HTTP in the Home: It is not just about PCs

Jeffrey Erman, Alexandre Gerber, Subhabrata Sen
{erman, gerber, sen}@research.att.com
AT&T Labs - Research, Florham Park, NJ, USA

ABSTRACT

HTTP (Hypertext Transport Protocol) was originally primarily used for human-initiated client-server communications launched from web browsers, traditional computers and laptops. However, today it has become the protocol of choice for a bewildering range of applications from a wide array of emerging devices like smart TVs and gaming consoles. This paper presents an initial study characterizing the non-traditional sources of HTTP traffic such as consumer devices and automated updates in the overall HTTP traffic for residential Internet users. Among our findings, 13% of all HTTP traffic in terms of bytes is due to non-traditional sources, with 5% being from consumer devices such as WiFi enabled smartphones and 8% generated from automated software updates and background processes. Our findings show that 11% of all HTTP requests are caused by communications with advertising servers from as many as 190 countries worldwide, suggesting the widespread prevalence of such activities. Overall, our findings start to answer questions about what is the state of traffic generated in these smart homes.

Categories and Subject Descriptors

C.2.3 [**Computer-Commmunication Networks**]: Measurements

General Terms

Measurements

Keywords

Traffic classification, home networks, consumer devices

1. INTRODUCTION

Over the past fifteen years, the content composition in the Internet has been changing drastically as new applications emerged. For example, Web (text/image), as the first killer Internet application, dominated Internet traffic usage in the first several years of the Internet [2, 9]. Users were spending their time sitting behind their PCs browsing on the Internet and searching for interesting content to read using the HTTP protocol. This time was then followed by the rise of P2P traffic, where users started downloading music and video content onto their computers [8, 4, 1]. Finally, more recent work analyzing Internet traffic on multiple continents [3, 7] highlighted the reemergence of HTTP as the number one application, accounting for 68% of residential broadband traffic. It has become the workhorse protocol for applications ranging from email to video streaming or file downloads.

Does that mean that, in 2010, HTTP traffic is simply generated by users that are spending their time sitting behind their PCs browsing for richer content like YouTube videos? Not necessarily. Previous studies have characterized HTTP traffic by studying what is being consumed (e.g. MIME types), but they left out other fundamental questions, such as how? By whom or by what? Or why? Indeed, over the years, the environment changed as much as the HTTP content type. First, residential users live in wireless homes full of consumer electronics connected to the Internet, such as video game consoles and smart phones offloading cellular networks by using WiFi connectivity. Second, human beings may no longer be the only HTTP consumers. Other smart devices or software might actually automatically generate and consume HTTP traffic today. For instance, antivirus software can no longer afford to wait for users to manually trigger a search for an update. Third, HTTP content doesn't have to be intentional pulled by human beings or machines. Advertising, a key part of the Internet nowadays, is also embedded into HTTP flows and pushed to end users. For example, when a user goes to the homepage of the New York Times website, 28 advertising objects are also retrieved at the same time. Is this potentially unwanted traffic a major contributor to the growth of HTTP traffic? Answering some of these questions will help us better understand the state of the smart homes, as well as better guide us when designing more efficient networks. In addition, focusing on these questions is a unique nature of this study as most traffic classification studies in the past have focused on the mime-type and application of the traffic generated and left these questions unanswered.

In this paper, we study the traffic generated by 17,000 DSL broadband users in the US over a month and attempt to provide answers about the actual nature of HTTP traffic.

HomeNets 2010, September 3, 2010, New Delhi, India.
Copyright 2010 ACM 978-1-4503-0198-5/10/09 ...$10.00.

We develop a methodology to identify the true nature of HTTP content that we then apply to that data set.

Our results show that:

- 13% of all HTTP traffic in terms of bytes is due to non-traditional sources.
- 5% is from consumer devices such as wifi enabled cellphones.
- Smartphones are in over 30% of all subscriber households.
- 7.9% of all HTTP traffic in terms of bytes is from automated software updates and background processes.
- We also found that 11% of all HTTP requests are caused from communications to advertising servers. The servers were spread across 190 servers around the world.

The remainder of the paper is organized as follows. Section 2 describes our data set and presents our measurement methodology. Section 3 highlights our results and Section 4 concludes the paper.

2. ANALYSIS METHODOLOGY

We first describe the data sets that we used for our analysis and then outline our approach to classifying the traffic.

