Monitoring the Impact of P2P Users on a Broadband Operator's Network

H.J. Kolbe
NEC Network Laboratories Europe
Kurfuersten-Anlage 36
69115 Heidelberg, Germany
kolbe@nw.neclab.eu

O.Kettig
Arcor AG & Co KG.
Alfred-Herrhausen-Allee 1
65760 Eschborn, Germany
oliver.kettig@arcor.net

E.Golic
University of Applied Sciences
Nibelungenplatz 1
60318 Frankfurt am Main, Germany

***Abstract*- Since their emergence peer-to-peer (P2P) applications have been generating a considerable fraction of the overall transferred bandwidth in broadband networks. Residential broadband service has been moving from one geared towards technology enthusiasts and early adopters to a commodity for a large fraction of households. Thus, the question whether P2P is still the dominant application in terms of bandwidth usage becomes highly relevant for broadband operators. In this work we present a method for classifying broadband users into a P2P- and a non-P2P group based on the amount of communication partners ("peers") they have in a dedicated timeframe. Based on this classification, we derive their impact on network characteristics like the number of active users and their aggregate bandwidth. Privacy is assured by anonymization of the data and by not taking into account the packet payloads. We apply our method to real operational data collected from a major German DSL provider's access link which transported all traffic each user generates and receives. We find that P2P users are still large contributors to the total amount of traffic seen. However, in comparison to data collected four years earlier, the impact from P2P on the bandwidth peaks in the busy hours has clearly decreased while other applications have a growing impact. Further analysis also reveals that the P2P users' traffic does not exhibit strong locality. We furthermore compare our findings to those available in the literature and propose areas for future work on network monitoring, P2P applications, and network design.**

I. INTRODUCTION

For broadband service providers and network operators there is a strong need to know what different types of customers exist and what their impact on network characteristics is. Many important business decisions such as designing appropriate tariff models, determining the scalability of network designs and dimensioning the network strongly depend on the demands of the customers. Since peer-to-peer (P2P) applications have recently been identified as having the biggest impact on broadband networks in terms of consumed bandwidth, the main questions addressed in this study are, what percentage of subscribers can be classified as P2P users at different times of a day, and how big their impact on the network characteristics is. It is furthermore important to get to know how these key numbers changed during the last years. Moreover, new applications that might become more dominant than P2P need to be identified early.

We have developed an application that can extract traffic characteristics from network links, correlate the data and generate reports on the impact of P2P on a per-subscriber-basis while maintaining their privacy by collecting only a small anonymized fraction of the overall data. We store the aggregated and anonymized data in a database that enables us to perform a detailed offline analysis. While the tool can be used for a variety of types of reports, our main goal was to determine the impact of P2P users on the overall network load of a major German ISP.

P2P users are defined in the scope of this work as customers using P2P software. The main application that uses P2P techniques is file sharing over the Internet. The most famous implementations of those types of applications are based on the BitTorrent [2] and the EDonkey protocols [5]. A recent report confirms that EDonkey and BitTorrent dominate the P2P traffic in Germany with more than 90% of the overall P2P traffic [11]. While in early studies that determined the impact of P2P users on broadband networks it has been sufficient to classify IP packets and flows based on the TCP and UDP port numbers [6],[9],[23], the situation has become more difficult in the past years. In order to prevent being detected by firewalls or rate limiting rules applied to the ISP's network elements, the current generation of P2P clients tries to hide their traffic by using randomly chosen port numbers [12]. Classification methods only using port-based approaches therefore lead to a growing fraction of "unknown" traffic types [6],[20]. An approach to solve this issue is to combine port-based methods with deep packet inspection of the payload to find signatures and the detection of heuristic communication patterns in order to classify P2P traffic [7],[10],[13], [15],[17],[24]. As this leads to quite sophisticated methods that have to adapt regularly to newly emerging applications, some further studies focus on classifying so-called "heavy hitters" by defining a threshold value for the bandwidth usage per customer [4],[8]. A completely different approach to all those passive measurements is to analyze P2P traffic by taking part in the network using crawling tools [22],[26].

In our practical approach, we classify the subscribers into P2P users and non-P2P users by counting the number of hosts (i.e., distinct IP addresses) a subscriber is exchanging IP packets with during a defined time frame. Subscribers having more peers than a defined threshold value will be classified as P2P users, the others will be put into the non-P2P group. This method comes with a lower accuracy compared to deep packet

978-1-4244-3487-9/09/$25.00 © 2009 IEEE

inspection methods, but its clear benefit is that it is a generic and simple method since it is based on a very unique characteristic of current P2P applications. Regular adaptations to new types of P2P clients or requiring to store and analyze parts of the packet payload which may result in possible conflicts with privacy do not apply to our methodology. It is worth noting that we indeed classify users, not traffic. Our interest is to count the whole traffic of P2P-users regardless if the individual packet belongs to a P2P application or not. This is reasonable since, first of all, a service provider has to deal with his customers as entities and, secondly, the consumed bandwidth of P2P users is expected to be dominated by the P2P applications they use as we will explain later in the text.

This paper is organized as follows: The following section describes how we retrieved the experimental data and processed it. After that, we show early results and then move on to the subscriber classification based on counting the peers. Having classified the users, we then show their impact on key figures like the consumed bandwidth during peak hours. In the last sections, we sum up our results, compare them to those available in literature and derive conclusions for network design, management and future work.

