Popularity of websites: HTTP Traffic Analysis

Javlonbek Abdujalilov

Tashkent University of Information Technologies
108, Amir Temur str., 100202 Tashkent, Uzbekistan
Email: javlontuit {at} gmail.com

ABSTRACT— ***Nowadays, Internet users exchange most of their content via HTTP. In this paper, we try to explore popularity of web sites, total traffic volume, content type popularity and error messages received during HTTP transactions. We used one hour long log file captured during 10am and 10pm HTTP traffic traces collected at the broadband network of ISP using Tstat passive monitoring tool and made comparison for those traffic traces.***

Keywords— HTTP, traffic volume, content type, host

1. INTRODUCTION

According to recent monitoring results on Internet traffic [3], the traffic generated by Hyper Text Transfer Protocol (HTTP) appears to have occupied the leading position comparing with other application protocols in current Internet environment. The reasons why HTTP traffic (or Web traffic) takes the leading position of volume in Internet traffic may cover several factors. Firstly, Web service is the most popular service on Internet and the amount of its users are much more than any other applications on Internet. Secondly, due to the influence of Web 2.0 the file transferred by HTTP now becomes larger than before. Besides, there are many new Internet applications implemented by HTTP, such as Instant Messengers, Updating Programs and P2P applications. As the importance of HTTP traffic specified above, the meaning of HTTP behavior analyzing and characterizing is obviously significant. The analysis on HTTP behavior can provide meaningful statistics to evaluate the performance of Web service, improve the network management and support the application development.

In this paper, individual HTTP Request and Response messages have been analyzed during an hour 10a.m and 10p.m data collected and made comparison of traffic fraction. For each Request-Response pair used only information from specific fields in the HTTP Header part of the message. From the Request header used these fields: (i) Host which provides the domain name of the server from which the content is being requested, e.g., Host: www.mediaset.it. However there are some hostnames are missing in hostname column, (ii) source and destination IP are used to define for each host name how many distinct request have been. From the Response Header used these fields (i) Content type, information to define fraction of content types(ii) response message status code is used to define most users suffered sites which responded with error messages (iii) Content-Length field is used to determine the total traffic volume of the corresponding object that is being downloaded by the subscriber. However, there are cases where the Content-Length field is either missing or set to zero – this is generally due to the data requested being dynamically generated so the HTTP server does not know the size of the object it is serving [4]. The network captures are first analyzed using a parser to extract the packet headers. These packet headers(log file) are processed by an awk script to reassemble flows defined by the 4-tuple(Client_IP, Client_Port, Server_IP, Server_Port) flow identifier.

2. RELATED WORK

The field of automatic Internet Traffic Classification (TC) and analysis has been extensively studied during the last decade [5]. Standard classification approaches rely on Deep Packet Inspection (DPI) techniques, using pattern matching and statistical traffic analysis [6]. Probably the most popular approach for TC exploited in recent years by the research community is the application of Machine Learning (ML) techniques [7]-[10].

In the specific case of HTTP traffic, classification and analysis has been the focus of many recent studies [8]-[9], [12]. In this work we used some fields of HTTP headers to recognize more popularity of web sites and their total traffic volume used during an hour and comparison between morning and night traffics. In [8], [9], authors use DPI techniques to analyze the usage of HTTP-based applications on residential connections, showing that HTTP traffic highly dominates the total downstream traffic volume. Authors in [13] study the extension of HTTP content caching in current Internet, characterizing HTTP traffic in 16 different classes using port numbers and heuristics on application headers. Recently, the authors of [11] provide evidence on a number of important pitfalls of standard HTTP traffic characterization techniques which rely exclusively on HTTP headers, showing for example that around 35% of the total HTTP volume presents a mismatch in headers like Content -Type, extensively used in previous studies.

