HTTPTag: A Flexible On-line HTTP Classification System for Operational 3G Networks

Pierdomenico Fiadino, Arian Bär, Pedro Casas
Telecommunications Research Center Vienna - FTW
{fiadino, baer, casas}@ftw.at

***Abstract*—The popularity of web-based services and applications like YouTube and Facebook has taken HTTP back to the pole position on end-user traffic consumption. We present HTTPTag, a flexible on-line HTTP classification system based on pattern matching and tagging. HTTPTag recognizes and tracks the evolution of more than 280 HTTP applications on the fly. This applications are responsible for about 70% of the HTTP traffic in the investigated operational 3G network. HTTPTag improves the network traffic visibility of an operator, performing tasks such as top-services ranking, long-term monitoring of the application popularity, and trend analysis among others.**

***Index Terms*—Traffic Classification, HTTP, Pattern Matching, 3G Networks.**

I. Introduction

HTTP is doubtlessly the dominating content delivery protocol in today's Internet. The popularity of services running on top of HTTP (e.g., video and audio streaming, social networking, on-line gaming, etc.) is such that they account for more than 75% of today's residential customers traffic [1]. In this scenario, understanding HTTP traffic composition and usage patterns is highly valuable for network operators, with applications in multiple areas such as network planning and optimization (e.g., content caching), traffic engineering (e.g., traffic differentiation/prioritization), marketing analysis (e.g., heavy-hitter applications), just to name a few of them.

In this paper we present HTTPTag, a flexible on-line traffic classification system for analyzing applications running on top of HTTP. The field of automatic Traffic Classification (TC) has been extensively studied during the last half-decade. Probably the most popular approach for TC exploited in recent years by the research community is the application of Machine Learning techniques [2]. Nevertheless, standard signature-based classification is widely employed for TC, complementing traditional approaches based on port matching. A good example of an automatic tool for TC is Tstat [3], a DPI-based (Deep Packet Inspection) system for traffic monitoring, using pattern matching and statistical traffic analysis.

Similar to [4], [5], HTTPTag focuses exclusively on HTTP traffic analysis. The approach adopted for the HTTP classification is based on *tagging*, i.e. associating a set of labels or *tags* to each observed HTTP request, based on the contents and service being requested. This association is performed by simple regular expressions matching, applied to the `host` field of the corresponding HTTP request's header. HTTPTag currently recognizes and tracks the evolution of more than 280 services and applications running on top of HTTP, including for example tags like `YouTube`, `Facebook`, `Twitter`, `Zynga`, `Gmail`, etc. Due to the highly concentrated traffic volume on a small number of heavy hitter applications, the current list of services spans more than 70% of the total HTTP traffic in the 3G network of a leading European provider.

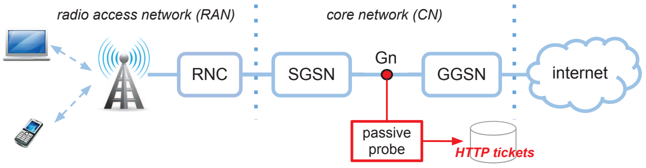


Figure 1. HTTPTag deployment in an operational 3G Network.

II. HTTPTag Overview

HTTPTag works with packet data, passively captured at the vantage point of analysis. Fig. 1 shows current deployment of HTTPTag in the network of a major European mobile operator, using the METAWIN passive monitoring system [6] for traffic capturing, filtering, and analysis. Packets are captured on the Gn interface links between the GGSN and SGSN nodes. HTTP packets are detected and analyzed on the fly: every new HTTP transaction is parsed and the contacted `hostname` (extracted from the URL) is compared against the defined regular expressions or *patterns*, see Fig. 2. If a matching pattern is found, the transaction is assigned to the corresponding service. To preserve user privacy, any user related data (e.g., IMSI, MSISDN) are removed on-the-fly, and payload content beyond HTTP headers is ignored.

HTTPTag uses TicketDB [7], a fast and scalable parallel database system tailored to meet the requirements of network monitoring in 3G networks. For every new HTTP transaction analyzed by HTTPTag, a summary ticket is stored and indexed in TicketDB, providing long term traffic analysis capabilities to the system. Each ticket contains a timestamp, the IP address of the contacted server, the requested URL, and volume stats (i.e., transferred bytes up/down). To improve pattern matching, patterns are ordered by probability of occurrence, which are computed from the history of successful matches. Many other different optimizations are performed at each of the steps of HTTPTag, including fast data imports within TicketDB, efficient indexing for fast data access, and flexibility in the data query to allow multiple types of analysis on the tagged services. Next section provides different analysis examples.

HTTPTag tagging approach is based on manual definition of tags and regular expressions, which might a priori impose

978-1-4799-0056-5/13/$31.00 ©2013 IEEE

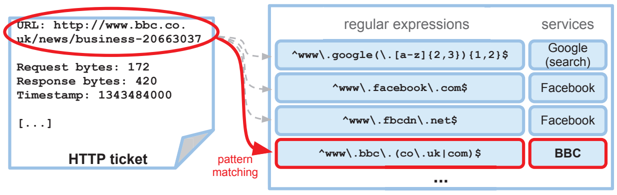


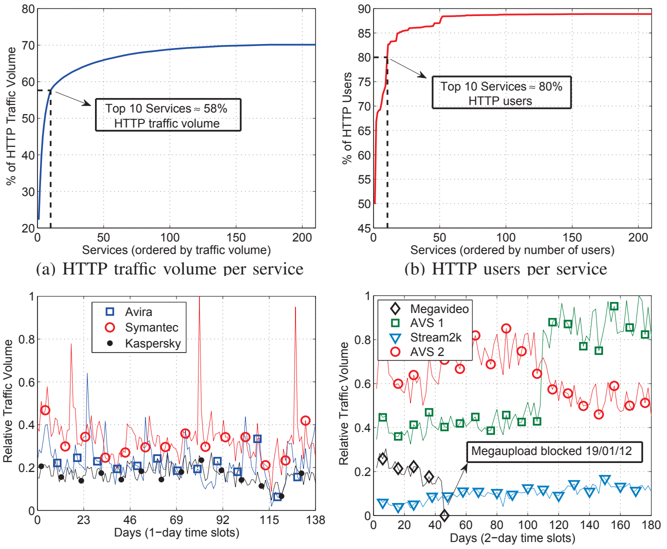
Figure 2. Matching URLs and Hostnames with patterns and services.

