Traffic analysis and characterization of Internet user behavior

Kihl, Maria; Aurelius, Andreas; Lagerstedt, Christina; Ödling, Per

Published in:
[Host publication title missing]

2010

Traffic analysis and characterization of Internet user behavior

Maria Kihl[1], Christina Lagerstedt[2], Andreas Aurelius[2] and Per Ödling[1]
[1] Dept. of Electrical and Information Technology, Lund University, Sweden
[2] Acreo AB, Kista, Sweden

***Abstract*— Internet usage has changed, and the demands on the broadband access networks have increased, both regarding bandwidth and QoS. Characterizing the traffic, as seen by an broadband access network, can help understanding both the demands of today and the demands of tomorrow. In this paper we analyze traffic measurements from a Swedish municipal broadband access network and derive corresponding user behavior models. The paper focuses on Internet usage in terms of traffic patterns, volumes and applications. Also, user activity characteristics, as session lengths and traffic rate distributions, are analyzed and modelled. Notably, the resulting models for user session lengths turn out different than traditionally assumed.**

I. Introduction

Internet usage is evolving from the traditional WWW usage (i.e. downloading web pages) to triple-play usage, where households may have all their communication services (telephony, data, TV) through their broadband access connection. The challenge then becomes to design the broadband access networks so that they can deliver services requiring strict QoS demands, such as IP-TV, at the same time as capacity for unpaid traffic (from the operator's perspective), for example file sharing, is demanded by the users.

One important part in meeting this challenge is to determine and understand Internet usage. Traffic patterns and applications need to be investigated and reported on. Here it is worth noting that traditional traffic measurements on aggregated traffic at the IP level do not serve this purpose. In order to capture user behavior and traffic patterns in broadband access networks, the measurements need to both be performed close to the users and be able to identify specific applications.

A number of papers have presented traffic measurements focusing on Internet applications. In most of the papers, the measurements have been performed on a high speed link in the backbone, and not in the actual broadband access network. In [1] several challenges were discussed for high speed network measurements and monitoring, for example the legal issues of storing data. The investigation in [2] presented measurements from the Sprint IP Backbone, while [3] presented measurements from seven major Japanese ISPs. These measurements were further analyzed in [4]. Also, [5] presented measurements from a core network connecting schools in Greece, and [6] discussed measurements from an OC-48 link in China. P2P traffic has been the focus of some papers, for example, [7] and [8]. In [9], [10] it was shown that port-based identification of Internet applications, used in many papers, will not give accurate results for P2P traffic, since many P2P applications use dynamic ports. This issue is also discussed in [11] which proposes an application signature based identification method. However, the problem with measuring aggregated traffic in a backbone is that the user behavior is not available. In order to capture detailed user behavior, the measurements need to be performed close to the users, that is in the broadband access network.

Some papers report on mesurements in broadband access networks. The volume of streaming media traffic has during the last years increased in the networks, in particular since the start of Youtube at the end of 2006. In [12], measurements were performed in a university campus network focusing on Youtube usage. Two papers that investigated streaming traffic generated by home users are [13] and [14]. In [15] traffic data from a Brazilian live streaming media server was analyzed, while [16] investigated statistical properties of aggregated ADSL broadband access traffic. Further, in [17] measurements with port-based identification from the Tsinghua University campus network was presented. Internet usage in wireless access networks was investigated in [18], where measurements from a campus wireless access network at the University of North Carolina were presented and analyzed.

However, only a few papers describe measurements that are similar to the measurements in this paper. In [19] user sessions in a Brazilian broadband access network were analyzed. Further, the investigations in [20], based on measurements from 1300 ADSL users in France, showed that most of the users have a low utilization of their bandwidth, mainly due to the P2P applications that limit the upload rate in the other end. However, 50% of the traffic was unidentified, due to the fact that port-based identification of traffic was used. In [21], with access network measurements from France, the focus was on P2P. A comparison of DSL and FTTH user traffic can be found in [4], where the measurements were performed in Japan. However, there are large differences in Internet usage behavior between countries [22], [23], and thus it is important that measurements from different countries and networks are reported on. Also, [4] is not reporting on specific applications, instead it focuses on the traffic volumes. Application focused measurements in a wireless broadband access network were presented in [24] and [25]

In Sweden, the Internet usage is increasing every year. Already, 90% of the adult Swedish population have access to Internet, and 79% of the population have access to Internet at home. Not less than 52% of the Swedes use Internet daily

and many people are active web 2.0 users, providing content as well as consuming content [26]. In this article, we present measurements from a Swedish municipal network including about 2600 households with FTTH broadband access. The measurements ere performed during May 2009. We have used an advanced commmercial monitoring tool, Packet Logic [27], which uses both payload-based and host behavior classification [25] of the traffic, which means that more than 95% of the traffic can be identified.

There are two main objectives with this article. First, we will give a detailed presentation and analysis of modern Internet usage, focused on applications and user behavior. The results will show daily traffic patterns for different application categories, and we also analyse the most popular applications. Both single hosts and households are used in the analysis. Also, we present some models for user activity, as session lengths and generated traffic volumes. Finally, we show how legal decisions, namely the Swedish enforcement of the European Union Intellectual Property Rights Enforcement Directive (IPRED), may effect the Internet usage behavior. The results in this paper will give a good view on what today's residential users are doing on the Internet. To our knowledge, this is the first time such a detailed characterization of household's Internet usaage has been published.

II. Relevant work

In this section, we present some relevant results from previuos work concerning traffic measurements focused on user behavior and applications.

In [4], a graph of the daily pattern for the aggregated traffic rate was shown. The traffic in the graph came from Japanese households with different broadband access technologies and the measurements were performed in 2004 and 2005. The traffic was asymmetric, with more outbound traffic than inbound traffic, and with a peak time from 9pm until 11pm. The weekends had a larger daytime traffic than the weekdays. The paper had no analysis of applications, but a port based analysis showed that 83% of the traffic used TCP dynamic ports, which is common for P2P applications. However, no definite conclusion can be made, since many other applications also use dynamic ports.

In [20], an analysis of access link saturation was performed with data from one day in 2006. In their measurements of French households with ADSL access, the peak time was during the day from 12 noon until 4pm. The paper has no analysis of top applications, only a list of the five applications that generated more than 5% of the total amount of bytes. These applications were eDonkey, applications using port 80/8080 (HTTP etc), Bit Torrent, email and telnet.

