# Content Sharing Dynamics in the Global File Hosting Landscape

Aniket Mahanti[1], Niklas Carlsson[2], Carey Williamson[1]
[1] Department of Computer Science, University of Calgary, Canada
[2] Department of Computer and Information Science, Linköping University, Sweden

***Abstract*—We present a comprehensive longitudinal characterization study of the dynamics of content sharing in the global file hosting landscape. We leverage datasets collected from multiple vantage points that allow us to understand how usage of these services evolve over time and how traffic is directed into and out of these sites. We analyze the characteristics of hosted content in the public domain, and investigate the dissemination mechanisms of links. To the best of our knowledge, this is the largest detailed characterization study of the file hosting landscape from a global viewpoint.**

***Keywords*-Measurement, file hosting services, traffic characterization, peer-to-peer.**

## I. Introduction

File hosting services enable convenient publishing and dissemination of content through a Web interface. In contrast to peer-to-peer (P2P) networks, no specific software is required. Instead, the system is typically built on top of HTTP. With file hosting services, a client typically uploads a file to an online file storage location in the cloud (hosted by the file hosting services). At the time of the upload the user is given a URL to the file, such that the file can be later downloaded using that URL. In some cases, at the time of a download, mandatory wait times, CAPTCHAs, and bandwidth throttling, are used to provide differentiated service to the different clients.

Today, there are a large number of file hosting services, which together are responsible for a significant fraction of the total Internet traffic. Recent reports [1], [2] suggest that up to 19% of the global Internet traffic volume are due to these services. While the future of some of these services are clouded by lawsuits, media scrutiny, as well as the controversial takedown of one of the more visible sites (as we show in this paper) the landscape is continually changing with new services being added and users migrating between services.

While file hosting services are easy to use, characterizing their usage is less trivial. The URLs to the content typically are not advertised by the services themselves, but instead are shared through other means. Furthermore, in contrast to highly characterized services such as YouTube, these services typically do not provide public information about the content or statistics related to the amount of sharing associated with each content. There are, however, many interesting aspects to the sharing of this content, as the usage is often channeled through forums, blogs, and various entertainment-related sites, where the content publishers may share and promote their URLs. There are also an increasing number of specialized search engines that help users find public download links.

In this paper we collect and analyze a number of datasets, targeted towards capturing the content sharing dynamics of the file hosting ecosystem. While previous works [3]–[5] have considered the usage and performance as observed by users on campus networks, *this paper takes a global viewpoint*. Using a combination of Web analytics and active measurements we study usage dynamics, content sharing, and content characteristic of these services. We primarily focus on the following sites: *RapidShare*, *Megaupload*, *MediaFire*, and *Hotfile*.

*Service popularity dynamics (Section IV)*: The first objective of this paper is to provide insights into the dynamics between some of the major file hosting players. To the best of our knowledge, there has been one longitudinal study of file hosting service usage [3]. In contrast to this work, which considered one year of network traffic on a campus network, we consider the usage as observed in the U.S. over a 45-month period. For comparison, we also contrast the service usage with other popular services, such as *Pirate Bay* (P2P), *YouTube* (video sharing), and *Hulu* (video-on-demand).

*Content discovery (Section V)*: The second objective of this paper is to provide insights into how users find their way to the content. With the file hosting services not advertising any links to the content themselves, file hosting service users typically must rely on other sites for the sharing of links to the content. While this complicates the characterization of the complete content sharing landscape, it also presents a unique social aspect that is not as apparent in other content distribution platforms. Through comparison of the sites visited before and after visiting different file hosting services, this paper provides a first glimpse into how users navigate to find content.

*Content sharing (Section VI)*: The third objective of this paper is to provide insights into how content links are shared. To this end, we present a case study for how content is shared across these services, analyze how much content is replicated across services, and take a closer look at downloading activity of the files uploaded by some publishers. We believe that these targeted experiments and analyses provide a unique perspective of these services. We find that content is released earlier than on P2P networks, large

number of duplicate contents are simultaneously available across services, and even though the users are not aware of the content downloaded by prior users, the popularity of the files uploaded by content publishers show signs of power-law skew.

*Content analysis (Section VII)*: The fourth objective is to provide a global view of the content shared through these services. An interesting aspect of these contents is the impact of file size limitations. This has resulted in publishers having to split contents into smaller files, as well as applying compression. To better understand the impact of these limitations, we identify multi-part contents and classify compressed files into categories associated with the original content. While the content characteristics are similar to those observed using only content downloaded at a campus network [3], we believe that our results validate prior observations at a much larger scale (because we analyze the largest set of files to date, which is not limited to a single geographic region).

The remainder of the paper is organized as follows. Section II reviews prior work. Section III presents our datasets and measurement methodology. Sections IV-VII each addresses one objective at a time. Finally, conclusions are presented in Section VIII.

## II. Related Work

Limited work has been done on characterizing the file hosting ecosystem. Two works have performed measurement and analysis of content sharing traffic for selected services [4], [5]. Mahanti *et al.* [3] analyzed the characteristics of the file hosting ecosystem as seen from a large campus edge network.

RapidShare service architecture, usage patterns, and content characteristics were studied by Antoniades *et al.* [4], with the traces collected from two academic networks. They used active measurements to compare RapidShare with BitTorrent in terms of user-perceived throughput and content similarity. Most RapidShare files on the academic networks were requested once. Through targeted experiments, they found that RapidShare evenly distributed load across storage servers.

Cuxart *et al.* [5] analyzed RapidShare and Megaupload traffic using traces collected from a research network. They studied traffic properties, usage, content distribution, and server infrastructure. They noted that RapidShare and Megaupload were responsible for a significant fraction of the total traffic, and relied on a huge server infrastructure. A non-negligible percentage of users paid for premium accounts.

Mahanti *et al.* [3] analyzed the usage behavior, infrastructure properties, content characteristics, and user-perceived performance of five services, namely RapidShare, Megaupload, zSHARE, MediaFire, and Hotfile. They observed positive growth trends for file hosting services as well as a large number of premium downloads. Most services had their servers in a centralized location, with the exception of Megaupload. Premium users achieved higher throughputs than free users, with both user types getting better rates than P2P transfers.

