Characterizing the File Hosting Ecosystem: A View from the Edge *

Aniket Mahanti[1], Carey Williamson[1], Niklas Carlsson[2], Martin Arlitt[1,3], Anirban Mahanti[4]

[1]University of Calgary, Calgary, Alberta, Canada

[2]Linköping University, Linköping, Sweden

[3]HP Labs, Palo Alto, California, U.S.A.

[4]NICTA, Locked Bag 9013, Alexandria, New South Wales, Australia

Abstract

We present a comprehensive, longitudinal characterization study of the file hosting ecosystem using HTTP traces collected from a large campus network over a one-year period. We performed detailed multi-level analysis of the usage behaviour, infrastructure properties, content characteristics, and user-perceived performance of the top five services in terms of traffic volume, namely RapidShare, Megaupload, zSHARE, MediaFire, and Hotfile. We carefully devised methods to identify user clickstreams in the HTTP traces, including the identification of free and premium user instances, as well as the identification of content that is split into multiple pieces and downloaded using multiple transactions. Throughout this characterization, we compare and contrast these services with each other as well as with peer-to-peer file sharing and other media sharing services.

1 Introduction

The Web has recently witnessed the emergence of file hosting services. These services provide users with a Web interface to upload, manage, and share files in the cloud. When a file is uploaded to a file hosting service, a unique URL is generated that can be used for downloading the file. The user may then make the link public for sharing content. Well-known file hosting services include RapidShare, Megaupload, and Hotfile, which are among the top 100 most visited Web sites in the world (according to `Alexa.com`).

File hosting services differ from traditional peer-to-peer (P2P) file sharing and other content sharing services. Many social media sites are restricted to sharing video files, while entertainment sites such as `Hulu.com` place geographic restrictions on its viewing audience. In contrast, file hosting services allow users to upload any file. Some of the advantages of file hosting services over P2P technologies are greater availability of active files, improved privacy for users, improved download performance, hosting popular and niche content, and economic incentive mechanisms for frequent uploaders [1]. File hosting services offer differentiated services for its *free* and *premium* users. Free users have limited download allowance, lower download speeds, and have their download requests queued and serviced after an imposed wait time. These restrictions are removed for premium users who pay a subscription fee.

Table 1 shows a comparison of the traffic growth rate of three file hosting services (RapidShare, Megaupload, and Hotfile) and two popular social media services (YouTube and Facebook) between May 2009 and May 2010. The comparison is performed based on two commonly used Web analytics metrics, namely, number of users and visits, with data from `Compete.com` that represents information about the usage pattern of over two million users in the U.S. We observe significant growth for the file hosting services, which is higher than YouTube and Facebook. These results show an increase in the uptake of file hosting traffic and motivate the need for studying these services.

The surging popularity of file hosting services has created a flourishing *ecosystem* composed of *content publishers*, *content consumers*, and a multitude of *source sites* that contain links to content (see Figure 1). While P2P file sharing [13, 14, 28] and online social media [4, 24] have been studied in detail, the usage, content characteristics, performance, and infrastructure of file hosting services have received little attention. The sole example we are aware of is by Antoniades *et al.* [1], who studied RapidShare usage and delivery infrastructure, as measured via passive and active experiments from two European research and educational edge networks. Using flow-level and HTTP header data, they studied RapidShare clients, traffic flows, file popularity, and server infrastructure at these edge networks.

**NOTICE: This is the authors' version of a work that has been accepted for publication in IFIP Performance 2011 conference to be held in Amsterdam from 18 October 2011 until 20 October 2011. The final version will appear in a special issue of the Performance Evaluation journal (Elsevier). Changes resulting from the publishing process, such as editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication.

Table 1: Percent change in users/visits for three file hosting sites and two social media sites between May 2009 and May 2010 (according to `Compete.com`)

Type	Site	Users (%)	Visits (%)
File Hosting Services	RapidShare	61	79
	Megaupload	128	205
	Hotfile	574	1348
Social Media	YouTube	28	68
	Facebook	43	69

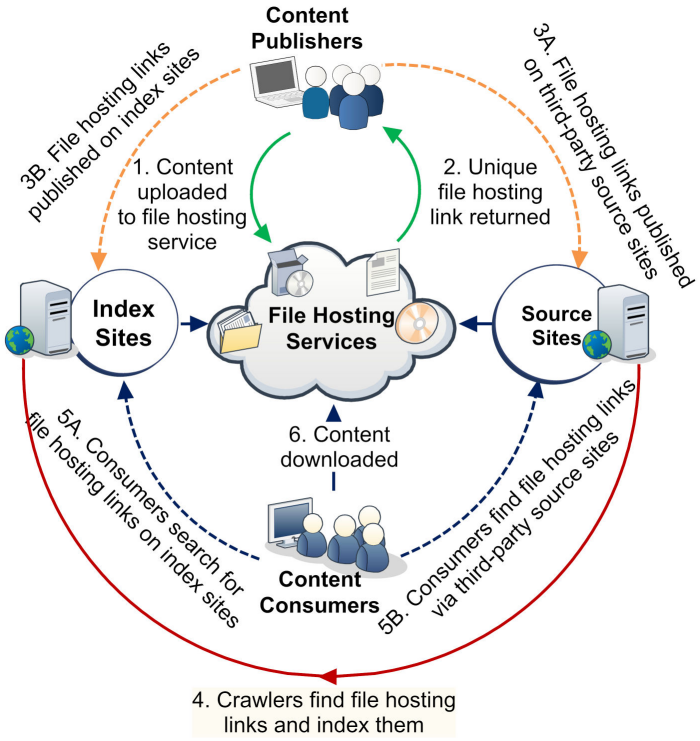


Figure 1: Dynamics of the file hosting ecosystem

In this paper, we present a comprehensive longitudinal characterization study of the file hosting ecosystem using traces collected from a large campus network over a one-year period. We performed detailed multi-level analysis of the usage behaviour, content characteristics, user-perceived performance, and infrastructure properties that illustrate the interaction within this ecosystem. We analyze in detail the top five services based on traffic volume, namely, RapidShare, Megaupload, zSHARE, MediaFire, and Hotfile. We believe our work complements the prior work [1] which focussed on RapidShare. Further, our trace durations are longer, capturing seasonal variations for a large heterogeneous demographic. Throughout the paper, we compare our findings with those of [1].

Our work makes three primary contributions. First, this is the largest and most detailed measurement study to date of the file hosting ecosystem, with focus on five popular hosting services, as observed from a large edge network. Second, we use detailed HTTP transaction logs that allowed us to study how the clients identify and select the content they download. For example, we identified signatures for user clickstreams[1] in the transaction logs to separate free and premium user instances. This has not been previously characterized, and provides a deeper understanding of the usage of these services, as well as the dynamics of new-age content sharing and distribution. Third, we compare and contrast these services with each other as well as with P2P file sharing and video sharing services. Our results have implications on caching, network management, content placement, and data centre provisioning, and are likely to be relevant for both network administrators and researchers.

Our main findings are as follows:

- Campus Usage Characteristics (Section 4): File hosting traffic exhibits positive growth trends for most ser-

[1] A clickstream is the sequence of user requests that generate HTTP transactions while browsing a Web site.

vices indicating that campus usage is tracking global popularity. Premium users dominated for two of the services, highlighting that consumers are willing to pay a subscription fee to acquire content. The usage pattern is skewed with most of the bytes transferred in the evenings. Users performed orders of magnitude more downloads than uploads. This categorizes users into two roles in the ecosystem - *content publishers* (who mainly upload the content to the file hosting sites) and *content consumers* (who primarily download content). Web browsers were the preferred method for downloading files, although we found instances of download managers being used. There was wide diversity in the files downloaded by users, with no apparent concentration of access. These observations indicate that caching at the edge may not be useful for reducing network bandwidth usage. In general, these observations agree with those in prior work [1].

- Server Properties (Section 5): The ecosystem is composed of both large and small services. Large services tend to have hundreds of host IPs spread over several /24 subnets. Most services are housed in large data centres with servers located at a few locations either in Europe or North America. File hosting services employ several upstream links to provide the best possible connection to their customers using Internet service providers (ISPs) that have peering arrangements with the customers' ISP. As reported previously in [1], RapidShare servers were found to be located in Germany. The zSHARE servers appear to be located in the New Jersey area, whereas both MediaFire and Hotfile appear to have servers located in the Houston/Dallas area. Megaupload, however, appears to distribute its servers between locations in the U.S., Canada, and the Netherlands.

- Content Characteristics (Section 6): Content in the ecosystem is dominated by video and audio, which is similar to P2P file sharing. Content sizes tend to be smaller than content sizes in P2P as file size limitations increase fragmentation. The file size chosen for splitting large content is influenced by reward incentives and the propensity to host on multiple services. File hosting links are sourced from multiple and diverse sites including forums, blogs, and search engines.

- User-perceived Performance (Section 7): Premium users tend to get an order of magnitude higher download rate than free users, with both types exceeding P2P transfer rates. Previous work [1] has reported similar results for RapidShare downloads. The wait times for RapidShare varied linearly with the file size, which followed a heavy-tailed distribution. This is in contrast to other services in the ecosystem that have a fixed wait time. Premium downloads often used concurrent TCP connections to quickly download files. File availability is higher in file hosting services compared to P2P file sharing. When files are deleted, the reasons include inactivity, copyright infringements, account expirations, and migration of users to new file hosting services.