scalability issues. Indeed, there are millions of websites on the Internet and it would be impossible to define enough patterns to classify every requested URL. However, the well known mice and elephants phenomenon also applies to HTTP-based services, and limiting the study to the most popular services already captures the majority of the traffic volume/users in the network. While the initial definition of tags is a time-consuming task, regular expressions identifying applications tend to remain stable in time, basically because they are associated to the name of the application itself and thus recognized by the end-user. This is specially true for popular services, which at the same time carry the most of the traffic. In the practice, an initial effort in classifying the top 50 sites combined with weekly updates ensures a high classification rate. HTTPTag provides a GUI-based exploring system to identify the top host names responsible for the largest non-classified traffic volume and number of visitors, easing the tagging of new services. HTTPTag does not recognize HTTPS traffic, since the requested URLs are encrypted. An on-going extension of HTTPTag to solve this issue is to rely on DNS queries analysis, similar to the approach introduced in [8]. HTTPS recognition is out of the scope of this paper.

III. Some Applications of HTTPTag

Figs. 3(a) and 3(b) depict the distribution of HTTP traffic volume and number of users covered by HTTPTag in a standard day. Using about 380 regular expressions and 280 tags (i.e. services) manually defined, HTTPTag can classify more than 70% of the overall HTTP traffic volume caused by more than 88% of the web users in an operational 3G network. As previously mentioned, a small number of heavy hitter services dominates the HTTP landscape: the top-10 services w.r.t. volume account for almost 60% of the overall HTTP traffic, and the top-10 services w.r.t. popularity are accessed by about 80% of the users. These results reinforce the hypotheses behind HTTPTag: focusing on a small portion of the services already gives a large traffic visibility to the network operator.

Figs. 3(c) and 3(d) depict two long-term tracking applications of HTTPTag. In 3(c) we track the traffic generated by three popular antivirus services (Symantec, Kaspersky, and Avira) over a four months period (from the 26/05/12 to 15/10/12). Analyzing the traffic patterns on a sufficiently long period gives a good image on the different approaches the three companies use to manage software and virus-definition updates. While `Kaspersky` shows a quite constant behavior, both `Symantec` and `Avira` present important peak volumes on specific update periods, which might heavily load the network. This information could be directly used by the network operator to define routing, load balancing, or prioritization/shaping policies. Fig. 3(d) depicts a comparison of four video streaming services on a 6-months period (from the 1/12/11 to the 25/05/12): Megavideo, Stream2k, and two adult video services (AVS 1 and 2). After 46 days from the starting tracking day, `Megavideo` traffic completely disappears, which correlates to the well-known shut-down of the Megaupload services on the 19/01/12. Part of the video streaming volume provided by `Megavideo` was taken by a direct competitor, `Stream2k`, which shows a slow yet constant growth on the following months. Finally, we observe a drastic shift in the consumed volume from the two AVS services after 3 months and a half of steady traffic. We do not have a direct answer for this shift, but a change in the charging policy to access the content (e.g., free to subscription-based access) could explain such a variation. Having a complete picture of these popularity/usage modifications gives the operator the chance to better react to them (e.g., by defining specific content caching policies to reduce the load on the core links).



(a) HTTP traffic volume per service

(b) HTTP users per service

(c) 5-months anti-virus services tracking

(d) 6-months video services tracking

Figure 3. HTTPTag classification coverage and some long-term tracking examples revealing different events of interest in an operational 3G network.

Acknowledgements

This work has been performed in the framework of the EU-IP project mPlane, funded by the European Commission under the grant 318627.

References

[1] G. Maier, A. Feldmann, V. Paxson, ans M. Allman, "On Dominant Characteristics of Residential Broadband Internet Traffic", in *ACM IMC 2009*, 2009.

[2] T. Nguyen and G. Armitage, "A Survey of Techniques for Internet Traffic Classification using Machine Learning", in *IEEE Comm, Surv. & Tut.*, 2008.

[3] A. Finamore et al., "Experiences of Internet Traffic Monitoring with Tstat", in *IEEE Network* 25(3), 2011.

[4] J. Erman, A. Gerber, M. Hajiaghayi, D. Pei, and O. Spatscheck, "Network-Aware Forward Caching", in *WWW 2009*, 2009.

[5] J. Erman, A. Gerber, and S. Sen, "HTTP in the Home: It is not just about PCs", in *ACM CCR* 41(1), 2011.

[6] F. Ricciato, "Traffic Monitoring and Analysis for the Optimization of a 3G network", in *IEEE Wireless Communications*, vol. 13(6), 2006.

[7] A. Bär, A. Barbuzzi, P. Michiardi, F. Ricciato, "Two Parallel Approaches to Network Data Analysis", in *LADIS 2011*, 2011.

[8] I. Bermudez et al., "DNS to the rescue: Discerning Content and Services in a Tangled Web", in *ACM IMC 2012*, 2012.