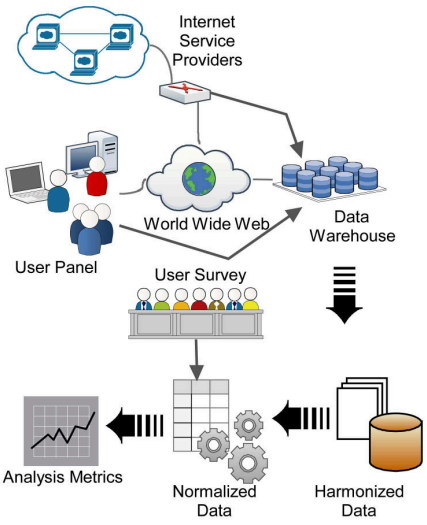



Figure 1. Web analytics data collection methodology

Our work complements these studies. While all of these works were performed with the data collected from academic edge networks, our datasets provide insights on the file hosting ecosystem from a global perspective. Academic networks have skewed demographics, which may affect how some of the file hosting services are used at these networks.

## III. Methodology

We employed a *three-pronged* approach to understand the file hosting ecosystem from a global viewpoint. First, we use `Compete.com`'s *analytics* framework to perform a longitudinal trend analysis of usage and popularity of four major file hosting services. We also undertake detailed analysis of sites that drive traffic in and out of these file hosting services. Second, we use active measurements through a *crawl* of links on a large file hosting index site (`Filestube.com`). We perform a detailed analysis of the content characteristics and file availability. Third, we collect *supplementary* data on server-side file popularity, as well as some targeted case studies of content replication and publishing.

*Analytics dataset:* This measurement is divided into four stages: *data collection*, *data harmonization*, *data normalization*, and *analysis metrics reporting*. Figure 1 presents an illustration of the methodology.

The analytics measurement is based on a user-centric, sample-based, multi-source panel [6]. The measurements are based on U.S. Internet users only. Compete has a core panel of 350,000 users who have installed Compete's measurement software on their machines. Compete also

gathers clickstream data from ISPs and application service providers. (A clickstream is the sequence of user requests that generate HTTP transactions while browsing a Web site.) Along with the core panel, Compete utilizes these data sources to accumulate clickstream data from 2 million users each month, representing 1% of the U.S. Internet population.

Clickstreams from the core panel and other data sources are aggregated into a unified online panel using proprietary harmonization algorithms, which takes into account the mapping of unique user identifiers to their profile consisting of age, income, gender, and geographic locale. Leveraging a monthly survey panel, these clickstreams are normalized to project the user behavior of the entire U.S. Internet population. The final outcome is a dataset that provides a representation of the characteristics of the monthly U.S. Internet population and its Internet usage. We use the final meta data to perform a comprehensive analysis of the file hosting ecosystem.

*Crawl dataset:* The second global dataset was collected by crawling a large file hosting index site. This site indexed over 100 million publicly available file hosting links at the time of the crawl. The index site offered an API that allowed us to crawl several file hosting links. Since file hosting services do not allow their hosted files to be searched, the site can only index files that are available in the public domain. The following information about the crawled files was collected: name, size, file extension, URL, tags, rating, and date added. The crawl was performed between March and July, 2010.

The crawl involved starting with a set of tags fetched from the index site's *Last added* section on the homepage. Each day our crawler would fetch the tags, put the tags into a queue, and initiate a multi-threaded crawl. Each crawling thread would pop a tag from queue, request *tag + file hosting site* to the index site, and then parse results.

The main thread would dump the crawled data into a text file every 20 seconds. This thread would also monitor request count; once the number of requests reached 20,000 in a day, the program would terminate. This restriction was imposed by the index site when using its API. We chose to use tags from the Last added section to reduce bias in our search results. Since the Last added section provides a snapshot of newly added links to the index site's database, it reduces the likelihood of using popular search tags.

We utilize this dataset to understand the characteristics of content hosted on the file hosting services. The crawl dataset consists of 920,775 files from the four file hosting services considered here, namely RapidShare (RS), Megaupload (MU), MediaFire (MF), and Hotfile (HF). These files represented about 1% of the total files indexed at the time of measurement.

Table I provides an overview of the characteristics of the dataset. RapidShare was the largest file hosting service accounting for about 44% of the total files. Hotfile was the second largest with 22% of the files. Megaupload and MediaFire had 18% and 16% of the total files, respectively. The sum of the file sizes was over 90 TB. In terms of file size, RapidShare represented 41% of the total bytes, while Megaupload and Hotfile each accounted for 26%, and MediaFire had 7% of the total byte count. The average file sizes of the different services are in line with the maximum file size restriction at that time. MediaFire files were the smallest on average, while Megaupload files were the largest.

Table I
OVERVIEW OF THE CRAWL DATASET

| Characteristic | RS | MU | MF | HF |
|---|---|---|---|---|
| Num. of files | 405,794 | 166,230 | 148,742 | 200,009 |
| Num. of content | 246,869 | 145,064 | 140,096 | 147,951 |
| Total bytes (TB) | 38.21 | 23.53 | 6.15 | 24.04 |
| Avg. file size (MB) | 98.74 | 148.43 | 43.377 | 126.03 |
| Avg. content size (MB) | 246.82 | 167.47 | 39.72 | 177.38 |
| Active links (%) | 81.92 | 94.75 | 93.26 | 78.72 |
| Avg. file age (days) | 254 | 210 | 218 | 51 |
| Files rated (%) | 0.64 | 0.89 | 0.66 | 0.32 |
| Fragmented files (%) | 62.34 | 18.29 | 11.52 | 46.59 |
| Fragmented bytes (%) | 79.81 | 19.98 | 29.86 | 57.84 |

Understanding the file activity is important to know how long files last in the file hosting ecosystem. We found that most of the files were active with Hotfile having the most inactive links at 13%. For active files, the average file longevity was about 8 months. Hotfile had the lowest file longevity because it was the newest file hosting service in the group.

