Daniel Morato
Departamento de Automática y Computación, Universidad Pública de Navarra,
Pamplona, Navarra, Spain.
Email: [luismiguel.torres, eduardo.magana, mikel.izal, daniel.morato]@unavarra.es

***Abstract*—In the field of traffic classification, previous efforts have been centered on identifying applications (HTTP, SMTP, FTP, etc) rather than the actual services that they provide (e-mail, file transfer, video streaming, etc.). Nowadays, however, a single application as HTTP can provide multiple services for the end-user. Some methods have been proposed to distinguish between these services but tuning and testing them remains a challenge as there is no easy way to obtain labelled HTTP traffic traces. In this paper we present a method to discover server IP addresses related to a specific website in a traffic trace. Our method uses NetFlow-type records which makes it scalable an impervious to encryption of packet payloads. By applying the method to a representative set of websites the resulting list of IP addresses can be used to label a sizeable number of connections in the trace.**

I. Introduction

When designing any type of traffic classification system, one of the biggest challenges to overcome is the difficulty to find correctly labelled traffic traces that can be used as ground truth in order to tune or test the system. These traces are so hard to come by because the labelling process is far from trivial. If we want to label the traffic of different Internet applications, some simple (albeit somewhat unreliable) techniques exist from port mapping [1] to signature-based classification [2]. New techniques, more complex and resilient, have been proposed in the last years [3]. However, in our case, we seek to label traffic from just one application (the web) depending on the session (connection to a website) during which it was generated. This is an interesting, complex and less studied scenario.

We have chosen to work at flow level [4] for simplicity and scalability. For the captured flows, we consider each client IP address, which belongs to our network, as a *user*. We make the assumption that the server IP addresses of those flows can be mapped to a website. For some of them, this will not be true: websites may share a server provided by a common hosting service or may even share content from a third party server. However, a sizeable number of addresses should be related to just one website, at least during a relatively short period of time. We will, then, try to label not each TCP connection but groups of connections to the same IP addresses.

We define a *session* as a collection of TCP flows generated by the web browser while the user is accessing a specific website. For example, a session to a webmail site would ideally span all the connections opened by the web browser from the moment the user opened the login webpage of the mail service until he or she closes the browser, the tab or opens a different website in the same tab. However, the beginning and ending of a session are difficult to infer from the captured traffic. In order to do so, we further simplify by considering a predefined set of websites whose connections we want to find and label. At the same time of the traffic capture, we resolve the main domain names of these websites and store the obtained IP addresses. We will use their appareance in the trace as a signal of the beginning of a session to the related website.

We expect that by studying different sessions from different users to the same sites, we will be able to obtain a list of IP addresses related to the site. In the following sections we will present some parameters that can be useful to do so in a reliable way. We believe that by selecting an appropriated number of popular websites we can obtain a sufficient set of labelled connections that can be used to tune and test other systems.

II. Data

For the different tests presented in this paper, we use three traffic traces (in pcap format) captured in the Internet link of the Public University of Navarre (table I). As we are only interested in web traffic we filtered all non-TCP packets and those TCP packets whose destination port was not 80 or 443 (we only consider outbound connections). In order to obtain flow records we subsequently use Argus [5] to store basic information (timestamps, IP addresses, ports, size) for each bidirectional flow. This greatly reduces the initial volume of data, making it more manageable.

In adition to the traffic traces, we have collected DNS information for a predefined set of websites. We have selected 40 sites, some of them worldwide known and others which are popular in Spain or even in our local community (e.g. Tuenti, a Spanish social network or a number of local newspapers). At the same time of the traffic capture, we resolve the domain names of these sites with automatic hourly DNS requests and store the obtained IP addresses. We observed that, for websites that belong to the same company, sometimes servers that offered a website would later offer another. In fact, in those cases, the sites used to share a lot of content making it difficult to differentiate them with our method. We decided to group the websites from the same companies in groups reducing our list to 28 websites or groups of websites. This

978-1-4673-1564-7/12/$31.00 ©2012 IEEE

TABLE I
TRAFFIC TRACES

Trace	Date of capture	# Flows	# Users
Trace 1	Jun 5 - 20	90M	1022
Trace 2	Sep 29 - Oct 10	85M	916
Trace 3	Dic 14 - 27	76M	857

affects Google websites, Microsoft's and groups national and international versions of others.

III. APPROACH

Modern websites present dynamic content that may be stored in various different servers. Some of these servers may even provide content for different websites as it can be the case with content distribution networks. As a consequence, the IP addresses that are accessed when loading a website change over time and some of them may be used by more than one site. Nevertheless, intuitively, if we capture enough sessions to the same websites, a number of IP addresses are bound to start appearing repeatedly. Also intuitively, these IP addresses must belong to servers that store the fundamental content of the site as opposed to some images, videos, advertisements and other rapidly-changing or third party content. In this section we put to test these assumptions with experimental data from the September-October traffic trace which we have used to tune our system. The results for the other two traces are similar.