The traffic in [17] comes from a Chinese university campus network, which is reflected in some of the results. The measurements were performed in 2005. The peak time is from 9pm until 10pm. The traffic is highly asymmetric, with much more outbound traffic than inbound traffic. The explanation for this behavior is, according to the authors, that the campus network contains a number of services that are used by students and employees outside the campus. However, their port-based identification of the traffic show two P2P applications, MAZE and Bit Torrent, among the top five applications when regarding traffic volumes.

The measurements in [18] were also performed in a university campus network, this time in the US during 2005. About 58% of the traffic volume belonged to web traffic (HTTP), and about 25% was P2P application traffic. Very little streaming traffic was seen in the network, however, the measurements were performed before Youtube was deployed. The paper compared these results, which came from a wireless access network, with measurements from comparable wired access networks. It was shown that the wireless network had less P2P traffic than the wired networks.

Measurements from a German wireless access network with about 250 households were presented in [24] and [25]. The papers mostly used payload-based classification of the traffic during July 2008. The results showed that the Internet applications generating the most traffic were P2P, web browsing and and streaming. Also, sessions were identified and modelled. Session durations and volumes could be fitted with lognormal distributions.

In [6], a peak time from 9pm until 10pm was reported in a Chinese high speed link. They used an application signature based identification and found that there were less P2P traffic than reported in other papers, since only 37% of the total traffic volume was due to P2P applications. Also, 25% of the traffic was generated by HTTP. Furthermore, streaming media generated 7% of the traffic, whereas a share of 3% belonged to online games. Finally, 5% of the traffic was VoIP traffic.

Also in [5], the two top application groups were HTTP and P2P applications. One interesting result is that the P2P applications generated about 50% of the traffic in the Greek school network. However, the results are not directly comparable with our results due to the difference in population. Our measurements are performed on residential users, whereas the measurements in [5] are performed on students and school employees.

P2P traffic was of course reflected in papers that have reported on measurements from other countries. An overview of reports on P2P applications was given in [8]. The measurements were performed during 2003-2004 in Germany and France, and in these reports the most popular P2P file sharing application was eDonkey. In [9], measurements on P2P traffic from 2003 was compared with measurements from 2004. It was shown that between 2003 and 2004, Bit Torrent bit rate increased with more than 100% at the same time as Fasttrack (used by e.g. Kazaa) bit rate dropped, probably due to legal reasons. In [11], the most popular P2P application was Fasttrack, however, the measurements were performed before the introduction of Bit Torrent. In [5], eMule and Bit Torrent generated the most P2P traffic, and in [21] eDonkey had the largest traffic volume.

To our knowledge, we are the first to present detailed results concerning which types of web sites users are visiting. Other investigations, see for example, [26], have asked users about their Internet activities, however, this is of course not the same as actually measuring what people are doing. Also, [19] contained an analysis of which e-business sites that users visit. They classified the requests according to the business

models defined in [28]. In their investigation, the most popular e-business sites were Brazilian content services and portals, Yahoo, Hotmail and Google.

To our knowledge, no other papers have presented detailed traffic measurements concerning online gaming, even though online gaming has become a major force of the evolution of Internet and computer technology. A few papers include some brief results for online games. In [18], 0.01% of the traffic was identified as gaming traffic. In [10], the games Age of Empires, Half-life and Quake were found in the packet traces. In [6], 3% of the traffic was identified as games, and the gaming traffic was generated by Battlefield 1942, Doom, Quake, Need for Speed, Unreal, Xbox live, and Counter-Strike.

III. Target Network and measurement procedures

The data presented in this paper was collected from a municipal IP access network in Sweden. The network offers broadband Internet access as well as other services such as IPTV to its customers. However, due to the special implementation of IPTV, which uses multicast addresses, we have excluded IPTV traffic in the analysis. The network is fiber-based, and customers can freely choose among the services offered by the different providers connected to the network. The network uses dynamic IP-address allocation with DHCP. The lease time for the IP-addresses varies with the service provider with the shortest lease time being 20 minutes. It should be noted, however, that we during the investigation did not have access to the actual IP addresses. All data was anonymized, where each IP address was coded with a unique identification number, using a hash function. This is necessary in order to comply with the Swedish laws on personal integrity.

A. Measurement tool

The measurements have been performed using PacketLogic (PL) [27], a commercial real-time hardware/software solution used mainly for traffic surveillance, traffic shaping or as a firewall. The PL is used in many commercial broadband access networks all over the world. Traffic is identified based on packet content (deep packet inspection and deep flow inspection) instead of port definitions. PL uses the self-developed Datastream Recognition Definition Language (DRDL) [29] to identify different application protocols. Currently, the PL can identify more than 1000 Internet application protocols, and the signature database is continously updated when new applications are deployed.

Since the PL is a commercial product, the details of its functions are propietary. However, the identification process is connection-oriented, which means that each established connection between two hosts is matched to a certain application protocol. When a new connection is established, usually detected by some hand shaking procedure, the identification of this connection begins. The identification algorithm searches for specific patterns, called application signatures, in the connection. The patterns are found in the IP header and application payload. The identification algorithm uses a tree-like structure of patterns. When the identification starts, the algorithm is in the root of the tree. When certain patterns occur in the traffic, the algorithm moves down in the tree. When it reaches one of the leaves, the full identification of the connection is completed.

Most connections are bidirectional. The PL uses the traffic in both directions in the identification process. Also, different types of connections need to be tracked for different periods of time. For example, a Bit Torrent connection only needs to be tracked when it is established. Once it is identified as a Bit Torrent connection, the tracking of the connection can stop. However, an HTTP connection needs to be tracked until it is finished, since HTTP may be used to tunnel other application protocols, e.g. some file-sharing applications and VPNs.

The PL can track and identify several hundred thousand simultaneous connections, storing statistics in a database. It records the short-time average amount of traffic in the inbound and outbound directions as well as the total traffic for all nodes in the network. The data is averaged over 5 minute periods.