The file hosting index site allowed its users to rate the links they found. It seems that social features such as ratings are not used frequently, with less than 1% of the file links being rated. This may be because the index site is not the source for the links. Rating and commenting are often used in P2P torrent discovery sites to monitor the quality of content, and for mitigating malicious or fake content.

## IV. SERVICE POPULARITY DYNAMICS AND TRENDS

To obtain a bird's-eye-view into the popularity dynamics of the most popular file hosting services, we analyze them using monthly trends for various usage characteristics. To this end, we leverage monthly analytics statistics spanning 45 consecutive months, from May 2008 until January 2012.

Figure 2 shows the number of unique monthly visitors to the four file hosting services and two P2P torrent discovery sites (*Pirate Bay* and *Mininova*). For the first half of the measurement period, the popularity of all services increased. After a temporal peak in mid-2010 about half of the services have seen increased usage and half have seen decreasing usage. More precisely, except for RapidShare and Mininova, the site usage of all the other considered sites peaked in May 2010. Mininova shows a declining trend starting November 2009, when (as a result of a court verdict) Mininova removed most of its indexed torrents during November 2009[1].

[1] http://torrentfreak.com/mininova-deletes-all-infringing-torrents-and-goes-legal-091126/

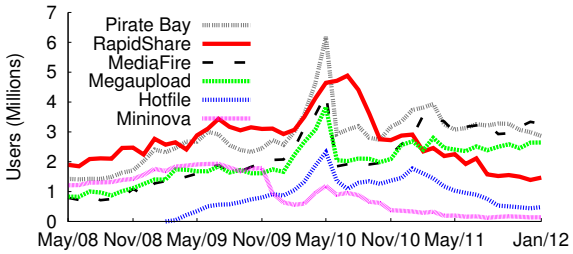



Figure 2. Number of monthly unique users

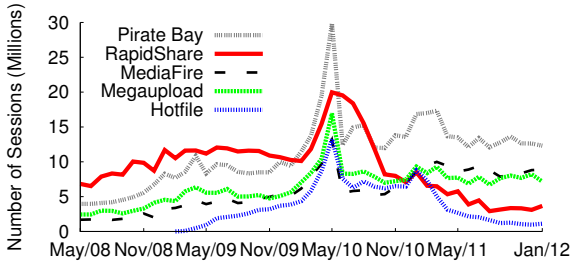



Figure 3. Number of monthly user sessions

Mininova now only hosts torrents from artists and producers who want to distribute their content for free. Apparently such content is not popular to sustain the user base, and hence Mininova has continued its declining trend since November 2009.

The RapidShare user count peaked in June 2010, and then started to decline, possibly because of the site's decision to end its rewards program[2]. Megaupload continued its growth until it was shut down towards the end of January 2012. It is interesting to observe that Hotfile's usage increased at the point when RapidShare stopped its rewards program. This could indicate RapidShare users migrating to Megaupload and Hotfile. Hotfile's usage declined starting February 2011, when in response to a lawsuit the site started terminating many user accounts[3]. MediaFire's usage increased during this turmoil. MediaFire does not operate a rewards program, which could mean that there may be limited overlap of users between MediaFire and other services. Pirate Bay's usage remained the highest among all the services between May 2010 and January 2012, with a slight dip towards the start of 2012.

We observed similar trends for other popularity-based characteristics. For example, Figure 3 shows the monthly trends for user sessions for the various services. These trends follow the monthly user count trends closely. The scale of the number of sessions is much larger, with the maximum reaching close to 30 million sessions for Pirate Bay, in comparison to a peak user count of about 7 million unique users in May 2010. These session counts are still orders of magnitude smaller than those of YouTube, which had 103 million unique users and over 800 million user sessions during that period.

Another example is the number of page views for the different services. While these results are omitted, the results follow that of the monthly users, but typically with steeper slopes. For example, between May 2010 and January 2012, the RapidShare user count reduced by 53%, while the page view count declined by over 83%. Page views are often used to measure engagement of users with a site, with a greater page view to user count ratio reflecting increased user interaction with the site. In the case of file hosting services, less page views could mean fewer content being hosted on the site.

Table II
SUMMARY STATISTICS FOR JAN 2012

| Characteristic | RS | MU | MF | HF | PB |
|---|---|---|---|---|---|
| Num. Users (millions) | 1.47 | 2.64 | 3.27 | 0.48 | 2.86 |
| Num. Page Views (millions) | 21.46 | 23.74 | 30.40 | 2.91 | 104.63 |
| Num. Sessions (millions) | 3.67 | 7.16 | 8.88 | 1.11 | 12.31 |
| Avg Session Length (min) | 4.37 | 3.41 | 3.70 | 2.93 | 5.10 |
| Num. Sessions per User | 2.50 | 2.71 | 2.72 | 2.31 | 4.30 |
| Num. Pages per Session | 5.85 | 3.31 | 3.42 | 2.63 | 8.50 |

Overall, these results show that the file hosting ecosystem is dynamic and that services are faced with heavy competition. Services that offer better rewards tend to be favored, and get more users. We also observe that popularity in the ecosystem is ephemeral, and users tend to move from one service to another frequently.

Table II shows summary statistics for our most recent measurement point, the month of January, 2012. For comparative analysis, we also show summary statistics for a popular P2P torrent discovery site, Pirate Bay (PB).

Today, MediaFire is the most popular file hosting service among the ones that we tracked over the 45-month period. In January 2012, it had 3.3 million users generating about 9 million sessions with over 30 million page views. While these numbers are smaller than those observed by RapidShare in June 2010, it appears that the landscape is continually evolving and new file hosting services are emerging.

When comparing file hosting services to P2P, we observe that Pirate Bay had a smaller user base of about 2.9 million users in January 2012, but these users generated more sessions and page views than MediaFire. These users also spent more time per session. These statistics show that although file hosting services were gaining acceptance among users, individually, they were smaller than Pirate Bay.