II. Data Collection

The results shown in this work have been derived from experimental data collected on four different days during one week in September 2007. The weekdays were a Tuesday, Thursday, Saturday and Sunday and will be referred to in the preceding text by their names. Network characteristics from past years revealed that those weekdays reflect the possible typical types of user behavior very well. Our experimental data is based on traffic captured for 120 seconds every two hours. Choosing this duration is a tradeoff between minimizing memory requirements of the processing systems and maximizing the available amount of data.

More data had been sampled to assure that the results were representative and that 120 seconds is a suitable timeframe. It had also been assured that none of these days were close to a public holiday or other events that have noticeable impact on the customer's usage of the Internet service.

All monitored traffic belongs to customers that have signed up to a PPPoE-based ADSL broadband internet service in combination with an ISDN telephone line. Since no IPTV or VoIP services had been bundled with it, our data reflects exclusively the Internet usage. Each customer obtains one official IP address from the ISP's address pools for the time the PPP session is up.

The number of customers connected to the whole network was about 2.5 million. The network link monitored served about 3200 connected customers, which was found to be sufficient to get statistically significant results.

A. Measurement

Monitoring traffic imposes high requirements on the performance of the monitoring systems and needs to be compliant to legal demands. Both needs had been addressed when defining the scope and the methods of the measurement. A recent paper from Ohm, Sicker, and Grunwald [16] gives very good guidance on how to assure privacy as good as possible.

In the operator's Ethernet access network, the subscribers connect through VLAN tunnels to the Broadband Remote Access Server (BRAS). Each customer sets up a PPPoE session spanning from his user equipment (a PC or as mostly seen a home gateway) to the BRAS. As the traffic is being tunneled, mirroring it on an aggregation node in the path between DSLAM and BRAS gives us the possibility to record the full amount of traffic generated and received by the customers. This is a big advantage compared to monitoring traffic inside a routed network (e.g., near an Internet peering point).

The available monitoring function of the aggregation node allowed us to restrict the data collection to the first 200 bytes of each frame on the wire. As this reduced the bandwidth of the sampled traffic, a standard PC was sufficient for capturing the packets with the well-known "dumpcap" tool, which is part of the "wireshark" network monitoring software suite [28]. A Perl script controlled the usage of this tool and took care of processing the data before storing it. As the measurement period was only 120s long for each run, we were able to process the whole collected data immediately after capturing it. Traffic that did not belong to residential customers, like keepalive messages and management traffic had been filtered out. All IP addresses from the subscribers had been replaced by anonymized ones starting with 1.0.0.1 and counting up with the occurrence of new addresses. As customers appear in a given trace in random order, there is no possibility to correlate data from two different measurements to one customer. The binding information between the real and the anonymized IP address had been deleted immediately after finishing this pre-processing step.

Before storing the data, the script extracted only the values from the packet headers that were of interest to us. No payload information had been recorded. The anonymized data had then been transferred to another computer system containing a database for offline analysis. The monitoring system was subject to the strong access restrictions that applied to all management systems within the network. Similar polices were applied to the system that contained the database containing the anonymized data.

B. Processing

We designed a database model for storing the data in an adequate way to perform our analysis. From the data stored in the table containing the attributes of each packet, it was possible to extract each IP address used by the customers and to collect all layer-4 flows for each customer by combining SQL commands and Perl scripts. Each anonymized IP address had been assigned to one customer. So the total amount of users was given by the number of distinct user-side IP addresses found in the trace. Due to the short timeframe of the measurement, it was highly unlikely that customers logged off

and on during that period, thus preventing them from appearing twice in the database.

The data model allowed to directly link the customers to the packets and the flows they were generating and receiving. In a second script run after filling the database initially, we derived additional customers' attributes like the number of flows by again combining SQL and Perl scripts. Classifying packets by used ports had also been realized by using SQL commands.

III. Results

The total amount of traffic transferred in the peak hour is of most importance to network operators since it defines how much bandwidth needs to be provided. We generated such a graph from our measurements and checked that it was compliant to the graphs from the operator's network monitoring system. We can derive a typical curve for a broadband network link on a weekday. By simply analyzing the slope of the graph in Figure 1, we can already make some assumptions on the underlying causes.

In Figure 1 we can clearly see that the peak hour is located somewhere between 6 p.m. and 10 p.m. Results from links with more users attached showed that the peak hour tends to be around 9 p.m. The used downstream bandwidth (towards the users) is in all cases higher than the upstream traffic's bandwidth. This is a consequence of the asymmetry of the access lines but also of the asymmetric nature of the most popular applications on the Internet.

While the downstream traffic shows very strong time dependence, the upstream traffic varies not so strongly. It rises between night-time and the busy hour by only about 50% compared to an increase of about 200% of the downstream traffic. Clearly visible, a gap opens up between both curves during daytime. This is a clear hint that interactive users are causing the increase in traffic. The shape of the curve reflects people coming home from work and using the Internet in the evening. This assumption is also being proven by the time dependence of the number of active users on the network (c.f. Figure 6). Non-interactive customers using P2P clients that are running autonomously without any user interaction would rather generate a behavior that is expected to be nearly constant over time. The shape of the upstream traffic resembles this behavior. The used bandwidth grows only slightly during peak hours. This rise is indirectly caused by the increased downstream traffic requiring a higher amount of TCP ACK packets to be sent in upstream direction. Also, there seems to be an underlying constant level of traffic within the downstream curve. An analysis of the IP packet size distribution derived from the IP headers of the truncated packets supports this reasoning. At daytime the average packet size in upstream direction decreases from the nocturnal value of 400 bytes to 300 bytes because of the higher fraction of small TCP ACK packets.

