Characterization of Content Delivery Applications

Aniket Mahanti

***Abstract*—We review the literature on Web traffic and workload characterization, P2P traffic characterization, and video sharing workload characterization. Section I describes related work on Web traffic characterization. Section II discusses P2P traffic characterization works. Section III discusses the literature on characterization of video sharing workloads. Section IV describes academic works that relate to file hosting traffic measurement. Section V discusses industry traffic reports.**

I. Web Traffic Characterization

There have been several studies on Web traffic characterization. We focus on studies that analyzed user-level and application-level characteristics of Web traffic. These studies have analyzed traffic as seen at servers [8]–[10], [12], [28], [41], [67], clients [6], [11], [29], [34], [64], and proxies [7], [29], [35], [40]. A survey of Web workload characterization studies prior to 1997 can be found in [50].

A. Server Workloads

Arlitt and Williamson [9], [10] performed one of the most comprehensive characterization studies of Web workloads. They collected and analyzed access logs of six different academic, research, and commercial Web servers between 1994 and 1995. They found that most HTTP requests were responded to successfully by the Web servers. Web documents (composed of HyperText Markup Language (HTML) pages and image files) were very small and their sizes followed a heavy-tailed distribution. A large proportion of the Web documents were requested only once, while most documents were requested multiple times. The file referencing pattern followed a Zipf-like distribution. There was heavy concentration of references, in that most of the requests and bytes transferred were for few files. Requests for popular files followed a Poisson arrival process. Remote hosts located on external networks were responsible for most of the requests.

In 2004, ten years after the original study, Williams *et al.* [67] revisited the work by Arlitt and Williamson. Most of the invariants in the original study held true, although the percentage of successful requests, and the fraction of HTML and image files transferred, were lower than before. They also commented on the drastic increase in Web traffic, which had grown by a factor of 30.

B. Client Traffic Traces

Smith *et al.* [64] used traces collected from a campus network between 1999 and 2000 to analyze Web traffic characteristics. Their analysis highlighted the impact of the HTTP/1.1 protocol on TCP's connection-level behavior. They found that the use of persistent connections had reduced the number of TCP connections by 50%. Multiple Web servers were being used to load a Web page, primarily for load-balancing reasons. Web pages consisted of more embedded objects than before. These objects were smaller than previously reported, mainly due to widespread use of icons and banner advertisements. There was an increase in the number of requests for large objects like Web email attachments, which caused an increase in large HTTP responses.

The advent of Web 2.0 applications has changed HTTP traffic patterns [23]. Web 2.0 applications are enabled by technologies such as Asynchronous JavaScript and XML (AJAX). Schneider *et al.* [60] studied the differences between conventional HTTP traffic patterns and Web 2.0 traffic. They collected full packet traces from networks in Germany and United States between 2005 and 2007, and all HTTP instances were extracted for four Web 2.0 applications. They found that the conventional HTTP characteristics significantly differed from AJAX-based Web 2.0 traffic. They reported that Web 2.0 applications consumed more bandwidth, made more requests, and actively pre-fetched data. These characteristics translated into more aggressive and bursty network usage compared to the overall HTTP traffic.

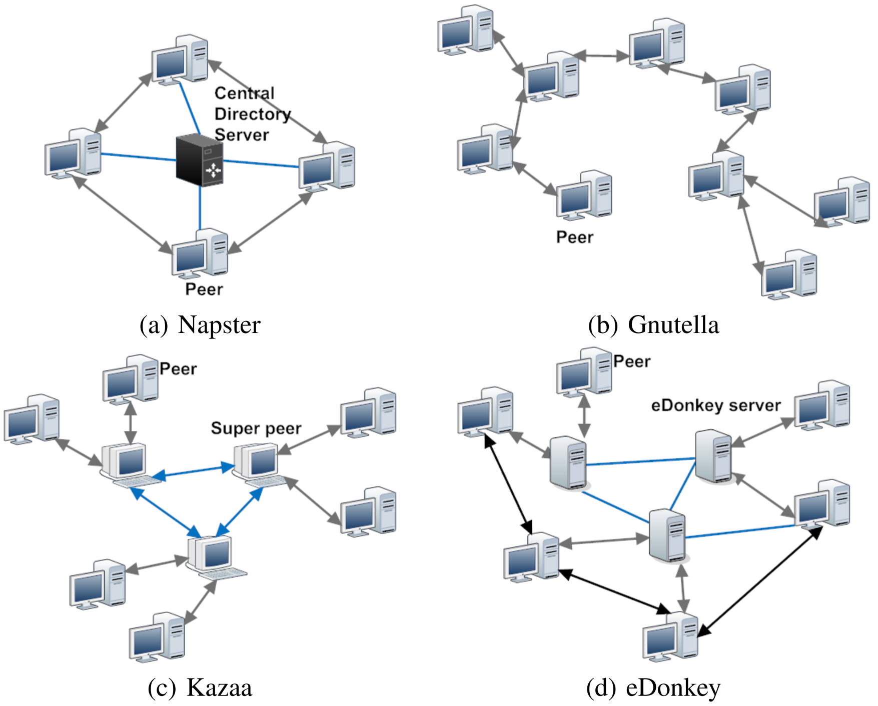
C. Proxy Traces

Gill *et al.* [29] analyzed HTTP traffic from a large enterprise (using proxy access logs), and from a large university (using packet capture at an edge router) to identify and characterize Web-based service usage. The two traces were collected simultaneously from the two network environments in 2008. The authors identified similarities and differences in Web traffic to those observed in previous studies. They remarked that while Web traffic had evolved, much of its underlying characteristics did not change significantly. The main changes observed were prevalent use of the HTTP `POST` method, increased use of scripts, greater number of large files, and popularity of new video formats. These differences could be explained by the contemporary changes in the Web usage. Web 2.0 encouraged greater user interaction, there was more video content posted, and there were new video formats such as Flash to enable better functionality of Web sites.