## V. BROWSING ANALYSIS AND SERVICE USAGE

In contrast to most other services, the content hosted by the file hosting services can typically be discovered through other sites and services than through the file hosting services

[2]http://torrentfreak.com/rapidshare-kills-reward-program-over-piracy-concerns-100620/
[3]http://torrentfreak.com/hotfile-goes-to-war-against-copyright-infringers-110219/

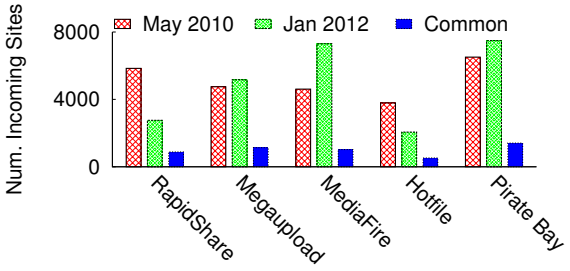



Figure 4. Distribution of unique incoming sites

Table III
INCOMING SITES FOR JANUARY 2012

| Incoming site | RS | MU | MF | HF | PB |
|---|---|---|---|---|---|
| Direct Traffic | 8.56 | 11.96 | 7.46 | 12.20 | 22.19 |
| File Hosting Search | 27.04 | 9.40 | 3.53 | 10.01 | 0.41 |
| Entertainment | 13.11 | 26.67 | 2.22 | 8.70 | 3.31 |
| File Hosting | 16.99 | 13.05 | 3.66 | 23.20 | - |
| Blogs/Portal | 10.78 | 10.72 | 11.09 | 15.67 | 8.85 |
| General Search | 6.94 | 6.53 | 37.79 | 8.02 | 38.52 |
| Social Media | 6.72 | 18.27 | 19.79 | 8.01 | 10.59 |
| Adult | 5.96 | 0.82 | 0.67 | 10.42 | 0.33 |
| Technology | 3.89 | 2.58 | 13.78 | 3.77 | 4.46 |
| Torrent | - | - | - | - | 11.33 |
| Top-100 Share | 66.35 | 66.88 | 72.78 | 50.60 | 77.85 |

themselves. For example, a publisher may provide links to the content through various media, including online forums, social networks, or email. We leverage statistics based on the analytics dataset to gain insights into how users obtain access to the content hosted on the four file hosting services.

The analytics dataset provides information about which sites the user may have been referred from, as well as where the user diverted to after accessing a particular site. We first look at *incoming* traffic sites. These are the external sites that the user was currently browsing before landing on the file hosting service. These sites are not necessarily referring sites, wherein the user clicks a link explicitly linking to the site of interest. We also look at *destination* sites, which are the sites the user lands on after they have navigated away from a file hosting service. A break down of the type of incoming and destination sites helps us understand how file hosting services receive traffic from other sources and also helps us understand the browsing habits of the users.

Figures 4 shows the distribution of incoming sites for May 2010 and January 2012, and the number of sites in common between these two months. We note that all services observe incoming traffic from a significant number of different sites, with only 14-32% sites in common between the two time periods. This may suggest that there is a high churn in the referral sites on which links to these contents are shared.

RapidShare has the highest number of sites in common as a fraction of the total incoming sites for January 2012. This may be due to significant decrease in the count of incoming sites for RapidShare, as well as many incoming sites continuing to be used by RapidShare users.

We found that in terms of scale, there are far fewer incoming sites for file hosting services than for YouTube and Hulu. YouTube had over 200,000 incoming sites, while Hulu had about 14,000 incoming sites. These differences are likely due to their widespread general usage, whereas content sharing still is seen as more of a niche activity. There was 112% decrease in the number of incoming sites for RapidShare, while there was a 45% decrease for Hotfile. All other services witnessed an increase, with the highest increase for MediaFire at 37%. Pirate Bay had a 13% increase in the count of incoming sites.

We now further analyze the incoming sites. We categorized the top-100 incoming sites for various services into 10 groups. Table III shows the percentage of user sessions as a result of referrals from the 10 incoming site groups in January 2012. The top-100 incoming sites resulted in over 50% of the user sessions for the services. We normalized the share of the 10 groups to sum to 100%.

We find that direct traffic accounts for over 8% of the file hosting user sessions, with the highest being for Pirate Bay at 22%. This is due to the organization of the services. File hosting sites do not have a built-in search engine to search for content, while Pirate Bay does, and hence more traffic is generated by users directly visiting the homepage of Pirate Bay. The direct traffic for file hosting services may indicate premium users accessing their hosted files or inquisitive users visiting the homepage. Interestingly, we find that file hosting search engines directed more traffic to these services in January 2012 as compared to May 2010. This could be due to file hosting search engines indexing more content from these services as they matured.

We also notice other file hosting sites from which users visit the four file hosting services. While other services may not be directly linking to these sites, it exemplifies the browsing pattern of the users. Users who access one file hosting site are likely to visit other similar sites looking for content. Other sites of interest to file hosting users are general entertainment sites, blogs and portals hosting links, adult content sites, and technology sites. We also analyzed the incoming traffic distribution for Hulu and found that direct traffic accounted for almost 30% of the sessions, while the top-5 incoming sites (general search engine and social media) alone accounted for over 65% of the sessions.