The rest of the paper is organized as follows. Section 2 presents our trace collection and analysis methodology. Section 3 provides an overview of distinguishing characteristics of the file hosting ecosystem. The next four sections characterize file hosting workloads in detail. Section 8 concludes the paper.

2 Methodology

Datasets: Our primary dataset is a trace of HTTP transactions (henceforth referred to as HTTP trace) collected over a one-year period (Jan-Dec 2009) from a large university's 400 Mbps Internet access link. The university has 33,000 students, faculty, and staff. The campus network spans academic buildings, student dormitories, WLAN, and meeting places.

The data was collected on a Sun Fire X4450, which was configured with four quad-core CPUs, 32 GB memory, and 1.2 TB disk. FreeBSD was used as the operating system, and `Bro`[2] version 1.3.2 was used to collect the data. Traffic from the 400 Mbps full duplex network link was mirrored by the network switch and sent over a half-duplex link to a 1 Gbps NIC on our network monitor. The data contains application-layer information such as HTTP headers (e.g., HTTP method, status code, `Host` header, etc.) and transport-layer information (e.g., bytes transferred, transfer duration, etc.). The HTTP traces were produced by a `Bro` script. We used `Bro`'s HTTP parsing capabilities (e.g., `http_request()` and `http_reply()`) to summarize the HTTP transactions (request-response pairs) on the university's Internet link in real time. User identifiable information such as IP addresses and cookies were not stored. Client IP addresses were replaced by a unique integer identifier. Each day at 4 AM local time, when the network utilization tended to be low, the `Bro` process was restarted. This would reset the client IP address to identifier mappings. This method allows for greater privacy for users; however, it limits long-term analysis of *user* characteristics. We aggregated the trace data for the transactions of interest. These transactions were identified based on the `Host` header field. Specifically, we extracted the transactions where the `Host` header field

[2] `http://www.bro-ids.org/`

Table 2: Service structure for free users

File Hosting Service	Max Upload Size (MB)	Max Download Size (MB)	Wait Time before Download (seconds)	Max Number of Downloads	File Expiry (days since last download)
RapidShare	200	200	Variable	1 per 15 minutes	60
Megaupload	500	1,024	45	Variable	90
zSHARE	500	500	50	Unlimited	60
MediaFire	200	200	None	Unlimited	60
Hotfile	400	2,048	60	1 per 30 minutes	90

was `rapidshare.com`, `megaupload.com`, `hotfile.com`, `zshare.net`, or `mediafire.com`. Our experiments showed that these `Host` names suffice since these services do not employ content distribution network (CDN) nodes, which would manifest a different `Host` name.

Discerning free and premium users: Table 2 shows the service limitations imposed by the five studied file hosting services on free users (during the trace collection period). For all sites, files uploaded by premium users are not deleted as long as their accounts are active (i.e., subscription is paid). We identified free and premium downloads in the trace by user clickstreams; in prior work [1], free and premium users were identified based on file download throughput. We performed extensive experiments (after the measurement period so as not to affect the HTTP trace) where we downloaded files from the five services as free and premium users. In both cases, we downloaded the files once using a browser and once using a download manager. We validated the experimental results on locally collected HTTP traces. Based on our experiments, we determined four indicators that allowed us to identify premium downloads:

- *The premium user logs into their account:* In this scenario, we observe the premium login Web page being accessed by the user. After the necessary Web scripts are loaded, the user submits his/her credentials. This action is followed by a HTTP `POST` method that submits the user information to the file hosting service. After a successful login, the download begins. Table 3 illustrates the clickstream for a premium user logging into Megaupload and downloading a file.

- *The premium user is already logged in:* We observe an HTTP status code 302 followed by a transaction containing the actual file download. This happens when the user has previously logged into the file hosting site and a cookie has been stored on the user's machine. When the user clicks on a download link, the cookie is checked and a HTTP status code 302 is returned. This is followed by a `GET` request to download the file. This scenario is similar to the aforementioned case; however, only transactions 13 and 14 in Table 3 are observed.

- *A download manager is used for downloading file:* When a premium user uses a download manager, several partial `GET` requests are made (HTTP status code 206). These requests allow the download manager to download several pieces of the file concurrently, resulting in an improved download rate [6]. A free user downloading a file using a download manager is restricted to a single connection. Additional connections initiated by the download manager are rejected by the server. In the case of RapidShare, HTTP status code 404 (file not found) is returned, while Megaupload returns the HTTP status code 503 (limit exceeded).

- *There is no wait before the download begins:* Premium users do not need to wait for their downloads to start. We leverage this fact to identify premium downloads. We calculate the wait time by subtracting the timestamp of the transaction when the Download option was clicked on the Web page and the transaction in which the Download button appears. If the calculated wait time is less than one second, we consider the download as a premium download.

All other downloads are labeled as free downloads. Once we have identified the download as free or premium, we tag the user identifier (assigned by `Bro`) associated with the download accordingly. Table 4 shows the clickstreams associated with a free download in RapidShare. Note that the wait time is calculated by subtracting the timestamps for transactions 28 and 15 (40.9 seconds).

Identifying Content: Since file hosting services impose limitations on the sizes of the files that can be uploaded or downloaded, users split large content into smaller files using an archiving program (e.g., `WinZip`, `WinRAR`). They then upload each of the smaller files, which can be downloaded separately, and joined using an archiving program to get the final content. After analyzing several publicly available multi-part (file hosting) content, we identified three patterns. The first pattern includes the string `partn` in the file name, where `n` is the part number (e.g., `contentname.part1.rar`). The second pattern includes the string `rn`, where `n` is the part number (e.g., `contentname.r00`, `contentname.r01`). The final pattern is noticed when content is split using the

Table 3: Transactions representing sanitized clickstreams for a Megaupload premium user

No.	Time offset	URI	Method	Type	Status Code
User clicks the download link in the browser					
1	0	`/?d=file_id`	GET	text/html	200 OK
2-5	*Images and javascripts are loaded after successful GET requests*				
User clicks on the Login button					
6	6.068	`/?c=login&next=d%file_id`	GET	text/html	200 OK
7-11	*Multiple images associated with this page are loaded after successful GET requests*				
User types in username and password and clicks on Login button					
12	12.417	`/?c=login&next=d%file_id`	POST	text/html	302 Found
The downloads starts after a GET request					
13	13.073	`/?d=file_id`	GET	text/html	302 Found
14	13.275	`/files/temp_string/output1.dat`	GET	application/octet-stream	200 OK

Table 4: Transactions representing sanitized clickstreams for a RapidShare free user

No.	Time offset	URI	Method	Type	Status Code
User clicks the download link in the browser					
1	0	`/files/file_id/output1.dat.html`	GET	text/html	200 OK
2	0.718	`/img2/styles.css`	GET	text/css	200 OK
3	0.765	`/img2/favicon.ico`	GET	text/html	200 OK
4-14	*Multiple images are loaded on the Web page after successful GET requests*				
User clicks on the Free User button					
15	14.025	`/files/file_id/output1.dat`	POST	text/html	200 OK
16-27	*Multiple images associated with this page are loaded after successful GET requests*				
The Download button appears after the wait time ends					
28	54.912	`/img2/download_file.jpg`	GET	image/jpeg	200 OK
User clicks on the Download button using POST method and the server responds with the file					
29	57.143	`/files/file_id/temp_string/output1.dat`	POST	application/octet-stream	200 OK

`HJSplit` program. This program appends a number to the content name (if `Content.AVI` has been split into 4 parts using `HJSplit`, then the files are named `Content.AVI.001` through `Content.AVI.004`).

Statistical Models: We selectively present statistical models for characteristics that are specific to the file hosting ecosystem. We tested the statistical models for accuracy using the Kolmogorov-Smirnov (K-S) goodness-of-fit test. We only show models that passed the K-S test at the 5% significance level.

Identifying P2P Traffic: We also collected information about BitTorrent downloads. We monitored all the peer-to-tracker communication between local peers and external trackers. Trackers maintain state information about the peers downloading each file. With peers periodically informing the trackers about their download progress, this supplementary dataset allowed us to measure the traffic characteristics of P2P file sharing in the network, including estimating the peers' download rate. This trace spanned the same period as that of the HTTP trace. Where possible, we use this data to compare characteristics of P2P with that of file hosting services.

3 Trace Overview

We analyzed over 500 GB of compressed HTTP header logs. The HTTP trace data contained about 5.5 billion transactions. We identified over 13 million transactions attributable to over 100 file hosting services, with the top five services accounting for about half of these transactions.

One issue we had to consider was that of transactions with inaccurate byte counts. These transactions accounted for 7% of all file hosting transactions and about 31% of all transactions related to file hosting downloads. The main reason for these transactions having incorrect byte volumes was due to the monitor dropping one or more packets and `Bro` being unable to parse the HTTP transaction.[3] The gaps occur more often for file hosting download transactions, which achieve higher download rates than other transactions, causing the monitor to miss some packets. For such transactions we use the `Content-Length` header for traffic volume related analysis. We believe these gapped transactions have a minimal effect on our analysis of file hosting workloads, since we rely primarily on information from the HTTP headers observed for each transaction (which are less likely to be dropped), and we focus on general behaviour (e.g., cumulative distributions) rather than specific values (e.g., means). However, for analyses like download rates we do need to observe the complete transaction. In these cases, we have fewer observations due to gapped transactions.

As mentioned earlier, our study concentrates on the top five file hosting services (generating over 60% of the file

[3] `Bro` 1.3.2 is single threaded; thus, even though we have additional CPU cores available, it cannot make use of them. The `Bro` developers are working on a multi-threaded version.