2.1 Data Description

For this study, a network monitor was placed on a BRAS (Broadband Remote Access Servers; an aggregation point of Digital Subscriber Lines, or DSLs) located in Texas. Approximately, 17,000 broadband subscribers, each receiving up to 6Mbps service downstream, are served off of this BRAS, and our analysis covers all the HTTP traffic generated by the subscribers over the entire February 2010. In total, our data represents 156 TB of traffic and 1.6 billion HTTP requests.

We first extract individual HTTP Request and Response messages, and then associate each Request with its corresponding Response message. For each Request-Response pair, we use only information from specific fields in the HTTP Header part of the message. From the Request header, we use the following fields (i) *Host* which provides the domain name of the server from which the content is being requested, e.g., *Host: www.cnn.com* (ii) *User-Agent* information about the user agent originating the request, e.g, *User-Agent: Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1)*. From the Response Header, we use the *Content-Length* field to determine the size 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. In these instances, we assigned these objects the average object request size of 93.0KB.

In Section 3.1, we show the overall application usage for these subscribers during the busy hour and on average during the month of February 2010. The data used is from another BRAS that is colocated with the first and serves approximately the same number and type of subscribers. This additional data set was required for this analysis as only HTTP traffic was collected at the first BRAS. Our application classification relies on application specific payload signatures, heuristics, and port number analysis to categorize each flows into an application class. We refer the reader to our previous work [3] where the methodology is described in detail.

2.2 Classifying the HTTP connections

The goal of our classification is to partition the traffic along a number of key dimensions to capture the various types of non-traditional applications that use the HTTP protocol.

- User-Initiated vs. Non-User-Initiated: distinguishes between traffic that is generated due to a human user interacting with a browser from that generated by programs in the background. We consider any traffic that belongs to the Update (e.g., Windows Update, Antivirus programs), RSS and Toolbars (e.g., Yahoo Search bar) categories to be Non-User-Initiated and the remaining traffic to be User-Initiated.
- Consumer devices vs. Traditional computer generated: separates all the traffic generated using non traditional subscriber devices like game consoles, phones etc from those generated using desktops and laptops. We have further categorized devices into the following categories: TVs, Phones, gaming devices such as Xbox 360.
- Traditional Web vs. non-Web: All traffic belonging to either the Browsers (e.g., IE, Firefox) and Multimedia category (e.g., Windows Media Player) are considered as Traditional Web, the rest is non-Web.

For each HTTP Request-Response pair, we use the User-Agent details reported in the HTTP Request header to identify the application and device used to generate the corresponding traffic. Many applications and devices use custom User-Agents which makes classification of traffic by this method effective. For example, a typical Internet Explorer User-Agent would look like this: *User-Agent: Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1; WOW64; SLCC2; .NET CLR 2.0.50727; Media Center PC 6.0)*. In comparison, an example of a User-Agent of the iPhone Stocks application would look like *User-Agent: Apple iPhone v3.1.2 Stocks v1.0.1.7D11* and the User-Agent of a Sony TV *User-Agent: SONY LocationFreeTV/LF-V30US DDNSClient/1.4*. In addition to devices, most applications (especially update requests) also use custom User-Agents such as *User-Agent: ZoneAlarm/7.0.408.000 (oem-1042; en-US) ZSP/2.2* used by popular antivirus program. As can be seen in these few examples the User-Agents can contain many pieces of information that can be useful to identify the type of application and computer or device making the HTTP requests.

Based on the above, we used domain knowledge to develop a set of regular expressions to classify each Request-Response pair traffic in one or more of the following categories of generating devices and applications, defined earlier: Browsers, Devices, Update programs, Multimedia applications, P2P, RSS and Toolbars. The initial set of hand-built rules for classifying User Agents left about 10% of the traffic by bytes that could not be classified into any of these categories. We next explored a number of traffic-characteristicbased heuristics to identify additional User-Agent associations for certain application types from the unclassified traffic. The heuristics are based on certain distinctive characteristics of machine generated traffic that is not initiated by a human user using a web browser.