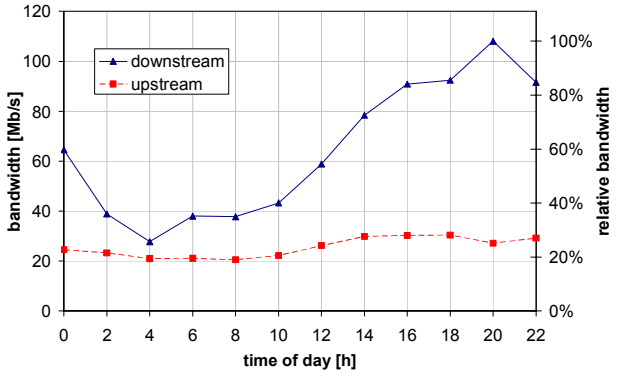

Fig. 1. Total bandwidth usage on Tuesday

The analysis so far already revealed that the traffic pattern can be separated into a static and a dynamic part. Using our new method, we will further investigate this.

A. Port-based Traffic Analysis

For each set of data we ran an analysis on the layer-4 ports used to assign each packet to the application that generated it. The list that correlates port numbers to applications was taken by combining several lists available on the Internet.

Table 1 shows the fraction of each traffic group on the whole transferred amount of traffic on one day. In downstream direction it is clearly visible that web-based traffic (HTTP/HTTPS) is dominating. The traffic consists mainly of pure web browsing, but spot tests of the remote IP addresses revealed also a high amount of YouTube (web-based videos) [29], RapidShare (file sharing using direct downloads) [21] and other file download traffic using HTTP. The traffic volume carried over the ports attributed to web applications over time exhibits a slope that reflects a behavior expected for interactive applications (not shown here). At night-time there is nearly no traffic visible, while it rises to a strong peak during the busy hours. The other quite dominant application group is what we assigned to the P2P group. The traffic found in this group mainly consisted of eDonkey [5], BitTorrent [2] and BearShare [3] traffic originating from different types of software clients. P2P applications other than file sharing tools could not be observed although their used port numbers were part of our list. The traffic using e.g. the eDonkey port 4662 revealed a rather constant behavior over time, reflecting the non-interactive behavior of the P2P "machines". A more detailed analysis on the two fundamentally different traffic types can be found later in this paper. The results of our port-based analysis are similar for all available days of measurement.

There remains still a huge amount of traffic that could not be classified at all. As already stated in literature, e.g. by Karagiannis, Broido, Faloutsos, and Claffy, P2P applications moved

TABLE I
Port-Based Analysis of Traffic Types

		P2P	Gaming	Chat	VoIP	Web	Other	Un-known
Down-stream	Tuesday	11.4%	3.6%	0.3%	1.0%	42.6%	4.6%	36.4%
	Thursday	9.8%	1.3%	0.3%	1.0%	47.4%	4.4%	35.8%
Up-stream	Tuesday	25.7%	1.5%	0.6%	0.4%	6.9%	1.6%	63.2%
	Thursday	24.6%	1.2%	0.9%	0.2%	7.0%	1.6%	64.5%

away from using fixed port numbers to "hiding" by choosing random port numbers [12]. In an unpublished port-based study performed in the same operator's network in 2003, only 20% of the traffic could not be assigned to an application by using the port numbers. In this former study, 64% of the traffic during peak hours had been identified as P2P traffic, while HTTP-traffic was only around 13%. From these findings we can already conclude that – at least during the peak hours – the fraction of interactive (web application-based) traffic has increased dramatically during the last years. To clarify the issue with the high amount of traffic that could not be classified, a different method of subscriber classification is needed. As we decided not to analyze any kind of traffic payload (due to privacy protection and complexity) we have developed a new method, which is described in detail in the following section.

B. Traffic Analysis Based on User Classification

All P2P clients exhibit at least one common characteristic. As they need to communicate with a high number of remote peers for signaling and data transfer, they usually have much more open flows to different IP addresses compared to other applications. Lab tests revealed more than 40 visible peers for BitTorrent and even more than 100 peers for EDonkey during a timeframe of only 120s. Thus, our approach for classifying users into P2P and non-P2P users will be based on exploiting this typical characteristic.

Peer-Based User Classification

For a first analysis, we create a viewgraph that shows the number of distinct peers in upstream direction versus the number of distinct peers in downstream direction per user. Each customer is represented by a dot in the graph. The traffic belonging to one distinct IP address on the user-side is classified as belonging to one customer, thus in the following, the term "customer" (or "user") is represented by the entity that uses this IP address. Figure 2 shows such a graph for experimental data from Tuesday. Nearly all the customers cause points close to the bisector, implying they received packets from as many peers as they had sent packets to. Furthermore, the majority of customers seem to have far less than 100 peers in either direction. Some customers deviate from the straight line. In case they receive packets from more peers they sent packets to, they might either be subject to an attack by port scans or have received an IP address that had been previously used by a customer running a P2P client. As in a P2P network it takes quite a long time to propagate the change of a clients IP address, they receive many packets from different sources that were destined for the customer that had used this IP address before.