The browsing behavior of over 70,000 global users was studied by Ihm *et al.* [35]. They analyzed five years of Web traffic from a globally-distributed proxy system between 2006 and 2010. They found that analytics sites used AJAX and JavaScript heavily to track users. These analytics firms had access to a large fraction of Web users, with some firms being able to track up to 65% of the users. There were regional differences in client bandwidth, browser popularity, and dominant content types, which had implications on design and deployment of systems. They found that almost half of the generated traffic was due to client-side interactions after the Web page had fully loaded. Although the Web pages had become more complex in the size and number of embedded objects, the page loading latency decreased. This decrease was attributed to increased use of simultaneous connections in browsers, and improved caching.

II. Peer-to-peer Traffic Characterization

P2P networks can be categorized based on their architectures. A centralized P2P network like Napster and DirectCon-



(a) Napster

(b) Gnutella

(c) Kazaa

(d) eDonkey

Fig. 1. P2P architectures

nect relied on a central file directory server to enable connection between peers. The central server made searching for files faster, however, the network was vulnerable because of a single point of failure (see Figure 1(a)). The next generation of P2P systems (e.g., Gnutella) eliminated the central server, and were completely decentralized. Peers connected to each other in an ad-hoc manner and formed an overlay network (see Figure 1(b)). File search was based on flooding requests with a time to live limitation, which made it inefficient.

Further improvements were made to P2P systems by introducing super peers in a two-tiered overlay network. In the first-tier, nodes connected to the super peers, while in the second-tier, super peers connected with each other. Super peers were well-provisioned hosts that made searching and sharing files more efficient. Examples of such systems were FastTrack and Kazaa. Figure 1(c) illustrates the two-tier architecture of the Kazaa network. The eDonkey network had a hybrid architecture, with a layer of central servers that maintained a list of files, and another layer of peers for sharing files. Figure 1(d) illustrates the interactions in an eDonkey network.

BitTorrent is currently the most popular P2P protocol. The BitTorrent network consists of three components, namely peers, discovery mechanisms, and BitTorrent index sites [68]. The BitTorrent network is a decentralized system, where end users exchange files among themselves without the use of any centralized servers. End users participating in the system are called peers. Peers that exchange a specific content among themselves form a swarm. Peers that have a complete copy of a given file are called seeders. Peers with incomplete copies are called leechers. Several BitTorrent clients exist such as μTorrent[1] and Vuze[2].

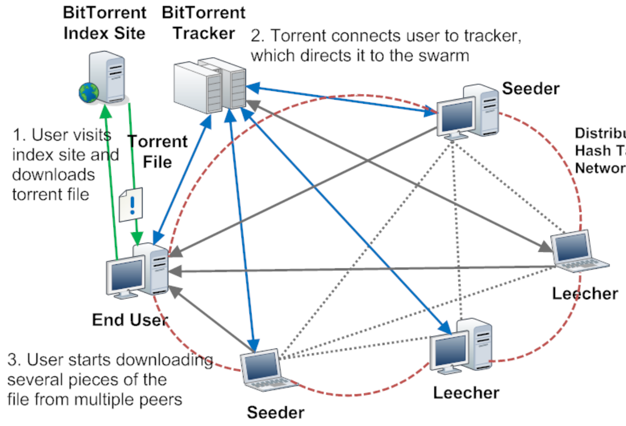


Fig. 2. Steps involved in file transfer on the BitTorrent network

Figure 2 illustrates the steps involved in file transfer on the BitTorrent network. An end user interested in downloading content from the network first visits a BitTorrent index site (e.g., Pirate Bay[3]). The index site hosts torrent files that contain important information about the content and from where it can be downloaded. The torrent file also contains the address of a tracker that assists in the communication among peers. The BitTorrent client calculates a hashcode using the torrent file and connects to a tracker that provides a list of peers with copies of the content. The end user can connect to these peers to download pieces of the content.

The centralized nature of tracker servers makes the BitTorrent network vulnerable to attacks, closure, or downtime. Many BitTorrent clients support distributed trackers through the use of distributed hash tables and peer exchange. In

[1] http://www.utorrent.com/

[2] http://www.vuze.com/

[3] http://thepiratebay.se/

a distributed hash table network, peers exchange peer lists directly with each other instead of a tracker [26]. Peer exchange is a gossiping mechanism in which peers leverage mutual knowledge of their respective peers to discover more peers who have a copy of the content.

Magnet links are another advancement in the BitTorrent protocol. These links do not require a tracker; rather, they directly connect with peers using the distributed hash table network. The magnet link contains the hash code for the torrent and does not require downloading a torrent file to initiate a file transfer. Pirate Bay recently switched entirely to magnet links [27].

The BitTorrent protocol implemented fair-sharing, due to the prevalence of free-riding in earlier P2P networks. BitTorrent networks are more file-focused, and they use swarms for sharing a file. Older generation P2P systems were based on a broad network of hosts who shared all files stored in their local publicly-available directory.

Researchers have analyzed traffic from various P2P systems such as Gnutella [37], [58], [59], [63], Kazaa [31], eDonkey [66], and BitTorrent [13], [32], [36], [51], [52], [68].

A. Napster, Gnutella, and Kazaa

Saroiu *et al.* [58], [59] analyzed Napster and Gnutella traffic by crawling their networks in 2001. They found that Gnutella users achieved higher throughputs than Napster users due to the flooding-based search mechanism in the Gnutella protocol that discouraged peers with low-bandwidth connections from participating. A large fraction of Gnutella clients had high-latency, owing to Gnutella's decentralized structure. On average, Napster peers participated more frequently in the network than Gnutella peers because of extra features (e.g., chat clients and MP3 player) bundled with the Napster client, which made users stay connected longer. There were more free-riders in Gnutella compared to Napster. Free-riders were generally low-bandwidth users who selected high-bandwidth peers for downloading content.

Using packet traces collected in 2001 from ISP border routers, Sen *et al.* [63] analyzed DirectConnect, Gnutella, and FastTrack systems. The study focused on network topology properties, traffic characteristics, and churn in the P2P networks. Significant increase in traffic volume and number of users were observed over the trace duration. Bandwidth consumption, peer connectivity, on-time, and mean throughput distributions were extremely skewed, but did not follow any power-law distributions. They also measured the P2P system characteristics at the network-prefix, Autonomous System (AS), and host-level. At the prefix and AS levels, bandwidth consumption and churn were more stable. To alleviate the churn at the host-level, local caching nodes were recommended.