B. Applications

In this article, we have classified the applications into the following categories. *Web Browsing (Web)* is traffic generated by HTTP. *Streaming multimedia* applications will be presented as a single category. Some applications that are included in this category are RTSP, HTTP media stream (HTTP ms), and RTP. The group *Peer-to-Peer (P2P)* consists of P2P applications that are mainly used for file sharing. Some examples of applications that belong to this category are Bit torrent (BT), Direct Connect (DC), eDonkey, Kazaa, Gnutella and PeerEnabler. The category *File transfer* consists of applications as FTP, used for raw download of files. *Online gaming (Game)* applications provide multiplayer online games, on all platforms. Many games are included in this category, where two popular examples are World of Warcraft and Half-Life. *Messaging and collaboration (M/C)* applications will enable people to send messages and talk over Internet. Applications for email, instant messaging, and Voice over IP belong to this category. Some examples are MSN messenger, IMAP, SIP, MGCP, Skype, and Ventrillo. *Other* applications are, for example, SSH, FTP, and SSL. Also, applications related to network management belong to this category.

Routers, Switches, Customer Premises Equipment (CPE), servers and management systems all periodically generate control data and run backups. This data is not consumer generated and it represents only a very small part of the total traffic (< 1 %). This type of data has been excluded from the analysis.

C. Measurements

The studied households have Fiber-To-The-Home (FTTH) broadband access. Approximately 2600 FTTH households were included in the measurements. The FTTH households are spread all over the town, representing many demographic groups and household constellations. Internet speeds range from 1 Mb/s to 100 Mb/s, depending on which service the customers have chosen.

The measurement equipment was connected to the network via optical 50/50 splitters, see Figure 1. The optical splitters

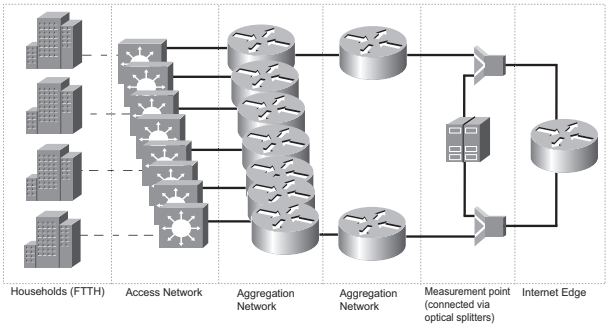

Fig. 1. Conceptual overview of the Municipal network architecture for FTTH. The dashed lines denote 100 MB/s links and the solid lines denote 1 Gb/s links

merely split the optical signal into 2 exact copies, so that the traffic in the network is not affected by the measurement device. The measurement point is the Internet Edge (IE) aggregation point, where the service providers are connected to the network. Since there are 2 redundant links to this node, a measurement hardware with 2 physical GBEthernet channels has been used. The measurements on were performed during the whole May 2009.

The main parts of the analysis were performed on hosts (IP addresses). Since one household can have several users, and thereby, several hosts, it is not possible to separate households in these analyses. Therefore, we also performed a detailed analysis of households for some ISPs in the network. In this study, comprising of a total of 1178 households, we also registered data from the DHCP server used to provide network addresses for users. The DHCP server logs data concerning date and time, IP address, service provider, access switch and access port. This data was combined with the IP address based data from the PL in a mySQL database, which meant that households with difference access speeds could be separated. It is worth noting that the online time of a household is measured in 5-minute periods. Thus, if the household has sent or received traffic during a 5 minute period, it is classified as active for 5 minutes. Also, data is truncated at 100 kbps average for this 5-minute period. If below this threshold, data is truncated to zero, and thus not included in the statistics.

IV. Aggregated traffic characteristics

This section discusses the aggregated traffic characteristics in the network. The aggregated traffic measurements show the general traffic patterns in the network. For a network operator it is important to understand how the traffic varies during the day. Service and maintenance work need to be performed during times of low traffic. For example, in the municipal network in this article, a service window is used between 2-4am with certain time intervals. During this service window, software and hardware updates and installations are performed.

A. Traffic volumes

Figure 2 shows the aggregated daily traffic pattern for the network, averaged over the measurement period. A total of 158 Tbytes of monitored data is included in this graph. The

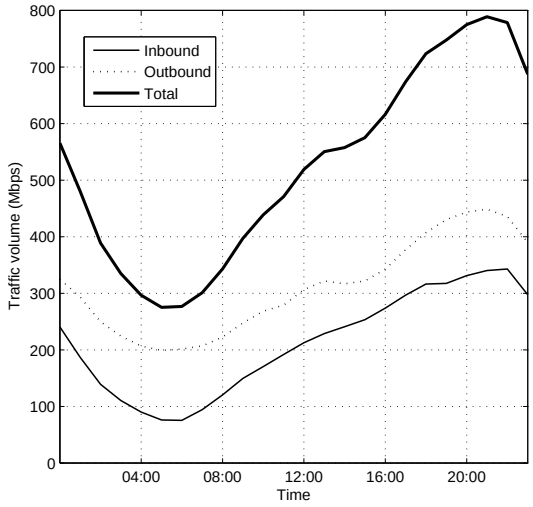

Fig. 2. Average daily total traffic volume (Mbps)

traffic patterns for weekdays and weekends are very similar. As can be seen, the peak time is from 7pm until 10pm. In the weekends, the peak time starts slightly earlier, however the average traffic volumes are about the same. Also, the traffic is asymmetrical, with more outbound traffic than inbound traffic. As will be shown below, this is completely due to P2P file sharing traffic. This asymmetrical behavior will have an effect on the design of broadband access networks, traditionally implemented as asymmetrical in the *other* direction, see [30].

Table I shows the aggregated traffic analysis when separating households with different broadband access speeds. As can be seen, households with higher bandwidth are heavier users of Internet than households with less bandwith, both when regarding traffic volumes and time. Households with the highest bandwidth (100/100 Mbps) are online and active in average 500 minutes per day, corresponding to more than 8 hours per day. Households with the lowest bandwidth (1/1 Mbps) are only online and active in average about 1.5 hours per day. Also, as can be seen in the table, households with the highest uplink bandwidth (100/100) do not use the uplink more than households with 100/10 subscriptions.