In this paper we present popularity of web sites among users based on our analysis. The main purpose of this study is not to provide a highly accurate analysis for HTTP traffic flows at the Internet-wide scale. Rather, the goal is to explore the possibility of using some fields of HTTP header for analyzing HTTP traffic flows, offering a practical and very flexible solution for traffic analysis .

3. EXPERIMENTAL SETUP

Using Tstat passive monitor tool we captured HTTP traffic. Structure of HTTP header in the log file is given below.

Table 1: HTTP request header structure

1	2	3	4	5	6	7
c_ip	c_port	s_ip	s_port	time_abs	method	host_ fqdn
12.132.45.178	51595	173.194.65.82	80	1369821872.821556	GET	crypto-js.googlecode.com

Table 2: HTTP response header structure

1	2	3	4	5	6	7	8	9
c_ip	c_port	s_ip	s_port	time_abs	HTTP	Status code msg	Content_length	Content_type
12.132.45.178	51595	173.194.65.82	80	1369821872.897267	HTTP	200	4267	image/png

We use primarily a trace that was collected at 10 am and 10 pm during one hour. During the analyze following statistics have been analyzed.

Table 3: Specification of analyzed data set

	log_10 a.m	log_10 p.m
Total numer of Flows in the capture	3494268	3754771
Total traffic volume (MByte) in the capture	1.1255e+06	2.77199e+06
Total number of distinct src_IP in the capture	5765	6490
Total number of distinct dst_IP(host) in the capture	32546	36078

4. FRACTION OF POPULARITY OF TOP 40 WEB SITES

To count unique user we used fraction of distinct source IPs for each hostname request to website = number of distinct source IPs for each hostname/ total distinct source IP in the capture.

In the Figure 1. has shown fraction of how many unique user visited a site during 10am and 10pm, while users enter to some website users are being forwarded to the advertisement sites when they enter to the site they will see ads on the webpage which are integrated into a website script. There are some advertisement websites and third party ad serving websites have higher fraction because when user visits a website like msn.com where the content of the website comes from the website's server, but the ads come from another server.

Google.com is high percentage as it's mostly used for searching, about 45% of users connecting with google.com server and some users uses google chrome browser. There are some domains like googlesyndication.com imworldwide.com, etc used to track user behavior when users go from website to website while surfing the Web. Non-user initiated application like anti-virus, RSS feed and update programs like Microsoft.com, windowsupdate.com, msn.com etc automatically runs when user is connected to the internet and updates program.

