On profiling residential customers

Marcin Pietrzyk[1], Louis Plissonneau[1],
Guillaume Urvoy-Keller[2], and Taoufik En-Najjary[1]

[1] Orange Labs, France
{marcin.pietrzyk,louis.plissonneau,
taoufik.ennajjary}@orange-ftgroup.com
[2] Université de Nice Sophia-Antipolis, Laboratoire I3S CNRS UMR 6070, France
urvoy@unice.fr

Abstract. Some recent large scale studies on residential networks (ADSL and FTTH) have provided important insights concerning the set of applications used in such networks. For instance, it is now apparent that Web based traffic is dominating again at the expense of P2P traffic in lots of countries due to the surge of HTTP streaming and possibly social networks. In this paper we confront the analysis of the overall (high level) traffic characteristics of the residential network with the study of the users traffic profiles. We propose approaches to tackle those issues and illustrate them with traces from an ADSL platform. Our main findings are that even if P2P still dominates the first heavy hitters, the democratization of Web and Streaming traffic is the main cause of the come-back of HTTP. Moreover, the mixture of applications study highlights that these two classes (P2P *vs.* Web + Streaming) are almost never used simultaneously by our residential customers.

1 Introduction

The research community has devoted significant efforts to profile residential traffic in the last couple of years. A large scale study of Japanese residential traffic [5, 4], where almost 40% of the Internet traffic of the island is continuously observed, has revealed specific characteristics of the Japanese traffic: a heavy use of dynamic ports, which suggests a heavy use of P2P applications and a trend of users switching from ADSL to FTTH technology to run P2P along with gaming applications. A recent study in the US [6], where the traffic of 100K DSL users has been profiled with a Deep Packet Inspection tool, has revealed that HTTP traffic is now the dominant protocol at the expense of P2P for the considered ISP, and probably for the US in general. This significant change in the traffic breakdown is not due to a decrease of P2P traffic intensity but a surge of HTTP traffic driven by HTTP streaming services like YouTube and Dailymotion. Similar results have been obtained in European countries. In Germany, a recent study [11] analyzed about 30K ADSL users and also observed that HTTP was again dominant at the expense of P2P traffic, for the same reason as in the US: a surge of video content distribution over HTTP. Early studies in France [15] for an ADSL platform of about 4000 users highlighted the dominance of P2P traffic in 2007 but a subsequent studies on the same PoP [14] or other PoPs under the control of the same ISP revealed similar

traffic trend of HTTP traffic increasing at the expense of P2P both for ADSL [12] and FTTH access technology [17]. In the above studies, the application profiling of residential traffic was used to inform network level performance aspects, e.g., cachability [6] of content or location in the protocol stack of the bottleneck of transfers performed on ADSL networks [15], [11]. The study in [11] further reports on usage of the ADSL lines with a study of the duration of Radius sessions.

The current work aims at filling the gap between the low-level (network) level performance study and high level (application) study by profiling ADSL users. We use hierarchical clustering techniques to aggregate users' profiles according to their application mix. Whereas many studies focus on communication profiles on backbone links, few ones dig into application mix at user level. In the analysis carried in [10], the authors take a *graphlet* approach to profile end-host systems based on their transport-layer behavior, seeking users clusters and "significant" nodes. Authors in [8], take advantage of another clustering technique (namely Kohonen Self-Organizing Maps) to infer customers application profiles and correlate them with other variables (*e.g.* geographical location, customer age).

Our raw material consists of two packet traces collected on the same platform, a few months apart from each other, that are fully indexed in the sense that both IP to user and connection to applications mapping are available. We use this data to discuss different options to profile both a platform and the users of this platform.

The remaining of this paper is organized as follows. In Sect. 2, we detail our data sets. In Sect. 3, we analyze high level traffic characteristics and, the contributions of users to the traffic per application. In Sect. 4, we discuss different options to profile users and come up with a specific approach that allows to understand application usage profiles.

2 Data Set

The raw data for our study consists of two packet level traces collected on an ADSL platform of a major ISP in France (Tab. 1). Each trace lasts one hour and aggregates all the traffic flowing in and out of the platform.

