ANALYSIS OF PEER-TO-PEER TRAFFIC AND USER BEHAVIOUR

Amir Alsbih[1], Thomas Janson[1] and Christian Schindelhauer[1]

[1]Department of Computer Science, University of Freiburg, Germany
alsbiha|janson|schindel@informatik.uni-freiburg.de

ABSTRACT

The quality of peer-to-peer networks depends highly on the user behaviour. High churn rates can destabilize the network structure. Egoistic users might try to download only. At this moment, only a small number of independent studies of Internet traffic is available. In this paper we analyse the complete traffic of 20,000 users in August 2009 of a German digital cable TV based Internet provider. Traffic was centrally monitored and its type classification given by deep packet inspection. We concentrate on BitTorrent traffic, since this is the most popular peer-to-peer network at the moment. We show that the upload traffic and download traffic are highly correlated. We show that there is only one strong periodic user behaviour. Many indicators of the traffic patterns obey piece-wise different Pareto (power law) distributions. These results may be used for more realistic models for peer-to-peer user behaviour and for prediction of peer-to-peer traffic.

KEYWORDS

Internet traffic, peer-to-peer Networks, user behaviour, BitTorrent

1. INTRODUCTION

No other communication network has ever grown as fast as the Internet regarding the number of users and amount of data transmissions. Because of the distributed structure of the Internet the global growth is not easy to measure. However, for local Internet Service Providers (ISP) it is possible and necessary to learn about the user behaviour, since these companies need to plan and adjust the network capabilities. While Internet users and network provides share a lot of interests – security concerns about, high availability and reliability of network services – there are many divergent objectives. Users want low prices for high download rates and user anonymity, while ISPs optimize their profit and need to keep track of the identities of their users and their behaviour.

In the recent years ISPs started to install special network monitoring systems, which allow analysing the behaviour of single users. This investment is primarily profit-driven. On the one hand, they want to improve their product for special applications like gaming, which requires a low ping. On the other hand, they want to limit high traffic of file sharers to reduce costs. Due to traffic classification for different services it is possible to regulate peer-to-peer network traffic, e.g. limit the bandwidth of that service.

Cooperating with a German Internet Service Provider we have been granted limited access to anonymous user data gathered by such a network monitoring systems using deep packet inspection (DPI). The main product of this ISP is digital cable TV, which they deliver to thousands of German households. As a by-product they also offer telephone and Internet service, which is our area of interest.

Technically, each user needs a digital cable modem, which encodes and decodes the data traffic. On the plus side the throughput rates are rather large ranging from 6-100 Mbit/s download (compared to DSL traffic ranges from 2 to 16 Mbit/s and the high end product of HSDPA ending at 7.2 Mbit/s). So these users do not face the network bottleneck like other users and the measured traffic behaviour (more or less) directly reflects the users' wishes. This is an ideal opportunity to find out what Internet users want: How long are users online? How much data do users download or upload? What are the network services they use?

In the last decade peer-to-peer network file sharing has been the dominant source of Internet network traffic, where BitTorrent has been the most popular protocol. We take a look into the upload and download behaviour of BitTorrent traffic. Finally, every user has his private course of the day, which includes sleeping, eating, working, and possibly Internet usage. We are interested in periodical behaviour which we will investigate using Fourier analysis.

2. Related Work

There are not many Internet traffic studies available, since only ISPs are able to measure all the traffic in their networks. The most recent study is [1] with data from anonymous global service providers, which does not publish sources or underlying data size. The results are given as relative shares. As the main result they report the traffic share for different categories with most important data e.g. HTTP 26%, online video 26%, P2P file sharing 25%, and other file sharing 19%.

BitTorrent [5] is the most successful peer-to-peer network protocol. Compared to other peer-to-peer networks BitTorrent encourages to upload data using incentives [4]. Egoistic users, so-called leeches, who are only downloading data, are punished by a reduced download rate following a tit-for-tat rule. These rules have limitations and several BitTorrent clients deviate from the original protocol. As a selfish example of a BitTorrent client, BitTyrant [3, 10] achieves a download gain up to 70 per cent without any change of the BitTorrent protocol; just by a strategic selection of peers in the swarm. BitThief [12] even takes things further. It is a free riding client which allows downloading without any upload and can achieve higher download rates than the official client.