Customers showing more peers in upstream direction might be originating port scans or – in case they are P2P users – trying to send packets to hosts that are no more reachable at the IP address that is still stored in the routing database of the P2P overlay network. This type of viewgraph looks similar for all

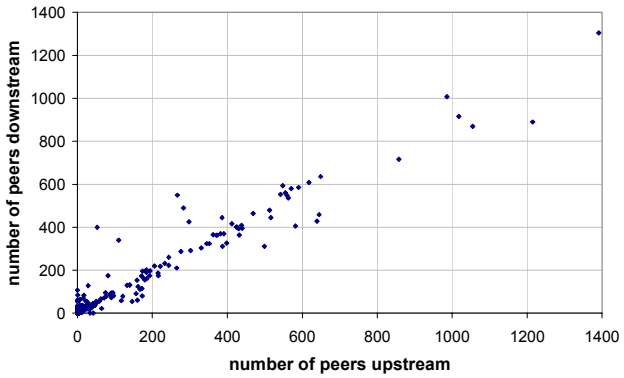

Fig. 2. Number of peers upstream vs. downstream on Tuesday, 10 p.m.

different daytimes and days of our experiments. A few customers that are not shown in Figure 2 have extraordinary high amounts of peers. It turned out that the customers having more than 6000 peers upstream but less than 700 peers downstream exhibited a traffic pattern that was common to a specific computer virus showing a "return rate" of 5-10% of the contacted hosts. One customer had more than 3500 peers in both directions which turned out to be a user running a BitTorrent server.

Although the majority of customers had far less than 100 peers in either direction, the plots do still not allow us to distinguish clearly between P2P and non-P2P users. Thus, we plot the data in a different way. We determine for each customer (i.e. IP address) the amount of peers in downstream and upstream direction and then define the minimum of those two values as the number of peers he has. All customers are then sorted according to the amount of peers they have and we plot the cumulative distribution function (CDF) of this data against the number of peers. Figure 3 shows the CDF on a logarithmic scale for experimental data taken on Tuesday.

We show only the graphs for three sets of data taken on the same day at night-time, during daytime and in the peak hour. The graphs for other daytimes fit in between the ones shown. CDF plots from other days exhibit the same characteristics.

From the graphs we can clearly deduce that at any given point in time there is a vast majority of more than 85% of the customers that do not exchange packets with more than 20

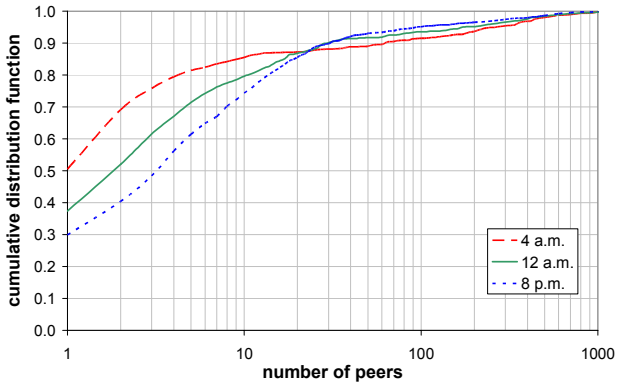

Fig. 3. CDF of customers ranked by number of their peers on Tuesday

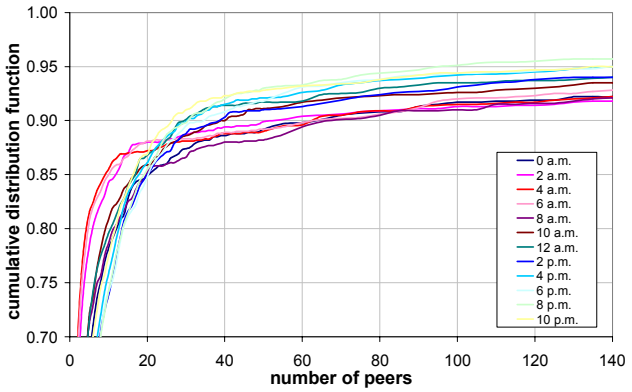

Fig. 4. CDF of customers ranked by number of their peers on Tuesday

peers during the measurement period. Only a few customers have more than 100 peers. It can also be seen that at night-time the fraction of customers having a high number of peers is increased. This reflects the assumption that at this time the fraction of P2P users is higher than in the peak hour and so – as a consequence of P2P applications communicating with many peers – more peers can be observed.

As we need to define a threshold value of peers to classify customers either into the group of P2P or non-P2P users, we magnify a part of the viewgraph and show it on a linear scale for all experimental data available from Tuesday.

We can see that between 20 and 40 peers the slope of the curve is strongly decreasing. This is a clear hint that the user type changes at those numbers. Beyond 100 peers the curves are already very flat. To decide on which value to take for the further studies we combine the deduced number of P2P users over time for different threshold values with an additional port-based classification. For the latter, we counted all customers that sent at least three packets upstream using one of the well known ports for P2P applications. Users just receiving P2P packets from the Internet while not running a P2P application themselves will thus not be counted in. The outcome of this comparison is shown in Figure 5.