In this platform, ATM is used and each user is mapped to a unique pair of Virtual Path, Virtual Channel, identifiers. As the packet level trace incorporates layer 2 information, we can identify users thanks to this ATM layer information. This approach allows for reliable users tracking. Indeed, 18% of the users change their IP address at least once, with a peak at 9 for one specific user. One could expect that the only source of error made when considering the IP address is that the session of the user is split

Table 1. Traces summary

Label	Start time	Duration	Bytes	Flows	TCP Bytes	TCP Flows	Local Users	Local IPs	Distant IPs
Set A	2009-03-24 10:53 (CET)	1h	31.7G	501K	97.2 %	30.7 %	1819	2223	342K
Set B	2009-09-09 18:20 (CET)	1h	41 G	796K	93.2 %	18.3 %	1820	2098	488K

onto several IP level sessions. However, we also noticed in our traces that a given IP could be reassigned to different users during the periods of observation. Specifically, 3% of the IPs were assigned to more than one user, with a peak of 18 re-assignments for one specific IP. Those results are in line with the ones obtained in [11] for a German residential operator.

Both traces are indexed thanks to a Deep Packet Inspection (DPI) tool developed internally by the ISP we consider. This tool is called ODT. In [13], we have compared ODT to Tstat (http://tstat.tlc.polito.it/), whose latest version features DPI functions. Specifically, we have shown that ODT and Tstat v2 offer similar performance (for most popular applications) and outperform signature based tools used in the literature. As ODT embeds a larger set of signatures than Tstat v2, we rely on the former to map flows and applications.

The classes of traffic we use along with the corresponding applications are reported in Tab. 2. Note that HTTP traffic is broken into several classes depending on the application implemented on top: Webmail is categorized as mail, HTTP streaming as streaming, HTTP file transfers as DOWNLOAD, etc. The OTHERS class aggregates less popular applications that ODT recognized. The DOWNLOAD class consists mainly of HTTP large file transfers from one-click hosting services [1], which are growing competitors of P2P file sharing services. The flows not classified by ODT (e.g. some encrypted applications) are aggregated in the UNKNOWN class.

We developed an ad-hoc C++ trace parser that relies on libpcap to extract the per user statistics from the raw traces. Users' data was anonymized prior to analysis.

Table 2. Application classes

Class	Application/protocol
WEB	HTTP and HTTPs browsing
UNKNOWN	–
P2P	eDonkey, eMule obfuscated, Bittorrent
	Gnutella, Ares, Others
MAIL	SMTP, POP3, IMAP, IMAPs
	POP3s, HTTP Mail
CHAT	MSN, IRC, Jabber
	Yahoo Msn, HTTP Chat
STREAMING	HTTP Streaming, Ms. Media Server,
	iTunes, Quick Time
OTHERS	NBS, Ms-ds, Epmap, Attacks
DB	LDAP, Microsoft SQL, Oracle SQL, mySQL
DOWNLOADS	HTTP file transfer, Ftp-data, Ftp control
GAMES	NFS3, Blizzard Battlenet, Quake II/III
	Counter Strike, HTTP Games
VOIP	Skype
NEWS	Nntp

3 Platform profile

In this section, we highlight high level platform traffic profiles, namely the traffic breakdown and the per users volume distribution.

3.1 Traffic breakdown

We report in Tab. 3 the bytes breakdown views of the two traces, where the DB, CONTROL, NEWS, CHAT and GAMES classes have been omitted as they do not represent more than 1% of bytes and flows in any of the traces. It has been observed in [6] and [11] that HTTP based traffic was again dominating at the expense of P2P traffic in residential networks in US and Europe. The traffic breakdown of our platform suggests the same conclusion. Indeed, when summing all HTTP-based traffic in sets A or B, namely Web, HTTP Streaming and HTTP Download, more than 50% of the bytes in the down direction is carried over HTTP. Clearly, HTTP driven traffic dominates at the expense of background traffic that is due to P2P applications.

3.2 Distributions of volumes per user

Understanding the relative contribution of each user to the total amount of bytes generated by dominating applications is important. Indeed these results, even if not surprising, justify the approach of focusing on heavy hitters in the last section of the paper.

In Fig. 1, we present the contribution of users to the total traffic aggregate per application, with users sorted by decreasing volumes for the considered application (sets A and B being similar we focus on set A here). Note that we sum up, for a user, her bytes in both directions. We also include in the graph the overall contribution by user without distinguishing per application.