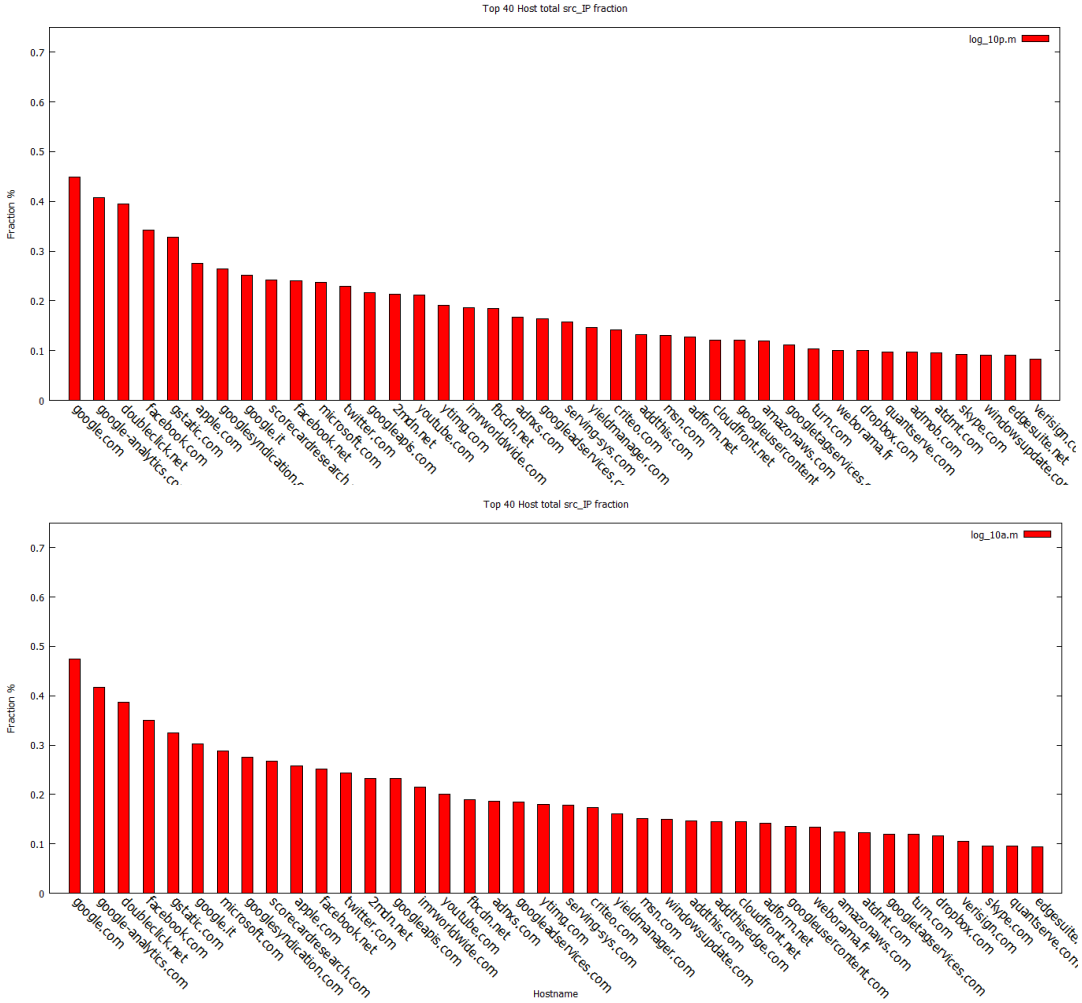
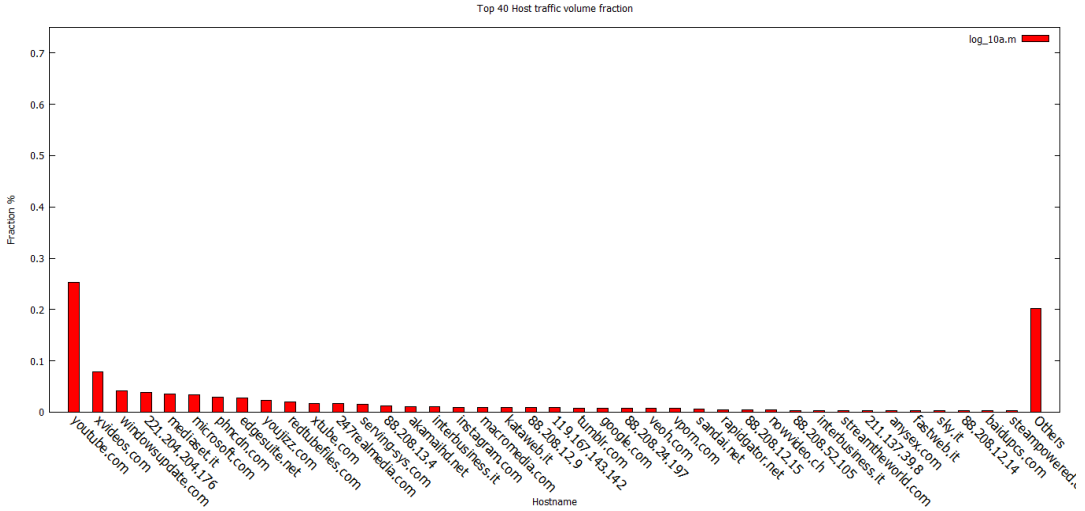


Figure 1: Fraction of distinct number of clients(src_IP) connected to the website

Cloudfront.net is a web service for content delivery and Amazonaws.com is cloud computing platform by amazon.com. Many advertisement and cookies are being made by websites and embedded to some websites and they have high fraction as users unwillingly connected to them when users enter to some website.