The graph shows the amount of identified P2P users using our method with threshold values of 30, 40 and 100 peers. It is obvious that if we lower the threshold, the derived number of P2P users rises. Based on the port-based approach, we added a curve for the amount of users that showed packets using the known port numbers from EDonkey and one that shows the amount of users that use any type of P2P application (mainly by adding the BitTorrent port numbers). The curve for 100 peers fits quite nicely with the one derived for EDonkey users, the curve for 40 peers fits to the one derived for all filesharing ports. This verifies our test results that EDonkey tends to have the highest amount of peers. Checking the database reveals that both methods identified the same customers with a hit rate higher than 95%. Thus, we decided to take 40 peers as threshold value for every further analysis. Moving further down to 30 peers increases the probability of "false positives". The above comparison had been done for all days with available experimental data and leads to the same results.

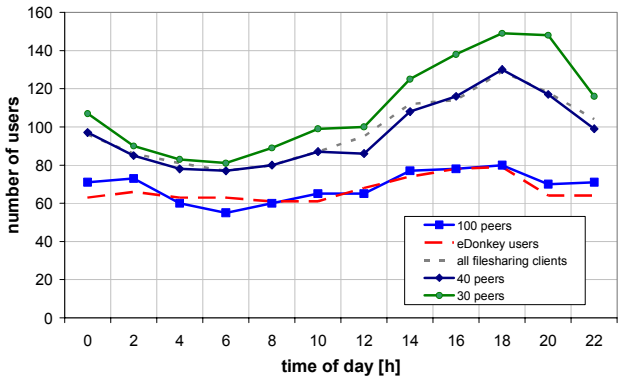

Fig. 5. Comparison of peer- and port-based analysis

Numbers of Users

Now that we can classify the users into P2P and non-P2P users, we take a closer look at the number of users over time. We classify users into inactive ones, active ones and active ones that run a P2P application. An active one is classified by a user that sends and receives packets during our measurement interval. Figure 6 shows the number of active users for Tuesday against time. The number of active subscribers during the peak hours is about three times higher than during night-time. Even at night-time more than 10% of the users are active. This high number might be due to some home gateways staying connected the whole day and sending scheduled packets like regular SIP registrations or just answering packets from the Internet.

While the total number of active users resembles quite well the bandwidth curve from Figure 1, the fraction of P2P users does not exhibit strong time dependence. At the peak hour the fraction of P2P users is only about 4% of the total number of users connected. During night-time, this number drops to about 2.5% of the total connected users. It seems that some users turn off their P2P clients at night-time, but the majority of P2P users stays connected throughout the whole day. This effect can be seen more clearly in Figure 5, where the curves show only moderate time dependence. As Gummadi et al. have shown in a detailed study [10], in P2P file sharing networks it takes quite long for downloads to complete. Thus, there is a strong need to keep the client running for a rather long time.

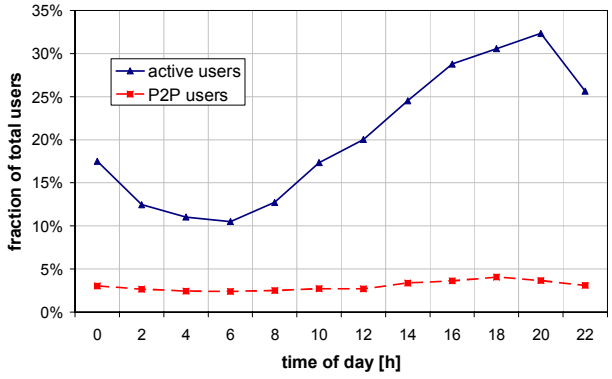

Fig. 6. Number of users in different states on Tuesday

Consumed Bandwidth per User

Now that we have learned how many subscribers are using P2P clients, we deduce their impact on the aggregate value that drives network cost, i.e. the total bandwidth. As the bottlenecks in current networks are in the downstream direction, we first focus on this direction of the traffic.

In Figure 7, each dot resembles one user. It shows the used downstream bandwidth versus the number of peers with a threshold value of 40 peers. P2P users are plotted in pink while non-P2P users are shown in blue. The plot already reveals a very important fact: It is not only the P2P users that consume huge amounts of bandwidth. In contrast, the users consuming the highest bandwidths are not the P2P users but ones that download from a few traffic sources. Popular sources have been identified as a network based video recorder service, RapidShare servers and YouTube videos.

In this study, we classify users and take the aggregate bandwidth. Thus, we do not collect only P2P traffic resulting from P2P users but all the traffic coming from those users. So in principle, we also count for example web based traffic of P2P users. But the traffic of those users is clearly dominated by P2P applications. Looking at the top 50 flows on Tuesday at daytime ranging from 640kbps down to 84kbps, reveals only 5 of them that could be identified as not resulting from a P2P application. At nighttime not a single one could be found. So in practice the difference between "P2P traffic" and the aggregated "traffic of P2P users" is negligible.

Figure 8 shows the total downstream bandwidth used, together with the fraction of it that was caused solely by the P2P users on Tuesday. This proves our findings from previous sections: The traffic profile can be separated into a rather constant part caused by P2P users who keep their computers running throughout the whole day and into a dynamic part that is caused by interactive non-P2P users. The curve for those interactive (i.e. non-P2P) users clearly follows the curve of the number of active users. At night-time there is only the steady (i.e. P2P) part left.