The fraction of users contributing to the majority of bytes in each application and even overall is fairly small. When looking at the global volumes generated, 90% of the bytes are generated by about 18% of users. For the same volume quantile, the fraction of users involved is even smaller when focusing on the applications generating most of the bytes (those represented in the graph). For the case of P2P traffic for instance, only

Table 3. Traffic Breakdown (Classes with more than 1% of bytes only).

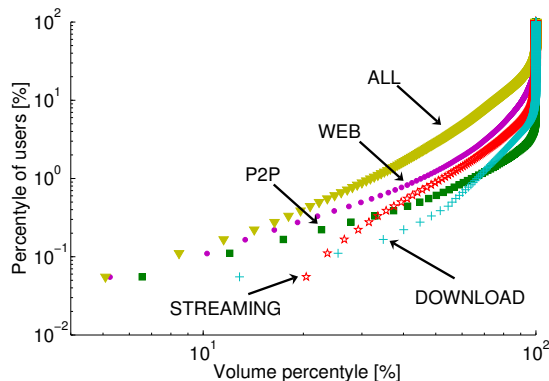| Class	Set A Bytes	Set B Bytes
WEB	22.68 %	20.67 %
P2P	37.84 %	28.69 %
STREAMING	25.9 %	24.91 %
DOWNLOAD	4.31 %	6.47 %
MAIL	1.45 %	0.54 %
OTHERS	1.04 %	0.44 %
VOIP	0.36 %	1.67 %
UNKNOWN	5.26 %	15.79 %

Fig. 1. Contribution of users to traffic aggregate (global and per application). Set A

0.3% of the users contribute to 90% of the bytes uploaded or downloaded. We confirm here the well known phenomenon explored for instance in [7, 3]. This also holds for the Streaming and Web classes, which are two key classes in the dimensioning process of links of ISPs (for example bulk of Streaming users is active in the evenings).

A consequence of these highly skewed distributions is that the arrival or departure of some customers on the platform can potentially have an important impact on the traffic shape. For instance, the first four heavy users of streaming are responsible for about 30% of all streaming traffic.

The above observations also motivates our approach in the next section which is on profiling customers (and especially heavy hitters) from their application usage perspective.

4 Users Profiling

In this section, we address the issue of building an application level profile of customers that would characterize their network usage. The problem is challenging as it can be addressed from many different viewpoints. Here are some questions that one might want to answer: Which amount of bytes or alternatively which number of flows should be observed to declare that a user is actually using a specific application? Can we characterize users thanks to the dominant application they use? What is the typical application profile of a heavy hitter? What is the typical application mix of the users?

We address the above questions in the next paragraphs. We discuss several options to map applications to users. Our first approach focuses on the dominating applications for each user, we further discuss the precise profile of the top ten heavy hitters in both traces. Last paragraph presents typical users application mixture using clustering technique.

4.1 Users dominating application

We present here a simple approach that provides an intuitive high level overview of the users activity: we label each user with her dominating application, the application that generated the largest fraction of bytes. Such an approach is justified by the fact that for both of our data sets, the dominating application explains a significant fraction of the bytes of the user. Indeed, for over 75% of the users, it explains more than half of the bytes. This phenomenon is even more pronounced when considering heavy users. Fig. 2 presents the distribution of the fraction of the bytes explained depending on which application dominates users activity.
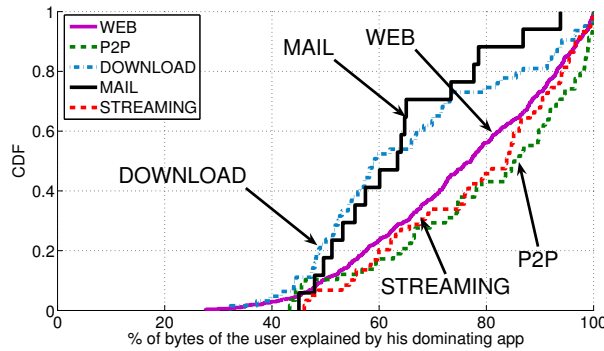

Fig. 2. CDF of the fraction of bytes explained by the dominant application of each user. Set B.