A. Web sessions

We have defined a web session as the set of connections generated by the web browser while the user is accessing a specific website. It should be noted that, even if we know when a user is accessing a website, we have no way to assess if the connections are truly caused by the load of that website. Users may open concurrent sessions and may use other applications that use the ports normally associated to HTTP(S). Because of that, our web sessions will comprise all the connections opened during a website access whether they are related to it or not.

We use the IP addresses that we obtained from the domain names of the websites as signals of the beginning of a session. We will call them *main IP addresses*. In other words, when we find a connection from a user to one of the main IP addresses, we consider that the user is visiting the corresponding website. The following connections initiated by that user will be considered part of the same session. Their server IP addresses will be then *candidate IP addresses* that, in the end, may or may not be associated with the website.

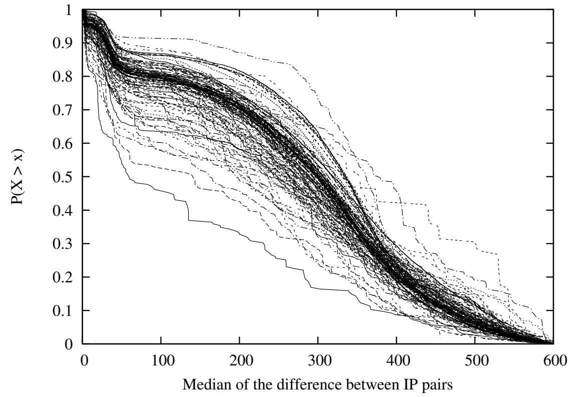


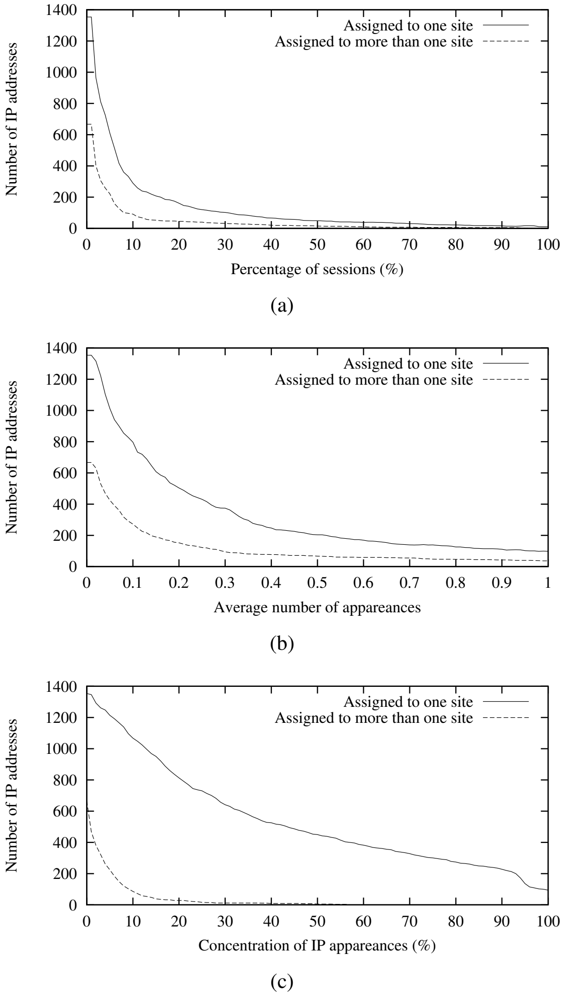
Fig. 1. Time differences between IP pairs

Choosing an indicator for session ending is not so simple. Sessions are variable in length depending on the type of service accessed and on user behaviour. We have chosen to set one fixed value for session duration. To set this value we have studied the time differences between the connections to a main IP and the next connections to candidate IP addresses made by the same user. We expect that connections to candidate IP addresses that do belong to a certain website will be on average closer to the main IP than those related to other websites. In this case the mean of the time differences proved to be too affected by extreme values so we opted to calculate the median for every main/candidate IP pair. In figure 1 each curve represents the distribution of these medians for a main IP limited to differences smaller than ten minutes. The flat area around 100 seconds suggests that connections to candidate IP addresses related to the main IP happen before, while the dots to the right of the graphic correspond to IP addresses that belong to other sessions. In view of this, we have chosen 120 seconds as a value for session length.