- *High Burstiness of Requests:* Motivated by the observation that certain non-user initiated requests like update operations become available only at certain times and often there is a flurry of update activity across users soon after a new update becomes available, we expect such update activity to be bursty across time. We calculate the hourly peak number of requests per user-agent and compare this to the average hourly number of requests across the entire month. A User-agent with a high peak to average ratio would be a candidate User-agent for Update applications. In our results, we used ratios greater than 10 as our threshold.
- *Small Number of Contacted Domains:* Many non-user-initiated applications like Update programs, anti-virus programs, etc. involve communications with a single or at most a handful of remote domains. For instance Windows update always goes to Microsoft domains. This is very different from web-browser initiated communications which could potentially connect to many external domains. Over the entire month, across all the subscribers, if a user-agent contacts a very small number of unique domains, we consider it to be a good candidate for non-user-initiated applications. In our evaluations we found a threshold of 10 domains to be good at separating the traffic.
- *Low Variance in Interarrival Times of Requests:* Various automated applications like Updates and RSS feeds tend to run periodically. We track the variance of the inter-arrival times of requests per user-agent per subscriber. If the standard deviation is very low then this is an indication that the user-agent is associated with such machine generated applications. We used values below 300 as being low variance to separate the traffic.

The above heuristics provide us with a list of candidate User-agents that could potentially belong to non-user-initiated applications. We then perform further manual inspection of the user agent information, the corresponding HTTP Request-Response pair and domain knowledge to both validate our approach, and arrive at a final determination of the device and application for that User-Agent and augment the User-Agent based classification ruleset. At the end, a smaller part of the traffic accounting for 2.9% of the flows and 2.4% of the bytes could not be classified and are marked as unknown.

2.3 Identification of Advertising Traffic

Today very often when a user interacts with a remote web server (e.g., a news site like www.cnn.com) , unbeknownst to the user, this action also triggers HTTP-based information exchanges between the user's machine and various third party advertisement services. Some of these sites are able to track a user across time and across multiple websites and build detailed user behavior profiles, giving rise to a host of privacy concerns [6]. We decided to explore the extent of such advertising traffic for our subscriber set.

Table 1: Application Mix During the Busy Hour

Class	Downstream		Upstream	
	Busy Hour	Average	Busy Hour	Average
HTTP Web	41.6%	42.9%	45.8%	35.5%
→/web/text-image	17.7%	17.1%	16.8%	13.2%
→/web/download	9.1%	9.8%	2.0%	1.7%
→/web/javascript	4.9%	4.8%	8.3%	6.0%
→/web/flash	3.1%	2.9%	1.1%	0.8%
→/web/https	1.2%	1.9%	4.4%	5.7%
→/web/iphone	1.2%	1.7%	0.4%	0.4%
→/web/other	3.4%	3.1%	5.0%	4.0%
→/web/xml	0.7%	0.8%	3.8%	1.7%
→/web/otherapp	0.2%	0.4%	3.9%	1.9%
→/web/rss	0.1%	0.1%	0.1%	0.1%
HTTP Multimedia	34.0%	31.1%	6.5%	4.6%
→ /web/video	32.2%	28.2%	6.0%	4.2%
→ /web/audio	1.8%	1.9%	0.5%	0.4%
Multimedia Other	11.2%	9.3%	4.2%	3.0%
FileSharing	3.5%	7.0%	13.6%	23.9%
Other Apps	9.6%	10.5%	30.0%	32.9%

To determine if traffic is advertising based, we used a set of rules provided from the adblockplus.org ad blocking subscription service. Our rule set used is dated from March 10, 2010. The rule set contains a set of regular expressions based on the domain and url and either identify the request as advertising or not. For example, the black list rule of *.com/ads/* categorized all traffic with a subdirectory of "ads" as being advertising.

3. RESULTS

Let us now apply our methodology, defined in the previous section, on the data set collected on the sample US broadband subscribers. This section presents our results and highlights the growing importance of non-traditional sources of HTTP traffic on wireline networks.

3.1 What are they generating?

Before drilling down into the true nature of HTTP traffic, let us first validate the significant share of HTTP in the overall traffic and analyze the content mix. As expected, Table 1 confirms that HTTP is by far the largest source of content. In our data set, collected in February 2010, it now accounts for almost three quarters of the total downstream traffic: 75.6% during the busy hour and 74.0 % bytes, on average. For Internet Service Providers (ISPs), the former number is the most important one, as it is used to drive capacity planning decisions.

Further drilling down into the application mix within the HTTP protocol, we see indeed that it is supporting a variety of applications. The three primary drivers are rich audio and video multimedia content consumption, file downloads and text/picture web browsing which account for 34.0%, 9.1%, and 17.7% of the busy hour bytes, respectively.