Table 5: Trace overview

Characteristic	RapidShare	Megaupload	zSHARE	MediaFire	Hotfile	Top-5 Services
Total HTTP Transactions	3,149,630	1,617,835	718,528	413,592	121,385	6,083,970
Number of Days of User Activity	342	345	338	342	243	349
Number of Files Downloaded	45,950	21,674	10,829	11,389	3,303	93,145
Premium User File Downloads (%)	51.0	63.0	–	–	13.8	40.3
Free User File Downloads (%)	49.0	37.0	100.0	100.0	87.2	59.7
Unique Content Downloaded	20,174	13,710	6,360	7,954	2,256	50,454
Total Size of Downloaded Files (GB)	3,839	3,896	1,009	551	366	8,662
Successful Downloads (%)	87.5	95.2	71.8	94.6	98.0	88.7
Number of Files Uploaded	253	41	–	72	30	393
Total Size of Uploaded Files (GB)	3.2	2.1	–	2.9	0.4	8.6
Avg. File Size (MB)	85.5	184.1	95.4	49.5	113.4	106.2
Avg. Content Size (MB)	184.6	270.9	88.9	67.4	146.1	175.8
Avg. Parts per Content	2.5	1.7	1.0	1.5	1.6	1.9
Avg. Premium Download Rate (KB/sec)	3,894	3,738	–	–	5,136	4,126
Avg. Free Download Rate (KB/sec)	130	527	95	223	154	183

hosting traffic volume) in the campus network: RapidShare, Megaupload, zSHARE, MediaFire, and Hotfile. Table 5 presents some high level characteristics of the five services. Over 90,000 files were downloaded using the top five services. In comparison, around 150,000 downloads (61 TB of P2P traffic volume) were done using BitTorrent in the campus network. The top five services were used almost every day of the year. About 89% of the (file hosting) files were successfully downloaded. The unsuccessful downloads were due to free users abandoning the download while in progress. Almost 60% of these file downloads were done by free users. The ratio of file uploads to downloads was negligible, indicating that content consumption substantially exceeded content production in our network. We observed differences between the sizes of files downloaded and uploaded. The median upload file size (considering the top five services) was lower than 10 MB. The HTTP trace did not contain names of the uploaded files; hence we were unable to perform content analysis on these files. We surmise that the smaller file sizes could mean that these files are being used for personal purposes such as email attachments or storage. Premium downloads achieved an order of magnitude higher transfer rates than free downloads.

RapidShare was different from other services in delivering files in response to the HTTP `POST` method. Approximately 50% of the files were downloaded in this manner. This method is frequently used when free users click on the download button after waiting for the link to appear. It is unusual since the `POST` method is generally associated with uploads and not download requests. Premium downloads, which are initiated without any waiting, are served using the HTTP `GET` request.

4 Campus Usage Characteristics

We investigate the file hosting usage behaviour in the campus network. Understanding the usage behaviour of services is important from a network management standpoint. It is also useful for file hosting service designers to improve their service to better suit the needs of users.

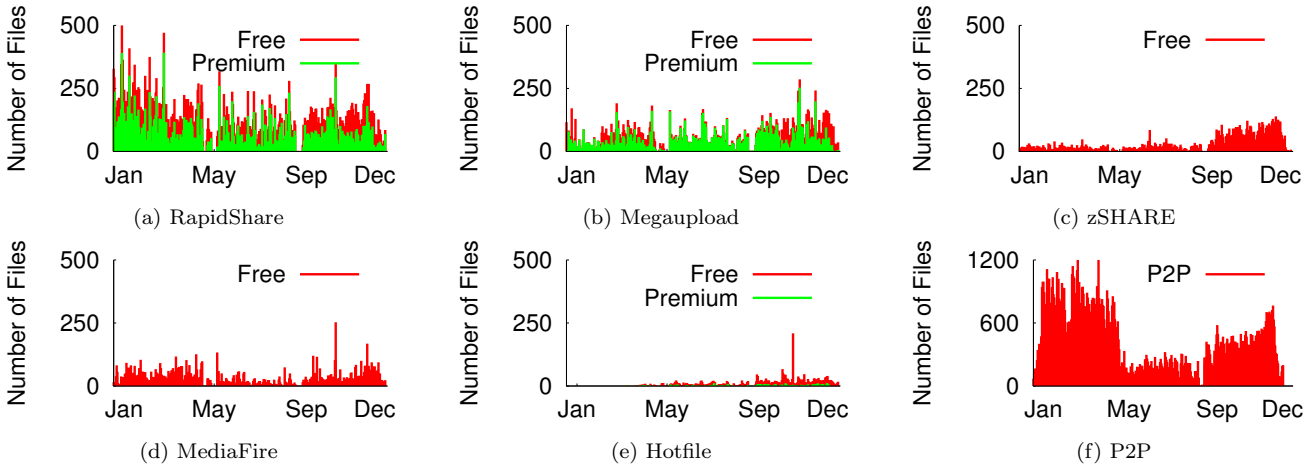
4.1 Growth

Figure 2 shows the number of free and premium daily file downloads over a one-year period for the five file hosting services and the BitTorrent P2P file sharing service. (Note the different y-axis scale on the P2P graph.) We find that the majority of files for Megaupload were downloaded by premium users, while the distribution of free and premium downloads was split almost equally among RapidShare users. Hotfile, being a relatively new service, had very few downloads using its premium service. The remaining two services saw no activity for their premium services. RapidShare and Megaupload are older services with established user communities, including numerous premium users. We measure the growth that happens between two four-month terms: winter (Jan-Apr) and fall (Sep-Dec).

RapidShare and MediaFire had 28% and 14% fewer file downloads in the fall semester compared to the winter semester. Megaupload had about 40% growth in premium and free file downloads. Hotfile and zSHARE had the highest growth rate at over 300%. P2P file sharing had 48% fewer file downloads in the fall term compared to the winter term, and potentially indicates user migration from P2P to competing offerings such as file hosting services.

4.2 Time Patterns

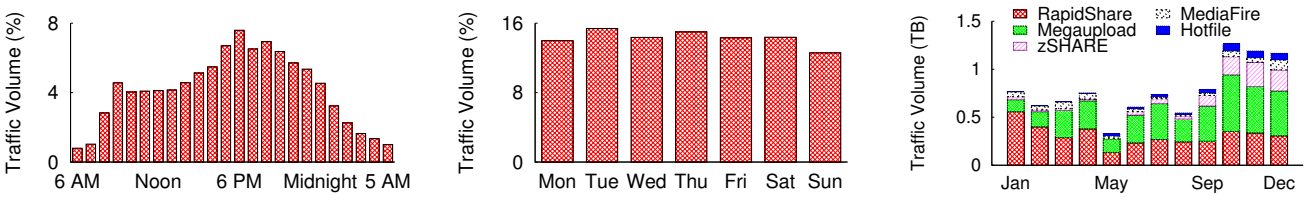
We wanted to observe whether file hosting user activity differed by time of the day or day of week. Figure 3 shows the downloading patterns of the users. Figure 3(a) shows that file hosting service usage is skewed towards the evening. Most of the file downloads (about 60%) happen during evening hours between 5 p.m. and 5 a.m.



(a) RapidShare (b) Megaupload (c) zSHARE

(d) MediaFire (e) Hotfile (f) P2P

Figure 2: Number of downloads per day



(a) Time of day (b) Day of week (c) Month of year

Figure 3: Downloading patterns

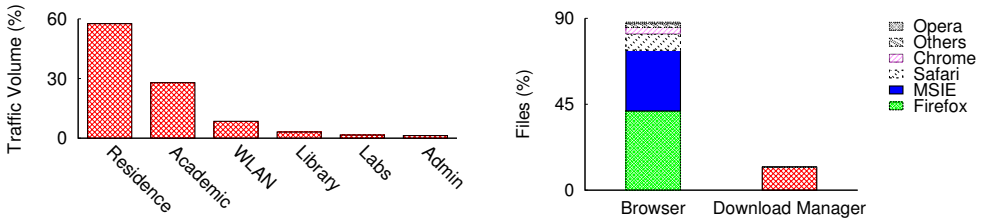
Similar usage patterns have been reported for residential networks [23]. The exception is MediaFire where more files (about 52%) were downloaded during the day. This, as will be shown later, is because of the type of files hosted on MediaFire. MediaFire has the smallest average file size among all services. MediaFire hosts many small files such as MP3 files and short duration video files. From Figure 3(b), we observe a uniform distribution of file hosting traffic during the week with a slight dip occurring on Sunday. These results show a deviation from the usual time patterns of Web and P2P usage. For example, Antoniades *et al.* found Web usage to be diurnal in research and educational networks, though P2P did not exhibit such behaviour [2]. Gill *et al.* found YouTube usage to be diurnal as well and found usage to be lower on weekends [11].

Figure 3(c) shows a clear increase in file hosting traffic volume towards the end of the calendar year. About 46% of the file hosting traffic was transferred during the winter term, fueled by increases in traffic from zSHARE, Hotfile, and Megaupload. We surmise that this is in part because of the increased availability of content on the ecosystem. Furthermore, several link indexing sites and specialized search engines have made these services more accessible.