TABLE I
Aggregated traffic analysis

Access In/Out Mbps	**1/1**	**10/10**	**100/10**	**100/100**
No. of households	124	910	101	43
Minutes/day	93	263	388	500
MB/day In	98	693	1879	2319
MB/day Out	32	928	3513	3241

B. Appplications

Table II shows the traffic volume divided into the different application categories. As can be seen, file sharing traffic dominates, with 74% of the total traffic volume.

TABLE II

TRAFFIC VOLUME RATIOS

	Total	**Inbound**	**Outbound**
Web Browsing	5.5%	13%	0.9%
Streaming media	7.6%	16%	1.9%
File sharing	74%	58%	84%
File Transfer	4.9%	3.1%	6.1%
Messaging	1.3%	1.0%	1.6%
Online Gaming	0.5%	1.0%	0.2%
Other traffic	6.2%	7.6%	5.3%

Table III shows the penetration of some well-known Internet applications when separating households with different broadband access speeds. A household is considered to use an application if there is registered data for this application anytime during the measurement period. Of course, the number of households is rather small, which means that the results should be used carefully. However, the results give some ideas on how different households use the Internet. For example, the households with the lowest bandwidth (1/1 Mbps) are not using as many Internet applications as other households. PPLive, mainly used for video and web-TV, is commonly used by households, which means that households are getting accustomed with using the web for video and TV applications.

TABLE III

PENETRATION OF APPLICATIONS

Access In/Out Mbps	**1/1**	**10/10**	**100/10**	**100/100**
Households	124	910	101	43
MSN messenger	25%	52%	77%	70%
Skype	13%	25%	38%	40%
Spotify	12%	18%	34%	37%
PPLive	7%	32%	44%	40%
Joost	6%	13%	21%	12%
iTunes	9%	13%	18%	26%
Google Earth	6%	7%	12%	7%

C. Cluster analysis

Clustering the end users into different groups is a way to analyze user types. In this article, we have performed a cluster analysis of 1446 households in the network, belonging to three ISPs. The unique number of applications, combined with the amount of data transferred, were the chosen parameters for the cluster analysis. The goal was to divide users into groups based on their habits. The results of the cluster analysis are shown in Figure 3, for inbound traffic, and Figure 4, for outbound traffic. The upper 10% of the households are colored red based on their high bandwidth consumption (Cluster 3). Similarly, the lower 10% of the households are colored blue (Cluster 1) based on low bandwidth consumption. The upper boundaries were calculated to 2.6 GB/day for inbound traffic, and 3.2 GB/day for outbound traffc. The lower boundaries were approximately 13 MB/day for inbound traffic, and 1.1 MB/day for outbound traffic. To put that scale in perspective, every household consuming more than 1 GB per day use at least the equivalent of data to download or upload one full length movie per day during one month.

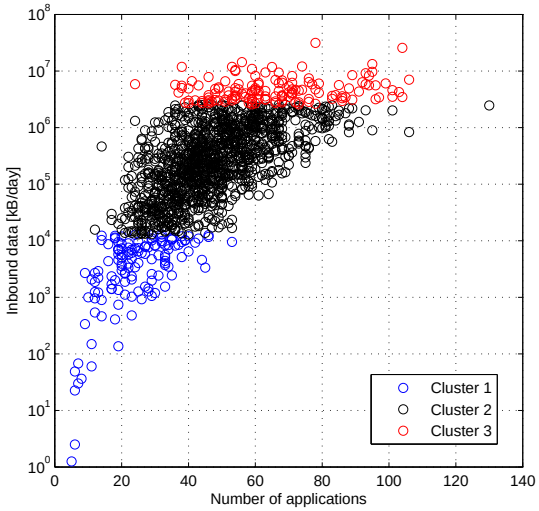

Fig. 3. Cluster analysis of inbound traffic

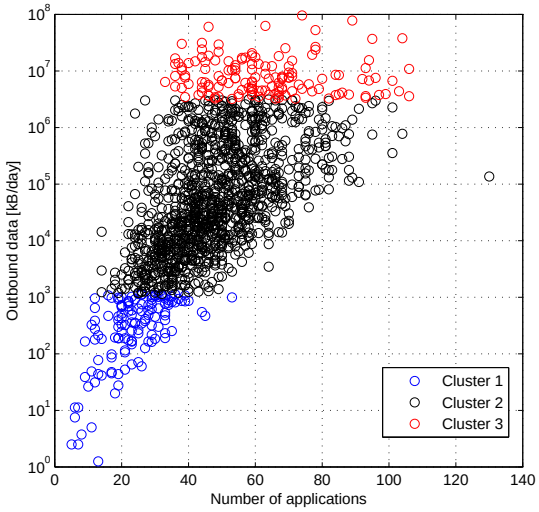

Fig. 4. Cluster analysis of outbound traffic

It can be seen that the application and protocol usage is quite high, the majority use more than 20 applications or protocols. One reason for such a high number is due to the fact that the PL separate protocols for one specific application. For example, Skype have seven different sub-protocols. Another example is HTTP, that can be divided into HTTP, HTTP media streaming, which means that with the use of SSL v2 and SSL v3, common web browsing is listed as four different applications.

In [31], our master student performed a cluster analysis for about 2000 households in the same network during September 2007. Comparing his results with the results in Figure 3 and Figure 4 reveals that the number of Internet applications per household is definitely increasing. Today, Internet users are

more active than only two years ago, in particular when it comes to the number of applications they are using. Looking at the 90% and 10% bounds in the above mentioned master thesis, we can see some differencies between our measurements and the measurements in 2007. While the incoming traffic has increased at both the upper and lower bounds (30% and 175% respectively), the outgoing traffic has decreased at the upper bound (by 26%) and increased at the lower bound (by 120%). Although the users are not exactly the same in the two measurement studies, the results imply that the traffic volume has increased in general (as expected), however the extremely heavy users are significantly less dominating.

V. File sharing

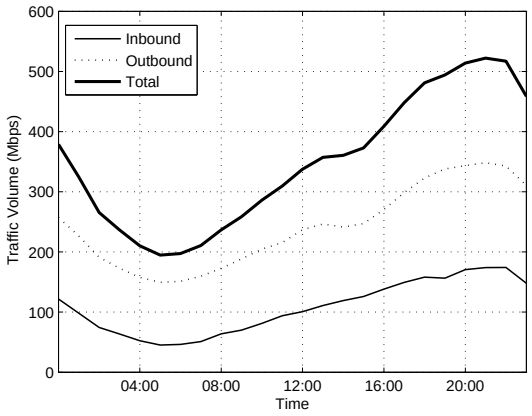

Fig. 5. Average daily traffic pattern for P2P File sharing.