The distribution of users per application with such an approach (dominant application) is reported in Tab. 4. As expected, the dominating class is Web. We have more Streaming than P2P dominated users. This complies with the intuition that every user, even if not experienced, can watch a YouTube video, whereas using a P2P application requires installing a specific software (P2P client). The remaining dominant applications correspond to clients that generate a small amount of bytes most of the time. For instance, users that have DB, Others, Control or Games as dominating application generate an overall number of bytes that is extremely low.

We present in Fig. 3 the users to application mapping for set B using the above dominant application approach. We adopt a representation in which each user is characterized by the total number of bytes she generates in the up and down direction and label the corresponding point in a two dimensional space with the dominant application of the user in terms of bytes. We restricted the figure to a list of 6 important applications: Web, Streaming, VOIP, Download and P2P. We further added the users having majority of bytes in the Unknown class to assess their behavior.

Most important lesson of Fig. 2 is that labeling a client with her dominating application is meaningful. Indeed, the dominating application in terms of bytes usually generates the vast majority of users' total volume. Customers with the same dominat-

Table 4. Users dominating applications breakdown. Each user is labeled with his dominant application in terms of bytes. (Only users that transfered at least 100B: 1755 users). Set B

Class	Fraction of Users	Fraction of Bytes explained
UNKNOWN	21%	12%
WEB	35%	19%
P2P	4%	35%
DOWN	5%	≤ 1%
MAIL	1%	≤ 1%
DB	9%	≤ 1%
OTHERS	8%	≤ 1%
CONTROL	7%	≤ 1%
GAMES	≤ 1%	≤ 1%
STREAMING	7%	25%
CHAT	1%	≤ 1%
VOIP	1%	2%

ing applications are clustered together, and exhibit behavior typical for this application, which we detail below.

We observe from Fig. 3 that:

– P2P heavy hitters tend to generate more symmetric traffic than Download and Streaming heavy hitters, which are far below the bisector.
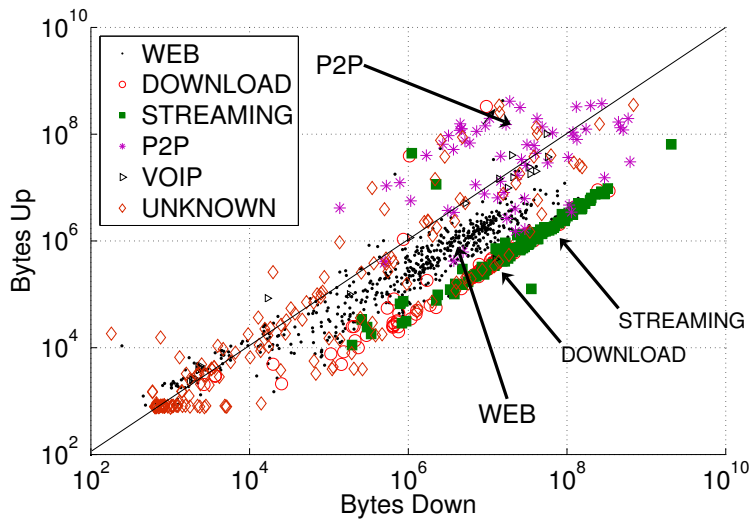

**Fig. 3.** Users bytes Up/Down. Dominating application marked. Set B

– Web users fall mostly in between the bisector and the heavy hitters from the Download and Streaming classes. This is also in accordance with intuition as Web browsing often requires data exchange from clients to servers, e.g., when using Web search engines. This is in contrast to Streaming or Download where data flow mainly from servers to clients.
– Concerning Unknown users, we observe first that a significant fraction of them generated almost no traffic as they lay in the bottom-left corner of the plot. As for Unknown heavy hitters, we observe that they are closer on the figure to P2P heavy users than to client-server heavy users. This might indicate that there exist some P2P applications that fly below the radar of our DPI tool. We further investigate this issue in the next section.

A last key remark is that the equivalent of Fig. 3 for set A is qualitatively very similar, emphasizing the overall similarity of users activity in the two data sets (even if several month apart and at a different time of day).