5. FRACTION OF TRAFFIC VOLUME FOR EACH HOSTNAME

Video site youtube.com has biggest traffic volume, interestingly vk.me social networking site has second highest traffic volume at 10pm but during the 10 am there is no in the to40 list, It seems this site is used by few users comparing to the night. 221.204.204.176 has one source and one server IP address at 10am and surprisingly it has high traffic volume 43 GByte traffic exchanged between a user and a server using 443 port, the server is located in Beijing, China[4]



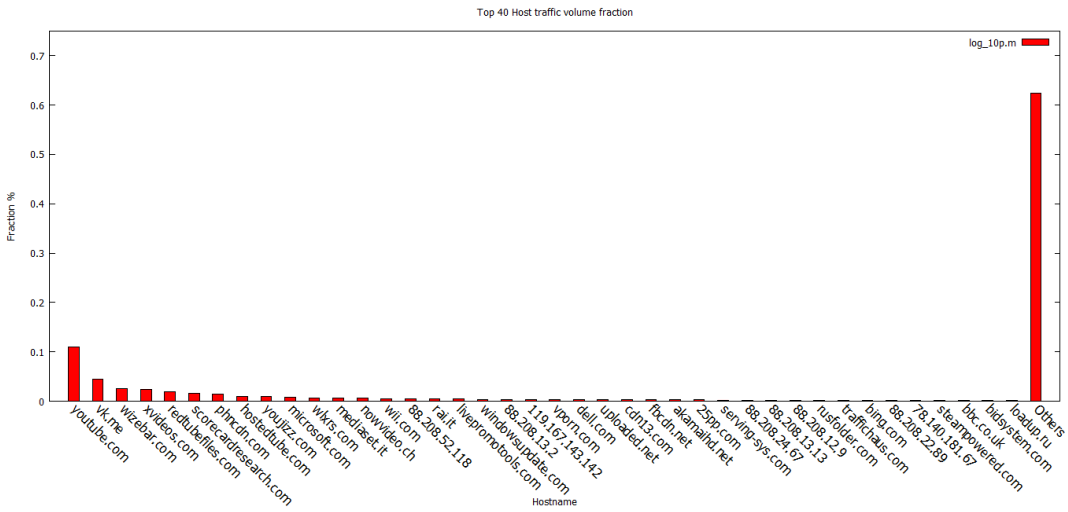


Figure 2: Fraction of traffic volume for each site

6. FRACTION OF CONTENT LENGTH

HTTP traffic volume can be measured, by (i) measuring the bytes transported on the wire, or (ii) using the Content-Length header of the HTTP response. The second option may introduce errors, caused by user interaction (e. g., interrupted downloads), software errors, or the lack of Content-Length headers. While canceling a download leads to larger values reported than actually downloaded, the lack of Content-Length headers will typically cause to ignore the request and therefore underestimate the volume [1].

Before analyzing content length (traffic volume) for each hostname, I made a filter for the log file and merged columns $3,$5,$6,$7 and $8, then sorted out with column 1. File made a list of request and response; by this way I avoided storing on the memory list of all GET requests, and then compared for each request corresponding response.

