Analysis of Peer-to-Peer Traffic on ADSL

Louis Plissonneau, Jean-Laurent Costeux, and Patrick Brown

France Telecom R&D
905, rue Albert Einstein
06921 Sophia-Antipolis Cedex - FRANCE
{louis.plissonneau,jeanlaurent.costeux}@francetelecom.com**

Abstract. Peer-to-Peer (P2P) applications now generate the majority of Internet traffic, particularly for users on ADSL because of flatrate tarification. In this study, we focus on four popular P2P systems to characterize the utilization, the performance and the evolution of P2P traffic in general. We observe and compare the influence of each P2P application over the traffic, and we evaluate the evolution of these P2P systems over a year. Our analysis is based on ADSL traffic captured at TCP level on a Broadband Access Server comprising thousands of users. Thus, we characterize the P2P traffic and users, and we draw interesting results on connectivity and cooperation between peers, localization of sources, termination of connections and performance limitations. The evolution of the traffic over the year allows us to see the dynamics of the use of P2P systems. The difference between week days and week-end days informs us about the behavior of P2P users.

1 Introduction

This study is based on TCP captures on ADSL, which are used to establish general characteristics of P2P systems. The fact that we take into account only ADSL traffic is important, because these users are predominantly present in P2P traffic. Indeed 24 hours per day, unlimited connection is proposed by ISPs to the ADSL customers. And as we shall see, P2P file sharing systems thus account for more than 60% of the total ADSL traffic.

The originality of our measures lies in the fact that, firstly, we analyze all the TCP flows of a regional ADSL concentrating point, secondly, we observe only ADSL traffic (excluding modems 56k) which is more representative of P2P utilization, and thirdly, the data collected is representative of general ADSL users and not restricted to a specific class of users or hosts (*e.g.* a University or a private network). Furthermore, our data include several thousands of users.

We shall differentiate systematically between the P2P users with the help of a unique ADSL user identification. As noticed in [7], an analysis based on IP addresses can have a negative influence on the interpretation of the traces because of NATs (Network Address Translator) and dynamic IP addresses. Indeed, we

** This work is partly supported by project Métropolis of RNRT (French Network for Research in Telecommunications).

noticed a significant qualitative difference between graphs based on IP addresses and those based on ADSL users.

In this study, we compare four popular P2P networks: eDonkey [2], BitTorrent [1], FastTrack and WinMX [3].

The evolution over a year shows that the popularity of P2P networks is very volatile, this popularity is also very dependent on the country.

Our flow level analysis enables us to describe volumetric properties, connection duration, traffic pattern over time, host connectivity and geographical location of peers. Then we map some of these experimental distributions into classical statistical laws. Our packet level analysis allows us to clearly identify beginning and termination of connections leading to some findings on performance limitations. We mention here two interesting results:

- about 40% of connections are only connection reattempts, and it concerns about 30% of peers;
- there are two main classes of peers: those contributing to most of the traffic volumes, and the other. The first class affects strongly the main characteristics of the P2P system, while the second one softly influences these characteristics.

The remainder of the paper is organized as follows: Section 2 details the methodology for our measurements. In Section 3, we elaborate some relevant characteristics of P2P traffic, such as proportion of signaling traffic, comparison of upstream and downstream volumes, connection duration, traffic pattern over time, geographical distribution of peers and termination of connections. Section 4 deals with the number of connections a peer establishes. We summarize the main results and conclude the paper in Section 5.

2 Capture methodology and P2P overview

2.1 Measurement details

First of all, we detail our experimentation protocol. As shown in Figure 1, the BAS[1] collects the traffic issued from the DSLAM[2] before forwarding it through the POP[3] to the France Telecom IP backbone. Our probe is located between a BAS and the IP backbone. We draw attention to the fact that we capture all TCP packets without any sampling or loss. We perform an analysis of the traffic over week days and week-end days of September 2004 and we compare these results with those computed over data recorded one year ago (in June 2003).

The identification of P2P protocols is done through a port analysis: a connection is classified as a P2P protocol if one of its TCP ports is a standard port of this protocol. We shall discuss the accuracy of this method in Section 2.2.