It is obvious from both our measurements and other investigations that the majority of the traffic in the access networks is generated by file sharing applications. Therefore, we detail the traffic patterns for file sharing. Figure 5 shows the aggregated daily traffic pattern for the file sharing application category group, averaged over the measurement period. A total of 117 Tbytes of file sharing data was recorded. The file sharing traffic is asymmetric with more outbound traffic than inbound traffic during all hours.

Table IV shows the most active applications, when considering traffic volume during the measurements periods. As can be seen in the table, BitTorrent dominates in the network. It is well-known that the top file sharing applications are country-specific, even if BitTorrent is growing in popularity in most parts of the world [22]. Also, the popularity of specific applications have changed during the years, both due to regulatory and legal issues, and due to the launch of new applications.

Table V shows the results for BitTorrent when separating households with different broadband access speeds. About 50% of the investigated households used BitTorrent during the measurememnt period. As can be seen, households with higher bandwidth use more bandwidth as well. This is clearly shown in the outbound traffic, where households with high bandwidth have high outbound traffic rates, probably due to the design of the BitTorrent protocol.

TABLE IV
File sharing applications

	Volume ratio
BitTorrent	94%
Direct Connect	5.4%
eDonkey	0.4%
Gnutella	0.1%
Thunder	0.1%
Other	<0.1%

TABLE V
BitTorrent

Access In/Out Mbps	**1/1**	**10/10**	**100/10**	**100/100**
Active households	23	461	71	32
Minutes/day	28	90	158	202
MB/day In	93	568	1279	1104
MB/day Out	76	1300	3386	2388

VI. Web browsing and multimedia streaming

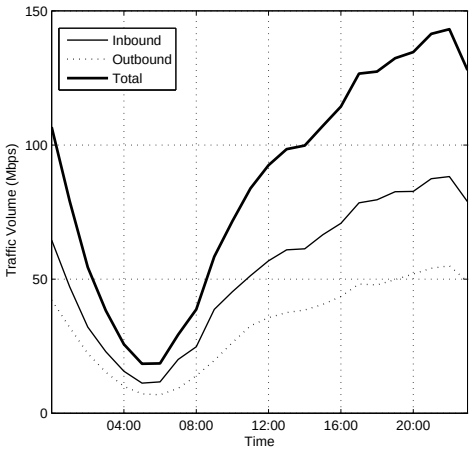

Fig. 6. Web traffic pattern.

Web browsing and multimedia streaming are activities related to more traditional World Wide Web (WWW) surfing. Figure 6 shows the average daily web traffic pattern. Both HTTP and streaming protocols are included in the traffic volume. As can be seen, web browsing is maninly performed during the day, which of course not is surprising.

Table VI shows the results for web browsing (HTTP) when separating households with different broadband access speeds. Since web browsing requires an active user by the computer, this table clearly shows that households with higher bandwidth use the Internet more than households with low bandwitdh. Households with the lowest bandwidth (1/1 Mbps) use the web in average about 1 hour per day, whereas households with the highest bandwidth use the web in average almost 4 hours per day.

Also, it is interesting to see what type of web sites people visit and download multimedia files from. Therefore, we

TABLE VI
WWW (HTTP)

Access In/Out Mbps	**1/1**	**10/10**	**100/10**	**100/100**
Active households	123	909	100	43
Minutes/day	59	155	211	228
MB/day In	17	63	85	108
MB/day Out	2.3	9.9	10	9.3

performed a deep-study analysis of the households, with measurements from May 1 to May 7 2009. We logged the names of the web sites visited. Also, we logged to amount of traffic that was generated to and from each web site. Note that the log files were totally anonymous, since only the amount of traffic to and from each web site was logged, without any information about the user that generated the traffic. This single week of measurements resulted in more than 100.000 different web addresses. 76 GBytes of traffic was generated (both inbound and outbound).

As many sites as possible were then classified into the following groups: *News/Media* sites are online news magazines, web radio and TV channel sites. Some examples are the Swedish broadcasting TV channels SVT and TV4, as well as BBC, CNN, and the online versions of numerous Swedish news papers. *Multimedia* sites contains images, video clips and audio. Users can share multimedia files on these sites. Examples are Youtube, ImageShack, FileFactory, and Dailymotion. On *Social* sites users can communicate with each other, for example with web pages, blogs, chats, discussion forums or email. Examples are Facebook, MySpace, Hotmail, Blogspot, and MSN. However, only the HTTP traffic to and from these sites was logged. The protocols for messaging, email and telephony were logged in the Messaging and Collaboration application category. The group *Information* sites contains search engines and other sites more related to information seeking. Examples are Google, Yahoo, Wikipedia, and tourist information sites. Some examples of *Commercial* sites are web shops and Internet banks. On *Software* sites users can download various computer software and security updates. The *Game* sites are related to computer games and other gaming activities such as poker or lottery. The group *Hobbies* contains sites related to more personal interests, as sports and fishing sites. The group *Adult* are all sites devoted to adult content. The group *3rd party* contains content delivery networks, proxies, and online advertisments, for example pop-up ads. Users connect to these sites through other sites. *WWW* contains various home pages, not classified in the other groups. All web addresses starts with "www.", "home.", or "homepage.". This group alone contains more than 25000 adresses. The reason for this is that there are numerous sites generating small amounts of traffic, whose names indicate more personal interests, as personal home pages, sports clubs, small web shops, and various non-profit organizations. *Images* is a group for addresses not classified above that starts with "pic" or "img".

Table VII shows the results of the analysis. The ratios are calculated in terms of traffic volume related to each group. About 15% of the traffic is ungrouped. This traffic belongs to numerous web addresses with very small ratio of the traffic volume.

TABLE VII
ANALYSIS OF WWW USAGE

News/Media	**Multimedia**	**Social**	**Information**
5.4%	24%	7.9%	4.4%
Commercial	**Software**	**Game**	**Hobbies**
3.8%	10%	5.6%	1.5%
Adult	**3rd party**	**WWW**	**Images**
2.3%	7.5%	11%	1.1%

Also, we performed a analysis on streaming, separating households with different broadband access speed. We decided to focus on Flash video, which currently is the dominating application for streaming data, used on, for example, Youtube. The results are shown in Table VIII. As can be seen, households with the lowest bandwidth (1/1 Mbps) use less streaming than other households, otherwise, there are no major differences.