Table 4: Sample view of log file for summing up the total traffic volume of each distinct site using content length

src_IP	abs_time	method	FQDN/respond_status	Cont_ length
173.193.202.116	1371242858.904404	GET	www.youtube.com	
173.193.202.116	1371242859.092829	HTTP	200	4635
173.194.40.2	1371244889.405586	GET	s.youtube.com	
173.194.40.2	1371244889.439925	HTTP	youtube.com 200	40

In order to estimate total value of content length for each distinct host we compared src_IP request and dst_IP respond transactions with corresponding absolute time value i.e., respond comes after the request manner. First checks the row and if there is not exist HTTP on $3, gets column $4(fqdn) and $1(dst_IP) and compares to the next rows(which exist HTTP), If dst_IP matches then from that row gets content length value from column $5 and writes to the list $4(fqdn) and content length . If same fqdn request with respond dst_IP comes again then adds content length value to the listed value.

To count the sum of the total traffic volume MByte on $8(content length) in the capture, I used simple awk sum function.

awk '/HTTP/{sum+=$4} END {print sum/ 1048576}' file.txt

Fraction of bytes for each hostname = total number of bytes for each hostname / total number of bytes in the capture.

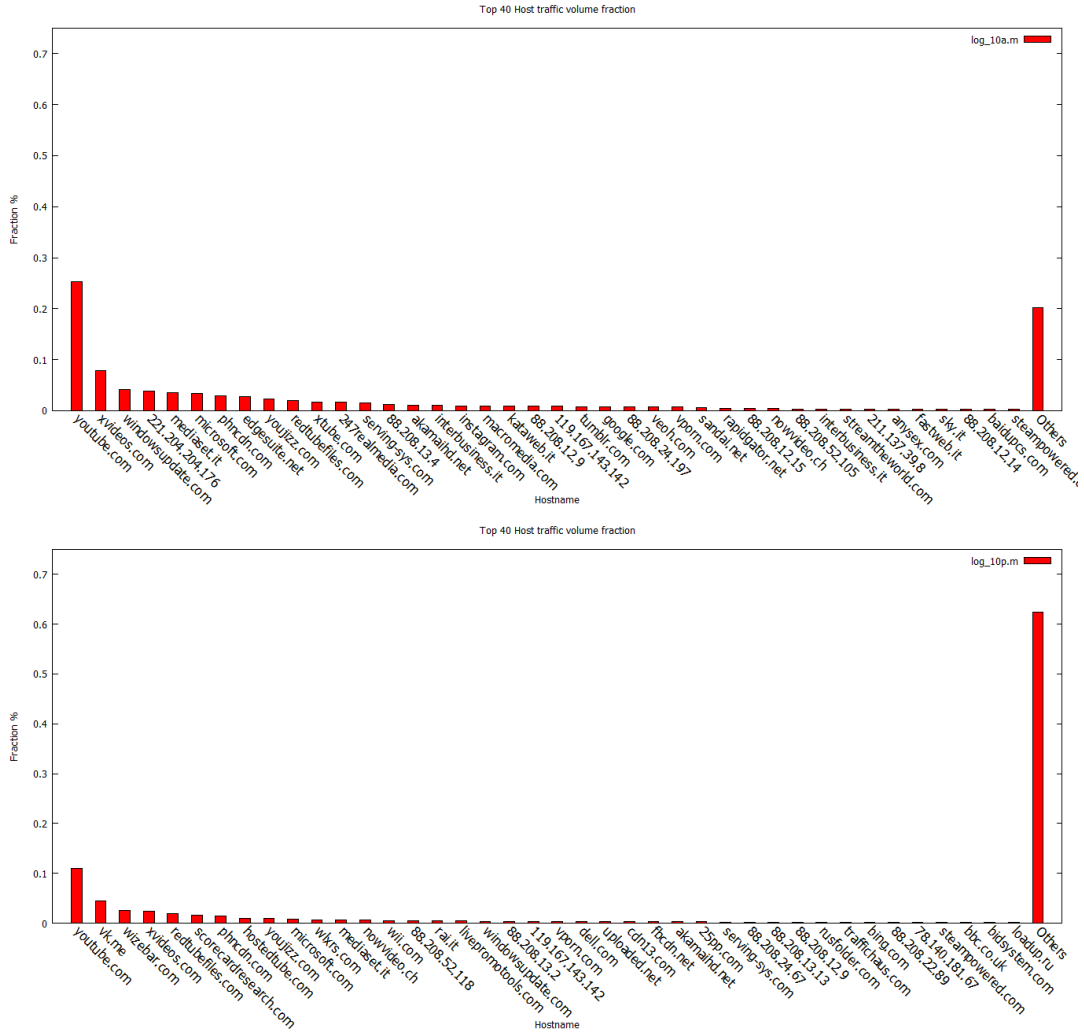


Figure 3: Fraction of Content types

Video site youtube.com has biggest traffic volume, interestingly vk.me social networking site has second highest traffic volume at 10pm but during the 10am it is not in the top40 list, it seems this site is used by few users comparing to the night. 221.204.204.176 has one source and one server IP address at 10am and surprisingly it has high traffic volume 43 GByte traffic exchanged between a user and a server using 443 port, the server is located in Beijing, China[4].

7. FRACTION OF HTTP ERROR STATUS

According to RFC 2616, there are some error message status codes from which users may suffer. During the analyse following status codes have been checked and made comparison two different hours traffic which sites responded with these codes and defined how many times these HTTP status codes occurred in response messages during the analyse time.

Table 5: Number of HTTP error messages captured during the analyse

Status	204	400	403	404	408	500	502	503
HTTP message	No Content	Bad request	Forbidden	Not found	Request Time-Out	Internal server error	Bad Gateway	Service Unavailable
Log_10a.m	127231	4899	3767	29965	8536	755	156	1169
Log_10p.m	158250	6320	3032	31638	9214	1406	684	4487