4.3 User Subnets and Download Clients

In this section, we analyze the distribution of file hosting traffic across campus subnets and how files are downloaded from these services. Figure 4 shows the user location and the client applications used for downloading files from file hosting services. We show the distribution of users based on the campus subnet such as the university residence, academic areas, wireless network (WLAN), university library, labs, and university administration. While some of the subnets are available for use by any type of users, some are restricted to a certain user population. For example, the WLAN may be used by faculty, staff, or students, but the residence subnet is used by students living in the dormitory. The academic subnet is for faculty, staff, and graduate students with assigned offices. In Figure 4(a), we observe that a majority of the traffic came from the university residences. There were some differences in the user locations of the five services. For Megaupload and zSHARE, over 60% of the traffic was due to students in the university residence. MediaFire traffic came from the Academic (23%) and Admin (13%) subnets as well as the WLAN (16%). RapidShare traffic was also prominent in the non-residence subnets, indicating a diverse user base. This user profile could explain the daily patterns observed earlier. For example, students in university residences are present on campus on both weekdays and weekends causing the consistent traffic volume in Figure 3(b).

Figure 4(b) shows the distribution of client applications used for downloading files from file hosting services.



(a) User location distribution (b) Download client distribution

Figure 4: User subnets and download clients

Web browsers are the main source for retrieving the files. It also means that users do not initiate bulk downloads through automated scripts (e.g., `curl`), rather than rely on the browser to execute the download. We further analyzed the traces to identify the prevalence of download managers. Since the use of the `User-Agent` HTTP header field is not standardized, we had to create an extensive list of download accelerator string identifiers. We noticed the use of download accelerators provided by the services (such as RapidShare Download Manager) as well as other commercial applications (such as Download Accelerator Plus). We estimate the prevalence of download managers using the number of TCP connections initiated to download a file. If a file is downloaded with more than six connections (most browsers restrict it to this value), then we assume it was downloaded using a download manager. Using this heuristic we found that about 12% of the file downloads happened using download managers. Note that browsers do not allow segmented downloading of a file; thus, this estimate can be considered as a lower bound. Often these download managers integrate with the user's browser. Once the user clicks on the link in the browser, the download manager takes over the download process. Some download managers also allow users to mask the `User-Agent` name, which obfuscates the actual client used.

4.4 Local File Popularity

We analyzed the popularity of file hosting downloads in our campus network. This property is useful for cache design and redundant traffic elimination. We used the filename in the transaction URI to count how many times a file is downloaded. Note that content publishers may upload the same content with different names. Unless the content has the same filename we consider it to be unique. Except for zSHARE, over 95% of files were downloaded only once. For zSHARE, about 83% of the files were downloaded once. No file was downloaded more than 10 times. zSHARE is different from other services because it allows its users to stream audio/video files instead of downloading them. In our traces, over 80% of zSHARE files (90% of zSHARE traffic volume) were streamed. This viewing pattern means that if a file is viewed again later it would be transferred again from the servers. As previously discussed in the context of RapidShare [1], these results suggest that there is no concentration in the accesses, indicating caching content near the users may not be helpful. As discussed later in the paper, the infrastructure used by these services is mostly centralized and not distributed, which further supports the observation regarding lack of locality in file requests at the network edge. Note that our conclusion is based upon analysis of filename per file hosting service; we do not know if there are multiple copies of the same content with different or similar names. There may be concentration of file references close to the server; however, we do not have the data to quantify this property.

Our results, together with that of prior work on RapidShare [1], indicate that campus users have diverse interests and that there are few files in common among the users. We compare these observations to YouTube campus usage results. YouTube video requests at network edges have been found to follow a Zipf-like distribution [11,37]. However, Gill *et al.* [11] also noted little concentration in requests, with the top 10% of the videos (as measured by the number of requests to these videos in the trace) accounting for less than 40% of the total YouTube video requests. Gill *et al.* hypothesized that the low concentration in the video requests was partly due to the diversity of content available from YouTube. In addition, we note that the delivery models of file hosting services and video sharing services such as YouTube potentially are different. Entertainment sites such as YouTube are geared towards users who would like to consume the content while it is being downloaded, which is achieved by Flash streaming. Thus, if a user decides to view content again, it needs to be requested again from the site. File hosting services are primarily targeted towards users who would like to download the content completely before consumption.

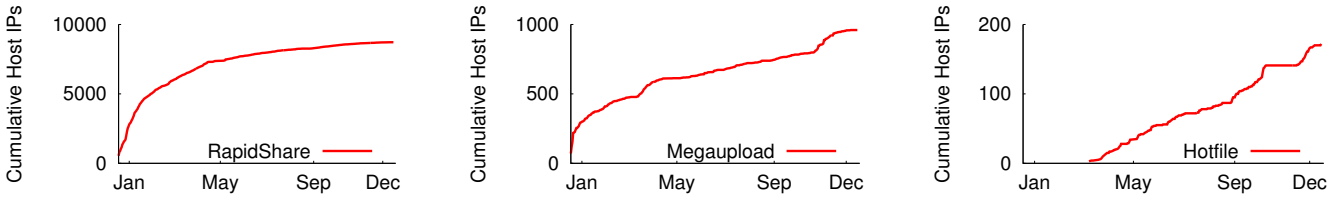
5 Server Properties

5.1 Infrastructure

We wanted to understand the infrastructure deployed by operational file hosting services, such as how many servers are used for these services, where they are located, and how the traffic is distributed across the servers. Table 6 shows

Table 6: Summary of file hosting service carriers

File Hosting Service	Hosting Company or Carrier	Location	/24 Subnets	Host IPs	Bytes Transferred (%)
RapidShare	Cogent Communications	Germany	4	738	9.8
	Deutsche Telekom		4	540	2.1
	Global Crossing		7	1,300	16.6
	Level 3 Communications		16	2,766	31.9
	TATA Communications		7	1,306	15.4
	TeliaSonera AB		10	2,050	24.2
Megaupload	Carpathia Hosting	U.S./Canada	15	651	96.2
	LeaseWeb	Netherlands	12	307	3.8
zSHARE	Choopa Hosting	U.S.	2	97	100
MediaFire	Cogent Communications	U.S.	3	523	78.4
	LinkRight LLC		2	334	21.6
Hotfile	Lemuria Communications	U.S.	3	114	41.4
	WZ Communications		3	58	58.6



(a) RapidShare (b) Megaupload (c) Hotfile

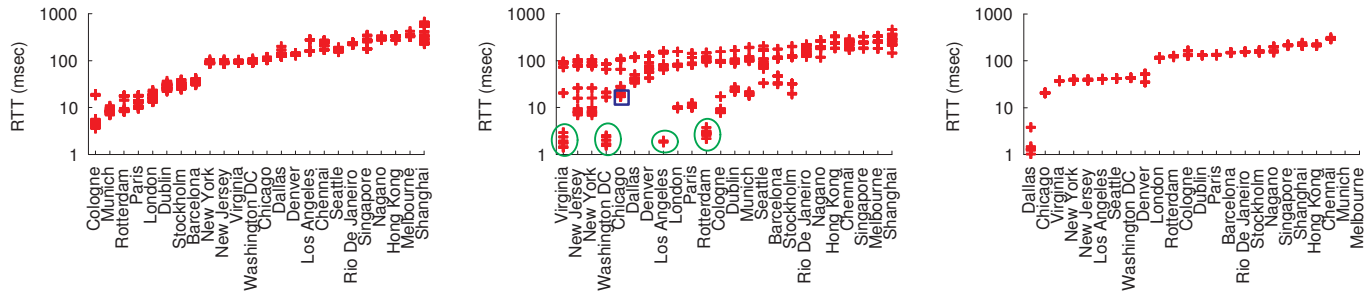
Figure 5: Cumulative host IPs seen over a year

the breakdown of the file hosting service hosts based on /24 subnets and their carriers (listed in alphabetical order) as observed in our trace. The carriers or hosting companies of the services (and their location) were determined by querying the Maxmind GeoIP database and cross-referenced using Internet Registries. We find that RapidShare and Megaupload are the largest file hosting services in terms of infrastructure. RapidShare servers are interconnected with six large Tier-1 ISPs. Carpathia and LeaseWeb provide a hosting solution for Megaupload in North America and Europe, respectively. Choopa hosts zSHARE servers. At least two /24 MediaFire subnets are owned by LinkRight, which appears to be the owner of MediaFire. Hotfile is an interesting case, since we found host IPs that were allocated to Limelight networks (a CDN company), WZ Communications (a Hosting company), and Lemuria Communications. It appears that Lemuria is the name used by Hotfile to acquire new address space. To understand how the address space of these services evolve, we performed ping requests for one host IP per /24 subnet. While there were some minor changes in the IP-name resolution for the four services, there were significant changes to the Hotfile address space. Hotfile now seems to have moved its host IPs to the Lemuria addresses. This highlights the transition of a small file hosting service expanding as the service becomes popular and demand grows.

Figure 5 shows the cumulative number of host IPs seen over a one year period. For brevity, we only show results for RapidShare, Megaupload, and Hotfile. Note the different y-axes on the figures. After an initial warm up period, we discover a growth pattern in the number of host IPs added to a file hosting service. For example, the hump in the RapidShare curve around May corresponds to RapidShare announcing a capacity increase. The curve for Hotfile is interesting as it shows how new host IPs are added as the service grows in popularity. In total, over 10,000 host IPs were identified from the trace. RapidShare had the largest deployment accounting for 80% of all the host IPs identified. In contrast to RapidShare, we found fewer than 1,000 servers for both Megaupload and MediaFire. These results provide clues on the setup of file hosting service infrastructure, but do not necessarily indicate physical servers. Typically, file hosting services are established in large data centres that interconnect with several Tier-1 ISPs, have redundant network connections, and provide customized hardware and bandwidth solutions.