B. Decision Parameters

By applying the previous definition to a traffic trace we obtain a set of sessions. Each session is related to a user which, as defined, is the source IP address in every connection, and to a website depending on the main IP address. By considering all the sessions related to a website, we gather a list of candidate IP addresses for that website. As a first filter, for each website, we only consider candidate IP addresses that appear in at least 1% of the sessions (or at least two different sessions for websites with less than 200 sessions). This allows us to eliminate a big number of candidate IP addresses that are not strongly related to the website. With this step, in the October trace we obtain a total of 2011 candidate IP addresses for all the 28 websites. Out of these, 1355 appear only in sessions of one website and 656 appear in sessions of more than one website.

It is clear that, at this stage, we cannot assign the doubtful 656 IP addresses to any website. But we cannot do it either for the 1355 candidate IP addresses that appear only in sessions of one website. Some of them will be related to that website but others may belong to websites that we did not consider in II. As a consequence, we need to find decision parameters that allow us to distinguish between these two cases. Ideally, with these parameters we will also be able to assign some of the doubtful IP addresses to the correct website if they appeared in the sessions of another just because of incidental overlapping of the sessions. We hypothesize that the doubtful IP addresses can act as an indicator of how good the chosen



(a)

(b)

(c)

Fig. 2. Assignment parameters

decision parameters are. We want to minimize that number as it is reasonable to think that reducing the number of IP addresses that are doubtful between the websites under study will also reduce the number of doubtful IP addresses with other websites. We now present three different decision parameters:

1) Percentage of sessions: As stated previously, we expect IP addresses that host important content of a website to appear in most of the sessions of that website. We will, in this case, assign a candidate IP to a website if it appears in more than x% of the sessions of that website. Those that do not meet the threshold for a website will then be eliminated from the website list. A high percentage should avoid confusion as it is extraordinary that users always open concurrent sessions to the same websites. In figure 2(a) we represent a sweep of the percentage of sessions in which candidate IP addresses appear. The solid line represents the number of candidate IP addresses only assigned to one website and the dotted one, the number of doubtful candidate IP addresses. The results are unexpected as IP addresses that appear in most sessions of a website are rare. In fact, normally only one or two addresses will appear in more than 50% of the sessions. Most of the content of the sites must be either dynamic or hosted dynamically. As a consequence, this is not a good parameter for labelling the trace as the number of labelled IP addresses (and thus, connections) would be small. Moreover, this parameter is not useful if we want to minimize the number of doubtful IP addresses: a sizeable number of them appear in multiple sessions of various websites. By analyzing the trace, it becomes clear that most of these conflictive IP addresses are related to web tracking services.

2) Average number of appearances in sessions: Still working with the same idea we propose a slightly different parameter. Connections to IP addresses related to a website not only may appear in multiple sessions of that website but may appear multiple times in each session. Figure 2(b) is similar to the previous one but now the parameter is the average number of appearances of a candidate IP per session. We sweep the threshold from 0 to 1 although its value could be higher. In any case, we see little improvement.

3) Concentration of IP appearances: Figure 2(c) represents a different approach. For every candidate IP in a website list, we have gone through the complete flow record and counted how many connections were made to that IP and how many of them were opened during the sessions of the website. We use the ratio of these two values as a parameter. It is important to notice that connections to a candidate IP may be opened outside the corresponding website sessions. For one thing, sessions, as they have been defined, may not encompass all the actual connections. Furthermore, a website may be accessed without a previous connection to a corresponding main IP when, for example, following a hyperlink. Nevertheless, a high value of this ratio strongly suggests that the candidate IP is related to the website. As shown in the figure, this parameter proves to be better suited to our purposes. The number of doubtful IP addresses decreases rapidly as we increase the threshold while the number of IP addresses assigned to only one site is still useful for labelling.

IV. Results and Discussion

Taking into consideration the information presented in section III we use the concentration of IP appearances decision parameter for candidate IP addresses in order to label our traces. We have chosen 50% as the assignation threshold. In other words, candidate IP addresses are assigned to a website if at least 50% of their appearances happen in sessions of that website. A 50% threshold ensures that the ratio between the candidate IP addresses assigned to more than one site and the candidate IP addresses assigned to only one site is under 0.05 for the three traffic traces we have studied. Although the doubtful IP addresses are simply not assigned and do not suppose a problem, this value gives an estimation of the probability of incorrect assignations of IP addresses that belong to websites which we are not studying. These mistakes will be, in any case, infrequent as the candidate IP would not only need to be equally popular in the sessions of the other website but the sessions of both websites would need to overlap or the candidate IP would not meet the threshold in both cases.

Using this assignation threshold we obtain a list of IP addresses that have been assigned to each website. In all, for the three traces, we are able to label 542, 446 and 654 IP addresses respectively which is a sizeable quantity given that we are only considering 28 websites or groups of websites.