General search engines such as Google or Bing did not provide much traffic to these sites, with MediaFire being an exception. MediaFire content publishers can make their hosted content searchable, and this could be the reason for such high traffic being generated from general search engines. Figure 5 shows the monthly trends for general search referral traffic for the various services. We observe that during some months, most of MediaFire's user sessions were due to search referrals. Pirate Bay received about

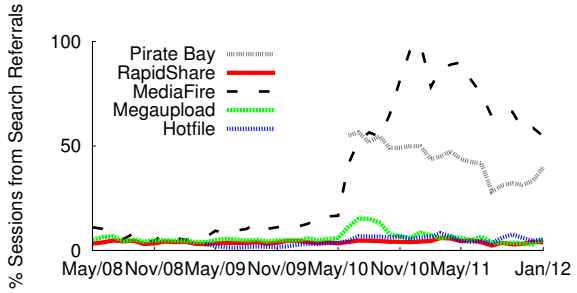



Figure 5. Percentage of user sessions generated due to search engines

Table IV
OUTGOING DESTINATION SITES FOR JANUARY 2012

| Destination site | RS | MU | MF | HF | PB |
|---|---|---|---|---|---|
| File Hosting Search | 18.29 | 4.39 | 2.56 | 8.03 | 0.29 |
| Entertainment | 10.84 | 15.51 | 3.48 | 7.20 | 5.46 |
| File Hosting | 19.50 | 14.19 | 3.35 | 28.80 | - |
| Blogs/Portal | 10.17 | 14.51 | 22.05 | 12.34 | 21.24 |
| General Search | 24.93 | 17.22 | 29.19 | 22.66 | 23.94 |
| Social Media | 8.48 | 21.74 | 29.47 | 8.29 | 12.58 |
| Adult | 6.45 | 0.64 | 0.30 | 9.27 | 1.01 |
| Technology | 1.04 | 11.56 | 9.43 | 3.40 | 8.46 |
| Torrent | 0.29 | 0.23 | 0.15 | - | 27.02 |
| Top-100 Share | 62.61 | 66.01 | 72.14 | 57.63 | 76.23 |

50% of its traffic from users searching for content. Other file hosting sites received negligible search referrals. Search service providers such as Google have lately started censoring names of some file hosting services from its search completion feature, which could impact how traffic is driven to these services[4].

We also analyzed the top-100 destination sites for users of file hosting services. This analysis allows us to understand what users do after they have used a file hosting service, and provide insights on the competition for these services. Table IV shows the outgoing traffic distribution for January 2012.

We observe that most users landed on other file hosting services, entertainment forums, and blogs/portals. Many of them also visited general search engines. In contrast, far more P2P users visited a general search engine. A non-negligible portion of file hosting service users visited torrent discovery sites, however, more Pirate Bay users visited other torrent sites. While our results suggest that there is strong competition between file hosting services, many users do not appear to alternate between using file hosting services and P2P within the same session.

Comparing with statistics collected for May 2010 (omitted due to space limitations), we can also note that more Pirate Bay users are visiting other torrent sites than previously. In contrast, there is not much change in outgoing traffic for file hosting service users. More users are now using file hosting search engines afterwards, as compared to May 2010.

[4]http://torrentfreak.com/google-starts-censoring-bittorrent-rapidshare-and-more-110126/

Table V
NUMBER OF DUPLICATE FILES ACROSS SERVICES

| Num. Services | Service combination | Dup. Files |
|---|---|---|
| One | RapidShare | 6,480 |
| | Megaupload | 5,760 |
| | Hotfile | 3,286 |
| | MediaFire | 797 |
| Two | RapidShare ∩ Hotfile | 15,969 |
| | RapidShare ∩ Megaupload | 2,424 |
| | RapidShare ∩ MediaFire | 1,290 |
| | Megaupload ∩ Hotfile | 1,655 |
| | Megaupload ∩ MediaFire | 755 |
| | Hotfile ∩ MediaFire | 438 |
| Three | RapidShare ∩ Megaupload ∩ Hotfile | 3,158 |
| | RapidShare ∩ Megaupload ∩ MediaFire | 945 |
| | RapidShare ∩ MediaFire ∩ Hotfile | 728 |
| | Megaupload ∩ Hotfile ∩ MediaFire | 307 |
| Four | RapidShare ∩ Megaupload ∩ MediaFire ∩ Hotfile | 461 |

## VI. CONTENT SHARING

### A. Duplication across services

Due to incentives and high content sharing success, some publishers share the same content across multiple file hosting services. We were interested in knowing how many files were replicated across the services. We utilized the crawl dataset and analyzed the file names across each service. Each file name was split into several tokens separated by special characters.

Next, we built two lookup tables. One table consisted of file to token mappings. The second table had a mapping of unique tokens to file names wherein for each token a list of files containing that specific token was kept. Next for every file we constructed a relative table of file versus weight. We processed all file tokens step-by-step and cross-referenced all the files that contained the tokens. These files were added to the relative table and a weight assigned based on the token length. Once all the token had been processed, we analyzed the relative table. Those files that had a relative table weight within 5% of the original file weight were labeled as duplicates. A lower threshold value would ensure better accuracy, albeit at the expense of lower coverage. We performed a manual inspection of the results on a sample of over 200 file combinations, and found the results to be accurate.

Table V shows the results of the classification for the dataset. There were many files that were replicated within each service. Such files could indicate multiple users uploading the same content on the same service. We found the greatest overlap between RapidShare and Hotfile. Hotfile was an emerging file hosting service, while RapidShare was already very popular during the time of measurement. This indicates some content publisher trying to move to a newer service, while keeping the content active on an established service. In general, we find that as a percentage of the total content hosted, these duplicate files formed a small fraction.

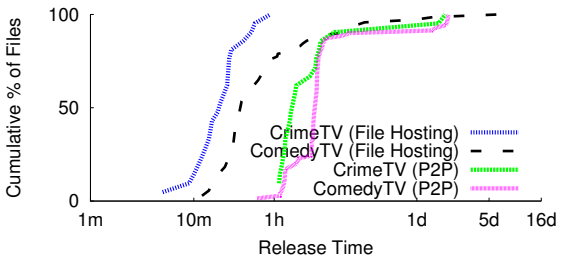



Figure 6. Release times of content

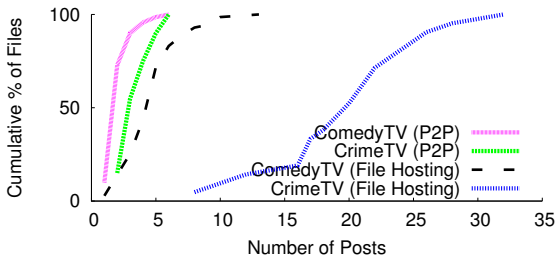



Figure 7. Number of posts

This may suggest that content publishers often host their contents on a limited set of sites. (Of course, they may still use multiple sites to promote the content links.)