When it comes to measuring and analysing the impact of BitTorrent on the Internet, there are different approaches. In [11] they compare the BitTorrent traffic with other protocols in different regions in the years 2008 and 2009. For Germany, they measure that BitTorrent amounts for 37% of the total traffic, 2.5 times more than HTTP traffic. This does not coincide with our results.

In [7], the authors concentrate on the analysis of single torrents in a BitTorrent system starting from their creation until they become unavailable. The underlying data is a tracker trace from Autumn 2003. The authors present distributions for a torrent describing the number of downloaders, the number of seeds, the download speed, and the share ratio. Also, distributions comparing different torrents with request number, peer number, and download speed are given. They also present how peers downloading multiple files connect different torrents and provide a randomized model for this. The authors of [6] analyse BitTorrent traffic recorded by an ISP in Sweden in 2004 filtered on torrents for Linux distributions. The analysis mainly focuses on client sessions and measures session duration and data volume. In contrast to our analysis many features like online time are Weibull or normal distributed. It is not clear whether the differences come from the focus on legal torrents, different culture and legal situation in Sweden, or different interconnection features. In [8], the authors present a technique to trace BitTorrent users from a single machine over a long time. For identification, they use three information sources recorded in Spring 2009: Websites providing torrent-files, tracker servers, and distributed DHT trackers providing peer IP addresses connected to the torrent. The analysis concentrates on finding the initial content providers of the torrents.

These models are used for predicting user behaviour. So, the authors in [9] create a prediction model for the availability of users in distributed systems. Their algorithm predicts hosts correctly to be online when they are online all the time. This analysis is refined in [2] and tested with real traces of the eDonkey P2P network in year 2007. They show that the distributed applications can benefit from such models. So it is crucial here to describe the behaviour of the users in the network as precisely as possible.

3. Analysing the Data Set

The data we are analysing has been collected by a German Internet service provider (ISP), which prefers to remain unnamed. The Internet connection is established here via digital TV cable and provides connections where the users have contracts with download bandwidth of 6–100 Mbit/s upload bandwidth in the range of 0.4–2.5 Mbit/s. The data collection contains information about 21,766 hosts in Germany from 1st to 31st of August 2009. We have also received a second data set for September 2009, which has lead to the similar observations, but is not shown in this paper.

The Internet Service Provider used a deep packet inspection system for analysing the type of traffic. After 15 minutes the DPI system reports the number of incoming and outgoing bytes for each protocol for each user. This information is collected in log files. We have received the data without IP addresses, which have been replaced by anonymous integer IDs to protect the privacy of the customers. For each interval of 15 minutes over a month we know for each anonymous user the number of open connections, the incoming and outgoing overall traffic, the incoming and outgoing unencrypted BitTorrent traffic. Besides this information we have received the sum of overall traffic in this month for each host for each service type. Furthermore, we have for each period of 15 minutes the sum of HTTP traffic of all users.

There are some shortcomings in this data set. The identification of each user by the IPv4 address (or its unique replacement) is not completely reliable. While many other German ISPs force a network reconnection every 24 hours, this ISP does not follow this practice. This means that the IPv4 address of a network user remains the same until the modem is rebooted, then the DHCP server of the ISP assigns a new address to this host.