TABLE VIII
FLASH VIDEO (STREAMING)

Access In/Out Mbps	**1/1**	**10/10**	**100/10**	**100/100**
No. of active households	107	865	96	42
Minutes/day	10	33	35	23
MB/day In	27	93	106	65
MB/day Out	0.6	1.8	1.7	1.0

VII. USER ACTIVITY MODELLING

A more detailed investigation of the user activity was performed during two weeks. From the measurements, we cannot, and wish not, identify the total activity of a single user. However, it is possible to find the traffic per active IP address (active host) during a certain time interval. Since dynamic IP addresses are used, a single IP address could belong to several users during a longer time period. However, during a short time period, in this case 5 minutes, we assume that the traffic generated by an active IP address corresponds to the activity of one user. The data in this section come from measurements between 2009-05-01 and 2009-05-14.

We have only analyzed *active users*, here defined as an IP address that generates more than 300 kbits (37.5 kB) during the 5-minutes interval. An online computer generates some keep-alive traffic all the time, even if there are no applications running. We have decided to not take the keep-alive traffic into account. A similar approach was used in [20], where an active client was defined as a user that generates at least 100 kB of data during a 30 minutes period. In [4] a time resolution of 2 hours was used, and, therefore, it did not contain a detailed analysis of user activity.

Also, we defined an *active session* as the time period during which one IP address generates more than 1kbps. We have made the realistic assumption that if one host disconnects

from the network during one 5 minute period, and thereby returns its IP-address to DHCP, this IP-address will not be distributed to another host during the same or next 5-minute period. Therefore, an active session represents the time period when at least one application is active on a host computer. Since all our data is averaged over 5 minutes, the length of an active session is a number of such 5 minute periods. In [19], a user session was defined as the time period when one host occupies one IP address, which means that sessions can contain periods with no active applications.

There were 6693 active IP addresses during the time period, with a total of about 2.5 million data samples where the traffic rate was at least 1 kbps.

A. Traffic rate distribution

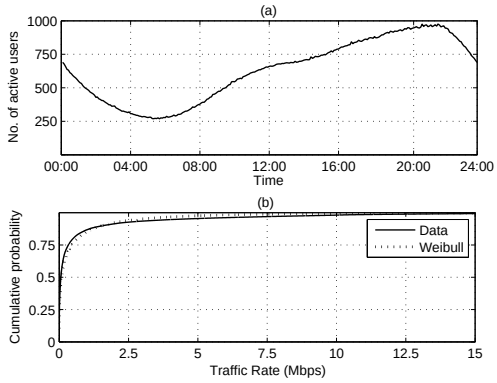

Fig. 7. (a) Average number of active users during a day. (b) Cumulative distribution function for traffic rates. The Weibull parameters are $\lambda = 0.2072$ and $k = 0.4140$

In Figure 7(a) the average number of active users during a day are shown. As can be seen, there are many active IP addresses also during night, mainly due to file sharing applications. Figure 7(b) shows the cumulative distribution function for the traffic rate per active user. The data samples are averages during the 5 minute measurement intervals. More than 90% of the data points have a traffic rate less than 2.5 Mbps. However, there is a very long tail, and the maximum measured rate was 122 Mbps (the maximum bandwidth was 200Mbps (100/100)). The mean traffic rate was about 850 kbps, whereas the median value was as low as 60 kbps. Earlier measurements, on DSL broadband access, see, for example, [20], showed a traffic rate of below 500 kbps for most users.

In the graph we have fitted the data with a Weibull distribution with a maximum likelihood estimation of the parameters computed using the data samples. The Weibull distribution is used in many areas due to its flexibility. The cumulative distribution function for a Weibull distributed stochastic variable X, $F(x) = P(X \leq x)$, is given by

$$1 - e^{-(x/\lambda)^k} \tag{1}$$

for $x > 0$. The distribution contains two parameters, the shape, $k > 0$, and the scale, $\lambda > 0$. The fit shown in Figure 7(b) has $\lambda = 0.2072$ and $k = 0.4140$. As can be seen, $k < 1$, which indicates a so called decreased failure rate. In our case, this means that there is a high probability for low bit rates, but there are also some data points with very high bit rates.

B. Active user sessions

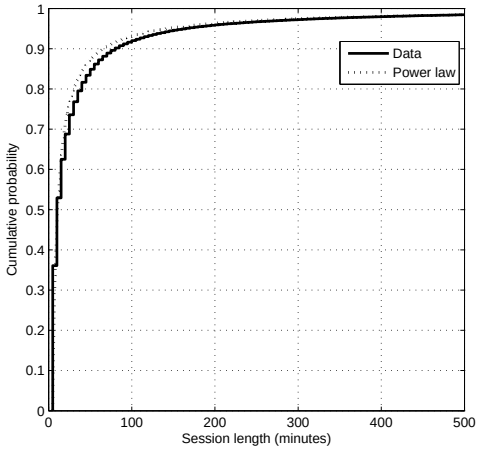

Fig. 8. Cumulative distribution functions for the length of an active session. Solid line: Data; Dashed line: Fitted power law distribution with $x_{min} = 5$ and $\alpha = 1.9$.

Some characteristics of active sessions have also been analyzed. During the two week period, there were 240206 registered active user sessions.

Figure 8 shows the cumulative distribution functions for the active session lengths. There were mainly short sessions, but some of the sessions lasted the whole measurement period (two weeks). The data shown in Figure 8 can be closely fitted to a power law distribution. The probability density function, $f(x)$, for a power law distribution is given by

$$f(x) = \frac{\alpha - 1}{x_{min}} (\frac{x}{x_{min}})^{-\alpha} \tag{2}$$

where $\alpha > 1$ and $x \geq x_{min}$. In our case, $x_{min} = 5$ (when counting in minutes) since all data is averaged over 5 minutes intervals. In Figure 8, a fitted power law distribution with $\alpha = 1.9$ is shown. The parameter α has been derived with the estimator equation

$$\hat{\alpha} = 1 + n[\sum_{i=1}^{n} ln \frac{x_i}{x_{min}}]^{-1} \tag{3}$$

where x_i are the n data points $x_i \geq x_{min}$ [32]. One characteristic of the fitted distribution is that all moments are infinite since $\alpha < 2$. However, this behavior can be explained in "real life", since we know that some users have P2P file sharing applications running constantly, and if a longer measurement period would have been used, longer sessions would have been registered. Also, we performed the same detailed analysis on sessions lengths for a two week period in September 2007. The results for that data were almost identical as the results shown in Figure 8, implicating that the model is accurate also in the long term.