[1] Broadband Access Server
[2] Digital Subscriber Line Access Multiplexer
[3] Point-Of-Presence

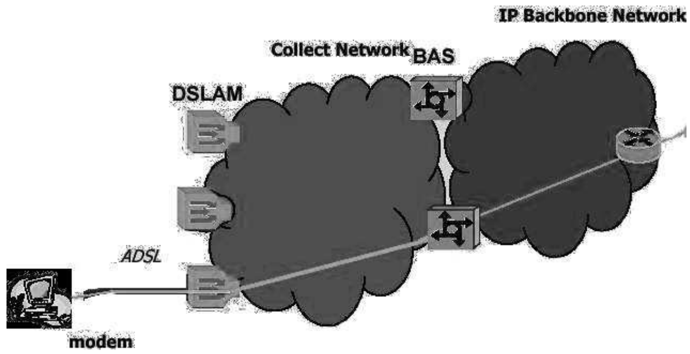


Fig. 1. ADSL architecture

We shall denote as *local* peers or users the ADSL hosts connected to the observed BAS, and as *non-local* or *distant* peers the remainder of the hosts. The *upstream traffic* will represent the packets transferred from local peers to the backbone, and the *downstream traffic* those transferred from the backbone to the local peers.

2.2 Overall P2P overview

In our data, about 60% of the traffic lies on P2P ports in September 2004. It represents a small drop compared to the proportion of traffic on P2P ports in June 2003 which was about 65%.

In Table 1, we reported the distribution of the main P2P protocols over P2P traffic. In September 2004, eDonkey is by far the most popular protocol in terms

Table 1. Distribution of protocol traffic over P2P traffic

Protocol	June 2003		September 2004	
	Volume	# Connections	Volume	# Connections
eDonkey	84%	96%	91%	93%
BitTorrent	0.8%	0.009%	6%	2.7%
Gnutella	0.8%	0.9%	1%	3.6%
WinMX	1.3%	0.06%	1%	0.08%
FastTrack	12%	1.8%	1%	0.01%
other protocols	1.1%	1.2%	0%	0.6%

of volume, BitTorrent is the second most popular and all the other protocols are almost negligible in volume as compared to eDonkey.

The popularity of each P2P file sharing system is very variable among location and time. According to [11] in October 2003, in Europe, eDonkey is overwhelmingly popular whereas in U.S. FastTrack is the most popular followed by WinMX. The evolution over time on our data shows that FastTrack lost its popularity in France (more than a year back, in June 2003, the proportion of volume of FastTrack traffic was the second most important).

In the remainder of the paper, we shall discuss only the protocols eDonkey, BitTorrent, FastTrack and WinMX, because of their popularity and the diversity of their working processes.

As reported by Karagiannis *et al.* in [9] and [10], some of P2P traffic might use non-standard port numbers so that we miss some traffic by restricting ourselves to a port analysis. In [12], Sen *et al.* reported that an identification of P2P traffic using application signatures could increase threefold the volume compared to a port based identification. But in this study, we remark that on the one hand, only Kazaa (using the FastTrack network) has a huge hidden traffic, and on the other hand, eDonkey and BitTorrent peers use mainly standard ports. Indeed, on the FastTrack network there is no limitation based on the port used by the P2P application, and some users (in fact many users) might change it. But on eDonkey network, the peers running their application on non-standard port receive a *Low ID* when they connect to an eDonkey server while other peers get a *High ID*. The High ID peers have no restrictions while Low ID peers can only download from High ID peers, so that eDonkey peers are strongly encouraged not to change the port number of their application. As we shall see, the main part of P2P traffic in France is on eDonkey network, and the port based identification of P2P protocols is relevant in this situation.

3 Characteristics of P2P traffic

3.1 Signaling traffic

P2P traffic can be split into two parts:

- the traffic generated strictly for downloading data,
- the traffic generated for maintaining the network and performing queries, that we shall denote as *signaling traffic*.

We separate these two kinds of traffic according to a threshold of the volume transmitted by each connection. In Figure 2, we plot the cumulative distribution function of the volume of connections for each P2P protocol. Note that Figure 2 *(a)* informs us on the frequency of a connection size, whereas Fig 2 *(b)* indicates the percentage of volume generated by the connections.