5.2 Server Location

We study the geographic distribution of the file hosting service servers. We started with the Maxmind GeoIP database for this purpose. Although it provided us with a starting estimate of the location, it was not useful on many occasions. The problem is complicated when large organizations distribute IPs across various geographic regions, but the database still maps the IPs to the place where the organization is headquartered. For example, RapidShare Level 3 IPs were mapped to the U.K., while Global Crossing IPs were mapped to the U.S. Poese *et al.* have found these commercial databases to be accurate at the country-level, but the bias in the results towards a few countries makes them unsuitable for general-purpose geolocation [27].



(a) RapidShare (b) Megaupload (c) Hotfile

Figure 6: RTT measurements for host IPs using the shortest ping method

To alleviate this problem, we instrumented a delay-based geolocation technique called *shortest ping* [25]. This method involves sending probes to an IP address from several known landmarks in the world and noting the round-trip times (RTTs). The location of the IP is mapped closest to the landmark with the lowest RTT. This method has been shown to be comparable with other complicated constraint and topology-based geolocation techniques [17][4]. We performed ping requests to one IP address per /24 subnet for each file hosting service from 74 landmarks in the Americas, Europe, Asia, and Australia[5]. To validate our results and gain additional insights on the location we also performed traceroutes from multiple landmarks to those IPs. We perused the last two hops for these requests to identify strings in the router names that could indicate the location of the IPs [25]. For example, we would observe the full name of a city (e.g., Frankfurt) or different abbreviations of the same city (e.g., FRA or FFM). A similar method was used in prior work to determine the location of RapidShare servers [1].

Figure 6 shows the results of our experiments. For brevity, we only show results for RapidShare, Megaupload, and Hotfile using 24 representative landmarks. Landmarks are sorted according to the RTT values. We find the RTTs for RapidShare hosts to be lowest for the Cologne landmark. The traceroute analysis showed instances of the city of Frankfurt in the last two hops for most of the subnets. It indicates that RapidShare utilizes a centralized architecture with the data centre located near the Frankfurt/Cologne region in Germany [1]. Megaupload is an interesting case as it shows how a hosting company manages its resources. We observed two subnets that were located near Los Angeles, seven near Virginia/Washington, and ten near Amsterdam, Netherlands. These are represented by a green circle in the figure. We also found two subnets with the lowest RTT from Chicago (indicated by a blue square). Upon analyzing the traceroute for these IPs, they were mapped near Toronto, Canada. Our server location results for Megaupload correspond to the various locations where Carpathia and LeaseWeb operate their data centres. zSHARE servers were mapped to New Jersey (Choopa's data centre is located in Newark). Our ping and traceroute results indicate that MediaFire and Hotfile servers were located near the Dallas/Houston area.

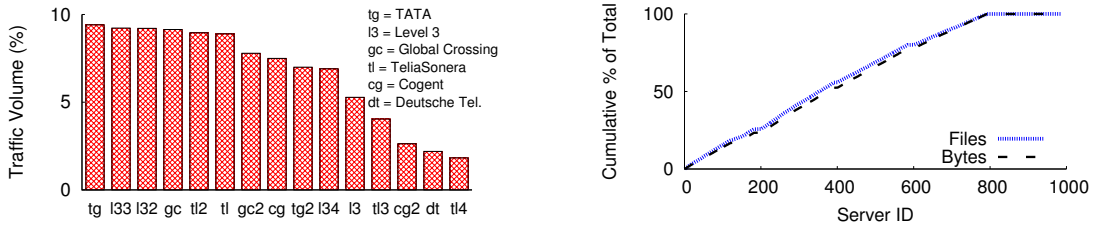
These findings highlight differences in the architecture of large file hosting services such as RapidShare and Megaupload. Our analysis suggests that file hosting service servers are located in a limited number of locations that are not geographically distributed. The file hosting architectures seem to differ from other large content providers such as YouTube that have their hosts distributed across several data centres in the world [34]. Large Web services such as Facebook and YouTube have CDNs located close to users, since these services have a lot of static content (e.g., images, text) that can be served efficiently from the CDNs. These real-time services require Web pages to be loaded quickly, and they benefit from CDNs. By contrast, file hosting services deliver large content that may not be popular among several users. This reduces the benefits of placing content near the users, although we found Megaupload serving static content from geographically distributed hosts.

5.3 Load Balancing

We next study the server load balancing characteristics of RapidShare. Table 6 shows the distribution of the bytes downloaded across the six Tier-1 ISPs used by the file hosting services. RapidShare is an interesting example since its server naming scheme allows us to understand how files were downloaded in the campus using an upstream network connection. The large number of RapidShare host IPs do not correspond to physical servers. As we show later, each file on RapidShare appears to be stored on a server (denoted by a *server id*) with 15 upstream network

[4]The shortest ping method suffices for our purpose since we intend to geolocate file hosting service servers that are likely to be housed in large data centres. This is different from geolocating end user IPs for providing localized content and security purposes, which may require greater accuracy.

[5]These landmarks were offered by two Web site monitoring services, namely, `site24x7.com` and `watchmouse.com`.



(a) Traffic volume distribution across RapidShare mirrors (b) Distribution of files/bytes across RapidShare servers

Figure 7: Load balancing of RapidShare servers

connections. We refer to these 15 upstream ports as *download mirrors*.

Figure 7(a) shows the distribution of the campus traffic volume across the 15 mirrors. RapidShare allows a user to choose any of these 15 mirrors or to rely on the built-in system to choose the best mirror based on the peering arrangement of the RapidShare mirror with that of the end user's ISP [26]. A large fraction of the download traffic is delivered from Level 3 and TeliaSonera. For each of these ISPs, RapidShare offers four mirrors each. Two connections each are offered for the remaining ISPs, except for Deutsche Telekom (1 mirror), which may be used for serving customers in Germany [30].

We next analyzed how RapidShare files were distributed across its servers. We used RapidShare's API to request information about the files observed in our traces. Each request about a file seen in the HTTP trace returned the server id, the best download mirror, and the file activity status, among other things. Figure 7(b) shows the distribution of files and bytes across the RapidShare servers. The largest server id observed is 986. RapidShare states that it has $1,000$ servers, which could mean that the server ids represent the number of physical servers [29].

The results also show that RapidShare servers are arranged in groups of 200 [1]. The older groups of servers have more files and bytes stored, while newer servers tend to have fewer. The older server groups have a slightly higher percentage of bytes stored than the newer ones. For example, server groups 0-200 and 201-400 accounted for approximately 30% of the total byte count each. Server groups 401-600 and 601-800 accounted for 25% and 15% of the bytes, respectively. The newly added server group 801-1000 accounted for a negligible byte count. These results indicate an efficient load balancing scheme adopted by RapidShare, with the objective to fill up old servers and then provision new servers to store new content.

Colocation data centres offer various bandwidth plans including metered and unmetered. Metered plans require the service to pay for the bandwidth consumed, while the unmetered plan involves capping the maximum data transfer rate to a specific speed, though the amount of data transferred is generally unlimited. File hosting services are more likely to choose the unmetered option since it is suitable for hosting large files.

6 Content Characteristics

We focus our analysis on what types of content are hosted on file hosting services, how they are fragmented, and the sizes of the file fragments and the content. Knowledge of content type can aid in several areas of system design such as file systems, data backup, scheduling policies, and disk layouts. We also analyze how users find content to download, and try to identify commonalities among the sources.

6.1 Content Type

Figure 8 shows the distribution of the type of content downloaded by file hosting service users. The content type was determined by inspecting the file extension of the content name. We observe differences in distribution of content type for the five services. For RapidShare and Hotfile, most of the downloaded content (in terms of byte and file count) was compressed archive files. Megaupload had a larger proportion (over 60%) of video content. MediaFire had a large number of MP3 audio files (34%); however, this content accounted for only 5% of its byte count. About 41% of the MediaFire content were compressed archives amounting to 48% of the byte count. Over 90% of zSHARE files were audio/video content accounting for 95% of the bytes.

The ubiquity of archived files is not surprising. It provides an easy method to split large content and upload to the hosting service. For example, it is more convenient to split large video content using a program like WinRAR, rather than use a video splitting program that would create multiple chunks of the video file that can be played separately. An archiving program, however, does not offer this feature. It requires all parts in order to reassemble the final content. Additionally, archiving allows the user to provide a password for the content. Users can then share the password with their intended audience. Moreover, content publishers can include a link in the archive to help advertise the content publisher's Web site.