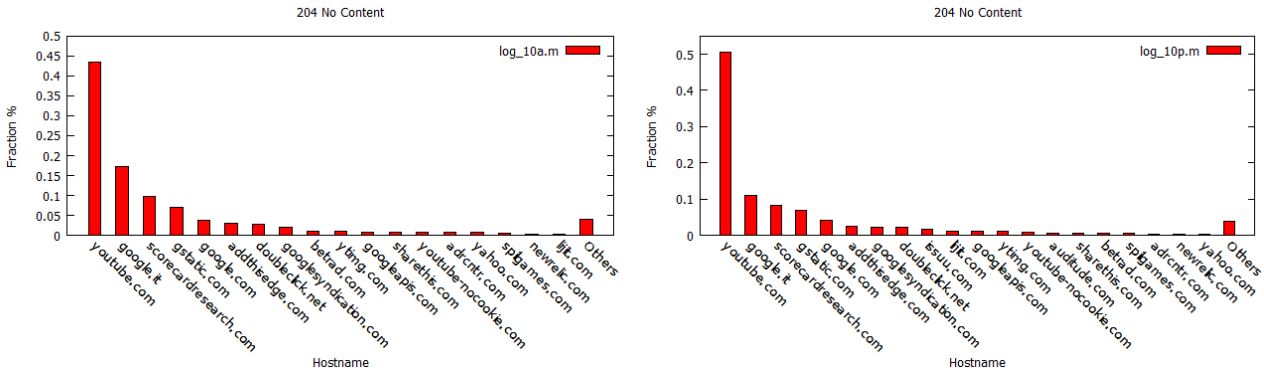


Figure 4: Fraction of HTTP 204 respond (No Content) from the sites

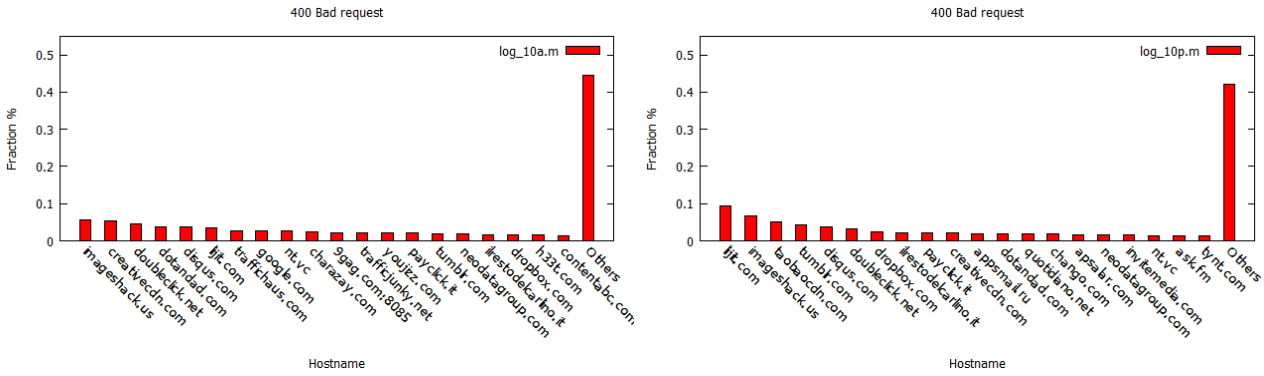


Figure 5: Fraction of HTTP 400 respond (Bad request) from the sites

Fbcdn.net has 0.05% (400 times) of request time out status response have been received by the users, this might related with internet connection while loading web script, image ect,.

8. CONCLUSION

In this paper the HTTP behavior was characterized by parsing header part of the HTTP messages using Tstat passive sniffer and collected log files have been analyzed using awk script program. According to our presented results: fraction of bytes for each website have shown that youtube and vk.me highest traffic volume requested from these servers and surprisingly during the capture we found that one IP address exchanged 43GB traffic using HTTP with 443 port. Fraction of distinct src_IP request for each website google plays dominating role in the analyse and there are domains like googlesyndication and imworldwide.com track users' behaviors while users go from one to another website while surfing throughout the Web.

9. REFERENCES

[1] Fabian Schneider, Bernhard Ager, Pitfalls in HTTP Traffic Measurements and Analysis, Springer-Verlag Berlin Heidelberg 2012.
[2] Gianluca Iannaccone, Always-on Monitoring of IP Backbones: Requirements and Design Challenges, SPRINT ATL RESEARCH REPORT