We choose a threshold of 20 kbytes for signalling traffic according to Figure 2 *(a)*. A direct identification of signalling connections, as in [14], leads to an average size of non-download streams of 16.7 kbytes, which is coherent with our data.

As also observed in [13] and [6], the overwhelming part (more than 90%) of P2P connections consists of signaling ones whereas they represent only a small

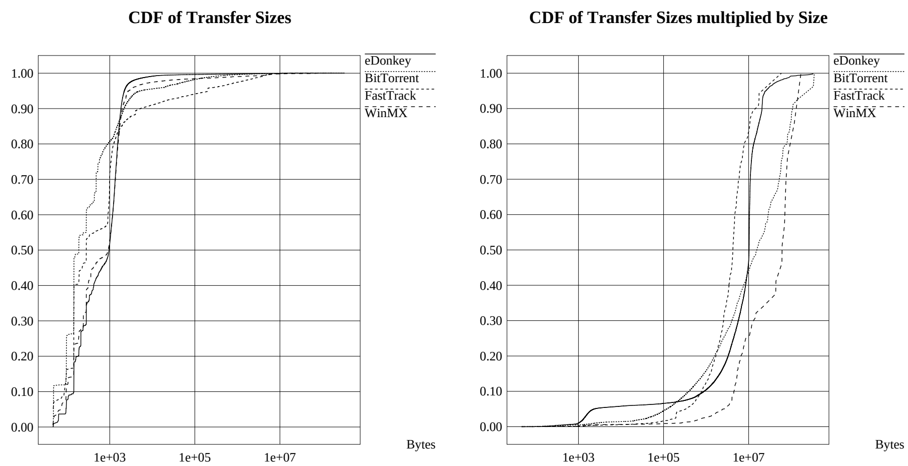


Fig. 2. Volume of P2P Connections

proportion of the volume transferred: eDonkey has the biggest proportion of volume for signaling traffic with 6% (see Figure 2 *(b)*).

3.2 Upstream *vs.* downstream volumes

User-based comparison In our data, the amount of downstream traffic is larger than the upstream traffic in terms of volume for each protocol. It means that local peers (*i.e.* several thousands of users) tend to download more than they upload on our observation point. This is a consequence of ADSL, which tends to offer much lower outbound than inbound capacity, whereas overall download and upload rates must be equal in a P2P file-sharing system .

On Figure 3, we have plotted a cloud of points representing for each eDonkey user its downstream volume versus its upstream one. On the figure we can identify the users contributing to big volumes (on the upper right corner), those generating small volumes (above the diagonal, in the middle), and a certain number of peers having no upstream volume whereas they have some downstream one (to be explained in section 3.6).

We analyze the behavior of these two kinds of users:

- peers having small volume, download files but upload very few, thus they can have downstream-to-upstream ratio up to 1000;
- peers contributing to big volumes have comparable upstream and downstream volumes with a downstream-to-upstream ratio of about 1.2.

This is the case of eDonkey peers which represents the vast majority of peers, and other P2P users have similar trends. The first class of users share few files,

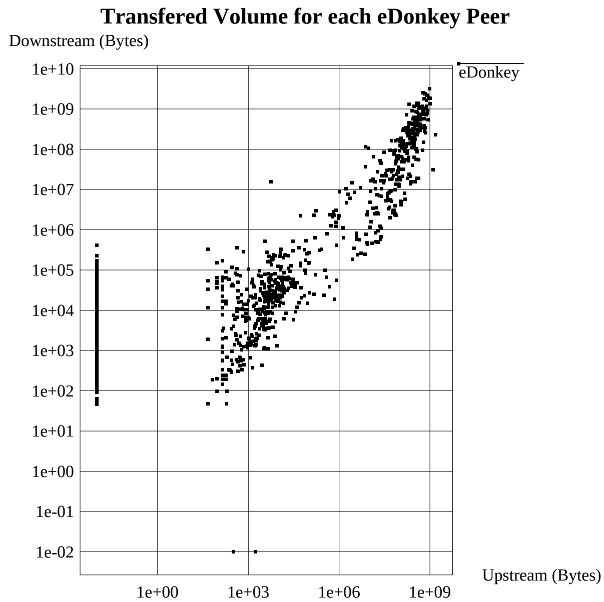


Fig. 3. Downstream volume versus upstream volume for each eDonkey peer

or disconnect themselves after download. And the second class of users have to stay connected to the P2P system for long periods to obtain high downstream volume, thus they share files (at least those being downloaded).