A. Validation

Validating the obtained results is a challenging process as identifying the assigned IP addresses manually is a time consuming task that may prove to be impossible in some cases. In order to do it, the tools that we can use are limited:

- Simply trying to access the web server (e.g. by typing the IP address in a web browser address bar) is not useful in most of the cases as servers expect to be asked for specific content and will provide a standard error page or simply reject the connection.
- Studying the application data of the packets of the connections to the IP address may sometimes help. In our case, the sniffer that captures the Internet traffic of our University limits the capture size to 100 bytes per packet. Because of that, we rarely see past the HTTP GET field and we have not found it very useful for identification purposes.
- In the end, we have primarily relied on information obtained via DNS and WHOIS protocols. We have used the tools provided by some websites [6], [7] that gather this information and complete it with data from BGP feeds and from other parties such as the Routing Assets Database (RADb), analytics companies (e.g. Alexa) or domain name providers.

In any case, trying all of these strategies to check the hundreds of labelled IP addresses is a daunting task. As a consequence, in order to validate our method we have applied a simple accuracy testing method inspired on the ones commonly used to validate classifications in which, as it is our case, it is difficult or time consuming to check if each particular element is correctly labelled [8]. Our approach is to sample around 10% of the labelled IP addresses, check if the assignation is correct for them and use the results to infer the accuracy of our classification.

The results of the validation are shown in table II. As we can see, of the sampled IP addresses, around 70% in the three traces are correctly labelled. This means that we have been able to establish a clear relationship between the IP address and the website to which it was assigned. A very small percentage of IP addresses are incorrectly labelled: of the five cases in which this happens, two are IP addresses related to malware, other two belong to web tracking services and the last one belongs to a different normal website. Finally, a sizeable percentage of IP addresses is marked as unknown. As we have previously stated, the labelling process is far from easy even manually and for these IP addresses it was impossible to know whether they were related to the website or were incorrectly classified. Most of these IP addresses belong to content distribution networks (Akamai, RedIRIS, Edgecast) or remote computing services (Amazon web services). We expect the majority of these unknown IP addresses to be related to their assigned websites as our thresholds are very restrictive and we have been unable to link them to any other website, but we cannot know for sure.

TABLE II
VALIDATION OF INDIVIDUAL IP ADDRESSES

Traffic Trace	**Sampled IPs #**	**Correct** #	%	**Incor.** #	%	**Unkn.** #	%
Trace 1	55	39	71	1	1.8	15	27.2
Trace 2	44	30	68.2	2	4.5	12	27.3
Trace 3	64	44	68.8	2	3.1	18	28.1

V. CONCLUSIONS

In this paper we have presented a method to label server IP addresses related to a predefined list of websites in a traffic trace. This is a far from trivial problem as users often access more than one website at the same time. Our initial motivation was to obtain labelled traffic traces that could be used to tune and test a web traffic classification system. Nevertheless, from the point of view of a network administrator, our system can also be used to monitor the traffic generated by specific websites or to identify the traffic directed to certain server IP addresses.

Our system labels individual IP addresses based on the number of times connections to them appear in sessions of a particular website against the total number of appareances in the trace. We have tested our system with three traffic traces of, at least, a duration of ten days. We have identified an average of more than 15 IP addresses per website in each of the traces. However, the difficulty of validating our results only allows us to give a lower bound for the accuracy of our system at around 70% although it probably is higher.

REFERENCES

[1] A. Moore and K. Papagiannaki, "Toward the accurate identification of network applications," in *Passive and Active Network Measurement*, ser. Lecture Notes in Computer Science. Springer Berlin/Heidelberg, 2005, vol. 3431, pp. 41–54.

[2] P. Haffner, S. Sen, O. Spatscheck, and D. Wang, "ACAS: automated construction of application signatures," in *MineNet '05: Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data*. New York, NY, USA: ACM, 2005, pp. 197–202.

[3] T. T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *Communications Surveys & Tutorials, IEEE*, vol. 10, no. 4, pp. 56–76, 2008.

[4] E. B. Claise, "RFC 5101: Specification of the IP flow information export (IPFIX) protocol for the exchange of IP traffic flow information," January 2008.

[5] "Argus: Audit Record Generation and Usage System," Apr. 2012, http://www.qosient.com/argus/.

[6] "Robtex swiss army knife internet tool," Apr. 2012, http://www.robtex.com.

[7] "revip.info," Apr. 2012, http://revip.info.

[8] R. G. Congalton and K. Green, *Assessing the accuracy of remotely sensed data: Principles and practices.* Lewis Publishers, Boca Raton, 1999.