Kazaa traffic was analyzed by Gummadi *et al.* [31]. In 2002, they collected a 200-day trace of Kazaa traffic at a large campus network. Kazaa users were found to be patient because requests for and transfers of large objects were completed over a long time. Older Kazaa clients made fewer requests than newer clients. The activity period of Kazaa users followed a heavy-tailed distribution. Most requests were made for small objects, but much of the byte traffic was due to requests for large objects. Objects downloaded by Kazaa clients obeyed the fetch-at-most-once principle, and object popularity was generally short-lived. Object popularity did not follow a power-law distribution.

B. eDonkey

Tutschku *et al.* [66] analyzed eDonkey traffic using flow-level traces collected from a campus network in 2003. Flows were classified into download and non-download flows (e.g., signaling flows) based on eDonkey protocol opcodes. On average, download flows were orders of magnitude larger than non-download flows. Due to file segmentation and flow concurrency, download flow sizes were limited in size, reducing eDonkey's ability to introduce small and large flows into the network. These flow sizes were well-modeled by a Lognormal distribution. Most of the flows came from local hosts instead of globally distributed hosts.

C. BitTorrent

Apart from characterizing BitTorrent traffic, recent works have focused on peer-assisted file hosting [39], file availability in BitTorrent [44], content publishing in BitTorrent [24], content popularity in BitTorrent [15], and private torrent trackers [19], [20], [33]. We discuss some of these works next.

In 2003, Guo *et al.* [32] collected data from a tracker site to analyze and model BitTorrent systems. The data had information on peer statistics, swarm information, and file size properties. The authors found that after the birth of a swarm, the number of requests from peers to the swarm decreased exponentially over time, quickly leading to reduced availability. Swarm size distribution was skewed, with most swarms being very small, and there was a limited number of large swarms. Peers in large swarms attained higher and more stable download rates than peers in smaller swarms. Download failures were less common in large swarms. Peers that achieved high download rates uploaded less often to other peers.

One of the earliest characterization studies of BitTorrent traffic was done by Pouwelse *et al.* [51]. They deployed a two-step data collection process. In the first step, they downloaded torrent files from torrent site mirrors, and parsed the torrent files to analyze the status of the trackers. In the next step, they tracked peers who downloaded torrent files from the mirrors. The data was collected between 2003 and 2004. The download rate was highly dynamic due to frequent torrent server, tracker, and mirror site failures. Content integrity was high in the system, with fake and corrupted files injected by the authors being quickly filtered by moderators of the torrent sites. The high content integrity was attributed to centralization, however, it conflicted with robustness of the system since decentralization can reduce failure rates.

The BitTorrent protocol has undergone several improvements. Neglia *et al.* [44] analyzed two such methods to improve BitTorrent availability, namely multiple trackers and distributed hash tables. They analyzed data from thousands of trackers and distributed hash table nodes over two months in 2006. Multiple trackers were found to improve availability. This improvement was mainly due to the choice of a single highly available tracker, however, the improvement was reduced by the presence of correlated failures. Multiple trackers reduced the connectivity in the peer overlay network.

Distributed hash tables led to greater file availability, but replies to peer queries took a longer time.

An extensive characterization study of the BitTorrent ecosystem was performed by Zhang *et al.* [68]. The data was collected by crawling several large torrent portals between 2008 and 2009. A multi-tracker crawler simultaneously crawled thousands of trackers with concurrent connections, and obtained peer lists of millions of torrents. They found that the ecosystem exhibited great diversity in terms of the operation of the major torrent index sites, user upload behavior, numbers of torrents and peers tracked by trackers, content type, and client implementations. Pirate Bay was a principal player in the ecosystem. Popularity of BitTorrent content depended on its age.

Cuevas *et al.* [24] studied content publishing activity on BitTorrent using data collected from large public torrent index sites between 2009 and 2010. A small fraction of users published and downloaded most of the content. Entertainment industry agencies carried out systematic attacks to deter downloading of copyrighted content. Malicious entities also utilized poisoning index attacks for propagating malware. About 30% of the content and 25% of the downloads were due to such attacks. Publishers were found to have financial incentives for posting content on BitTorrent portals. The authors remarked that attrition of financially-motivated publishers could significantly affect the popularity of the portals as well as the BitTorrent ecosystem.

Carlsson *et al.* [15] analyzed the differences in BitTorrent download characteristics and popularity dynamics from two vantage points: locally at a campus network, and globally using a torrent discovery site. Between 2008 and 2009, passive measurements captured all non-encrypted peer-to-tracker communication on the campus network, while simultaneous active measurements were performed to collect information on files from the site. They found that campus users typically downloaded larger files such as movies and TV shows. These files were downloaded before these files peaked in popularity globally. Music files were an exception, in that the campus users downloaded them after they had peaked in popularity at the global level. There was high churn in the weekly popularity of files both locally and globally.

III. Video Sharing Workload Characterization

Video sharing sites enable easy hosting, searching, and distribution of user-generated videos. YouTube[4] is the most popular video sharing site. Other examples are DailyMotion[5] and Metacafe[6]. Large video sharing sites rely on Content Delivery Network (CDN) nodes to deliver popular content to their users, while other content is provided from their own servers. Since YouTube's acquisition by Google, YouTube videos are delivered through a network of Google data centers [30], [42].

Figure 3 illustrates the steps involved in the delivery of a video requested by an end user [65]. In the first step, the user finds the video of interest by searching for it on the YouTube homepage or directly through the video link. The frontend server returns the video information using a plugin and name of the video server. The content server name is resolved using Domain Name System (DNS), and the end user queries the content server to get the video. Recent research has shown that the selection of the content server is based on round trip times between the user and the data centers [65]. YouTube also widely deploys cache servers across different locations in the world as well as inside ISPs [2].