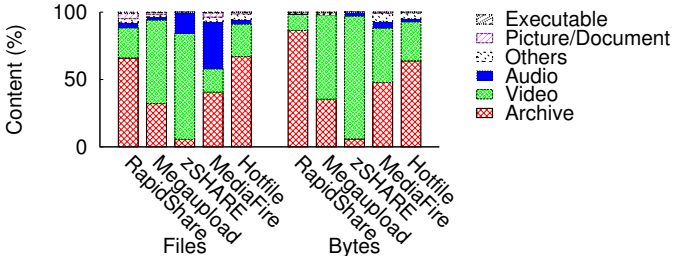


Figure 8: Content type distribution

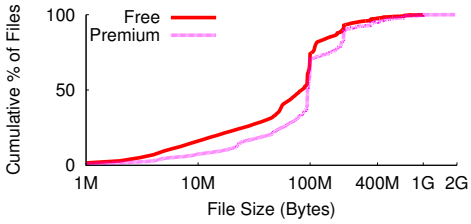
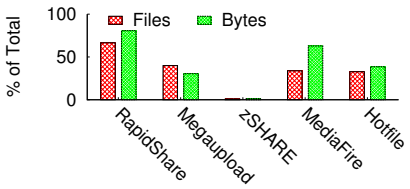
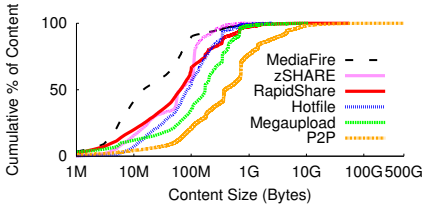


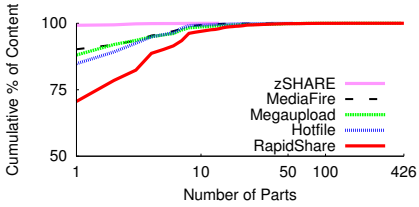
Figure 9: File size distribution



(a) Part file distribution



(b) Content size distribution



(c) Total parts per content

Figure 10: Content properties

We further analyzed the archive contents to understand their file types. We implemented the Centroid algorithm [15] that was originally used for automatically classifying Web pages from HTML tags. The algorithm was modified to classify objects based on their file name. We trained the algorithm on an extensive list of file names with known file types. The algorithm classifies a content into audio, video, document, executable, or archives (if unclassified). The algorithm was tested on a sample of 250 randomly selected files that we manually classified; we found that 80% of these files were correctly classified. After applying the algorithm on archive content, we found that the majority of these files were either audio, video, or executables.

P2P file sharing also had a similar makeup of content types [35, 36], which seems to confirm the hypothesis by Labovitz *et al.* that P2P users were migrating to the file hosting ecosystem [21]. These results point to the evolution of Web traffic; much of the traffic in 2000 was due to images and HTML text, while video and binary downloads dominate most of the bytes downloaded today [8, 10, 31].

6.2 File Download Size

Figure 9 shows the file size distribution of files downloaded from the five file hosting services by free and premium users. The file size distributions for free and premium users are similar, indicating that file uploaders use file sizes that can be downloaded by many users. Many of the file sizes are distributed between 90 and 100 MB. This corresponds to the old file size limit for RapidShare. It appears that uploaders use a file size that can allow them to upload files to any file hosting service. The second region is around 200 MB, which is the new size limitation for RapidShare. It shows that uploaders are slowly adopting the new file size limit. The tail for the premium files is slightly longer than that for free users, which shows the presence of some large files exclusively for (Megaupload and RapidShare) premium users.

File size limitations are necessary for managing the operating costs and maintaining quality of the file hosting services. These size limitations also allow the file hosting services to continue operating their free service and keep their bandwidth usage under control. It is also used by file hosting services as an incentive to recruit new subscribers. We find that the file sizes are larger than those hosted on video sharing sites; for example, Cheng *et al.* [5] show that most YouTube files are smaller than 25 MB. File hosting file sizes are also significantly different from those of Web traffic where objects 10 KB or smaller account for most of the transfers [10], although the average file size of Web objects has increased by 30% between 2000 and 2007 [31].

6.3 Content Fragmentation and Size

Figure 10(a) shows the distribution of content fragments per file hosting service. RapidShare had the largest set of fragmented content, accounting for over 80% of the byte count, followed by MediaFire. Megaupload and Hotfile had about a third of the files hosted as fragmented content. zSHARE had negligible fragmentation because users preferred its streaming option over downloading. Such a high level of fragmentation has implications for users. The file fragments are useless as stand-alone files. Thus, the user has to download all the fragments to reassemble the

final content. This may be a time-consuming process for free users.

Figure 10(b) shows the size distribution of file hosting and P2P content. We find that, on average, file hosting services host much smaller content than P2P. While average size of file hosting content was less than 200 MB, the average P2P content was an order of magnitude larger at over 1.3 GB. P2P file sharing, being a decentralized system, allows users to create file bundles of any size for downloading. Among the file hosting services, we notice differences in the content sizes. For example, services that offer higher upload limits, such as Hotfile and Megaupload, tend to have larger content sizes.

Figure 10(c) shows the number of fragments into which large content is split. Note the y-axis scale. Except for RapidShare, most content from other file hosting services were composed of a single file. About 30% of RapidShare content had more than one part. We also observe a small percentage of the files were composed of many parts. For example, RapidShare hosted a high-definition video content that had 426 parts, while Hotfile hosted a software program with over 50 parts.

There are many reasons for creating multiple parts for given content. The most obvious is the upload/download file size limitation placed by file hosting services. The reward schemes likely contribute to files being split as well. These reward systems allow users to acquire points when their files are downloaded. Large content such as high-definition videos and games often exceed the file size limitations. Users may split a large content into smaller sizes (significantly lower than the upload size limit) to have more parts per content. This increases the number of files, and they can earn more points. Because of the disparity between the upload file size limits among these services, the users may choose the smallest file hosting service size limit, thus increasing the number of parts. Another reason may be convenience. Users may choose an easy-to-use fragmentation size. While premium users have no restrictions on the upload size, uploading large files restricts their base of downloaders, since only premium users may be able to download such files.

6.4 Content Sources

Since files hosted on typical file hosting services are not searchable, the links to the files are found using third-party sources. We use the HTTP `Referer` header for this analysis. The `Referer` field provides the URL of the Web site from which the user navigated to the file hosting Web page. Figure 11 show the distribution of sources of the file downloads. We observe that blogs, forums, and community portals are the dominant sources for file hosting link retrieval.

Often, content publishers post links to their content on several forums and blogs (using automated scripting methods). We also observed the use of search, especially file hosting dedicated search engines that crawl file hosting content posted on blogs, forums, portals, and other sites. Mirroring services are also used, but to a lesser extent, where users upload content to the mirroring site, which automatically uploads it to several well-known file hosting services. The user shares the link to the mirror site containing all the links to other services. Anonymizers are used to shorten the URLs and hide them from crawlers. The use of social media or email is limited in exchanging file hosting links.

The `Referer` analysis also highlighted the type of content users download from each of the file hosting services. For example, by manually visiting these sites, we found that Megaupload and MediaFire are popular for hosting Asian community content as well as music albums, while zSHARE is used for hosting TV episodes. Many of the zSHARE links are used as embedded videos in blogs and portals. We also analyzed site referral information collected from `Compete.com` of real-time entertainment sites (YouTube and Hulu) and P2P file indexing sites (Mininova). For real-time entertainment sites, the referrals mostly came from direct traffic, search, and social networking, while for P2P file indexing sites, most referrals came from other torrent sites as well as search and direct traffic.

We wanted to understand how many file download links are provided per source site. Figure 12 shows a complementary cumulative distribution of the percentage of file downloads per `Referer`. The distribution has a heavy tail and fits a power law with exponential cutoff well. While there were few sites that provided many download URLs, the power law shape indicates a large pool of source sites from which download URLs are gathered. The results indicate that users rely on a wide variety of sources to obtain links for download.

7 User-perceived Performance

We focus on performance aspects of the file hosting ecosystem. We study the advantages offered by premium service over free service or P2P in terms of download rates. Such a comparison is important for understanding the performance of file hosting services, as P2P file sharing has been subject to traffic shaping by ISPs worldwide [7]. We also compare file availability in file hosting services and P2P.

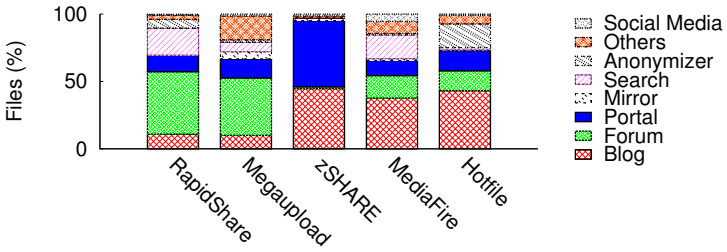


Figure 11: Distribution of download sources

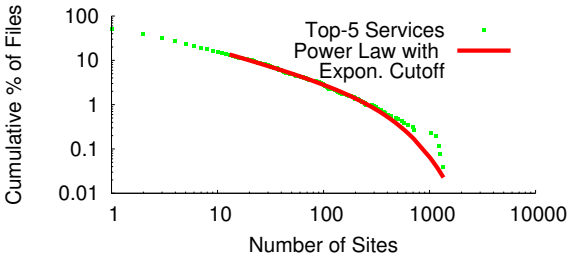
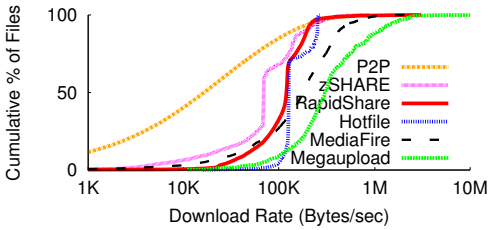
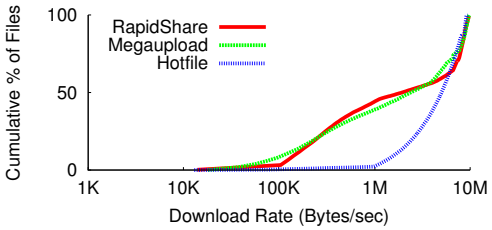


Figure 12: Complementary cumulative distribution of download referrals



(a) Free downloads



(b) Premium downloads

Figure 13: Distribution of download rates

7.1 Download Rates

File download rates are calculated using the bytes transferred field and the transfer duration. When a file is downloaded using multiple connections, the download rate is calculated as the sum of download rates of all the parallel connections. Figure 13 shows the distribution of the achieved throughput rates for file downloads. We observe that premium downloads achieved an order of magnitude higher download rates than free downloads, which in turn were significantly higher than P2P download rates. The average P2P download speed we observed was 55 KB/sec (the median was 16 KB/sec), compared to an average download speed of $19-22$ KB/sec reported by Iliofotou *et al.* [16]. We observe that RapidShare and Hotfile throttle free downloads to a maximum rate of 190 KB/sec. MediaFire and Megaupload provided higher download rates for free users compared to other services.