Another interesting number is the ratio of the bandwidth usage at night-time compared to the one in the peak hour. In our measurements we see a ratio of about 1:3. Four years ago it had been about 1:2 throughout the whole network. This implies that within the last years the DSL bandwidth consumption curves have shifted towards a more time-dependent behavior resembling interactive usage. Although the impact of P2P traffic in the peak hour has decreased, still only a small fraction of the users, approximately 4%, are responsible for more than 40% of the traffic volume in the peak hour.

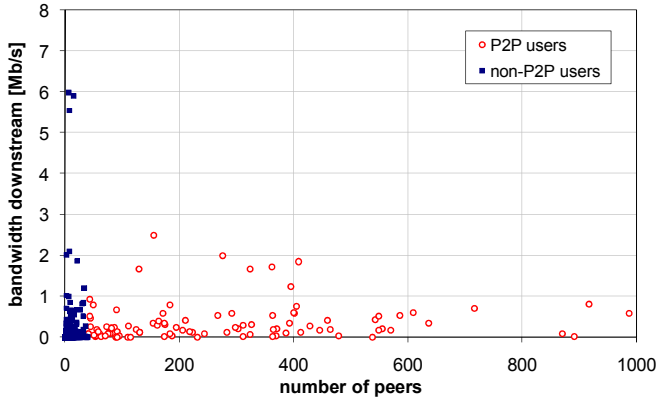

Fig. 7. Bandwidth used downstream versus number of peers downstream on Tuesday at 10 p.m.

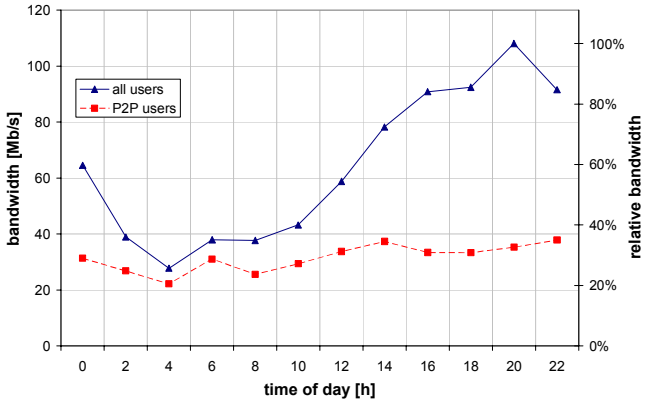

Fig. 8. Total bandwidth usage downstream on Tuesday including the fraction of P2P traffic

As shown in Figure 9, the upstream traffic is completely dominated by P2P users. Only during the evening hours the total upstream traffic deviates notably from the P2P curve.

The average bandwidth per active customer is not strongly dependent on time for any status in which the users can be. So, we calculate the average values only for two parts of the day, night-time (2-6 a.m.) and the evening hours (6-10 p.m.).

Table 2 shows the results of this analysis based on the data collected on all four days. The table shows the mean values calculated over all days since no big differences of the average values have been identified between weekdays and weekends. Active users on average need around 100 kb/s in downstream direction. In contrast, the small fraction of P2P users consumes more than 320 kb/s on average. But even this value is far below the technically possible maximum bandwidth of 6 or even 16 Mb/s, which shows that even P2P applications cannot utilize all the potential bandwidth offered by the network. This is due to the fact that the sources of one user are

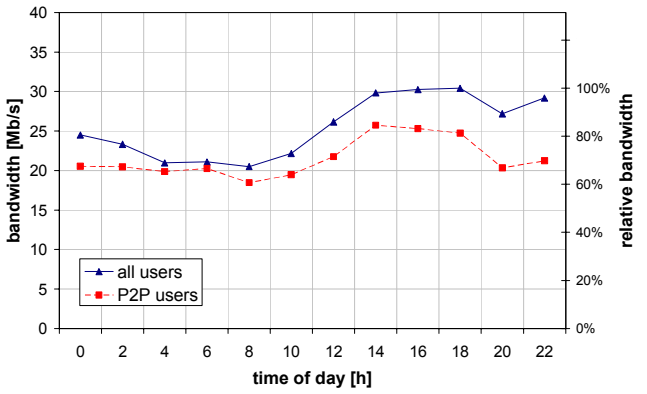

Fig. 9. Total bandwidth usage upstream on Tuesday including the fraction of P2P traffic

TABLE II
AVERAGE BANDWIDTHS PER USER TYPE

	BW Upstream [kb/s]		BW Downstream [kb/s]	
	Night	Evening	Night	Evening
Average b/w per active user	61.4	33.8	107.9	103.6
Average b/w per P2P user	263.3	202.9	327.6	321.5

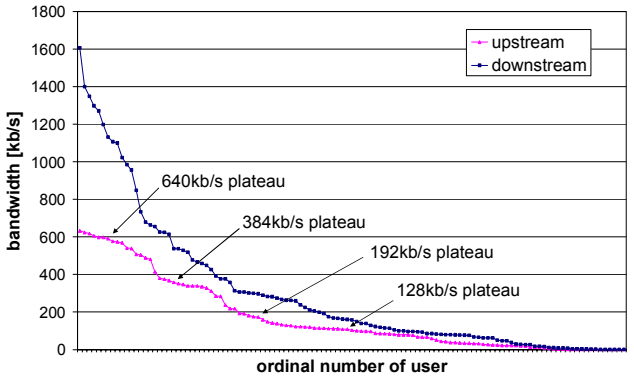

Fig. 10. Histogram of bandwidths used by P2P customers on Tuesday, 20h

other Internet users, who have limited upstream only.