By recalling that less than 10% of the users contribute to the most significant part (98%) of the traffic (see also [13]), we can see that the non-cooperative behavior of small volume users doesn't disturb the balance of P2P system. The downstream-to-upstream ratio for overall P2P traffic is about 1.2 over our pool of users, whereas the mean ratio is of 38.

In our analysis, we identified that about 20% of the peers are probably *free-riders* (*i.e.* peers that do not share files) over the eDonkey network.

In [4] and [5], the number of *free-riders* on Gnutella network is evaluated at 70% and 42% in 2000 and 2001 respectively. Mechanisms like multi-part download (now used by most P2P applications) allow peers to share already downloaded file chunks. This explains the reduction of the number of peers who do not share data compared to previous studies.

Connection-based comparison Most of the connections bring in a very small proportion of the volume of each P2P network. Indeed, the connections transferring less than 100 kbytes represent less than 8% of the traffic volume, whereas they account for more than 90% of the connections (see Figure 2).

We explain this overwhelming number of small connections as follows:

– signalization generates a lot of small transfers,

– many transfers are interrupted,
– many peers attempt to connect to offline peers (see Section 3.6).

For BitTorrent, the distribution of transfer sizes is different. BitTorrent generates a higher proportion of *big* transfers, indeed there is no search process included in this protocol, only the coordination of transfers is taken care of.

We have approximated the observed distribution of transfer sizes by classical statistical laws using Kolmogorov-Smirnov (K-S) goodness-of-fit test, and we conclude that:

– eDonkey can be approximated by a lognormal distribution;
– FastTrack by a lognormal one, but the tails of the distribution (*i.e.* big transfers) fits better a Pareto one;
– BitTorrent by a Weibull one.

To conclude this section, we mention that not only is the median volume per connection very small (less than 1 kbyte) due to numerous small connections, but so is the median volume per user (10 kbytes). The huge proportion of signaling traffic induces a mean volume per connection of 10 kbytes. On the contrary, due to some users contributing to large traffic volumes, the mean volume per user per day amounts to 70 Mbytes.

3.3 Connection duration

We present the cumulative distribution function of the connection duration in Figure 4.

Connection durations are very long in view of their size. Indeed, more than 85% of the connections stay open for more than 10 seconds while more than 90% of the connections comprise less than 20 kbytes. This reveals long idle periods during the connections. These idle periods often encourage authors to consider as finished a connection after 5 seconds of inactivity as in [6] or filter out these periods as in [14].

eDonkey and FastTrack connections are of shorter duration than BitTorrent connections. BitTorrent encounters longer connection durations and also bigger throughputs. We give the mean value of connection durations in Table 2.

Table 2. Mean value of connection duration

	eDonkey	*BitTorrent*	*FastTrack*	*WinMX*
Signaling traffic	12s	22s	7s	13s
Non-signaling traffic	1436s	1670s	356s	2721s

As far as BitTorrent is concerned, a peer distributing or downloading a file has a pool of 5 upload slots, *i.e.* only 5 other peers can download the file from

