Measurement and Analysis of Cyberlocker Services

Aniket Mahanti
University of Calgary
Calgary, AB, Canada
mahantia@ucalgary.ca

ABSTRACT

Cyberlocker Services (CLS) such as RapidShare and Megaupload have recently become popular. The decline of Peer-to-Peer (P2P) file sharing has prompted various services including CLS to replace it. We propose a comprehensive multi-level characterization of the CLS ecosystem. We answer three research questions: (a) what is a suitable measurement infrastructure for gathering CLS workloads; (b) what are the characteristics of the CLS ecosystem; and (c) what are the implications of CLS on Web 2.0 (and the Internet). To the best of our knowledge, this work is the first to characterize the CLS ecosystem. The work will highlight the content, usage, performance, infrastructure, quality of service, and evolution characteristics of CLS.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Measurement Techniques

General Terms

Measurement, Performance

1. INTRODUCTION

User-generated content has transformed the way people disseminate and share information over the Web. Today, an ordinary user has the ability to create and publish content, and both the user and the user-generated data are key in so-called "Web 2.0" applications [6]. Recently, the Web has witnessed the emergence of Cyberlocker Services (CLS), also known as One Click Hosting services. These services were originally designed for file backup purposes and for uploading files that were too big to be sent as email attachments. Some of the well-known CLS are RapidShare, Megaupload, and MediaFire[1]. CLS allow their users to easily upload a file to their servers. Once the file has been successfully uploaded, the site generates a unique URL that can be used for downloading the file. The user may then make the link public for sharing content.

Over the years, the method for sharing content has evolved. Almost a decade ago, Napster was a popular application for sharing music files among users. This application was the precursor to decentralized Peer-to-Peer (P2P) file sharing applications such as KaZaA and BitTorrent. P2P file sharing was not restricted to music files, but included all sorts of content. The popularity of P2P file sharing surged and according to estimates was responsible for up to 60% of the total Internet traffic in some regions during its peak [23]. Recently, the popularity of Web 2.0 applications has caused an increase in Web traffic, and P2P file sharing traffic appears to be on the decline [12, 23, 24].

CLS differ from traditional P2P file sharing and other new-age content sharing services. Many social media sites are restricted to sharing video files, while entertainment sites such as `Hulu.com` place geographic restrictions on its viewing audience. In contrast, CLS allow users to upload any file. CLS offer differentiated forms of service for their users. CLS offer several advantages over P2P technologies such as greater availability of active files, improved privacy for users, hosting both popular and niche content, and economic incentive mechanisms for frequent uploaders [1].

CLS have recently received attention from networking researchers. Maier *et al.* [15] found that CLS account for about 16% of total HTTP traffic in a large residential network. Ipoque suggests that CLS account for almost 10% of the total Internet traffic, while [9] contemplates that some CLS contribute to more Internet traffic volume than Facebook. Labovitz *et al.* [12] report a decline in P2P traffic, but growth in traffic for CLS. Sandvine [20] found that CLS traffic volume share has increased by over 40% in 2008-09. They also report that a CLS is among the top-10 applications (downstream bytes) in Europe during peak hours [21].

Figure 1 shows the number of unique U.S. visitors to two CLS (RapidShare, Megaupload) and a well-known P2P filesharing indexing site (`Mininova`) (according to `Compete.com`). We observe a proliferation in CLS users, while a subsiding pattern in P2P file sharing usage[2]. The apparent decline of P2P file sharing points to a paradigm shift in how users share content [18]. Despite the wide adoption of CLS, not much is known about their infrastructure, content characteristics, and user-perceived performance.

We will study the following research questions:

- ***What is a suitable measurement infrastructure for gathering CLS workloads?*** A measurement framework for data collection from multiple viewpoints [3] will be designed. Undertaking a large measurement study in-

[1]RapidShare, Megaupload, and MediaFire are listed in `Alexa.com`'s global top 100 Web site rankings.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

[2]As a result of a court ruling `Mininova` removed most of its indexed torrents during November, 2009. `Mininova` now only hosts torrents from artists and producers who want to distribute their content for free.

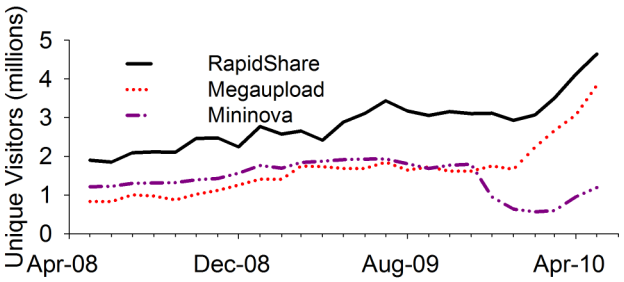

Figure 1: Number of unique visitors

volves several technical challenges. Efficient data collection strategies and processing schemes will be analyzed and applied. A distinguishing feature of the work is the use of active and passive measurements from multiple observational viewpoints.