The above analysis has again underlined the crucial role of (per application) heavy hitters. In the next section, we will focus on the top 10 heavy hitters in each trace. Each of them generated at least 0.6 GB of data and up to 2.1 GB and, overall, they are responsible for at least 1/4 of the bytes in each trace. We profile these users by accounting simultaneously for all the applications they use.

4.2 Top ten heavy hitters

In this section, we focus on the top 10 heavy hitters for sets A and B. Note that these are distinct sets of users. It is a small, but very important group of customers from the ISP perspective, and better understanding of this group (aggregating 1/4 of total volume) might have significant impact on network provisioning and dimensioning. Fig. 4(a) and 4(b) show the fraction of bytes they have generated in the up (positive values) and down direction (negative values) for each application. For sake of clarity, we put in the figure only the labels of the significant applications for each user. We do observe from Fig. 4(a) and 4(b) that heavy hitters, for the most part, use P2P applications. Streaming and (at least for one user) download activities seem also to give birth to some heavy hitters.
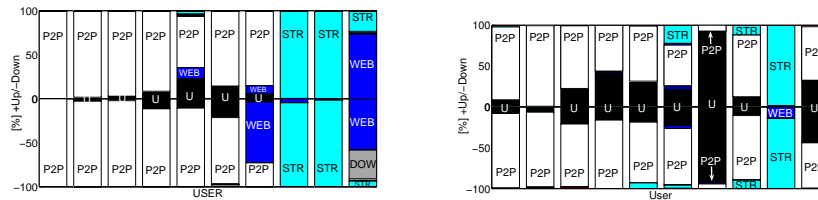
(a) Set A (Users generating 31% of the bytes in the trace) (b) Set B (Users generating 24% of the bytes in the trace)

Fig. 4. Top 10 heavy hitter users. Application usage profiles expressed in bytes fractions. (U stands for UNKNOWN)

We also observe that unknown traffic seems to be associated mostly with P2P users (which is in line with Fig. 3). This is an important finding from the perspective of the traffic classification, which often relies on per flow features. This user level information could be used as a feature in the classifier. It is also in line with the findings in [12] where it is shown that a significant fraction of bytes in the unknown category (we use the same DPI tool but different traces) is generated by P2P applications. In the present case, 67 % and 95 % of unknown bytes are generated by the users having in parallel peer-to-peer activity for set A and B respectively. The reason why some of the P2P traffic might be missed by our DPI tool is out of the scope of the paper. We note that there are at least two possible explanations: either we missed in our trace the beginning of a long P2P transfer and the DPI tool might not have enough information[3] to take a decision, or these users run currently unknown P2P applications in parallel.

4.3 Users application mix

In the previous sections, we analyzed our users profile taking only bytes into account. This approach is informative and makes sense from a dimensioning viewpoint. However as the per applications volumes are very different – e.g., P2P applications tend to generate much more bytes than Web browsing – we miss some usage information with this purely byte-based approach. In this section, we explore a different perspective. We associate to each user a binary vector, which indicates her usage of each application. We take advantage of clustering techniques to present typical application mixtures.

"Real" *vs*. "fake" usage We represent each customer with a binary vector: $A = [appli_1 \cdots appli_n]$ where n is the number of applications we consider. Each $appli_i \in \{0, 1\}$ is a indication weather the customer used application i or not. We define per application heuristics to declare that a customer actually uses a class of application. To do that, we define minimal thresholds for three metrics: bytes up, bytes down and number of flows. Depending on the application any or all of the three thresholds need to be matched. We summarize the heuristics in Tab. 5. The values were derived from the data as it is exemplified in Fig. 5 for P2P and WEB traffic.

---

[3] Application level information are often at the onset of transfers [2].
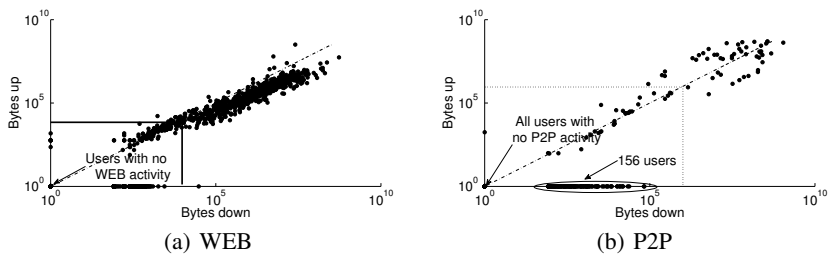