### B. Case study: Publishing and replication

With many of the links being shared over transient discussion platforms, including speciality blogs and forums, these services may appeal more to a special category of users. This was the case when looking at the user profiles observed in the analytics dataset. For example, we found file hosting users to be younger than the average U.S. Internet user. Our results show that up to 56% of the file hosting users are below the age of 35, while 37% of the U.S. Internet users are below this age. YouTube's share of this age-group is 40%.

We next present a case study comparing the dissemination of popular content via file hosting services and P2P. Specifically, for both services, we analyze how quickly TV content is made available once the content has been broadcast, and how many copies of the content are available on the Web. We tracked the postings of a popular weekly crime television drama and a daily comedy show on file hosting services and on P2P for one season in 2009-10. In total, we collected the release times of 180 episodes from a large forum with over 1.9 million registered members listing file hosting links, and from Pirate Bay.

Figure 6 shows the distribution of release times of content for file hosting services and P2P. File hosting links for the content are posted more promptly when compared to P2P. For the crime TV program, about 20% of the content is released within 10 minutes of the program ending its first broadcast. The median release time was around 20 minutes. All file hosting links are posted within 55 minutes. It takes much longer to post the content on P2P, with 20% of the links being posted within 71 minutes after the program ended airing. For the comedy show, 50% of the links were posted within 28 minutes. The earliest release time was 12 minutes. We also notice a much longer tail to the distribution with the latest links being published 16 days after the original airing. The distribution for P2P was similar to that of the crime TV program, but with slightly greater delays.

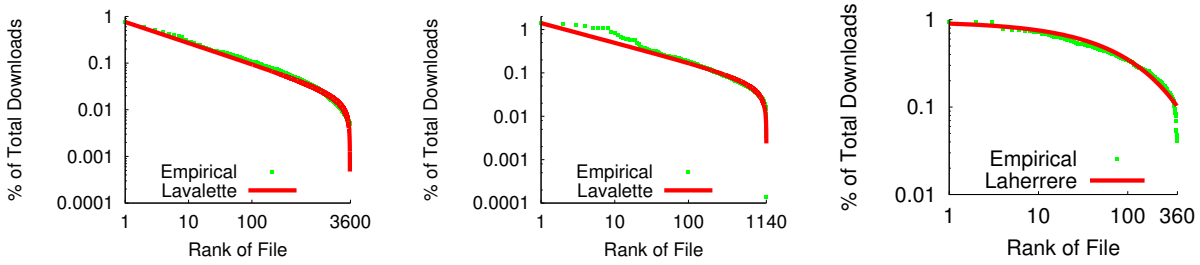
Figure 7 shows the number of posted links for the TV programs. We notice far more file hosting links are available for the programs than on P2P. This translated into greater range of available times when these links could be posted. Content publishers would like to publish the links as early as possible to increase their download count, which would result in greater rewards. In marketing research, it has been observed that there is a significant first-mover advantage. This translates to content that is published earlier being more likely to be downloaded than content that is posted later. File hosting publishers may compete to be the first to publish popular or highly sought-after contents to maximize their incentives. Additionally, the large number of posts indicate greater choice of file hosting services. It means greater opportunity for content publishers to provide multiple downloading options for content consumers.

### C. Content popularity: The publisher's perspective

An important aspect of designing good content delivery systems is to understand the popularity of different contents. Unfortunately, file hosting services typically do not provide public information about the number of downloads for various contents. Therefore, it is difficult to assess the skew in popularity of contents. It is also unclear if the fact that content consumers often are not aware of the total number of downloads of a particular content (as it may not be shown in the forum or elsewhere) may affect the probability of the content being downloaded. In this section, we look at content popularity as observed from the perspective of a set of publishers.

For this analysis we used three datasets consisting of file names and their corresponding download counts from three anonymous RapidShare content publishers. These publishers had, on request, voluntarily provided us with this data. The first dataset contained 3,525 files, the second dataset contained 1,140 files, and the last dataset had 354 files.

Figure 8 shows the normalized download counts for the three publishers in the *rank-frequency* domain. Note the different scales across the three figures. The first rank is assigned to the file with the largest share of the total download count. We analyzed these distributions for the best *power-law* fit. None of the distributions fit the *Zipf* distribution, which is represented by a straight line on a




(a) Large publisher ($b = 0.4519$)

(b) Medium publisher ($b = 0.4533$)

(c) Small publisher ($n_0 = 99.6626, b = 0.6292$)

Figure 8. Server-side file popularity

log-log plot. We fitted the distributions to several variations of the power-law distribution, including *Lavalette*, *Zipf-Mandelbrot*, and *Tsallis*, as well as the stretched exponential *Laherrere* distribution [7]. These distributions offer better flexibility in fitting, without introducing greater parameter complexity than the Zipf distribution. We find that the file downloading activity of the large and medium publishers (see Figure 8(a) and (b)) are well-modeled by the Lavelette distribution, signifying popularity in files downloaded, as seen from the server-end. The file download popularity of the small publisher (see Figure 8(c)) fitted the Laherrere distribution much better than any power-law distribution. Formally, the Lavelette distribution is defined as $f(r) = c\left(\frac{Nr}{N-r+1}\right)^{-b}$, and the Laherrere distribution is written as $f(r) = c\exp(-\left(\frac{r}{n_0}\right)^{b})$. $f(r)$ is the frequency function, $r$ is the rank, $N$ is the number of ranked items, $c$ is a normalizing constant, and $b$ is the shape parameter, and $n_0$ is the scale parameter. The fitted parameters are included with the figure.