3.2 Traditional Web Browsing vs Everything Else

Next, we quantify the share of traditional web browsing: human users sitting behind their desktops or laptops and using their web browsers to retrieve content. The results presented in Table 2 show that 87% of the HTTP bytes and 80% of the HTTP flows are generated by traditional web browsing activities.

Table 2: Breakdown by Classification Category

Category	% Bytes	% Flows
Traditional Web Browsing	87.4%	80.2%
Everything Else	12.8%	19.8%
Computers	95.5%	98%
Consumer Device	4.5%	2.0%
User-initiated	93.6%	93.2%
Non-user-initiated	6.4 %	6.8 %

The share of non-traditional web varies during the day ranging from 8.4 % to 14.6 %, on average during the hours of the week. Traditional web browsing has an average daily peak to valley ratio of 7.1 vs. 5.8 for the remaining HTTP traffic.

3.3 What devices are generating HTTP ?

The next natural question is what devices are consuming HTTP traffic if it isn't desktops or laptops. Consumer devices other than these traditional computing platforms account for 4.5% of the total bytes generated as shown in Table 2. Our results found an average of 1.3 consumer devices other than a PC in each household.

A significant component, as seen in Table 3, is Smartphone traffic which accounts for 3.5% of the total HTTP bytes generated. During the month, 32% of the subscribers used at least one Smartphone. This highlights that the rise of cellular networks has not only transformed our Internet connectivity outside our homes, but also within them. Why consume Internet content on your desktop or laptop, if you already have a cell phone in your pocket that will do the job? This also indicates that all these households have WiFi enabled.

Another growing component of this consumer device traffic is from gaming devices such as the XBOX 360, Wii and Playstation 3. These devices, are observed in our data to be generating over 1.0% of the total bytes using streaming video and audio such as Netflix. Over the years, the manufacturers of these gaming platforms have all developed a strategy to connect their users to other gamers or to the Internet. While they don't account yet for a large source of traffic, our data shows that they have the potential to transform homes over time due to their large penetration rates. For instance, one game console was found in 40.6% of the homes.

Finally, it is interesting to note that other devices that connect the TV to the Internet are entering homes such as Internet enabled TVs (labeled as TV in chart) and multimedia devices such as the Apple TV (labeled as Multimedia in the chart). These two categories were in over 5 % of the homes.

The peak consumer device traffic is during the overall busy hour as shown in Figure 1. This is consistent with the way consumer devices are typically used; e.g., a Smartphone is generating traffic while it is actively being used and then goes into inactive mode to save power. This behaviour is reflected by the much higher peak to valley ratio of 19.6 for these consumer devices vs. 7.1 for traditional computer traffic.

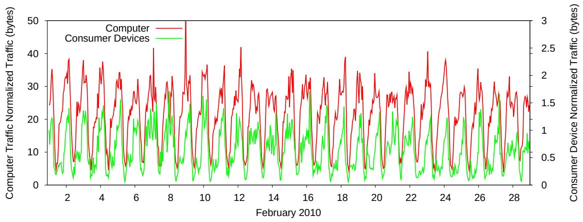

Figure 1: PC Traffic vs Consumer Devices

Table 3: Consumer Device Usage

Class	Bytes	Flows	Subscribers
Smartphone 1	3.4%	0.9%	29.7%
Gaming 1	0.6%	0.5%	3.5%
Gaming 2	0.2%	0.2%	12.5%
Gaming 3	0.1%	0.1%	11.6%
Smartphone 2	0.1%	0.0%	1.8%
Gaming 4	0.1%	0.0%	4.3%
Music Device 1	0.0%	0.0%	2.8%
Gaming 5	0.0%	0.1%	40.6%
TV 1	0.0%	0.0%	0.1%
Multimedia 1	0.0%	0.0%	2.9%
Multimedia 2	0.0%	0.0%	0.2%
TV 2	0.0%	0.0%	1.0%
Music Device 2	0.0%	0.0%	8.5%
Router 1	0.0%	0.0%	1.4%
Router 2	0.0%	0.0%	3.3%
Gaming 6	0.0%	0.0%	9.4%
Multimedia 3	0.0%	0.0%	2.8%
TV 3	0.0%	0.0%	0.0%

3.4 How much is generated by humans ?

Another way to analyze non-traditional consumption of HTTP content is to separate human requested content from content requested by machines. A typical example of these automated applications is when a computer downloads a software update. Prior work studied specific update applications, such as the familiar Windows Updates [5]. Our results in Table 2 indicate that, over a month, these automated applications accounted for 2.7% of the HTTP bytes downloaded and 2.1% of the requests.