(a) WEB   (b) P2P

Fig. 5. Example of how the threshold is selected. Set A

Table 5. Ad-hoc, per application and user minimum thresholds to declare application usage

Class	Volume		Number of Flows	Policy
	Down	Up		
WEB	300kB	500kB	20	All
P2P	1 MB	1 MB	10	Any
STREAMING	1 MB	1 MB	–	Any
DOWNLOAD	2 kB	1 kB	–	Any
MAIL	30kB	3 kB	–	All
GAMES	5 kB	5 kB	–	Any
VOIP	200kB	200kB	–	All
CHAT	10kB	10kB	–	Any

Heuristics are necessary to separate real application usage from measurements artifacts. For instance, large fraction of users of the platform have a single flow which is declared by the DPI tool as WEB browsing. It is hard to believe that this flow is a real web browsing activity, as current web sites tend to generate multiple connections for a single site (single search without browsing on google.com shows up to 7 connections). Similar problems might occur with other applications, for instance peer-to-peer user that closed his application, might still receive file requests for some time due to the way some P2P overlays work.

Choice of clustering We have considered several popular clustering techniques to be able to understand the application mix of each user, see [9] for a complete reference on main clustering techniques. As explained in the previous paragraph, we have discretized the user's characteristics according to some heuristic threshold in order to keep only "real" application usage.

We have first tried the popular k-means clustering algorithm, and observed that the resulting clusters are difficult to match to applications. Moreover the choice of the number of clusters can dramatically change this representation.

Hierarchical clustering offers an easily interpretable technique for grouping similar users. The approach is to take all the users as tree leaves, and group leaves according to their application usage (binary values). We choose an agglomerative (or down-up) method:

1. The two closest nodes[4] in the tree are grouped together;
2. They are replaced by a new node by a process called linkage;
3. The new set of nodes is aggregated until there is only a single root for the tree.

With this clustering algorithm, the choices of metric and linkage have to be customized for our purpose.

We want to create clusters of users that are relatively close considering the applications mix they use. Among comprehensive metrics for clustering categorical attributes the Tanimoto distance [16] achieves these requirements. It is defined as follows: $d(x,y) = 1 - \frac{x^t \cdot y}{x^t \cdot x + y^t \cdot y - x^t \cdot y}$. This means that users having higher number of

---

[4] at first occurrence, nodes are leaves

common applications will be close to each other. For example, consider 3 users having the following mix of applications[5]:

User	Web	Streaming	Down	P2P
A	1	1	0	0
B	1	1	1	0
C	1	1	0	1

With Tanimoto distance, users B and C will be closer to each other because they have same total number of applications even if all 3 users share same common applications.

We use a complete linkage clustering, where the distance between nodes (consisting of one or several leaves) is the maximum distance among every pair of leaves of these nodes. It is also called farthest neighbor linkage.

Due to the chosen metric, and as we chose not to prune the resulting tree, the hierarchical clustering leads to as many clusters as there are applications combinations: $\sum_{i=1}^{n} \binom{n}{i}$. In our case, we restrict the set of applications we focus only to Web, Streaming, P2P and Download.

Applications mix We present in Fig. 6 and 7 the clustering results for the top 50 and second 50 most active users respectively. In total, the first one hundred users of the platform are responsible for 75% of the volume. We first consider only the classes generating most of the traffic, as described by Tab. 3 namely: Web, P2P, Streaming, and Download

Each barplot represents a single user and expresses his total volume share. Barplots (thus users) are grouped into the sorted clusters. Each cluster, indicated by a different color groups the users that had the same applications. Thus close clusters in the graph are similar with respect to their application mix.

Considering only four applications, we have 15 possible combinations. What we observe is that some combinations are clearly more popular than others, while a few of them never occurred in our data. We present below a more precise analysis that reveals some insights about the typical users profiles.

Looking at the top 50 active users, we see that the P2P cluster clearly dominates (34 users). The most popular cluster is due to the "pure" peer-to-peer clients, followed by the users that use both P2P and Web. The P2P related clusters (P2P only, P2P + Web, P2P + Web + Streaming) aggregate 40% of the total volume of the trace. Pure Web + Streaming profiles are a minority, although the biggest heavy hitter of the trace (over 5% of the whole traffic) is a Streaming user.