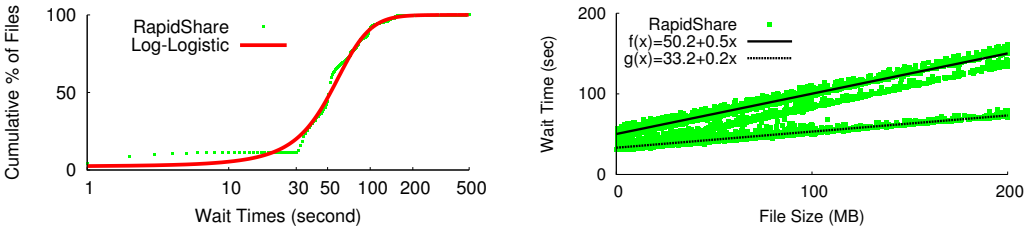
Megaupload has a distinct service structure, offering free users download rates comparable to premium users for limited downloads. MediaFire seems to offer premium services at no cost and comparable to other premium file hosting services. The higher download rates offered by file hosting services mean that users can quickly share, disseminate, and consume content with other (premium) users. Our results from passive measurements confirm the findings of Antoniades *et al.* [1] on the greater incidence of high download rates for RapidShare premium users versus BitTorrent download rates, obtained via active measurements for a small sample of files.

7.2 Wait Times

Requests for file downloads by free users are queued and serviced after a pre-determined wait time. Most file hosting services have fixed wait times ranging between 45 and 60 seconds, while RapidShare is the only service that has a variable wait time. Often file hosting sites display advertisements while the user waits for the download to begin. We compare the wait time characteristics of RapidShare to other services. These results helps us understand the file hosting economic model and latencies in service times of user requests.

Figure 14(a) shows the distribution of the wait times for free downloads. The median wait time was 50 seconds, which is higher than all other file hosting services. The wait time distribution is well modeled by the Log-Logistic distribution, highlighting the heavy tail present in the data. We observe that most users waited between 16 seconds and 160 seconds for their download link to appear. For the occasional downloader this wait may not be significant; however, repeatedly waiting for downloads may be frustrating enough that free users purchase a subscription.

Figure 14(b) illustrates the relationship between wait time and file size. We find that wait times are proportional to the file sizes; likely a design decision by RapidShare. The figure shows two regions where the wait time increases linearly by a factor of 0.5 and 0.2 per MB of the file size. This model of serving smaller files to free users earlier than larger files allows RapidShare to manage their traffic and bandwidth costs. The lower region represents "happy hour" periods when RapidShare relaxes the wait times for its free users. These periods correspond to the late night



(a) Wait time distribution

(b) File size versus wait time

Figure 14: Wait times for RapidShare free users

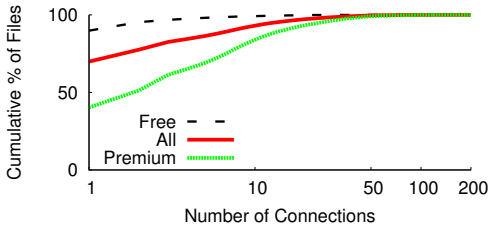


Figure 15: Distribution of concurrent connections

hours (on server-side) when there is reduced activity. Wait time is an essential tool for RapidShare to serve free users efficiently from its queue. It also serves as a good marketing tool to convert free users into premium users.

7.3 Simultaneous Downloads

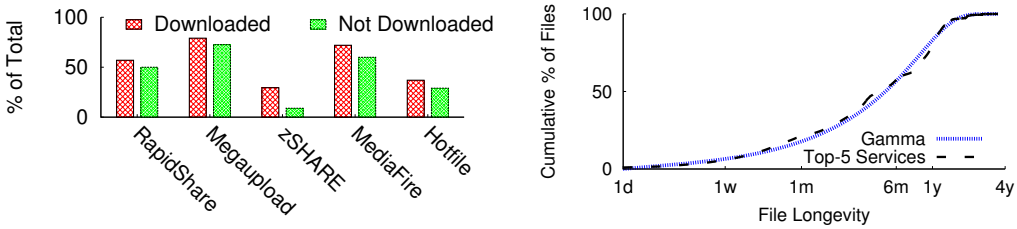
We next study the download concurrency characteristics of file hosting clients. We investigate how many TCP connections are used to download a file. Figure 15 shows the distribution of number of connections used to download a file. Over 50% of premium downloads used more than a single connection, while most free downloads happened using a single connection. The distribution has a pronounced tail indicating that several premium users utilized multiple connections to download a file. Such downloads happen using download managers, which initiate several partial `GET` requests for different pieces of the same file. These pieces are downloaded simultaneously and the aggregate download rate increases substantially. Generally, using more connections results in faster downloads. The right tail shows extreme values for abnormal downloads. These downloads represent scenarios where downloads were initiated at some point in time, interrupted, and resumed again. Also, when downloads are initiated with a browser and then transferred to the download manager, the process results in additional connections being initiated, interrupted, and resumed. Typically, download managers allow users to choose up to 20 connections per download. Our results indicate that about 93% of the premium downloads used fewer than 20 connections. Since most file hosting services (except MediaFire) do not allow free users to download a file using multiple connections, an overwhelming majority of free downloads happened using a single connection.

The use of parallel connections and parallel downloads is quite common in P2P file sharing. BitTorrent clients employ multiple connections from several peers to retrieve pieces of the same content to increase the download rates [3]. BitTorrent clients also allow downloading multiple files (each using multiple parallel connections) in the same session. While Web clients often initiate multiple connections to obtain different objects from a Web site, the use of multiple connections to retrieve a single object is limited. The results have performance implications on network management. Download managers have the ability to quickly adjust the download speed to that of the user's network line speed. With more users using these programs to speed up their downloads, the additional overhead may cause performance degradation for users [12, 19].

7.4 File Availability

We characterize file availability in file hosting services using two metrics, namely, retention and longevity. We quantify file retention by the fraction of file hosting links that are active. File retention is directly proportional to the fraction of active links (higher file retention means higher percentage of active links). For all the files that were observed in our traces, we queried the file hosting site to check their status. We queried for two types of links from the HTTP trace: files that were downloaded by users, and those that were not.

Figure 16(a) shows the percentage of active links for each service. We observe a larger proportion of non-downloaded files had inactive links indicating a reason as to why these files were not downloaded in the first place.



(a) Percentage of active links

(b) Distribution of file longevity

Figure 16: File availability characteristics

There are several reasons for links to be inactive. One reason is that the file is unpopular and the URL was not accessed for a long time. Another reason is that the link was reported as a case of copyright infringement, and removed by the service. In some cases, the user removes the files for unknown reasons. Files may also be deleted if the premium account expires. Megaupload and MediaFire tend to have greater file retention than the other services.

We did not have the upload time of the files observed in the HTTP trace. To understand file longevity in the file hosting ecosystem, we crawled about 500,000 files from a large file hosting index site[6]. The index site records the date the link was added to its database. We use this date as an approximation of the time the file was uploaded to the service. We removed all dead links and calculated the time elapsed since the upload time of each file. Figure 16(b) shows this distribution of file availability. The median file availability is around four months, while 80% of the files were available for about a year. RapidShare and Megaupload had files that were older than a year. We also observed a file that was traced back to 2006, when Rapidshare was established. File availability in file hosting sites is well-modeled by a Gamma distribution.

Our results suggest that the median file availability in file hosting services is higher than that of active files on BitTorrent [35]. File hosting services rely on centralized servers for stored content, and due to this limitation their file availability is limited. The difference is that in case of P2P, a file may persist for a long time, but it may not have enough peers for the content to be downloaded fully. Kaune *et al.* found that a lack of seeders often resulted in file unavailability (86% of the time) [18]. Private torrent sites institute a minimum upload-to-download ratio to increase file availability, however, these schemes have been shown to be subject to collusion [22]. File hosting sites do not suffer from this problem. Additionally, private forums and boards that post links to content on file hosting services do not impose any explicit rules for participation. There is significant competition among file hosting sites with each service attracting users with different incentives. This causes users to choose a different service after some period of time. As the user moves to a new service, the files stored on the old service are deleted since the user's subscription is not renewed. The files that have been available for more than a year belong to users that are faithful to a service. File availability in file hosting services is lower than that for video sharing sites (such as YouTube) since these sites do not impose an expiration time for the hosted videos [11].

8 Concluding Remarks

We presented a longitudinal characterization study of the file hosting ecosystem using traces collected from a large edge network. Using HTTP transaction logs, we developed signatures to distinguish free and premium service instances, which we used to understand the usage and dynamics of these service classes. Our study highlighted the salient features of the file hosting ecosystem and identified similarities and differences among the underlying services in usage, server architecture, content, and performance.

The file hosting ecosystem appears to be flourishing. There are hundreds of file hosting services at the disposal of users, which gives them enough choice to select a service of their liking. Our results indicate that there are a significant number of premium users, suggesting that the economic model based on advertisement and subscription revenue is sustainable.