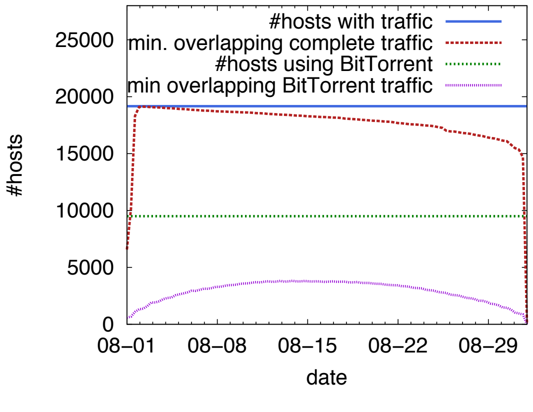

Figure 1. Minimum overlap for BitTorrent and all services

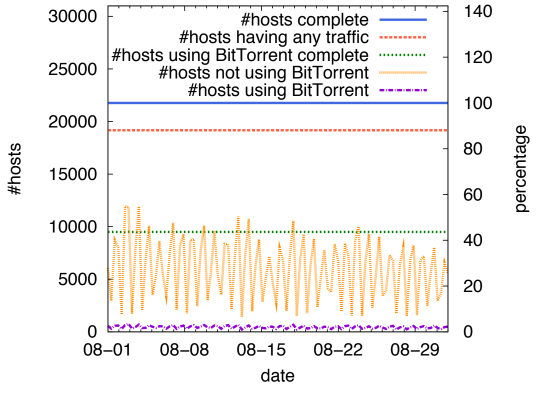

Figure 2. Time course of active hosts with traffic in any service or only BitTorrent

So, two types of errors might occur. First, a user might occur within the month under several IP addresses, which leads to an overestimation of users. Second, a different user might reuse a free IP address and different user behaviour is wrongly combined. We neglect this second case since the ISP assured us that IPv4 addresses are rarely reused. For the first case, note that, if two IP addresses are used at the same time, then this is a proof that these addresses represent different users. Now, we look at the intervals when an IP address is used and count the number of such simultaneous time intervals. This number gives us a lower bound of the number of distinct users. Figure 1 shows the sum of these intervals for the full time period of the full month.

In fact the lower bound based on all services matches the complete number of hosts up to 36 hosts. So, this error type can be neglected, too. We also plot the same function using only the BitTorrent service data. These curves show, that users do not use BitTorrent as regularly as all other types of Internet service. For the complete traffic and BitTorrent traffic displayed in Fig. 2 we see some hosts not producing any Internet traffic. Furthermore, the number of hosts varies over the month in a day and night schedule. Only a small percentage of all hosts use BitTorrent traffic at all.

The underlying DPI system can classify more than 1,000 different categories including the encrypted and unencrypted BitTorrent traffic. Fig. 3 shows the average download and upload rate for the top 10 services, as well as the sum of residual traffic. Most downloads traffic of the average host is produced by HTTP, which is around five times larger than BitTorrent, the second largest service regarding downloads. For the upload BitTorrent is the leading service before eDonkey. So, peer-to-peer file sharing services account for most of the upload traffic while HTTP accounts for downloads.

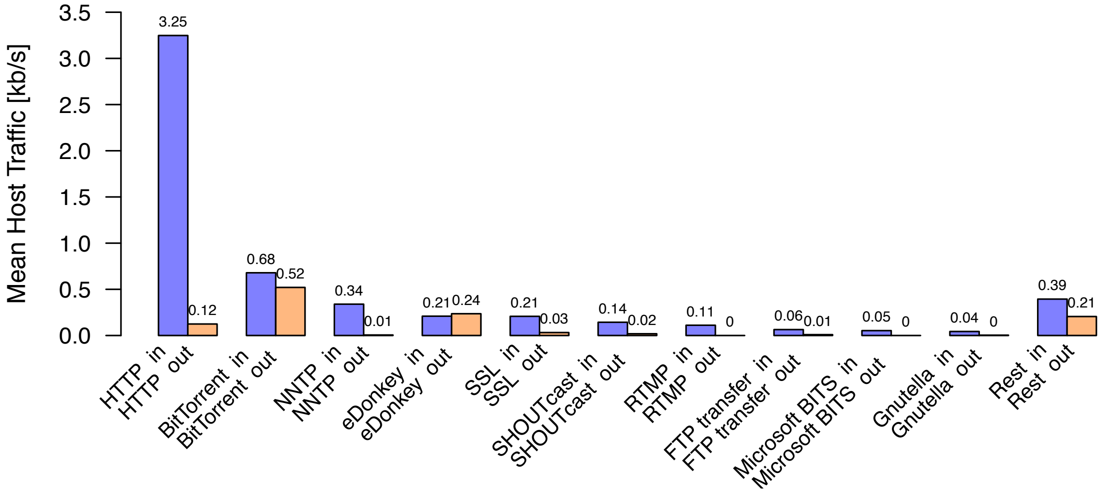

Figure 3. Mean host traffic for different services/protocols.