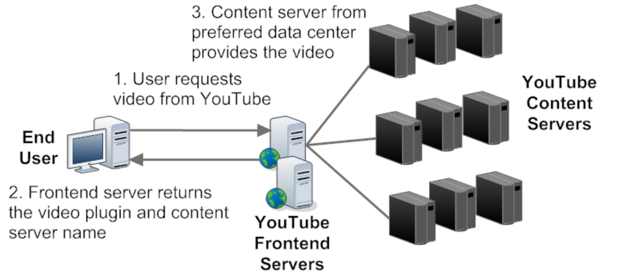

3. Content server from preferred data center provides the video
1. User requests video from YouTube
YouTube Content Servers
End User
2. Frontend server returns the video plugin and content server name
YouTube Frontend Servers

Fig. 3. Steps involved in video download from YouTube

The popularity of video sharing sites has led to several research works characterizing usage [16], [17], [21], [30], [69], popularity of content [14], [18], [43], uploading behavior [25], and content delivery infrastructure [1], [65]. Most of the work in this area is focused on YouTube, mainly due to its immense popularity and large content delivery infrastructure.

A. Video Traffic at Edge Network

Gill *et al.* [30] examined usage patterns, file properties, popularity and referencing characteristics, and transfer behavior of YouTube videos. The study utilized HTTP traces collected from a large campus network in 2007. They observed that a small number of YouTube video requests accounted for most of the traffic volume from the site. Video file sizes were orders of magnitude larger than files of other content types. Video access patterns were strongly correlated to human behavior, with traffic volumes varying significantly by time-of-day and day-of-week. These and other characteristics led the authors to suggest that caching could improve the performance and scalability of Web 2.0. The diversity in available content reduced the concentration of references in accesses, which would abate the efficacy of caching and prefetching strategies.

Zink *et al.* [69] analyzed how content distribution in YouTube was realized through a measurement study of YouTube traffic in a large university campus network. Based on these measurements, they studied the duration and the data rate of streaming sessions, the popularity of videos, and video access patterns from the campus users. The analysis of the traffic showed that trace statistics were relatively stable over short-term periods while long-term trends could be observed. They used simulations to study the benefits of alternative distribution infrastructures to improve the performance of YouTube-like video sharing services. They found that P2P-based distribution and proxy caching could reduce network traffic significantly and provide faster access to video clips.

B. Video File Popularity and Uploading Behavior

Mitra *et al.* [43] analyzed workload data from four video sharing services (Dailymotion, Yahoo! video[7], Veoh[8], and

[4] http://www.youtube.com/

[5] http://www.dailymotion.com/

[6] http://www.metacafe.com/

[7] http://video.yahoo.com/

[8] http://www.veoh.com/

Metacafe). Their traces were collected in 2008 by crawling the sites, and contained metadata on millions of videos. They identified invariants related to video popularity distribution, Web 2.0 feature usage, and the uploading of new content. The number of uploaders to a service was an order of magnitude smaller than the number of uploaded videos, and several orders of magnitude smaller than the number of views to these videos. There were significant differences across the services. For example, while the number of video uploads by users followed the Pareto principle, the fraction of repeat uploaders for Veoh was twice that of Yahoo!. They also considered implications for system design based on the identified invariants. Lifetime video popularity measures were useful in case of large cache sizes. However, this benefit decreased as cache size shrunk, owing to video popularity dynamics.

The scale and uploading behavior of YouTube users was studied by Ding *et al.* [25]. They used a depth-first-search of the related uploader graph to collect data on tens of millions of users in 2010. They found that these users uploaded over 400 million videos (totaling 2,600 years in video length), and were viewed over a trillion times. Uploading behavior followed the Pareto principle. The top 20% of the most active uploaders attracted most of the views. Users belonging to a social network were more active and uploaded more videos than others.

C. Video Sharing Infrastructure

YouTube infrastructure (before acquisition by Google) was studied by Adhikari *et al.* [1]. Using flow-level data collected at multiple points of presence of a large backbone ISP, they located YouTube data centers and their connections to ISPs. They investigated how load balancing strategies and routing policies affected the traffic dynamics. They found that user location did not play a role when serving video content. Instead, YouTube employed proportional load balancing among its data centers to service user requests from all locations.

Torres *et al.* [65] analyzed YouTube's content delivery mechanism using traces collected from five edge networks. They found that the YouTube infrastructure had been redesigned compared to the one previously analyzed in the literature [1]. Most YouTube requests were directed to a preferred data center, usually based on the round-trip delay between users and YouTube data centers. There were, however, many instances when videos were served from non-preferred data centers. Such non-preferred server accesses were due to load balancing issues, variations across DNS servers within a network, and availability of unpopular video content in a given data center, among other things.

IV. File Hosting Traffic Characterization

File hosting sites offer a simple Web-based solution for hosting files that can be accessed conveniently using a URL. After a file is uploaded to the site, a URL is generated by the site to access this file. These sites offer two levels of service - free and premium. The free service has limitations on the number of downloads and the maximum throughput achieved for the download. Premium users have to pay a subscription fee, and the access restrictions are removed for such users. A free user has to go through a series of steps before the download can begin. Most often the user has to wait for a pre-determined amount of time before the link is clickable. Premium users do not have to wait for their download to start.

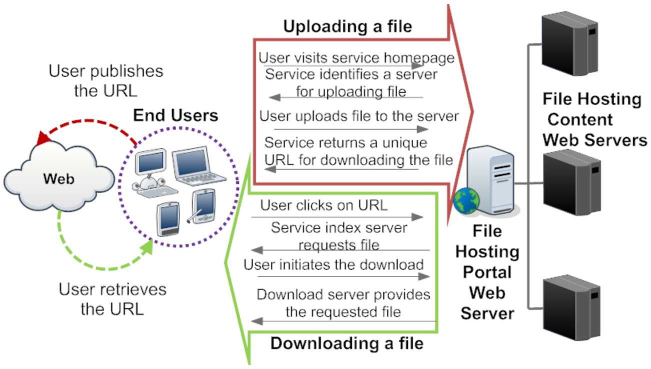

Uploading a file
User visits service homepage
Service identifies a server for uploading file
User uploads file to the server
Service returns a unique URL for downloading the file
User publishes the URL
End Users
Web
File Hosting Content Web Servers
User clicks on URL
Service index server requests file
User initiates the download
Download server provides the requested file
Downloading a file
User retrieves the URL
File Hosting Portal Web Server

Fig. 4. Steps involved in content transfer on file hosting services

All sites impose limitations on the maximum size of the uploaded file, regardless of the user type. However, a user can split large media content into smaller parts and upload them separately. Consumers who download these parts can use an archiving program to join the parts and obtain the original content.