One of the drivers of file hosting service growth is the incentive schemes instituted by the services to attract content publishers. As more content is uploaded, it causes more consumers to download the content, which in turn increases traffic. These incentive schemes have become controversial lately. RapidShare suspended its incentive scheme in July 2010 because of accusations that these schemes induce the uploading of copyrighted content [30]. Other services such as Hotfile and Megaupload are facing lawsuits for allegedly hosting copyrighted content. In response, Hotfile started terminating accounts of publishers suspected of repeated copyright infringement. Recent measurements from a Polish and a Hungarian ISP suggest a migration away from popular services such as Rapid-

[6] `http://www.filestube.com/`

Share [9]. This migration, potentially due to changes in the reward policies, may suggest that users are migrating to other file hosting services and/or to new variants of P2P services, including UDP-based BitTorrent applications [9]. This behaviour is reminiscent of the stickiness property of Web sites, where users will stick to a site that serves them adequately until another competitive site offers better alternatives [32]. Some file hosting index sites have been targeted by the U.S. government with the seizure of the domain name of the sites [20]. While some of these index sites ceased to operate, others moved to a different domain name registered in another country [20]. Search service providers such as Google have started censoring names of some file hosting services from its search completion feature [33]. Such actions raise issues of network neutrality and have impact on content consumers.

9 Acknowledgements

The authors are grateful to our shepherd Tim Brecht and the anonymous reviewers for their constructive suggestions, which helped improve the clarity of the paper. This work was supported by funding from the Natural Sciences and Engineering Research Council (NSERC) of Canada, Alberta Innovates Technology Futures (AITF) in the Province of Alberta, CENIIT at Linköping University, and National ICT Australia (NICTA).

References

[1] D. Antoniades, E. Markatos, and C. Dovrolis. One-click Hosting Services: A File-sharing Hideout. In *Proc. ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 223–234, Chicago, U.S.A., 2009.

[2] D. Antoniades, M. Polychronakis, N. Nikiforakis, E. Markatos, and Y. Mitsos. Monitoring three National Research Networks for Eight Weeks: Observations and Implications. In *Proc. IEEE Network Operations and Management Symposium Workshops (NOMS)*, pages 153–156, Salvador da Bahia, Brazil, 2008.

[3] N. Basher, A. Mahanti, A. Mahanti, C. Williamson, and M. Arlitt. A Comparative Analysis of Web and Peer-to-Peer Traffic. In *Proc. Conference on World Wide Web (WWW)*, pages 287–296, Beijing, China, 2008.

[4] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *Proc. ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 1–14, San Diego, U.S.A., 2007.

[5] X. Cheng, C. Dale, and J. Liu. Statistics and Social Network of YouTube Videos. In *Proc. IEEE International Workshop on Quality of Service (IWQoS)*, pages 229–238, Enschede, Netherlands, 2008.

[6] Conceiva. Download Managers - A Better Downloading Experience. Technical report, 2008.

[7] M. Dischinger, M. Marcon, S. Guha, K. P. Gummadi, R. Mahajan, and S. Saroiu. Glasnost: Enabling End Users to Detect Traffic Differentiation. In *Proc. USENIX Conference on Networked Systems Design and Implementation (NSDI)*, pages 405–418, San Jose, U.S.A., 2010.

[8] J. Erman, A. Gerber, M. T. Hajiaghayi, D. Pei, and O. Spatscheck. Network-aware Forward Caching. In *Proc. Conference on World Wide Web (WWW)*, pages 291–300, Madrid, Spain, 2009.

[9] A. Finamore, M. Mellia, M. Meo, M. Munafo, P. Torino, and D. Rossi. Experiences of Internet Traffic Monitoring with Tstat. *IEEE Network*, 25(3):8–14, 2011.

[10] P. Gill, M. Arlitt, N. Carlsson, A. Mahanti, and C. Williamson. Characterizing Organizational Use of Web-based Services: Methodology, Challenges, Observations, and Insights. *ACM Transactions on Web*, 11(1):2:1–2:26, 2011.

[11] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube Traffic Characterization: A View from the Edge. In *Proc. ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 15–28, San Diego, U.S.A., 2007.

[12] C. Gkantsidis, M. Ammar, and E. Zegura. On the Effect of Large-scale Deployment of Parallel Downloading. In *Proc. IEEE Workshop on Internet Applications (WIAPP)*, pages 79–89, San Jose, U.S.A., 2003.

[13] K. Gummadi, R. Dunn, S. Saroiu, S. Gribble, H. Levy, and J. Zahorjan. Measurement, Modeling and Analysis of a Peer-to-Peer File-Sharing Workload. In *Proc. ACM Symposium on Operating Systems Principles (SOSP)*, pages 314–329, Bolton Landing, U.S.A., 2003.

[14] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang. Measurement, Analysis, and Modeling of BitTorrent-like Systems. In *Proc. ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 35–48, Berkeley, U.S.A., 2005.

[15] E.-H. Han and G. Karypis. Centroid-Based Document Classification: Analysis and Experimental Results. In *Proc. European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*, pages 424–431, London, U.K., 2000.

[16] M. Iliofotou, G. Siganos, X. Yang, and P. Rodriguez. Comparing BitTorrent Clients in the Wild: The Case of Download Speed. In *Proc. International Workshop on Peer-to-Peer Systems (IPTPS)*, San Jose, U.S.A., 2010.

[17] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe. Towards IP Geolocation using Delay and Topology Measurements. In *Proc. ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 71–84, Rio de Janeriro, Brazil, 2006.

[18] S. Kaune, R. Rumin, G. Tyson, A. Mauthe, C. Guerrero, and R. Steinmetz. Unraveling BitTorrent's File Unavailability: Measurements and Analysis. In *Proc. IEEE Conference on Peer-to-Peer Computing (P2P)*, Delft, Netherlands, 2010.

[19] S. Koo, C. Rosenberg, and D. Xu. Analysis of Parallel Downloading for Large File Distribution. In *Proc. IEEE Workshop on Future Trends of Distributed Computing Systems (FTDCS)*, pages 128–135, San Juan, Puerto Rico, 2003.

[20] C. Labovitz. Takedown. `http://asert.arbornetworks.com/2010/07/takedown/`, Jul 2010.

[21] C. Labovitz, S. Johnson, D. McPherson, J. Oberheide, and F. Jahanian. Internet Inter-domain Traffic. In *Proc. ACM SIGCOMM Conference*, pages 75–86, New Delhi, India, 2010.

[22] Z. Liu, P. Dhungel, D. Wu, C. Zhang, and K. W. Ross. Understanding and Improving Ratio Incentives in Private Communities. In *Proc. IEEE International Conference on Distributed Computing Systems (ICDCS)*, pages 610–621, Genoa, Italy, 2010.

[23] G. Maier, A. Feldmann, V. Paxson, and M. Allman. On Dominant Characteristics of Residential Broadband Internet Traffic. In *Proc. ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 90–102, Chicago, U.S.A., 2009.

[24] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proc. ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 29–42, San Diego, U.S.A., 2007.

[25] J. A. Muir and P. C. V. Oorschot. Internet Geolocation: Evasion and Counterevasion. *ACM Computing Surveys*, 42(1):4:1–4:23, 2009.

[26] I. Poese, B. Frank, B. Ager, G. Smaragdakis, and A. Feldmann. Improving Content Delivery Using Provider-aided Distance Information. In *Proc. ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 22–34, Melbourne, Australia, 2010.

[27] I. Poese, S. Uhligh, M. A. Kaafar, B. Donnet, and B. Gueye. IP Geolocation Databases: Unreliable? *ACM SIGCOMM Computer Communication Review*, 41(2):53–56, 2011.

[28] J. Pouwelse, P. Garbacki, D. Epema, and H. Sips. The Bittorrent P2P File-sharing System. In *Proc. International Workshop on Peer-to-Peer Systems (IPTPS)*, Ithaca, U.S.A., 2005.

[29] RapidShare Company Profile. `http://tinyurl.com/RapidShareProfile`, Apr 2011.

[30] RapidShare News. `http://tinyurl.com/RapidShareNews`, Apr 2011.

[31] R. Sadre and B. R. Haverkort. Changes in the Web from 2000 to 2007. In *Proc. IFIP/IEEE Workshop on Distributed Systems: Operations and Management: Managing Large-Scale Service Deployment (DSOM)*, pages 136–148, 2008.

[32] A. Savoia. Web Page Response Time 101. *Software Testing and Quality Engineering Magazine*, pages 48–53, 2001.

[33] TorrentFreak. Google Starts Censoring BitTorrent, RapidShare and More. `http://tinyurl.com/GoogleCensor`, Jan 2011.

[34] R. Torres, A. Finamore, J. R. Kim, M. Mellia, M. Munafo, and S. Rao. Dissecting Video Server Selection Strategies in the YouTube CDN. In *Proc. IEEE International Conference on Distributed Computing Systems (ICDCS)*, Minneapolis, U.S.A., 2011.

[35] C. Zhang, P. Dhungel, D. Wu, and K. W. Ross. Unraveling the BitTorrent Ecosystem. *IEEE Transactions on Parallel and Distributed Systems*, 22(7):1164–1177, 2011.

[36] S. Zhao, D. Stutzbach, and R. Rejaie. Characterizing Files in the Modern Gnutella Network: A Measurement Study. In *Proc. of SPIE/ACM Multimedia Computing and Networking (MMCN)*, San Jose, U.S.A., 2006.

[37] M. Zink, K. Suh, Y. Gu, and J. Kurose. Characteristics of YouTube Network Traffic at a Campus Network-Measurements, Models, and Implications. *Computer Networks*, 53(4):501–514, 2009.