All three publishers observe high skew in their file popularity, with the two bigger publishers observing power-law-like popularity. Similar skew have been observed in other systems, such as YouTube, for which the download statistics of file contents (a measure of past popularity) is accessible to the consumer at the time of the download. The presence of a high skew, combined with HTTP-based content delivery, provides significant caching opportunities which may be used to offload the origin servers.

## VII. Content Analysis

With many services imposing size limitations, finding the file size distribution is a non-trivial task. A prior work [3] had considered the file size distribution and other content related properties on a campus network. However, as there is no guarantee that users download all pieces of a file, this paper takes a closer look at the content properties observed globally. For this analysis we utilize our crawl dataset. Our analysis is expansive in terms of number of files and geographic scope.

Table VI
File type distribution by file count

| Category | Example types | RS | MU | MF | HF |
|---|---|---|---|---|---|
| Archive | rar, zip | 88.14 | 64.64 | 59.28 | 82.82 |
| Video | avi, mp4 | 5.03 | 22.07 | 7.84 | 10.41 |
| Audio | mp3 | 2.42 | 4.99 | 20.28 | 2.12 |
| Executable | exe | 0.28 | 0.78 | 0.87 | 0.26 |
| Document | pdf | 0.83 | 1.48 | 2.38 | 1.09 |
| Other | - | 3.31 | 6.04 | 9.36 | 3.28 |

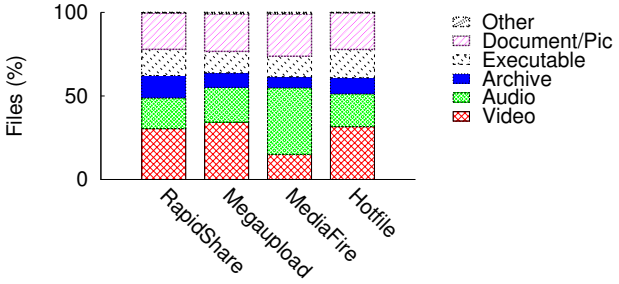



Figure 9. File type based on classification

### A. File type

We wanted to understand what type of content are hosted on the services. Table VI shows the distribution of file counts using the file extension. We notice a preponderance of archive files with WinRAR being the favored choice. RapidShare had the highest number of archive files due to its low file upload size limitation. Video files were the next most frequent file type. The AVI file type was the most common video file type; it is a common container to keep video file sizes small without loss of quality. MediaFire had a large fraction of audio files in comparison to other services. Other files types such as executables and documents were not prevalent.

Given the large percentage of archive files, we were interested in further investigating their underlying file types. The Centroid algorithm [8] had been proposed to classify Web pages using HTML tags. We apply a variation of the algorithm to classify files based on their file name. We

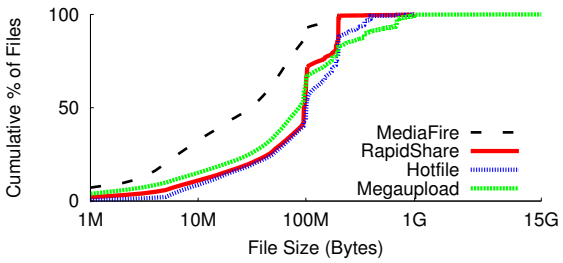



Figure 10. File size distribution

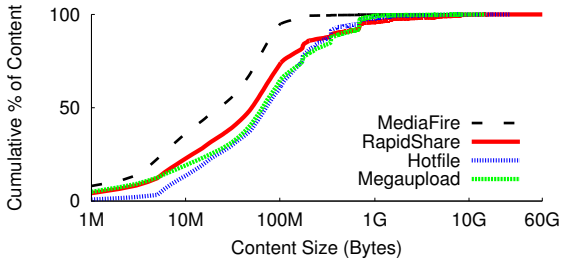
started by creating an exhaustive list of patterns, which were associated with a certain file type. For example, text patterns such as *720p*, *dvdscr*, and *bdrip* are associated with video files, and *serial*, *trial*, and *windows* are typically associated with applications. We next trained the algorithm on this list, and classified files into audio, video, document, or executable. Archive files that remained unclassified, are put in the archive category. We tested the algorithm on a sample of 243 randomly selected files. The algorithm was able to correctly classify about 71% of the files. After applying the algorithm on the crawl dataset, we found majority of them were multimedia files. Figure 9 shows the results. This category distribution is similar to those reported for P2P files [9].

With much of the content being shared through both P2P and file hosting services, hybrid download managers are possible [10]. While much of the file hosting traffic consists of large flows (aka elephant flows) such download managers may split the traffic into many smaller flows, making this traffic class more difficult to identify and manage.

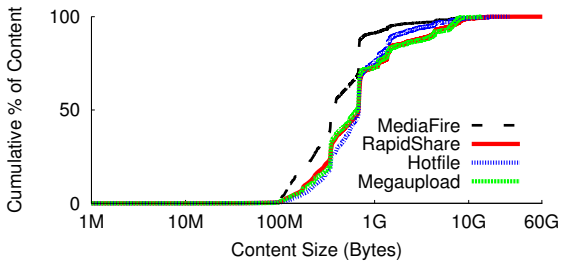
### B. File and content size

Figure 10 shows the file size distribution for the four file hosting services. Except for MediaFire, the other file hosting services had similar median file sizes. RapidShare had a median file size of 96 MB, Megaupload had a median file size of 85 MB, and Hotfile had the largest median file size at 100 MB. RapidShare had the lowest upload size limit of all the services, and hence we see that file sizes tend to follow this limit. RapidShare was the largest file hosting service at that time, and content publishers would follow its file size limitations even when files were hosted on other services. The file size distribution for Megaupload has a noticeable tail, which means that there were some extremely large files hosted for premium users only. Based on our file type classification, we found that video files had the largest median size.