For HTTP the upload traffic is about 4 per cent of the download traffic. BitTorrent has nearly a balanced incoming and outgoing traffic where the upload is 76 per cent of the download traffic. BitTorrent has a download/upload of 0.68 kb/s|0.52 kb/s averaged over all users. If we average only over all 9,495 BitTorrent users of all 21,766 hosts the BitTorrent download/upload average rate is 1.47 kb/s|1.16 kb/s. We have noticed that the encrypted BitTorrent traffic is rather small compared to the unencrypted BitTorrent traffic. So, we analyse only the unencrypted BitTorrent traffic throughout this paper.

We could not spot any large peak in the Internet traffic caused by global or local events. This can be seen from the average host traffic of the 15 minutes period in Fig. 4. The peaks result from the regular day and night schedule. If one averages over a period of 24h Fig. 5 shows a slight decrease in the overall traffic. Most interestingly this decrease does not take place for the BitTorrent traffic or the complete upload traffic.

Since the observed data rates are far below the available bandwidth, the observed behaviour does not stem by the network limitations of this ISP but from the limitations of the corresponding partners and the user behaviour which gives interesting insights into unrestrained Internet and peer-to-peer network user behaviour.

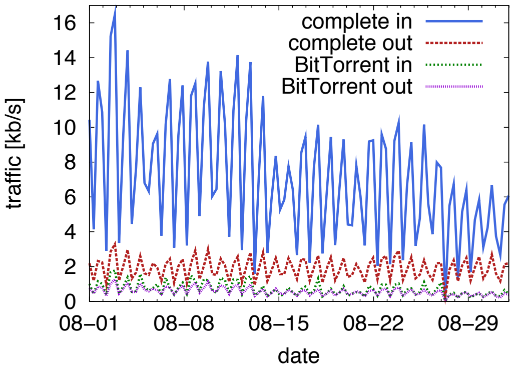

Figure 4. Average host traffic (15 minutes resolution).

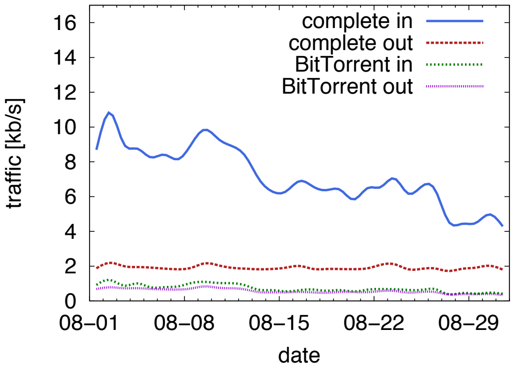

Figure 5. Average host traffic (24 hours resolution).

3. BitTorrent Fairness and Periodicity

Much of the popularity of BitTorrent has been attributed to its built-in incentives [4], which should improve fairness. Some peer-to-peer networks suffer from masses of so-called leeches. Our data shows that BitTorrent clients behave very well in our data set. First we consider the overall average upload speed and average download speed for each user. For each user this defines a point in Fig. 6. The correlation coefficient is 0.58, which is a positive correlation.

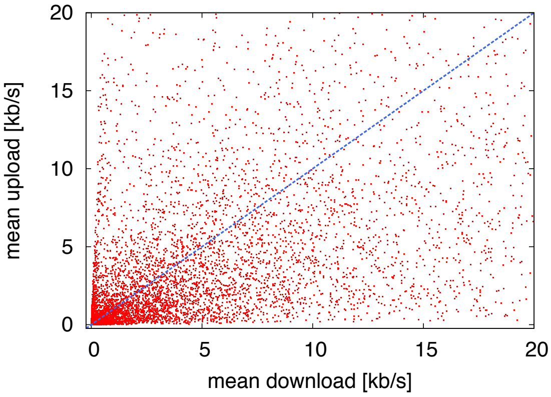

Figure 6. Scatter plot with (upload, download) points for all hosts (Speed range 0-20 kb/s).

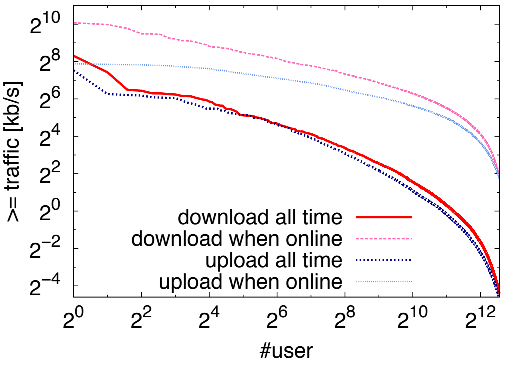

Figure 7. Log-Log plot of average users' BitTorrent upload and download speed.