The log-normal distribution has since long been known to well characterize connection sizes and durations [33], [34]. In [19] a detailed analysis of residential user sessions was included. In the paper, they also found that the session lengths could be fitted with a log-normal distribution. The same result was derived in [15] where live streaming media sessions were analyzed, as well as in [24]. Our analysis concerns active sessions, during which several connections can be established. The characteristics of active sessions are of course of importance for access network operators, since these represent the time periods when users generate traffic. To only model user sessions are not enough, since many users today are online constantly. Our investigations showed that the length of an active session cannot be accurately described by a log-normal distribution.

VIII. Effects of legal decisions

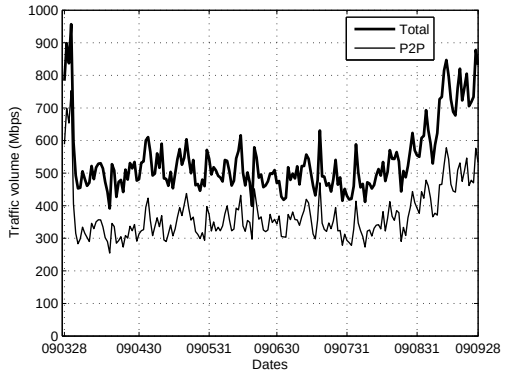

Fig. 9. Average traffic volumes per day from 2009-03-28 until 2009-09-28.

The scope of this paper has been to characterize user behavior regarding Internet usage. One aspect regarding Internet usage is its dynamics and sometimes rapid changes, making trend analysis an extremely important research topic. This, however falls outside of the scope of this paper, and will have to be regarded as future work.

One aspect that needs to be mentioned, though, is the change in usage due to the Swedish enforcement of the European Union's Intellectual Property Rights Enforcement Directive (IPRED). The law, which came into force on 1 April 2009, makes it possible for copyright holders to get a court order requesting ISPs to provide IP addresses associated with computers which have downloaded copyrighted material without paying for it.

Figure 9 shows the traffic volumes from 2009-03-28 until 2009-09-28, measured as an average value per day. Both the total traffic volumes and the P2P file sharing traffic volumes are shown. As can be seen, the traffic volume dropped by approximately 50% on April 1st, mostly due to the drop in file sharing traffic. When looking at the total traffic volume in the end of September, it can be determined that the network has lost roughly the equivalent of one year of traffic volume increase. It is worth noting here that the number of users did not drop in the same dramatic way, but rather by approximately 15-20%.

Another change in user behavior that can be noted is the increased usage of the PPTP protocol since April 1st 2009. The PPTP protocol is used by many anonymisation services. From basically not being used at all in the beginning of 2009, the preliminary results show that PPTP is used by 5-10% of the IP addresses at the end of 2009. This may have a significant effect on Internet traffic patterns, since the traffic is forced outside of the operator network, to a third party provider. However, these results can still be regarded as preliminary, and the data needs further analysis.

IX. Conclusions

Internet usage has evolved from mainly web browsing, file transfer and email, to a wide variety of applications, including many that incorporate multimedia content provisioning by the users. With new applications beeing deployed, e.g. web-TV, the demands on the broadband access networks increase. Therefore, it is important to characterize Internet traffic in these access, in order to support the understanding of the coming demands.

In this article, we have presented an analysis of traffic measurements from a Swedish municipal network with Fiber-To-The-Home (FTTH). We have presented detailed usage characteristics regarding traffic volumes, applications, and user activity. The measurements were performed with a commercial monitoring tool, Packet Logic, which is used by network operators in many countries. The advantage with this tool is that it can give a very detailed identification and classification of the Internet applications.

Some general conlusions can be made. First, many households generate much file sharing traffic, and more than 70% of the traffic volume is generated by file sharing applications, mainly BitTorrent. The traffic is asymmetric, with more outbound traffic than inbound traffic. This is mainly due to BitTorrent that many users run on their computers all the time.

Other application groups that have been analyzed are web browsing, multimedia streaming, and online gaming. Multimedia sites are generating the most traffic when it comes to web browsing. However, rather much traffic is also generated by news sites, commercial sites, social networks, and gaming sites.

Furthermore, we have characterized households based on their Internet usage. Minutes of use per day for certain applications, number of applications used, penetration of applications as well as volume generated per application was analyzed for households with different broadband subscriptions, ranging from 1 Mb/s to 100 Mb/s. Also, some models for user activity, as session duration, have been derived.

Finally, we showed how legal decisions can affect the Internet usage behavior. In Sweden, the enforcement of IPRED had dramatic effects on the traffic volumes, where the BitTorrent traffic dropped considerably from March 31 until April 1st when the law was enforced. Also, some preliminary results show that the use of anonymization services increase in the network.

X. Acknowledgements

The work has partly been financed by the CELTIC project TRAMMS, with the Swedish Governmental Agency for Innovation Systems (VINNOVA) supporting the Swedish contribution. Maria Kihl is financed in the VINNMER programme at VINNOVA. Andreas Aurelius and Christina Lagerstedt are partly financed by the VINNOVA project Broadband Behavior.

References

[1] R. Clegg, M. Withall, A. Moore, I. Phillips, D. Parish, M. Rio, R. Landa, H. Haddadi, K. Kyriakopoulos, J. Auge, R. Clayton, and D. Salmon, "Challenges in the capture and dissemination of measurements from high-speed networks," *IET Communications*, vol. 3, no. 6, pp. 957–966, 2009.

[2] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and C. Diot, "Packet-level traffic measurements from the sprint ip backbone," *IEEE Network*, vol. 17, no. 6, pp. 6–16, 2003.