Studying the time of day patterns of user initiated content can help us understand if applications already try to offload networks by running their automated downloads during the least busy hour. Figure 2 shows the percentage of automated/background during the month and the peak to valley ratio of automated traffic is 5.4 compared to 7.9 for user initiated content. While the share of automated content is the highest during the least busy hours, this rise is mostly due to the change in the user initiated content than to an increase in automated requests during the off-peak hours. Therefore, we can observe little attempt is currently made to spread out the load. Another reason this happening is that not all people leave their computers on all the time.

Table 4 shows the install-base of various software and antivirus applications installed per subscriber. We found that subscribers had on average 2.5 different antivirus/antispyware programs looking for updates during the month. This probably speaks to the heterogeneity of computers, laptops, and devices used by each subscriber during a given month as well as to many computers having multiple antivirus and anti-spyware software installed.

Table 4: Top 20 Software and AntiVirus Updates Install-base

Class	Bytes	Flows	Subscribers
OS 1	0.7%	0.3%	95.7%
Software 1	0.7%	0.0%	69.1%
Software 2	0.0%	0.0%	63.1%
Antivirus 1	0.6%	0.3%	59.9%
Hardware 1	0.0%	0.0%	53.3%
Antivirus 2	0.5%	0.1%	39.2%
Antivirus 3	0.1%	0.1%	32.3%
Software 3	0.0%	0.0%	31.2%
Computer 1	0.0%	0.4%	26.0%
Antivirus 4	0.0%	0.4%	25.2%
Antivirus 5	0.0%	0.3%	24.2%
Software 4	0.0%	0.0%	21.2%
Hardware 2	0.0%	0.0%	20.8%
Antivirus 6	0.0%	0.0%	16.3%
Computer 3	0.0%	0.0%	16.1%
Antivirus 7	0.0%	0.0%	11.7%
Software 5	0.0%	0.0%	11.3%
Antivirus 8	0.0%	0.0%	10.6%
Antivirus 9	0.0%	0.0%	10.2%
Computer 4	0.0%	0.0%	9.9%

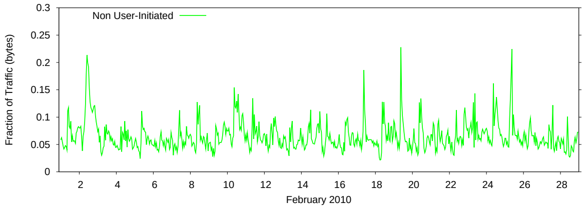

Figure 2: User-Initiated Traffic vs Automated/Background Traffic

3.5 How much is Advertising overhead ?

Another question we explored is how much HTTP traffic is generated due to the advertising on web pages. A single user request can spawn tens or hundreds of HTTP requests in order to retrieve all the objects on a webpage. For example, loading the homepage of the NYTimes.com website requires 150 objects be downloaded. Many of these requests are from images but substantial amounts (28) are also from advertising. As seen in Figure 3 only a small 0.2% of total bytes downloaded are from advertising but a substantial 11.0% of all requests are. In addition, 3.0% of all fully qualified domain names accessed in the urls is done so to access ads.

Krishnamurthy describes in [6] that as users visit various websites they are triggering data gathering operations of third party aggregation that is then used to track user behaviours and intrude on their privacy. Table 5 shows the top 10 domain extensions used by advertising domains. Of the domain extensions we were able to resolve to countries during the month, we found 190 different countries from 6 continents were observed serving advertising traffic. With 11% of all HTTP requests going to advertising, the infrastructure to track user's behaviours seems to be pretty pervasive and entrenched across the web.

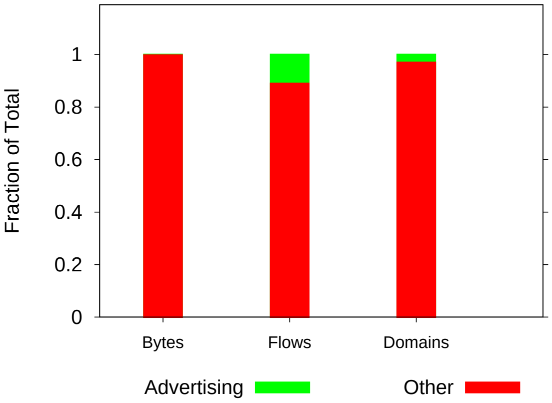

Figure 3: Advertising Traffic

Table 5: Domain Extensions of Advertising Domains

Domain Extension	# Unique Domains
com	11122
net	1075
org	535
uk	176
co	160
de	145
info	87
tv	70
us	66
au	61
Below Top 10	1289

3.6 Detecting Automated Traffic

In our hand classification of the User-Agents we used a variety of heuristics as described in the methodology section to help classify User-Agents as being from an automated source. For future studies, if these heuristics could be used alone to automatically detect traffic that is automated this would save much of the manual work. As shown in Table 6, the most effective heuristic is simply looking at the number of unique domains a User-Agent is making requests to.