If one relates BitTorrent download traffic (given by the average download bandwidth) with all other Internet download traffic of users one observes a negative dependency with a correlation coefficient of −0.38. Similarly for the BitTorrent upload and residual Internet traffic upload the correlation coefficient −0.53 can be found. This means that either the BitTorrent traffic blocks other traffic from this host or that active peer-to-peer network users do not use other Internet services to the same extent.

From the scatter plot we have already seen that there is no sharp distribution for BitTorrent traffic. In fact, if one considers the difference of download and upload traffic one observes a piecewise power law (Pareto) distribution, see Fig. 8.

$$P_{d-u}\left[\text{share-difference } x\right] \approx \begin{cases} 0.68 \cdot (x+1.1)^{-2.04} & \text{for } x \geq 0, \qquad (\sigma = 0.0008) \\ 4.33 \cdot (3.07 - x)^{-2.33} & \text{for } x < 0 \qquad (\sigma = 0.0006) \end{cases}$$

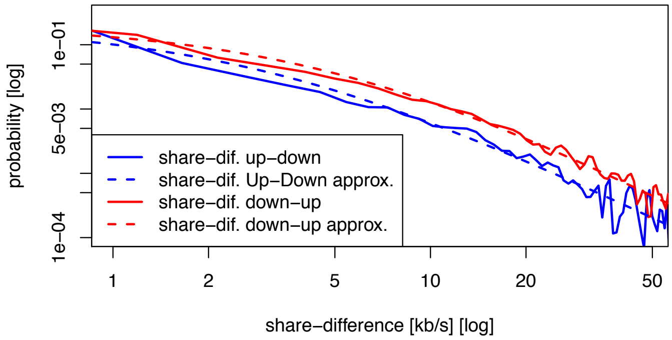

Figure 8. Probability distribution of share difference (upload-download).

To our knowledge, we are the first to observe a heavy-tail distribution in the BitTorrent traffic difference. An explanation may be the power law distribution of the overall BitTorrent upload and download, see Fig. 7 and the power law distribution of the continuous online period, see Fig. 9.

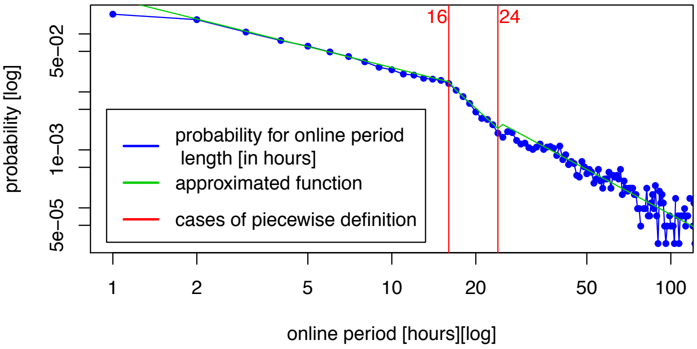

Figure 9. Distribution of online period lengths for all uninterrupted online periods of all hosts.

The online period can be very well described by a piecewise defined power law. The interval bounds are at 16h and 24h, which reflect the users' decisions, whether to let the hosts run over night.

$$P\left[\text{online period } t\right] \approx \begin{cases} 0.18 \cdot t^{-0.82} & \text{for } t \geq 16, \qquad (\sigma = 0.013) \\ 2782 \cdot t^{-4.40} & \text{for } 16 < t \leq 24, \quad (\sigma = 0.00006) \\ 11 \cdot t^{-2.58} \text{ for } t > 24 & \qquad (\sigma = 0.000015) \end{cases}$$

If one considers the ratio of upload and download (instead of the difference) one observes a super-exponential decline and decrease on the left and right borders, which shows to a high concentration around the optimum ratio of 1. We could not find a reasonable approximation of these functions, while we observe a similar function of the cumulative daily online time, i.e. the number of hours a user is producing any traffic.