[3] K. Fukuda, K. Cho, and H.Esaki, "The impact of residential broadband traffic on japanese isp backbones," *ACM SIGCOMM Computer Communications Review*, vol. 35, no. 1, 2005.

[4] K. Cho, K. Fukuda, H. Esaki, and A. Kato, "The impact and implications of the growth in residential user-to-user traffic," in *Proc. of ACM SIGCOMM'06*, 2006.

[5] C. Kattirtzis, E. Varvarigos, K. Vlachos, G. Stathakopoulos, and M. Paraskevas, "Analyzing traffic across the Greek school network," in *Proc. of the 14th IEEE Workshop on Local and Metropolitan Area Networks*, 2005.

[6] G. Xie, G. Zhang, J. Yang, Y. Min, V. Issarny, and A. Conte, "Survey on traffic of Metro Area Network with measurements on-line," in *Managing Traffic Performance in Converged Networks*, vol. 4516 of *Lecture Notes in Computer Science*, pp. 666–677, Springer-Verlag Berlin Heidelberg, 2007.

[7] S. Sen and J. Wang, "Analyzing peer-to-peer traffic across large networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 2, pp. 219–232, 2004.

[8] G. Hasslinger, "ISP platforms under a heavy peer-to-peer workload," in *Peer-to-Peer Systems and Applications*, vol. 3485 of *Lecture Notes in Computer Science*, pp. 998–1010, Springer-Verlag Berlin Heidelberg, 2005.

[9] T. Karagiannis, A. Broido, N. Brownlee, K. Claffy, and M. Faloutsos, "Is P2P dying or just hiding?," in *Proc. of IEEE Globecom*, 2004.

[10] T. Karagiannis, A. Broido, M. Faloutsos, and K. Claffy, "Transport layer identification of P2P traffic," in *Proc. of the 4th ACM Conference on Internet Measurement*, pp. 121–134, 2004.

[11] S. Sen, O. Spatscheck, and D. Wang, "Accurate, scalable in-network identification of P2P traffic using application signatures," in *Proc. of the ACM 13th International Conference on World Wide Web*, pp. 512–521, 2004.

[12] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: A view from the edge," in *Proc. of ACM Internet Measurement Conference*, 2007.

[13] L. Guo, S. Chen, Z. Xiao, and X. Zhang, "Analysis of multimedia workloads with implications for internet streaming," in *Proc. of ACM International World Wide Web Conference*, 2005.

[14] L. Guo, E. Tan, S. Chen, Z. Xiao, O. Spatscheck, and X. Zhang, "Delving into internet streaming media delivery: A quality and resource utilization perspective," in *Proc. of the ACM Internet Measurement Conference*, 2006.

[15] E. Veloso, V. Almeida, W. Meira, and A. Bestravos, "A hierarchical characterization of a live streaming media workload," *IEEE/ACM Transactions on Networking*, vol. 14, no. 1, 2006.

[16] G. Hasslinger, J. Mende, R. Geib, T. Beckhaus, and F. Hartleb, "Measurement and characteristics of aggregated traffic in broadband access networks," in *Managing Traffic Performance in Converged Networks*, vol. 4516 of *Lecture Notes in Computer Science*, pp. 998–1010, Springer-Verlag Berlin Heidelberg, 2007.

[17] J. Zhang, J. Yang, C. An, and J. Wang, "Traffic measurements and analysis of tunet," in *Proc. of the 4th IEEE International Conference on Cyberworlds*, 2005.

[18] M. Ploumidis, M. Papadopouli, and T. Karagiannis, "Multi-level application-based traffic characterization in a large-scale wireless network," in *Proc. of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, pp. 1–9, 2007.

[19] H. M. Neto, L. Rocha, P. Guerra, J. Almeida, W. M. Jr., and V. Almeida, "A characterization of broadband user behavior and their e-business activities," *ACM Performance Evaluation Review*, vol. 32, no. 3, 2004.

[20] M. Siekkinen, D. Collange, G. Urvoy-Keller, and E. Biersack, "Performance limitations of ADSL users: A case study," in *Passive and Active Network Measurement*, vol. 4427 of *Lecture Notes in Computer Science*, pp. 145–154, Springer-Verlag Berlin Heidelberg, 2007.

[21] L. Plissonneau, J.-L. Costeux, and P. Brown, "Analysis of peer-to-peer traffic on ADSL," in *Passive and Active Measurement*, vol. 3431 of *Lecture Notes in Computer Science*, pp. 69–82, Springer-Verlag Berlin Heidelberg, 2005.

[22] "Internet study 2007," tech. rep., 2007.

[23] "2008 analysis of traffic demographics in North-American broadband networks," tech. rep., 2008.

[24] R. Pries, F. Wamser, D. Staehle, K. Heck, and P. Tran-Gia, "On traffic characteristics of a broadband wireless internet access," *IEEE Next Generation Internet Networks*, pp. 1–7, 2009.

[25] R. Pries, F. Wamser, D. Staehle, K. Heck, and P. Tran-Gia, "Traffic measurement and analysis of a broadband wireless internet access," in *IEEE 69th Vehicular Technology Conference*, 2009.

[26] "World Internet Project home page." http://www.worldinternetproject.net/.

[27] "Procera networks." http://www.proceranetworks.com.

[28] M. Rappa, "The utility business model and the future of computing services," *IBM Systems Journal*, vol. 43, no. 1, 2004.

[29] "Packetlogic drdl signatures and properties 10340 (beta)," tech. rep., 2007.

[30] M. Forzati and C. Larsen, "On the symmetry requirements for tomorrow's fibre access networks," in *11th International Conference on Transparent Optical Networks*, 2009.

[31] T. Bonnedahl, "Traffic measurements and analysis in fixed and mobile broadband access networks," Master's thesis, Lund University, Sweden, Sweden, 2009.

[32] A. Clauset, C. Shalizi, and M. Newman, "Power-law distributions in empirical data," 2007.

[33] V. Paxson, "Empirically derived analytic models of wide-area TCP connections," *IEEE/ACM Transactions on Networking*, vol. 2, no. 4, 1994.

[34] S. Floyd, "Difficulties in simulating the Internet," *IEEE/ACM Transactions on Networking*, vol. 9, no. 4, 2001.