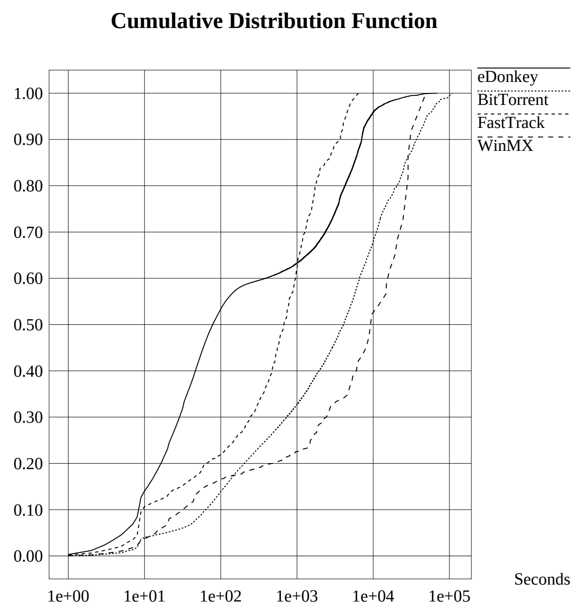


Fig. 4. Connection Duration

him. To ensure a certain fairness, these slots are reallocated periodically, but the TCP connections are not closed. As a result, BitTorrent also encounters idle times during the transfers and long connections.

The case of eDonkey is different: a file is splitted into chunks. With the most popular eDonkey client (eMule), when a peer has finished to download a chunk, it should wait for another chunk in a queue, but the TCP connection between the peers stays open.

The difference between BitTorrent on the one hand and eDonkey and FastTrack on the other hand is confirmed by the connection duration distributions. We find again that a lognormal law fits better eDonkey and FastTrack, whereas BitTorrent's connection durations fail to be well approximated by a classical law (the K-S test doesn't give a satisfactory answer).

3.4 Traffic pattern over time

In figure 5, we represent the traffic volume and the number of peers (transferring more than one packet during the considered hour) for every hour of a week day for eDonkey. The overall volume does not vary much throughout the day for eDonkey: a small drop (about 20%) of transferred volumes is observed between 12pm (midnight) and 9am. This naturally correspond to a diminution of the number of peers. This reduces the importance of *time-of-day* effect observed in [13], [7] and [8], the following ideas explain why:

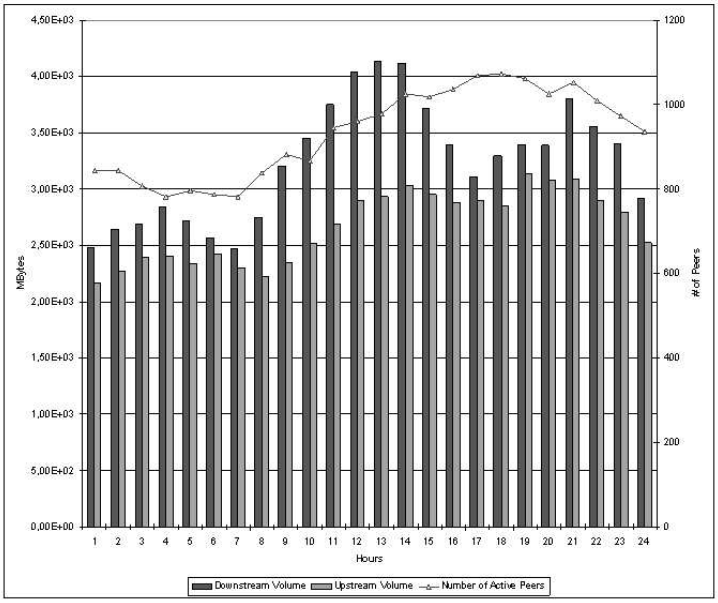


Fig. 5. eDonkey Traffic Volume and Number of Peers over Time during a Week Day

- the main part of the traffic volume is generated by a small proportion of users (10%, see Section 3.2), and these users are permanently connected to the P2P network,
- peers download very large files which accounts for long durations of connections to the P2P network.

A closer look at Figure 5 shows that the upstream traffic is more stable than the downstream one. This is also a consequence of the nature of peers involved in the eDonkey network. The peers contributing to big volumes and permanently connected are responsible for the main part of downstream and upstream volumes, which are comparable. The other peers introduce a lot of downstream traffic whereas they contribute in a smaller proportion to upstream traffic.

For BitTorrent, the behavior is completely different. The connections are actually longer and the receiver sees many *on* and *off times* during its transfer, so the traffic pattern is a bit skewed. It is interesting to note that one year ago, the time-of-day effect was clearly observable on BitTorrent's traffic, as most of the volume (85%) was usually transferred between 3p.m. and midnight in France. With the increase of BitTorrent users, the traffic pattern is more regular over the day.