Consequently, P2P users require also a huge amount of upstream bandwidth. At daytime the average upstream bandwidth for these users exceeds 200 kb/s which is more than 60% of the average downstream bandwidth. As discussed above, this simply resembles the fact that the upstream traffic of one P2P user is the downstream traffic of other P2P users. At night-time their average bandwidth consumption increases. This might be due to the fact that no other applications are using the same lines at that time. Ranking the users of the two classes in decreasing amount of bandwidth and plotting a histogram for downstream and upstream direction into one viewgraph, leads to the results shown in Figure 10 for data taken on Tuesday at 8 p.m. using a threshold of 40 peers.

The upstream traffic clearly exhibits plateaus that reflect the maximum bandwidths that are available in different tariff models. P2P tends to saturate the upstream bandwidth to the maximum available rate. Plotting the same histogram for non-P2P users does not exhibit these steps. In downstream direction we see a ramp up, but the curve does not get into the range of the available 16 or 6 Mb/s. Since the sources for P2P downstream traffic are expected to be mainly provided by the residential customer's uplinks, the available uplink bandwidths seem to limit also the downstream.

Traffic Locality

Based on the IP addresses of the customer's communication partners, we have performed an analysis on how local the P2P users' traffic is exchanged. The addresses could be classified as either being part of the address range used in the same area (serving around 8-10% of all customers of the ISP), being part of the ISPs network or not residing in the ISPs network at all. It turns out that only 1.5-3% of the P2P users' traffic is exchanged at the local exchanges (first IP hop in the aggregation network). Overall, only 7.5-13.5% of the traffic, either in up- or downstream direction was exchanged within the operator's network. The ratio of both fractions reflects the relative size of the area compared to the whole network. Reasons for this behavior can be either that a huge amount for content does not exist on clients inside the operator's network or that the metrics of the P2P overlay network does not take into account the local network topology.

IV. RELATED WORK

A recent study on traffic characteristics in broadband networks published by LightReading claims that HTTP traffic now contributes nearly 39% to the peak bandwidth, while the contribution by P2P traffic has decreased to 37% [27]. This finding reflects our results. Port-based studies in the DSL network of France Telecom [19], [20] revealed a similar timely dependence of the P2P traffic and also showed that EDonkey clients connect to much more peers than other P2P applications like e.g., implementations of BitTorrent.

Only a few of the papers in literature focus on a broadband subscriber-based traffic analysis. In a studies done in a Hungarian Operator's network [7, 17], the overall traffic characteristics exhibits the same ratio of 1:3 between the load at night-time and the peak load that we observed.

Fukuda, Cho, and Esaki [8] have classified subscribers in Japanese ISPs' backbone networks by the amount of data they transferred in a given time period into heavy and normal users. As they had not been classifying customers into P2P and non-P2P users, the average bandwidth per user of 230 kb/s is not directly comparable to our results. Further work of this group included an analysis on the number of peers each customer shows [4], but since the authors did not count all remote peers, the results from their study are again not directly comparable to ours.

The locality of current popular P2P services has been topic of many studies, which all agree that locality is very poor [4],[8],[9],[10],[14]. Our data confirms these findings.

V. SUMMARY

We have developed a simple, practical, and robust method of classifying P2P users by focusing on a key characteristic they exhibit, namely the huge amount of peers they communicate with. This method does not require any kind of payload or port-based analysis. Privacy of all user-related data had been assured.

During the peak hours, around 30% of the DSL lines are being actively used by the customers. Only up to 4-5% of all customers are using P2P clients at the same point in time. Most of them stay online during the whole day. Those customers contribute 40% to the total downstream bandwidth during peak hours.

The total downstream bandwidth profile consists of a rather stable background created by the P2P users which is superimposed by a time dependent traffic component caused by the non-P2P users. The downstream bandwidth during peak hours is more and more being driven by interactive applications like videos and direct downloads. Although the

contribution of P2P is still very high, it becomes less important as interactive applications tend to consume more and more bandwidth. In contrast to this, in upstream direction the P2P users' traffic is still by far the dominant driver. The P2P applications tend to use the full available uplink speed. Only a small fraction of the P2P users' traffic is exchanged within the operator's network.

VI. Implications and Future Work

Identifying P2P traffic has become more complex than it has been in the early years. In case there is a need for network operators to identify and manage this type of traffic in real-time this will become a big challenge from many technical viewpoints.

The big question remains on what an operator is supposed to do once he has detected P2P traffic. As we have seen in this study, the impact of the P2P users' traffic during the peak hours has clearly declined. Thus, throttling P2P traffic in order to save bandwidth will not have such a big effect on the overall bandwidth needed as it would have had years ago. The drawbacks of doing so would have to be weighted against the decreasing possible gained bandwidth. Still, new broadband access technologies like VDSL or FTTx might open the race again since they would offer much higher maximum uplink speeds to the applications, resulting in more offerings which then would cause the downstream bandwidth to increase again – even in other operator's networks.

Assisting P2P networks with local caches [14] is another approach to alleviate the impact of P2P traffic. Some possible downsides of this method are legal aspects and the possible loss of the transparency of the Internet service. As already much work on deploying IP TV content using P2P techniques has been started [25], combining this with P2P metrics that take into account the operator's network architecture [1] might then turn P2P networking into a "win-win" solution.