[3] Jeffrey Erman, Alexandre Gerber. HTTP in the Home: It is not just about PCs. In ACM SIGCOMM Computer Communication Review. Volume 41, Number 1, January 2011
[4] www.iplocation.net
[5] A. Dainotti, A. Pescape, and K. C. Claft), "Issues and Future Directions in Traffic Classification", in IEEE NelWork, 2012.
[6] A. Finamore et aI., "Experiences of Internet Traffic Monitoring with Tstat", in IEEE Network 25(3), 2011.
[7] T. Nguyen and G. Armitage, "A Survey of Techniques for Internet Traffic Classification using Machine Learning" , in IEEE Comm. Surv. & Tut., 2008.
[8] J. Ennan, A. Gerber, and S. Sen, "HTTP in the Home: It is not just about PCs", in ACM CCR 41(1),2011.

[9] G. Maier, A. Feldmann, V. Paxson, ans M. Allman, "On Dominant Characteristics of Residential Broadband Internet Traffic", in ACM fMC, 2009.
[10] P. Casas, J. Mazel, P. Owezarski, "MINETRAC: Mining Flows for Unsupervised Analysis & Semi-Supervised Classification". in ITC, 2011.
[11] F. Schneider et al., "Pitfalls in HTTP Traffic Measurements and Analysis", in PAM, 2012.
[12] P. Fiadino, A. Har, P. Casas, "HTTPlag: A Flexible On-line HTTP Classification System for Operational 3G Networks", in IEEE INFOCOM, 2013
[13] J. Emman, A. Gerber, M. Hajiaghayi, D. P ei, and O. Spatscheck, Network-Aware Forward Caching", in WWW, 2009.