In Figure 6, we expose the traffic volume and the number of active peers for every hour of a week day for BitTorrent. From midnight to 5pm, the number of peers is quite stable and so is the upstream volume. But it is surprising to see

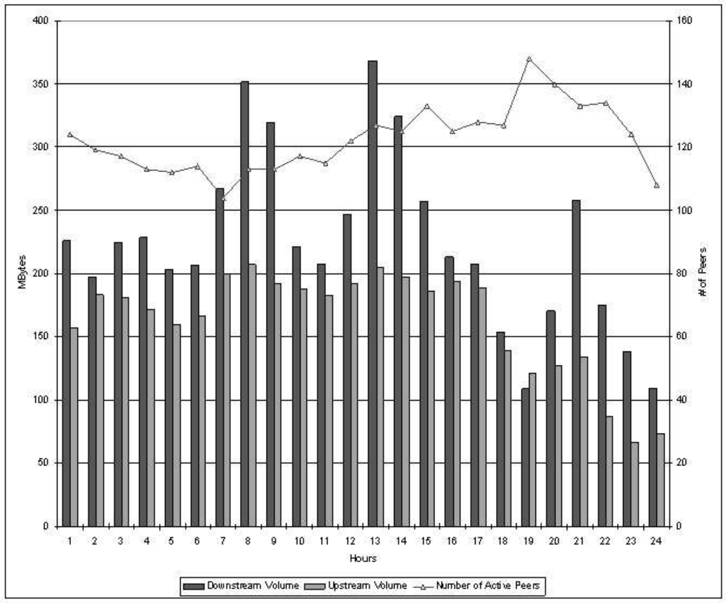


Fig. 6. BitTorrent Traffic Volume and Number of Peers over Time during a Week Day

downstream volume having high fluctuations. We explain this by the functionnig system of BitTorrent: in [15], the model corresponding to a BitTorrent-like file-sharing system states that the service capacity (*e.g.* the total throughput) grows with the peers involved in the file-transfer, resulting in a very good response to flash crowds (when many peers simultaneously ask for the same file). The two peaks of downstream volume probably result from a sudden increase of the number of distant BitTorrent peers. We can add that the file(s) corresponding to this growth of peers should not be very big, indeed the peaks indicates that the download time is less than two hours.

For FastTrack and WinMX, we observe a dynamic traffic pattern over the day: users connect to these P2P networks mainly between 12am (noon) and 12pm (night) for week days and all day long during week-ends. We explain this phenomenon as follows:

- the afternoons and evenings are high activity periods for ADSL activity (during week days),
- users download shorter files with Kazaa or WinMX, hence the P2P applications can be used for shorter periods of time.

For week-end days, the traffic pattern is completely different. The whole saturday and sunday morning are high activity periods for eDonkey transfers, but they end at sunday noon! BitTorrent exhibits the same pattern but the traffic distribution is more erratic.

3.5 Geographical distribution of peers

We locate the destinations of transfers over a week. We use a description field of IANA database to establish this geographical information.

The geographical distributions show that most of the traffic (about 30%) heads to and from France, followed by U.S. for eDonkey and BitTorrent. On the FastTrack and WinMX networks, U.S. is the primary source and destination of transfers.

We have to remark that these distributions are very sensitive to the activity of the week. Indeed in June 2003, we reported a majority of eDonkey traffic coming from Germany.

The distinction of signalling traffic allows us to determine if the geographical distribution of servers follows the distribution of peers. Our first analysis indicates that the distributions are similar, but small differences can give interesting conclusions: *e.g.* Belgium encounters more small connections than big ones: indeed the most popular eDonkey indexing server (with 800,000 peers connected to *Razorback 2*) lies in Belgium.

3.6 Termination of connections

In order to characterize the termination of connections, we expose in Table 3 the connections with a normal TCP-ending (four-way handshake, denoted as "NORMAL"), and for the other connections, we observe the TCP-Flag of the last packets.