Since the methodology used in this project is irrespective of the P2P protocols and relies on an intrinsic characteristic of all P2P applications, we expect the monitoring framework to be able to track future changes in customer behavior. Thus, we will be able to examine the behavior of the ISP's customers with our new method over a larger timeframe and perform more sophisticated studies like Perényi, Gefferth, Dang, and Molnár. did for Skype traffic [18]. Still the main boundary condition for network traffic and user impact monitoring that needs to be obeyed is privacy of user data.

Acknowledgment

We would like to thank W. Haeffner, U. Trick, J. Aráuz and R. Winter for many fruitful discussions. We would also like to thank M. Ranaudo and X. Liu for great technical support.

References

[1] V. Aggarwal, A. Feldmann and C. Scheideler. Can ISPs and P2P systems cooperate for improved performance? ACM SIGCOMM Computer Communications Review, 37 (2007), pages 31-40.

[2] BitTorrent, http://www.bittorrent.com

[3] BearShare, http://www.bearshare.com

[4] K. Cho, K. Fukuda, H. Esaki, A. Kato. The impact and implications of the growth in residential user-to-user traffic. SIGCOMM, 207-218, 11-15 September 2006, Pisa, Italy

[5] EDonkey, http://en.wikipedia.org/wiki/EDonkey_network

[6] 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, 17(6):6–16, Nov. 2003.

[7] T. D. Dang, M. Perényi, A. Gefferth, and S. Molnár. On the Identification and Analysis of P2P Traffic Aggregation. In Proceedings of 5thInternational IFIP-TC6 Networking Conference, Coimbra, Portugal, May, 2006.

[8] K. Fukuda, K. Cho, and H. Esaki. The impact of residential broadband traffic on Japanese ISP backbones. SIGCOMM CCR, 35(1):15–21, Jan. 2005

[9] A. Gerber, J. Houle, H. Nguyen, M. Roughan, and S. Sen. P2P The Gorilla in the Cable. In National Cable & Telecommunications Association (NCTA) 2003

[10] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In SOSP-19, pages 314–329, Bolton Landing, NY, Oct. 2003.

[11] Ipoque P2P Survey 2006, http://www.ipoque.com/userfiles/file/p2p_survey_2006.pdf

[12] T. Karagiannis, A. Broido, N. Brownlee, kc claffy, and M. Faloutsos. Is p2p dying or just hiding? In GLOBECOM 2004, pages 1532–1538, Dallas, TX, Dec. 2004.

[13] T. Karagiannis, A.Broido, M. Faloutsos, and kc claffy. Transport layer identification of P2P traffic. In ACM SIGCOMM IMC, 2004.

[14] T. Karagiannis, P. Rodriguez, and D. Papagiannaki. Should Internet service providers fear peer-assisted content distribution? In IMC 2005, pages 63–76, Berkeley, CA, Oct.2005.

[15] M. Kim, H. Kang, J. W. Hong. Towards Peer-to-Peer Traffic Analysis Using Flows. DSOM 2003: 55-67.

[16] P. Ohm, D. Sicker, and D. Grunwald. Legal Issues Surrounding Monitoring During Network Research, In IMC 2007, pages 141–148, San Diego, CA, Oct. 2007

[17] M. Perényi, T. D. Dang, A. Gefferth, S. Molnár. Identification and Analysis of Peer-to-Peer Traffic. Journal of Communications, Vol. 1, Issue 7, November/December 2006

[18] M. Perényi, A. Gefferth, T. D. Dang, S. Molnár, Skype Traffic Identification, GLOBECOM 2007, Washington, DC, USA, 26-30 November, 2007.

[19] L. Plissonneau, J.-L. Costeux, and P. Brown. Analysis of peer-to-peer traffic on ADSL. In PAM2005 (LNCS3431), pages 69-82, Boston, MA, Mar. 2005.

[20] L. Plissonneau, J. L. Costeux, P. Brown. Detailed analysis of eDonkey Transfers on ADSL. In: Proceedings of EuroNGI. (2006)

[21] RapidShare, http://www.rapidshare.com

[22] S. Saroiu, P. K. Gummadi, and S. D. Gribble. A measurement study of peer-to-peer file sharing systems. In MMCN'02, pages 156–170, San Jose, CA, Jan. 2002.

[23] S. Sen and J. Wang. Analyzing peer-to-peer traffic across large networks. In ACM SIGCOMM IMW, pages 137-150, Marseille, France, Nov. 2002.

[24] S. Sen, O. Spatscheck, and D. Wang. Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures. In 13th International World Wide Web Conference, New York City, 17-22 May 2004

[25] A. Sentinelli, G. Marfia, M. Gerla, and L. Kleinrock. Will IPTV Ride the Peer-to-Peer Stream? IEEE Commun. Mag., June 2007.

[26] D. Stutzbach, R. Rejaie, and S. Sen. Characterizing unstructured overlay topologies in modern p2p file-share systems. In IMC2005, pages 49–62, Berkeley, CA, Oct. 2005.

[27] Surveys: Internet Traffic Touched by YouTube, January 2007. http://www.lightreading.com/document.asp?doc_id=115816

[28] Wireshark, http://www.wireshark.org

[29] YouTube, http://www.youtube.com