The second set of 50 most active clients reveals an inverted picture. Here, over 30 users show different profiles consisting of mainly Web and Streaming (DOWNLOAD is minority), while the group of P2P users is very small.

It is interesting to see that P2P and Streaming/Web users form very distinct groups as only 10 of 100 most active users mix these 2 applications. This is also the case with Download whose profile almost never overlaps P2P. This shows that there is a set of

---

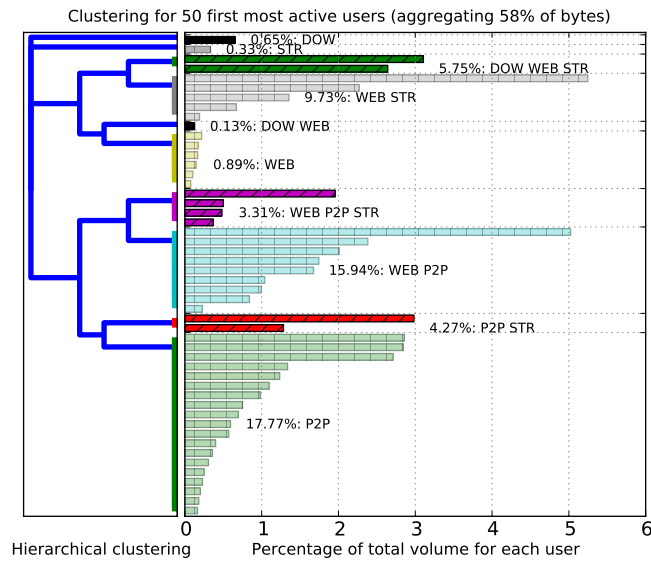[5] 1 means application usage and 0 means no application usage.

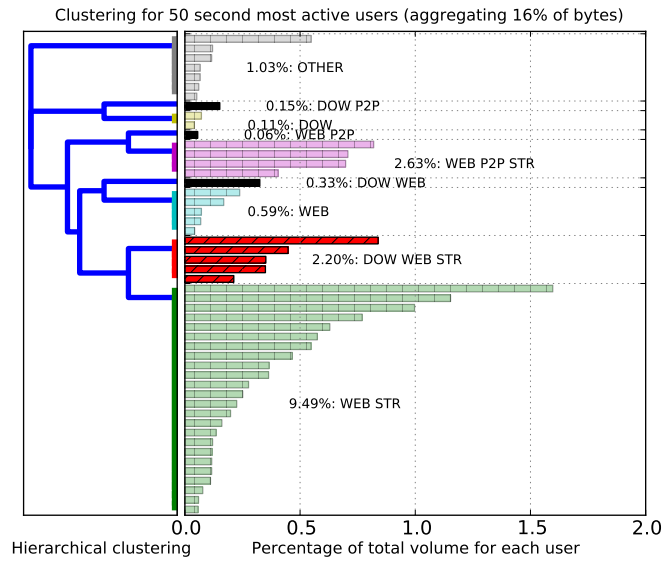**Fig. 6.** Application clustering for top 50 most active users. Set A.

Fig. 7. Application clustering for top 51-100 most active users. Set A.

clients that prefer classical P2P and another set of clients that use one click hosting to download contents.

Application mix - discussion Focusing on the first heavy hitters we observe that this family of users is dominated by P2P heavy-hitters. Even if streaming activity can also lead a user to become a heavy user, the main part of the volume generated by this class comes from a majority of medium users.

We conjecture that this situation will persist as the popularity of streaming continues to increase. Indeed, this increase of popularity is likely to translate into more users streaming more videos rather than a few users streaming a lot. If the main content providers switch to High Definition video encoding (which has bit-rates up to 4 times larger than standard definition), this could have a dramatic impact for ISPs.

5 Conclusion

In this paper, we have proposed and illustrated several simple techniques to profile residential customers, with respect to their application level characteristics.