Figure 4 shows a simplified illustration of uploading and downloading a file using a generic file hosting site. A user uploads a file to the file hosting server through a simple Web interface. The site provides the user with a unique URL for the file. The user may publish this link on the Web for other users to download the file using the link. When a user visits the download URL, the file hosting site requests the file from the server where the file is stored.

There has been limited work on characterizing the file hosting ecosystem. Two works have performed measurement and analysis of RapidShare [5] and RapidShare/Megaupload traffic [57]. Nikiforakis *et al.* [45] analyzed the privacy features of file hosting services. We also discuss a recent study by Kim *et al.* [38] on modern Usenet newsgroup traffic characterization. Usenet is a collection of distributed servers and clients used for posting and reading information. It uses the Network News Transfer Protocol (NNTP) for transporting data. Usenet existed long before the advent of Web and P2P, and has been used for exchanging files.

A. RapidShare and Megaupload Traffic

RapidShare service architecture, usage patterns, and content characteristics were studied by Antoniades *et al.* [5]. The analysis was based on up to 9-month long traces collected from two academic networks between 2008 and 2009. They also used active measurements to compare RapidShare with BitTorrent in terms of user-perceived throughput and content availability. They found that most RapidShare files on the academic networks were requested only once. Only a few files were requested more than five times. Upon exploration of selected index sites, they found that most of RapidShare's content was uploaded by few users. Using targeted experiments with limited files they found that free RapidShare users achieved similar throughput as BitTorrent users, while premium RapidShare users obtained an order of magnitude higher download rate. Their experiments also showed that RapidShare evenly distributed content across storage servers, with the most recent file uploads being directed to new servers.

A parallel work by Cuxart *et al.* [57] analyzed RapidShare and Megaupload traffic using traces collected from a research network over three months in 2009. 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. They remarked that these services relied on a huge server infrastructure, and a non-negligible percentage of users paid for premium accounts[9]. They mentioned that users could benefit from using download managers that exploited TCP's congestion control mechanism, and by parallelizing downloads.

B. *File Hosting Privacy*

Nikiforakis *et al.* [45] investigated the privacy features of several file hosting services. Their results showed that a large fraction of the services generated download links in a deterministic fashion. These file hosting services were vulnerable because their file links could be generated using sequential enumeration or easily guessed by malicious entities. Using different enumerators, they crawled hundreds of thousands of unique files. They used the Bing search engine to determine whether the file was public or private. A little more than half of the crawled files were labeled private, highlighting the apparent lack of privacy for uploaders of the hosted content. They devised an experiment to show that attackers were aware of these vulnerabilities and were leveraging them for malicious purposes. To alleviate the problem, they proposed a client-side protection mechanism to protect user files when uploaded to insecure file hosting services.

C. *Usenet Traffic*

Kim *et al.* [38] analyzed Usenet newsgroup traffic using traces collected from a residential Digital Subscriber Line (DSL) service provider between 2008 and 2009. They observed that most of the traffic volume was due to binary data transmissions. Transaction sizes were bimodal. Much of the traffic was exchanged with fee-based commercial servers. The achieved throughput of NNTP connections was an order of magnitude higher than P2P systems like BitTorrent and eDonkey. They also noted the similarities in content (most of them were archives) and user-perceived performance with that of file hosting services.

V. Industry Traffic Reports

Over the past few years, many Internet traffic management vendors have issued white papers on the global Internet traffic trends. These reports present statistics on Internet application usage from various geographic vantage points, and help in understanding the evolving nature of Internet traffic. We present a succinct summary of these reports with regards to usage and bandwidth share of Web applications, P2P filesharing protocols, and file hosting services. This discussion is organized based on the type of network studied.

A. *Internet Service Providers*

Ipoque analyzed Internet traffic collected from ISPs and universities in Australia, Eastern Europe, Germany, the Middle East, and Southern Europe between August and September 2007 [61]. Internet applications were identified using deep packet inspection and behavior analysis. They found that P2P produced, on average, 49-83% of all Internet traffic with evening peaks of over 95% (about 20% of P2P traffic was encrypted.). They included 62 file hosting services in the study, and found these services were responsible for about 9% of the Internet traffic (17% of users) in the Middle East and over 4% in Germany (9% of users). RapidShare was the most popular service, responsible for 55% of the total file hosting traffic. They made an interesting observation for P2P, where they found about 20% of the Internet users were responsible for up to 70% of the overall bandwidth consumption. In contrast, they noted that file hosting service users did not disproportionately produce more traffic. This was due to the smaller average size of files hosted by these services; many of them were picture files.

Ipoque followed up with another Internet study in 2009 where they analyzed traffic from eight regions of the world (the five regions studied in the previous report, plus Southern Africa, South America, and Southern Europe) [62]. They found that P2P still generated the most traffic in all monitored regions, ranging from 43% in Northern Africa to 70% in Eastern Europe. The regional differences were attributed to varying subscriber access bandwidth, availability of localized content, and cultural habits. The proportion of P2P traffic had decreased in comparison to 2007, while Web traffic had increased, primarily due to file hosting services. BitTorrent was the most popular protocol followed by eDonkey. File hosting traffic had increased to 45% of all Web traffic. Flash-based streaming accounted for up to 80% of all streaming traffic, indicating the popularity of video sharing sites such as YouTube.

The *Sandvine* 2008 Global Broadband Phenomena report [53] presented traffic analysis results collected from 26 networks distributed across 18 countries between July and September 2008. They found that P2P file sharing still dominated upstream networks, and accounted for more than 61% of all upstream traffic. Web dominated in the downstream traffic at 43%, while P2P accounted for 22% of the total traffic. They observed that file hosting services, newsgroup, and tunneling services accounted for 13% of the downstream traffic.