We are also interested in describing the periodicity of the user behaviour. While the 24h frequency is obvious when one takes a look at Figures 2 and 3 one might assume that also longer frequencies like a week frequency (168h) or possibly smaller frequency might reflect user behaviour. We analyse the incoming and outgoing traffic of all users using a Fourier analysis, see Fig. 12. For this we compute the absolute value of the complex Fourier coefficient for each frequency *f* and divide this value by the frequency. We observe a strong peak at 24h and a smaller one at 12h. While there is a local maximum at 168h its size is rather small. To verify whether the 24h and 12h frequency actually describe some periodic behaviour we overlay all traffic in the 12h period in Fig. 13 and 24h period in Fig. 14. Obviously the 12h term does not show much periodicity while the 24h period roughly resembles a sine curve. So, the 12h period is a harmonic of the 24h period, which is the main periodic behaviour. Users switch off their computers at night and turn them on during the day peaking at 8pm. A strong weekly periodicity cannot be observed as one can see in Fig 15.

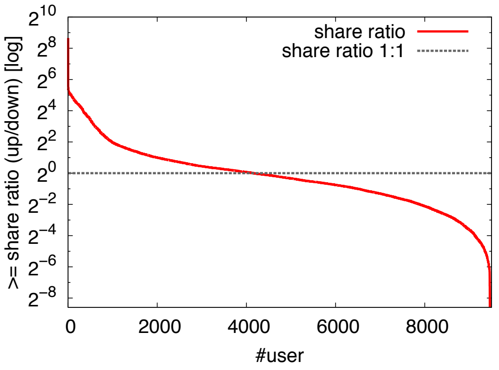

Figure 10. Cumulative share ratio distribution (upload/download) in linear-log scale.

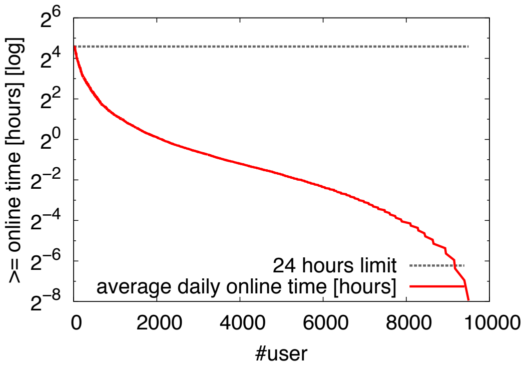

Figure 11. Cumulative daily online time distribution in linear-log scale.

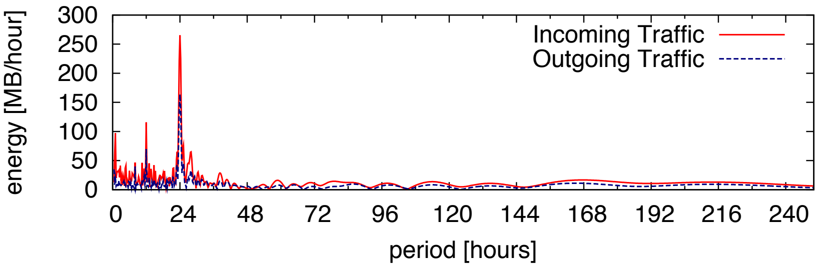

Figure 12. Fourier analysis of traffic.

5. Conclusions

We analyse the web traffic of 21,766 hosts of an Internet service provider (ISP) in Germany and concentrate on BitTorrent traffic in the month of August in 2009. A surprising large share of 50% used BitTorrent in this month. However, at any time only at most 40% of these BitTorrent users were online at the same time, see Fig. 1. Installing BitTorrent on a private host seems to be no burden and we show that BitTorrent clients are wide spread. Many users participate in this peer-to-peer network only for some short time periods. Most Internet traffic in this month is caused by HTTP and we assume mostly from video streaming and client-server based file sharing. For the upload BitTorrent is producing most of the traffic followed by eDonkey. The little HTTP upload is caused by the nature of the client-server based web traffic and the little presence of web server hosts within the ISP's network.

For the overall BitTorrent traffic we see a power law distribution for the download and upload. The difference of upload and download traffic is power law distributed. However, for the ratio of upload and download the picture is completely different. Here the probability of finding an unfair player decreases faster than exponentially. Hence, BitTorrent users with a large difference of upload and download still have nearly a constant download/upload ratio.