Table 3. Distribution of the TCP-Flag of the last packet of connections

Flag of last packet	*eDonkey*	*BitTorrent*	*FastTrack*	*WinMX*
NORMAL	20%	9%	15%	20%
SYN	21%	42%	39%	27%
FIN	6%	5%	4%	5%
PUSH	20%	16%	5%	20%
RESET	7%	11%	11%	9%
OTHER	23%	16%	25%	17%

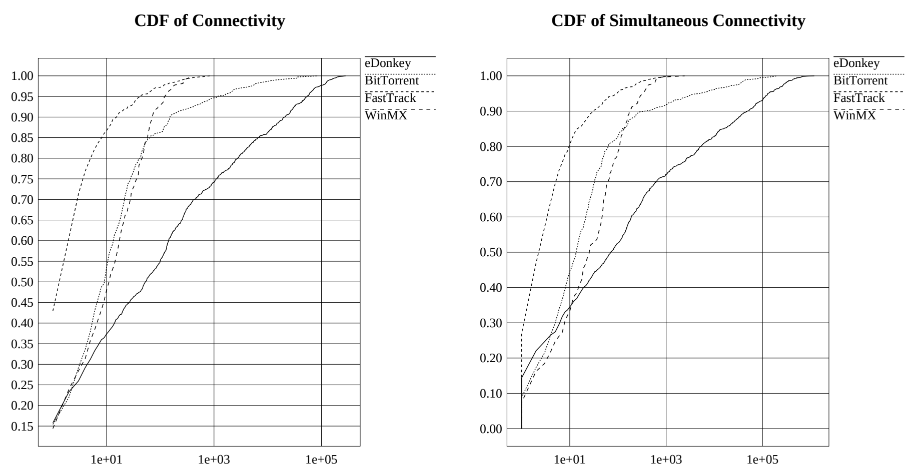
Here all four protocols have similar trends: only few connections end normally. We have identified that a P2P client disconnecting from the P2P network results in sending a RESET packet, this accounts for many connections. We also observe a high percentage of connections that end abnormally, *e.g.* by a PUSH.

But the main remark here is that 20 to 40% of connections are only connection attempts (the last packet of these connections is a SYN). The peers involved in these connections (about 30% of peers) receive connection requests whereas they are no longer connected to the P2P network. This is observable in Figure 3 by looking at users who have no upstream volume whereas they have some downstream one (the vertical line at the left side of the figure). We explain this by the delay in forwarding the information on peer availability across the P2P network. At present, we are further investigating this issue because it can have some influence on P2P networks.

4 Connectivity of peers

4.1 Local peer connectivity

Now, we study the number of distant hosts contacted by a local user during a day, *i.e.* the *connectivity* of a local peer. Figures 7 *(a)* and *(b)* show these data over a complete day and over periods of two minutes spanning the day, respectively. The results of the two minutes period graph represent the number of simultaneous connections by user.



(a) over the whole day (b) over periods of 2 minutes

Fig. 7. Number of different IP addresses (X axis) connected to a local user

Firstly, the connectivity of eDonkey peers is greater. Indeed this protocol generates a lot of connections between peers. Only 18% of the local eDonkey peers contact a single peer, almost 50% of the local peers contact more than 10 other peers. Some eDonkey peers connect to more than 100,000 other peers, which shows that we have a small number of eDonkey indexing servers among the set of local peers.

The similarity of Figures 7 *(a)* and *(b)* tend to prove that peers establish connections with other peers within periods of 2 minutes. Indeed, when a peer begins to download a file, it tries to connect to all the peers having this file.

For BitTorrent, the downloading traffic presents a higher connectivity than the signaling traffic: this is due to the information management, which is done by a single *tracker* for each single file but the download is made from several sources.

4.2 Distant peer connectivity

Now we investigate the connectivity of a non-local peer, *i.e.* the number of local hosts connected to each distant peer. For this section, we can only distinguish distant peers by their IP addresses. These results are computed over a pool of 1 Million distant peers, mainly eDonkey ones.

Only few local peers are connected to the same distant peer. We don't see any accumulation of connections on a distant IP address: traffic and requests are well distributed over distant hosts.

eDonkey peers still have the densiest connectivity with 20% of the distant peers being contacted by more than three local peers. For the three other protocols, about 80% of the distant peers are contacted only once by local peers.