4. SUMMARY

We explored how non-traditional sources of HTTP are contributing to the Internet use of residential Internet subscribers. We showed that 13% of the HTTP traffic was generated from non-traditional sources such consumer devices and non-user initiated events such as software updates. Consumer devices, contribute 4.5% of total bytes generated with a 30% penetration rate for smartphones. Another element explored was the overhead of advertising traffic in the HTTP requests. Our findings show 11% of all HTTP requests are caused by communications with advertising servers from as many as 190 countries worldwide, suggesting the widespread prevalence of such activities.

Our results show that content going to the home over broadband wireline connections may not be consumed on just a PC monitor but also on TV's and small screens of various sizes. The average household had at least 1.3 consumer device types and 1 PC in the household. The IP address analysis done to separate users does not allow us to identify household with multiple PCs or multiple units of the same device type and so the number of actual devices of different types could be larger. With the proliferation of the wide variety of consumer devices at home this traffic mix is

Table 6: Heuristics for Detecting Automated Traffic

Method	Accuracy	Precision	Recall
Peak vs Average (bytes)	54.4%	34.7%	4.0 %
Peak vs Average (flows)	51.3%	23.3%	3.2 %
Number Domains (bytes)	97.1%	54.6%	82.6%
Number Domains (flows)	95.6%	55.9%	67.5%
Interarrival Variance (bytes)	14.0%	95.6%	5.4%
Interarrival Variance (flows)	18.0%	93.9%	6.6%

likely to become richer than what one would expect for traffic generated from a single user behind a PC. Such device and traffic heterogeneity suggests the potential for resource management techniques such as class of service and traffic prioritization within the home.

Our current study is limited to subscribers in a single state (Texas) in a single country (US) and to residential DSL services receiving up to 6 Mbps. One question is how the results might look for subscribers in different geographical regions or for those with greater access speeds. A broader study involving much larger set of subscribers is planned to understand this. Additional future work will also extend this characterization to include all protocols in addition to HTTP. As our work ultimately relied on well over 120 derived rules to classify the traffic, another area for future work is the opportunity to coordinate future sharing of rulesets for classifying new devices and applications.

5. REFERENCES

[1] K. Cho, K. Fukuda, H. Esaki, and A. Kato. The Impact and Implications of the Growth in Residential User-to-User Traffic. In *SIGCOMM'06*, Pisa, Italy, 2006.

[2] K. C. Claffy. *Internet Traffic Characterization*. PhD thesis, University of California, San Diego Supercomputer Center, San Diego, 1994.

[3] J. Erman, A. Gerber, M. T. Hajiaghayi, D. Pei, and O. Spatscheck. Network-Aware Forward Caching. In *WWW'09*, Madrid, Spain, 2009.

[4] A. Gerber, J. Houle, H. Nguyen, M. Roughan, and S. Sen. P2P, The Gorilla in the Cable. In *National Cable and Telecommunications Association(NCTA) 2003 National Show*, Chicago, USA, June 2003.

[5] C. Gkantsidis, T. Karagiannis, and M. Vojnovic. Planet scale software updates. In *SIGCOMM'06*, Pisa, Italy, 2006.

[6] B. Krishnamurthy and C. E. Wills. Privacy Diffusion on the Web: A Longitudinal Perspective. In *WWW'09*, Madrid, Spain, 2009.

[7] G. Maier, A. Feldmann, V. Paxson, and M. Allman. On Dominant Characteristics of Residential Broadband Internet Traffic. In *IMC'09*, Chicago, USA, 2009.

[8] L. Plissonneau, J.-L. Costeux, and P. Brown. Analysis of Peer-to-Peer Traffic on ADSL. In *Proc. PAM'05*, Boston, USA.

[9] C. Williamson. Internet Traffic Measurement. *IEEE Internet Computing*, 5(6), November 2001.