The Sandvine 2009 report [54] was based on 20 cable and DSL networks from North America, Europe, Caribbean and Latin America, Asia-Pacific, and Africa. The data was gathered for a period of 22 days in September 2009. They found that file hosting services were transferring 56% more traffic volume per user compared to 2008. Globally, P2P traffic declined to 20% of total Internet traffic volume. However, the decline was not consistent in every region: North America experienced a 20% relative decline, while the Caribbean and Latin America region experienced an increase of more than 30%. They also found that RapidShare was primarily being used for data acquisition (limited upstream traffic), and was generally not popular with average broadband subscribers. The downstream traffic for heavy and average YouTube users was found to be similar. There was a significant variation in the upstream traffic; heavy users had a small fraction of the

[9] They identified premium users using cookies.

upstream YouTube traffic compared to average users. This discovery was shown as evidence that average users were more likely to be interactive (via video uploads or comments) than heavy users.

The Sandvine 2010 report [55] found that real-time entertainment services such as Netflix[10] were the largest contributor to data consumption on both fixed (43% of peak period traffic) and mobile access (41%) networks in North America. In Europe, Web traffic accounted for almost 45% of the total volume during the peak evening hours. BitTorrent was the dominant P2P protocol and represented almost 30% of upstream peak period traffic and slightly more than 8% of downstream peak period traffic. zSHARE displaced Megaupload and RapidShare as the top file hosting service in Europe and accounted for 3% of downstream traffic during the peak period.

The Sandvine 2011 report [56] found that the four largest Internet services in North America, by daily downstream traffic volume, were Netflix (27.6%), HTTP (17.8%), YouTube (10.0%), and BitTorrent (9.0%). Real-time entertainment applications were the primary drivers of network capacity requirements, accounting for 60% of peak downstream traffic. Cisco had made similar findings in its 2010 report [22]. Furthermore, usage was becoming increasingly concentrated in the evening hours, driving up network costs despite relatively constant per-user monthly data consumption. Megaupload was among the top-10 applications used in Brazil and Africa. zSHARE was the most popular file hosting service in Eastern Europe.

B. Mobile Networks

The *Allot* 2009 report [3] presented application and bandwidth usage and growth in mobile broadband networks, with the data collected between July and December 2009. They found that YouTube was responsible for 10% of global mobile data bandwidth. HTTP streaming applications were the fastest growing application with a 99% increase, while file hosting traffic grew by 73%. File hosting traffic accounted for 16-21% of the total traffic volume depending on the region.

In a followup report [4], Allot reported that YouTube traffic had increased to 13% of total mobile broadband traffic in 2010. Social media services traffic grew over 200%. File hosting traffic growth was about 40%. While streaming applications were dominant (35% of traffic volume), P2P traffic and file hosting traffic consumed almost the same amount of bandwidth at about 15% of total traffic volume each.

C. Enterprise Networks

Palo Alto Networks published several reports on application usage in enterprises around the globe. In the Fall 2009 report [46], they found that file hosting services were being used more often than P2P applications, although bandwidth consumption by file hosting services was 18% of what P2P applications consumed. The Spring 2010 paper [47] reported that 87% of the users visited a file hosting site as compared to 77% of users who used P2P applications. They noted the high number of file hosting variants and the bandwidth consumed, and showed that some of these services were essential for business needs.

The December 2011 Palo Alto Networks report [49] stated that over the last 3 years the number of file hosting services being used in enterprises had more than tripled, growing from 22 to 71 variants. They divided these services into two groups, namely, productivity and entertainment. The services in the former were geared towards work-related activities, while the latter were mostly for personal enjoyment. The entertainment category included services such as Megaupload and MediaFire. Megaupload was found in 57% of the participating organizations, however, it consumed the highest amount of file hosting bandwidth, indicating that the shared files were large. P2P traffic consumed 2% of the total bandwidth, while file hosting traffic consumed 1% of the volume, slightly lower than social networking applications.

D. University Campus Networks

Palo Alto Networks [48] analyzed 326 university networks between 2009 and 2011. They found a diverse mix of social networking, entertainment, and education applications in use. P2P file sharing accounted for 22% of the total bandwidth consumed, while file hosting services only accounted for 4% of the total traffic volume. In total, 34 P2P applications were in use with the top-3 applications accounting for 19% of the bandwidth. The most common P2P variant was BitTorrent, which was noticed in 90% of the university networks. There was increased popularity of file hosting services: 58 variants were noticed, up from 24 in 2009. They also noted that file sharing was much higher (by a factor of three) on university networks compared to enterprise networks. While productivity file hosting services were very frequently used, entertainment file hosting services consumed the most bandwidth. They also suggested that file hosting services are yet to supplant P2P file sharing, since P2P is still the most prevalent choice among university users.

In summary, these industry reports suggest that Internet traffic is constantly changing. The reports cover various types of networks, multiple demographics, and multiple geographic regions. We find that in many regions P2P is no longer the dominant application in terms of bytes transferred. HTTP is the most dominant protocol mainly because of video streaming and various services offered on top of HTTP. The growth in HTTP traffic has also been attributed to file hosting services. For file hosting services, dominance is ephemeral. A file hosting service can be popular one year, the next year it could be a different file hosting service. We also observe a growing number of file hosting service variants each year. We also notice differences in usage across geographic regions and network types. Regions where there is easier access to paid content (e.g., Netflix, Hulu), P2P usage is lower. In university networks, users preferred BitTorrent over file hosting services because BitTorrent is free. In North America, we notice greater bandwidth being consumed in the evening despite the monthly bandwidth consumption remaining unchanged.

References

[1] V. Adhikari, S. Jain, and Z. Zhang. YouTube Traffic Dynamics and its Interplay with a Tier-1 ISP: An ISP Perspective. In *Proc. ACM SIGCOMM Conference on Internet measurement*, Melbourne, Australia, November 2010.

[10] http://www.netflix.com/

[2] V. Adhikari, S. Jain, and Z. Zhang. Where Do You "Tube"? Uncovering YouTube Server Selection Strategy. In *Proc. International Conference on Computer Communication Networks*, Maui, USA, July/August 2011.

[3] Allot. Allot MobileTrends: Global Mobile Broadband Traffic Report. White paper, Allot Communications, 2009. http://www.allot.com/mobiletrends.html.