The previous analysis allows us to identify highly connected peers. In this paragraph, we characterize the traffic of eDonkey peers which connect to more than 10,000 other peers during the day. We report about 130 peers with this property. These peers encounter same mean of connections sizes than other peers. But the mean downstream volume transfered amounts to 500MBytes per peer and the upstream one to 340MBytes per peer. As most of the traffic is generated by these peers, their cumulative distribution function of the tranfers sizes is very similar to those exposed in Figure 2.

5 Conclusion

In this paper, we compare the performance and characteristics of four P2P applications. Our measurement methodology allows us to deeply analyze a complete set of traffic traces stemming from all the users of a regional ADSL area. We derive many results dealing with characteristics of connections (volume, duration, termination of connections), localization of peers, traffic pattern over the day and connectivity of peers.

Our study indicates firstly that, even for P2P traffic, most of the connections are very short and represent a small volume, and secondly that very few users contribute to the most significant part of the traffic volume. The two kinds of peers involved in P2P networks (*i.e.* those conributing to big volumes and the other) strongly influence the functionning system of P2P file-sharing. We reveal that even on a whole regional concentration point, users tend to download more than they upload. We also find that local peers tend to contact many different distant peers. Focusing on the packets sent at the termination of connections, we detect that unsuccessful connection attemps represent a lot of connections, and concern many users.

The persistency of our measures has allowed us to see a change of popularity in P2P applications (FastTrack being overtaken by BitTorrent) and some changes in the location of sources over a year.

6 Acknowledgements

We would like to thank our colleagues Anne-Marie Bustos and Denis Collange for their contributions to the geographical analyses and the utilization of measurements. A lot of improvements in the final paper is due to anonymous reviewers, we thank them for their detailed comments.

References

1. BitTorrent. www.bittorrent.com.
2. eDonkey. www.edonkey2000.com.
3. WinMX. www.winmx.com.
4. E. Adar and B. Huberman. Free riding on Gnutella. *First Monday*, 5(10), October 2000.
5. A. Asvanund, K. Clay, R. Krishnan, and M. Smith. An Empirical Analysis of Network Externalities in P2P Music-Sharing Networks. In *The 23rd Annual International Conference on InformationSystems (ICIS 02)*, Barcelona, Spain, December 15–19 2002.
6. N. Ben Azzouna and F. Guillemin. Analysis of ADSL traffic on an IP backbone link. In *Proceedings of Globecom 2003*, San Fransisco, CA, USA, December 2003.
7. R. Bhagwan, S. Savage, and G. Voelker. Understanding Availability. In *Proceedings of the 2nd International Workshop on Peer-to-Peer Systems*, Berkeley, CA, USA, February 2003.
8. J. Chu, K. Labonte, and B. Levine. Availability and Locality Measurements of Peer-to-Peer File Systems. In *Proceedings of ITCom: Scalability and Traffic Control in IP Networks*, July 2002.
9. Thomas Karagiannis, Andre Broido, Nevil Brownlee, K. C. Claffy, and Michalis Faloutsos. Is P2P dying or just hiding? In *IEEE Globecom 2004 - Global Internet and Next Generation Networks*, Dallas, Texas, USA, December 2004.
10. Thomas Karagiannis, Andre Broido, Michalis Faloutsos, and K. C. Claffy. Transport Layer Identification of P2P Traffic. In *Internet Measurement Conference (IMC)*, Taormina, Sicily, Italy, October 2004.
11. Sandvine Corporation. Regional characteristics of P2P, October 2003. www.sandvine.com.
12. Subhabrata Sen, Oliver Spatscheck, and Dongmei 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.
13. Subhabrata Sen and Jia Wang. Analyzing Peer-to-Peer Traffic Across Large Networks. *IEEE/ACM Transactions on Networking*, 2004.
14. K. Tutschku. A Measurement-based Traffic Profile of the eDonkey Filesharing Service. In *Proceedings of Passive and Active Network Measurement (PAM 2004)*, Antibes Juan-les-Pins, France, April 2004.
15. Xiangying Yang and Gustavo de Veciana. Service capacity in peer-to-peer networks. In *Procedings of IEEE INFOCOM 2004*, Hong Kong, March 2004.