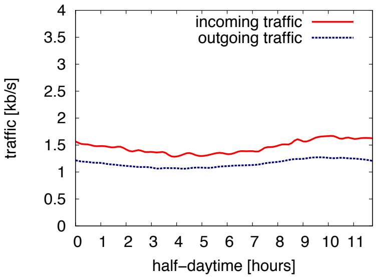

Figure 13. Half-Daily traffic (12 hour peak in Fig. 15).

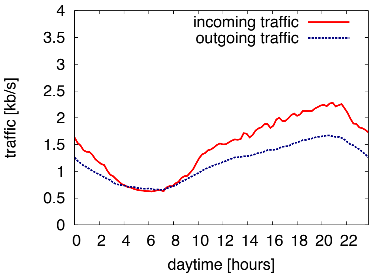

Figure 14. Daily traffic (24 hour peak in Fig. 15).

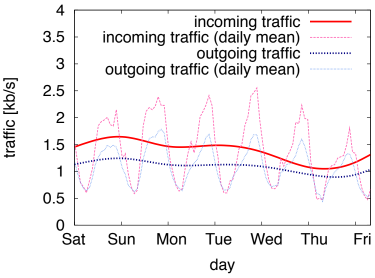

Figure 15. Weekly traffic course.

For the online durations we see a power law distribution when it comes to the overall online-time within the month. Most surprisingly we do not see a power law, a geometric distribution, a Weibull nor a Gaussian distribution when it comes to online-time over the 24h limit. It is part of future research do understand the nature of this distribution.

References

[1] Cisco visual networking index: Usage. white paper, http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/Cisco_VNI_Usage_WP.html, October 2010.

[2] Stevens Le Blond, Fabrice Le Fessant, and Erwan Le Merrer. Finding good partners in availability-aware p2p networks. In *International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS'09)*, 2009.

[3] Damiano Carra, Giovanni Neglia, and Pietro Michiardi. On the impact of greedy strategies in BitTorrent networks: The case of BitTyrant. In *Proceedings of the 2008 Eighth International Conference on Peer-to-Peer Computing*, pages 311–320, Washington, DC, USA, 2008. IEEE Computer Society.

[4] Bram Cohen. Incentives build robustness in BitTorrent, 2003.

[5] Bram Cohen. The BitTorrent protocol specification. http://www.bittorrent.org/beps/bep_0003.html, 2008.

[6] David Erman, Dragos Ilie, and Adrian Popescu. BitTorrent traffic characteristics. In *Proceedings of the International Multi-Conference on Computing in the Global Information Technology*, pages 42, Washington, DC, USA, 2006. IEEE Computer Society.

[7] Lei Guo, Songqing Chen, Zhen Xiao, Enhua Tan, Xiaoning Ding, and Xiaodong Zhang. Measurements, analysis, and modeling of BitTorrent-like systems. In *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*, IMC '05, pages 4–4, Berkeley, CA, USA, 2005. USENIX Association.

[8] Stevens Le Blond, Arnaud Legout, Fabrice Lefessant, Walid Dabbous, and Mohamed Ali Kaafar. Spying the world from your laptop: identifying and profiling content providers and big downloaders in BitTorrent. In *Proceedings of the 3rd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more*, LEET'10, pages 4–4, Berkeley, CA, USA, 2010. USENIX Association.

[9] James W. Mickens and Brian D. Noble. Exploiting availability prediction in distributed systems. In *Proceedings of the 3rd conference on Networked Systems Design & Implementation - Volume 3*, NSDI'06, pages 6–6, Berkeley, CA, USA, 2006. USENIX Association.

[10] Michael Piatek, Tomas Isdal, Thomas Anderson, Arvind Krishnamurthy, and Arun Venkataramani. Do incentives build robustness in BitTorrent. In *NSDI'07*, 2007.

[11] Hendrik Schulze and Klaus Mochalski. Internet study 2008/2009. Master's thesis, 2009.

[12] Michael Sirivianos, Jong Han, Park Rex, and Chen Xiaowei Yang. Free-riding in BitTorrent networks with the large view exploit. In *IPTPS '07*, 2007.