We have first presented an approach where the focus is on the dominant application of a user, which is justified by the fact that the dominant application explains a large majority of bytes for most users (in our data sets at least). This approach enables us to observe overall trends among moderately heavy and heavy users in a platform. We have next focused more deeply on the heavy hitters. Those heavy hitters are mostly P2P users, even though the global trend of traffic shows that Web and Streaming classes dominate. It is however understandable as P2P applications naturally tend to generate a few heavy hitters, while Web and Streaming tend to increase the volume of traffic of the average user.

We also devised an approach that seeks for common application mixes among the most active users of the platform. To this aim, we defined per application thresholds to differentiate real usage of an application from measurement artifacts. We use hierarchical clustering, that groups customers into a limited number of usage profiles. By focusing on the 100 most active users, divided in two equal sets, we demonstrated that:

- P2P users (pure P2P or mixed with other applications) are dominant in number and volume among the first 50 most active users;
- whereas in the second set of 50 most active users, the killer application is the combination of Web and Streaming.

Moreover while almost all P2P bytes are generated by the first 50 most active users, the Web + Streaming class is used by many users, and generates a fraction of bytes comparable (or higher) to P2P.

Our study sheds light on the traffic profile of the most active users in a residential platform, which has many implications for ISPs. However, we have only scratched the surface of the problem.Application at a larger scale of similar techniques, e.g., on much longer traces, would bring more insights than the snapshots we analyzed. As part of our future work, we plan to further extend the analysis, by tracking the evolution of users profiles on the long term.

We strongly believe that hierarchical clustering on discretized attributes is a good approach because it greatly eases interpretation of the resulting clusters. Still, we plan to extend the discretization process from binary to (at least) ternary variables to take into account low/medium usage of an application *vs.* high usage.

References

1. D. Antoniades, E. P. Markatos, and C. Dovrolis. One-click hosting services: a file-sharing hideout. In *IMC*, 2009.
2. L. Bernaille, R. Teixeira, and K. Salamatian. Early application identification. In *CoNEXT*, 2006.
3. L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and Zipf-like distributions: evidence and implications . In *INFOCOM*, 1999.
4. K. Cho. Broadband Traffic Report. *Internet Infrastructure Review*, 4:18–23, August 2009.
5. K. Cho, K. Fukuda, H. Esaki, and A. Kato. The impact and implications of the growth in residential user-to-user traffic. In *SIGCOMM*, 2006.
6. J. Erman, A. Gerber, M. T. Hajiaghayi, D. Pei, and O. Spatscheck. Network-aware forward caching. In *WWW*, 2009.
7. A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True. Deriving traffic demands for operational IP networks: methodology and experience. *IEEE/ACM Trans. Netw.*, 9(3):265–280, 2001.
8. F. Fessant, V. Lemaire, and F. Clrot. Combining Several SOM Approaches in Data Mining: Application to ADSL Customer Behaviours Analysis. In *Data Analysis, Machine Learning and Applications*. 2008.
9. J. Han. *Data Mining: Concepts and Techniques (Second Edition)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2006.
10. T. Karagiannis, K. Papagiannaki, N. Taft, and M. Faloutsos. Profiling the end host. In *In PAM*, 2007.
11. G. Maier, A. Feldmann, V. Paxson, and M. Allman. On dominant characteristics of residential broadband internet traffic. In *IMC*, 2009.
12. M. Pietrzyk, J.-L. Costeux, G. Urvoy-Keller, and T. En-Najjary. Challenging statistical classification for operational usage: the ADSL case. In *IMC*, 2009.
13. M. Pietrzyk, G. Urvoy-Keller, and J.-L. Costeux. Revealing the unknown ADSL traffic using statistical methods. In *COST*, 2009.
14. L. Plissonneau, T. En-Najjary, and G. Urvoy-Keller. Revisiting web traffic from a DSL provider perspective : the case of YouTube. In *ITC Specialist Seminar on Network Usage and Traffic*, 2008.
15. M. Siekkinen, D. Collange, G. Urvoy-Keller, and E. W. Biersack. Performance Limitations of ADSL Users: A Case Study. In *PAM*, 2007.
16. T. Tanimoto. An elementary mathematical theory of classification and prediction. In *IBM Program IBCLF*, 1959.
17. G. Vu-Brugier. Analysis of the impact of early fiber access deployment on residential Internet traffic. In *ITC 21*, 2009.