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

[5] 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.

[6] M. Arlitt. Characterizing Web User Sessions. *ACM SIGMETRICS Performance Evaluation Review*, 28(2):50–63, September 2000.

[7] M. Arlitt, R. Friedrich, and T. Jin. Workload Characterization of a Web Proxy in a Cable Modem Environment. *ACM SIGMETRICS Performance Evaluation Review*, 27(2):25–36, September 1999.

[8] M. Arlitt and T. Jin. A Workload Characterization Study of the 1998 World Cup Web site. *Network, IEEE*, 14(3):30 –37, May/June 2000.

[9] M. Arlitt and C. Williamson. Web Server Workload Characterization: The Search for Invariants. In *Proc. ACM SIGMETRICS Conference*, Philadelphia, USA, May 1996.

[10] M. Arlitt and C. Williamson. Internet Web Servers: Workload Characterization and Performance Implications. *IEEE/ACM Transactions on Networking*, 5(5), October 1997.

[11] P. Barford, A. Bestavros, A. Bradley, and M. Crovella. Changes in Web Client Access Patterns: Characteristics and Caching Implications. *World Wide Web*, 2(1-2):15–28, January 1999.

[12] L. Bent, M. Rabinovich, G. Voelker, and Z. Xiao. Characterization of a Large Web Site Population with Implications for Content Delivery. *World Wide Web*, 9(4):505–536, December 2006.

[13] A. Bharambe, C. Herley, and V. Padmanabhan. Analyzing and Improving a BitTorrent Networks Performance Mechanisms. In *Proc. IEEE INFOCOM Conference*, Barcelona, Spain, April 2006.

[14] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti. Characterizing and Modelling Popularity of User-generated Videos. *Performance Evaluation*, 68(11):1037–1055, November 2011.

[15] 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.

[16] 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 Conference on Internet Measurement*, San Diego, USA, October 2007.

[17] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. Analyzing the Video Popularity Characteristics of Large-scale User Generated Content Systems. *IEEE/ACM Transactions on Networking*, 17(5):1357–1370, October 2009.

[18] G. Chatzopoulou, C. Sheng, and M. Faloutsos. A First Step Towards Understanding Popularity in YouTube. In *Proc. IEEE INFOCOM Conference on Computer Communications Workshops*, San Diego, USA, March 2010.

[19] X. Chen, X. Chu, and Z. Li. Improving Sustainability of Private P2P Communities. In *Proc. International Conference on Computer Communications and Networks*, Maui, USA, August 2011.

[20] X. Chen, Y. Jiang, and X. Chu. Measurements, Analysis and Modeling of Private Trackers. In *Proc. IEEE P2P Conference*, Delft, Netherlands, August 2010.

[21] X. Cheng, C. Dale, and J. Liu. Statistics and Social Network of YouTube Videos. In *Proc. International Workshop on Quality of Service*, Enschede, Netherlands, June 2008.

[22] Cisco. Cisco Visual Networking Index: Usage. White paper, Cisco Systems, October 2010. http://tinyurl.com/CiscoNetworks2010.

[23] G. Cormode and B. Krishnamurthy. Key Differences between Web 1.0 and Web 2.0. *First Monday*, 13(6), 2008.

[24] R. Cuevas, M. Kryczka, A. Cuevas, S. Kaune, C. Guerrero, and R. Rejaie. Is Content Publishing in BitTorrent Altruistic or Profit-driven? In *Proc. International Conference on Emerging Networking Experiments and Technologies*, Philadelphia, USA, November/December 2010.

[25] Y. Ding, Y. Du, Y. Hu, Z. Liu, L. Wang, K. Ross, and A. Ghose. Broadcast Yourself: Understanding YouTube Uploaders. In *Proc. ACM SIGCOMM Conference on Internet Measurement Conference*, Berlin, Germany, November 2011.

[26] Ernesto. BitTorrent's Future? DHT, PEX and Magnet Links Explained. TorrentFreak, November 2009. http://tinyurl.com/DHT-PEX-MagnetLinksExplained.

[27] Ernesto. The Pirate Bay, Now Without Torrents. TorrentFreak, February 2012. http://torrentfreak.com/the-pirate-bay-dumps-torrents-120228/.

[28] A. Faber, M. Gupta, and C. Viecco. Revisiting Web Server Workload Invariants in the Context of Scientific Web Sites. In *Proc. ACM/IEEE Conference on Supercomputing*, Tampa, USA, November 2006.

[29] 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*, 5(4):19:1–19:23, October 2011.

[30] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube Traffic Characterization: A View from the Edge. In *Proc. ACM SIGCOMM Conference on Internet Measurement*, San Diego, USA, October 2007.

[31] 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*, Bolton Landing, USA, October 2003.

[32] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang. Measurements, Analysis, and Modeling of BitTorrent-like Systems. In *Proc. ACM SIGCOMM Internet Measurement Conference*, Berkeley, USA, October 2005.

[33] D. Hales, R. Rahman, B. Zhang, M. Meulpolder, and J. Pouwelse. BitTorrent or BitCrunch: Evidence of a Credit Squeeze in BitTorrent? In *Proc. IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises*, Toulouse, France, June 2009.

[34] F. Hernandez-Campos, K. Jeffay, and F. Smith. Tracking the Evolution of Web Traffic: 1995-2003. In *Proc. IEEE MASCOTS Symposium*, Orlando, USA, October 2003.

[35] S. Ihm and V. S. Pai. Towards Understanding Modern Web Traffic. In *Proc. ACM SIGCOMM Conference on Internet Measurement Conference*, Berlin, Germany, November 2011.

[36] M. Izal, G. Uroy-Keller, E. Biersack, P. Felber, A. Al-Hamra, and L. Garces-Erice. Dissecting BitTorrent: Five Months in Torrent's Lifetime. In *Proc. Passive and Active Measurement Conference*, Antibes Juan-les-Pins, France, April 2004.

[37] T. Karagiannis. Filesharing in the Internet: A Characterization of P2P Traffic in the Backbone. Technical report, University of California, Riverside, November 2003. http://research.microsoft.com/apps/pubs/default.aspx?id=71488.