Figure 11 shows the content size distribution. Figure 11(a) shows the content size distribution of all content. We notice that the distribution has a pronounced tail indicating some content that is very large. The maximum content size was about 60 GB. Except for MediaFire, all other services had




(a) Content size of all content




(b) Content size of multi-part content

Figure 11. Content size distribution

similar content sizes. This could indicate that these services host similar type of content. The median content size was small with MediaFire at 22 MB and Hotfile at 77 MB. RapidShare had a median content size of 50 MB, while Megaupload's was 65 MB. The average content sizes were significantly larger. RapidShare had an average content size of nearly 247 MB, Megaupload and Hotfile had similar average content sizes (around 170 MB), and MediaFire had the lowest (40 MB).

Figure 11(b) provides a closer look at the size distribution of multi-part content. The curves are shifted to the right because multi-part content is larger. This shows that content publishers do not split content if the content is within the file hosting upload size limit. The RapidShare tail is longer than other services. We also notice steep increases at specific content sizes. The first increase is visible at the 350 MB mark, which is a typical size for smaller video content. Another increase is visible at around 700 MB, which is a suitable size for content that can be copied onto a CD. The final distinguishable increase is at around 4 GB, which is the size of a DVD. Carlsson *et al.* [9] report similar results for size distribution of P2P content. Large content sizes can significantly impact the quality of service of other traffic. An increased presence of flows transporting these contents may require better network provisioning algorithms.

### C. Content fragmentation

The file size limitations imposed by the services often cause content publishers to split content into several fragments or parts and upload them individually. We now look at how much of the hosted content is fragmented.

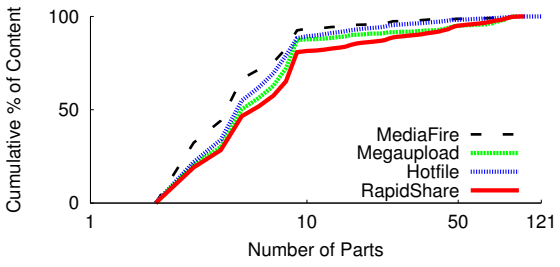



Figure 12. Content fragmentation of multi-part content

We analyzed the distribution of number of parts per content. We find that over 80% the contents consisted of a single file. RapidShare had the highest fragmentation because it had a lower file upload size limit. Megaupload had the lowest fragmentation, which can be attributed to its large upload size limit. Figure 12 provides a closer look at the distribution of the number of parts for multi-part contents. On average, the content on the services are split into 6 (for MediaFire) to 12 (for RapidShare) parts. The median parts per content were similar for all services (5 parts), except RapidShare (6 parts). Some services tend to have higher fragmentation such as Hotfile, which had the highest number of parts per content at 121 parts.

As the files belonging to the same content are likely to see similar popularity, the fragmented contents may be used to improve load balancing among servers. In fact, (in analysis omitted from this paper for brevity) we found that RapidShare already spreads files belonging to the same content across multiple servers, helping them effectively balance their load.

## VIII. Conclusions

We presented a comprehensive longitudinal characterization study of the dynamics of content sharing in the global file hosting landscape. We utilized measurements collected from multiple vantage points to understand service popularity dynamics, content discovery, content sharing, and content characteristics of the file hosting ecosystem.

We find that the file hosting ecosystem is highly dynamic, with popularity being ephemeral as users tend to move from one service to another frequently. Content publishers are key in the ecosystem, driving traffic to services where they post their content. Incentive programs often play an important role in attracting publishers, which in turn drives traffic to the file hosting sites. Traffic was directed to file hosting services through various channels including specialized search services, blogs, forums, technology sites, and social media. There was some duplication of content across the services indicating that publishers hosted their content on a limited set of services. Content was published on file hosting services earlier than P2P. Content popularity on the server-end had power-law behavior. Most of the content hosted were multimedia files.

Our results show a vibrant file hosting ecosystem with several services available at the disposal of users. While the ecosystem is subject to perturbation from publisher migration, HTTP-based content sharing is not affected. The apparent ease of publishing and downloading content and replication of content across services makes the ecosystem nimble to cope up with drastic changes.

## Acknowledegements

The authors are grateful to 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, and CENIIT at Linköping University.

## References

[1] Cisco Systems, "Cisco Visual Networking Index: Usage," White Paper, 2010, http://tinyurl.com/CiscoNetworks2010.

[2] Allot Communications, "MobileTrends: Global Mobile Broadband Traffic Report," White Paper, 2010, http://www.allot.com/mobiletrends.html.

[3] A. Mahanti, C. Williamson, N. Carlsson, M. Arlitt, and A. Mahanti, "Characterizing the File Hosting Ecosystem: A View from the Edge," *Perform. Evaluation*, vol. 68, no. 11, pp. 1085–1102, November 2011.

[4] D. Antoniades, E. Markatos, and C. Dovrolis, "One-click Hosting Services: A File-sharing Hideout," in *Proc. ACM SIGCOMM Conference on Internet Measurement*, Chicago, USA, November 2009.

[5] J. Sanjus-Cuxart, P. Barlet-Ros, and J. Sol-Pareta, "Measurement Based Analysis of One-Click File Hosting Services," *Journal of Network and Systems Management*, pp. 1–26, May 2011.

[6] Compete, "Overview of Data Methodology and Practices," White Paper, December 2011, http://tinyurl.com/CompeteMethodology.

[7] I. Popescu, "On a Zipf's Law Extension to Impact Factors," *Glottometrics*, vol. 6, pp. 83–93, 2003.

[8] E. Han and G. Karypis, "Centroid-Based Document Classification: Analysis and Experimental Results," in *Proc. European Conference on Principles of Data Mining and Knowledge Discovery*, London, U.K., September 2000.

[9] N. Carlsson, G. Dan, A. Mahanti, and M. Arlitt, "A Longitudinal Characterization of Local and Global BitTorrent Workload Dynamics," in *Proc. Passive and Active Measurement Conference*, Vienna, Austria, March 2012.

[10] P. Dhungel, K. Ross, M. Steiner, Y. Tian, and X. Hei, "Xunlei: Peer-Assisted Download Acceleration on a Massive Scale," in *Proc. Passive and Active Measurement Conference*, Vienna, Austria, March 2012.