[38] J. Kim, F. Schneider, B. Ager, and A. Feldmann. Today's Usenet Usage: NNTP Traffic Characterization. In *Proc. IEEE INFOCOM Workshops*, San Diego, USA, March 2010.

[39] F. Liu, Y. Sun, B. Li, B. Li, and X. Zhang. FS2You: Peer-Assisted Semipersistent Online Hosting at a Large Scale. *IEEE Transactions on Parallel and Distributed Systems*, 21(10):1442–1457, October 2010.

[40] A. Mahanti, D. Eager, and C. Williamson. Temporal Locality and its Impact on Web Proxy Cache Performance. *Performance Evaluation*, 42(23):187 – 203, October 2000.

[41] A. Mahanti, C. Williamson, and L. Wu. Workload Characterization of a Large Systems Conference Web Server. In *Proc. Annual Communication Networks and Services Research Conference*, Moncton, Canada, May 2009.

[42] R. Miller. Google-YouTube: Bad News for Limelight? Data Center Knowledge, October 2006. http://tinyurl.com/YouTube-Limelight.

[43] S. Mitra, M. Agrawal, A. Yadav, N. Carlsson, D. Eager, and A. Mahanti. Characterizing Web-Based Video Sharing Workloads. *ACM Transactions on Web*, 5(2):8:1–8:27, May 2011.

[44] G. Neglia, G. Reina, H. Zhang, D. Towsley, A. Venkataramani, and J. Danaher. Availability in BitTorrent Systems. In *Proc. IEEE INFOCOM Conference*, Anchorage, USA, May 2007.

[45] N. Nikiforakis, M. Balduzzi, S. Van Acker, W. Joosen, and D. Balzarotti. Exposing the Lack of Privacy in File Hosting Services. In *Proc. USENIX Conference on Large-scale Exploits and Emergent Threats*, Boston, USA, March/April 2011.

[46] Palo Alto Networks. The Application Usage and Risk Report: An Analysis of End User Application Trends in the Enterprise. White paper, Palo Alto Networks, Fall 2009. http://www.paloaltonetworks.com/literature/whitepapers/.

[47] Palo Alto Networks. The Application Usage and Risk Report: An Analysis of End User Application Trends in the Enterprise. White paper, Palo Alto Networks, Spring 2010. http://www.paloaltonetworks.com/literature/whitepapers/.

[48] Palo Alto Networks. Academic Freedom or Application Chaos: An Analysis of End-User Application Traffic on University Network. White paper, Palo Alto Networks, March 2011. http://www.paloaltonetworks.com/literature/whitepapers/.

[49] Palo Alto Networks. The Application Usage and Risk Report: An Analysis of End User Application Trends in the Enterprise. White paper, Palo Alto Networks, December 2011. http://www.paloaltonetworks.com/literature/whitepapers/.

[50] J. Pitkow. Summary of WWW Characterizations. *World Wide Web*, 2(1-2):3–13, January 1999.

[51] J. Pouwelse, P. Garbacki, D. Epema, and H. Sips. The Bittorrent P2P File-Sharing System: Measurements and Analysis. In *Proc. International Workshop on Peer-to-Peer Systems*, Ithaca, USA, February 2005.

[52] D. Qiu and R. Srikant. Modeling and Performance Analysis of BitTorrent-like Peer-to-Peer Networks. In *Proc. ACM SIGCOMM Conference*, Portland, USA, August/September 2004.

[53] Sandvine. 2008 Global Broadband Phenomena. White paper, Sandvine Incorporated, October 2008. http://tinyurl.com/Sandvine2008.

[54] Sandvine. 2009 Global Broadband Phenomena. White paper, Sandvine Incorporated, January 2010. http://tinyurl.com/Sandvine2009.

[55] Sandvine. Fall 2010 Global Internet Phenomena Report. White paper, Sandvine Incorporated, October 2010. http://tinyurl.com/Sandvine2010.

[56] Sandvine. Fall 2011 Global Internet Phenomena Report. White paper, Sandvine Incorporated, September 2011. http://tinyurl.com/Sandvine2011.

[57] 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*, pages 1–26, May 2011.

[58] S. Saroiu, K. Gummadi, and S. Gribble. A Measurement Study of Peer-to-Peer File Sharing Systems. In *Proc. Multimedia Computing and Networking Symposium*, San Jose, USA, Jaunary 2002.

[59] S. Saroiu, K. Gummadi, and S. Gribble. Measuring and Analyzing the Characteristics of Napster and Gnutella Hosts. *Multimedia Systems*, 9(2):170–184, August 2003.

[60] F. Schneider, S. Agarwal, T. Alpcan, and A. Feldmann. The New Web: Characterizing AJAX Traffic. In *Proc. Conference on Passive and Active Network Measurement*, Cleveland, USA, April 2008.

[61] H. Schulze and K. Mochalski. Internet Study 2007. White paper, Ipoque Gmbh, 2007. http://tinyurl.com/Ipoque2007.

[62] H. Schulze and K. Mochalski. Internet Study 2008/09. White paper, Ipoque Gmbh, 2009. http://tinyurl.com/Ipoque2008-09.

[63] S. Sen and J. Wang. Analyzing Peer-to-Peer Traffic across Large Networks. *IEEE/ACM Transactions on Networking*, 12(2):219–232, April 2004.

[64] F. Smith, F. Campos, K. Jeffay, and D. Ott. What TCP/IP Protocol Headers can tell us about the Web. In *Proc. ACM SIGMETRICS Conference*, Cambridge, USA, June 2001.

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

[66] K. Tutschku. A Measurement-Based Traffic Profile of the eDonkey Filesharing Service. In *Proc. Passive and Active Network Measurement Conference*, Antibes Juan-les-Pins, France, April 2004.

[67] A. Williams, M. Arlitt, C. Williamson, and K. Barker. Web Workload Characterization: Ten Years Later. In X. Tang, J. Xu, S. T. Chanson, and Y. Zhang, editors, *Web Content Delivery*, volume 2 of *Web Information Systems Engineering and Internet Technologies*, pages 3–21. Springer, 2005.

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

[69] 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, March 2009.