- ***What are the characteristics of the CLS ecosystem?*** A comprehensive multi-layered characterization of the CLS ecosystem will be performed. We will study and compare the characteristics of user behaviour, content popularity, content delivery, content dissemination, infrastructure, and performance of several CLS.
- ***What are the implications of CLS on Web 2.0 (and the Internet)?*** The implications of the CLS workload analysis on the performance of Web 2.0 will be analyzed. We will study the longitudinal evolution of CLS and its impact on content sharing. We will also analyze what are the implications of CLS on caching, copyright issues, and contrast with other content distribution mechanisms.

The rest of the paper is organized as follows. §2 provides background information on CLS. §3 discusses related work. §4 states the objectives of the proposed research. §5 describes the methodology for CLS trace collection and analysis, and the datasets used. §6 presents some preliminary results from our analysis. §7 concludes the paper.

2. CYBERLOCKER SERVICES

CLS 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 these 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.

All sites impose limitations on the size of the uploaded file regardless of the user type. However, a user can split a large media content into smaller parts and upload them separately. On the other end, users who download these parts can use an archiving program to join the parts and obtain the original content. Figure 2 shows a simplified illustration of uploading and downloading of a file using a generic Cyberlocker site.

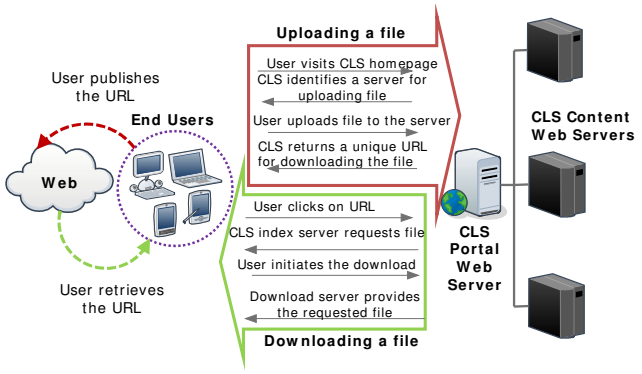

Figure 2: CLS file upload and download

3. RELATED WORK

Previous research on Web workload characterization has focussed on proxy caches [2, 14], AJAX traffic [22], video sharing sites [5, 8, 26], and online social networks [16]. P2P systems have been studied in the context of file sharing workloads [10, 11, 19, 25], incentive schemes [13], and content availability [17]. As CLS are a relatively new phenomena, there has been limited work on understanding the usage of these services, their infrastructure, and the workloads they experience. The sole example we are aware of is [1].

Antoniades *et al.* [1] studied RapidShare usage and delivery infrastructure, as measured through a number of independent and targeted experiments. In contrast to Antoniades *et al.*, we analyze and characterize the overall usage of CLS as observed from multiple observation viewpoints and take a closer look at the dominant services observed. This allows us to compare and contrast the properties observed for different CLS. We also use longer duration workload traces for our analysis. We collect data-rich HTTP traces that allow us to study how the clients identify and select the content they download. Furthermore, we utilize user clickstreams[3] to distinguish free and premium CLS users. This has not been previously characterized, and provides a deeper understanding of the usage of these services, as well as the dynamics of new-age content sharing and distribution. Understanding this linkage can provide important insights into today's content sharing trends. This discussion is further augmented by a comparison with the usage of P2P systems and video sharing sites.

4. RESEARCH GOALS

Design multi-level measurement framework: We apply a multi-level measurement approach for collection of the datasets. First, we study the activity of the CLS ecosystem from a local viewpoint using HTTP transaction summaries. The HTTP data provides detailed information about various features of CLS transaction summaries, which is not possible with any other data. Second, we analyze the activity of the CLS flows at the network-layer using connection summaries. Finally, to gain a global perspective of the CLS ecosystem, we use crawling techniques and gather data from third-party Web analytics sources. We constructed a

[3]A clickstream is the sequence of user clicks or HTTP transactions while browsing a Web site.

measurement framework that allows us to assay the CLS ecosystem from multiple viewpoints. Our experiences will aid researchers in improving their measurement framework when performing workload characterization.

Workload characterization: We are presently conducting a comprehensive analysis of the content being distributed on CLS. We are analyzing what content is being hosted on CLS, how it is being hosted, what is its size, which methods are used for uploading and downloading the content, how the original content is fragmented, where are links to the content being indexed, what is the dissemination rate of the content, whether content is being replicated inside and across individual CLS, and if some content is more popular than others. We perform this analysis both at the local and global levels. When possible, we will contrast the results to similar systems such as P2P systems and Web-based video sharing services. Since not much is known about the CLS ecosystem, this analysis will help us understand the implications of increased CLS use in the future. The results may also assist CLS developers in designing better content distribution and incentive schemes.

Quality of service characterization: We are doing a comprehensive analysis of quality of service measures in CLS. We are studying the availability of hosted files, service differences between free and premium users, throughput rates of free and premium file downloads, wait times for downloads, success rate of file downloads, background traffic associated with actual content downloads, time of day and day of week effects, and downloading of files using multiple concurrent connections. We try to understand how different CLS systems guarantee quality of service and how they differ from other systems. This analysis will help us understand the functionality of CLS and allow CLS operators to further improve their service guarantees.

Network infrastructure characterization: We will study infrastructure and network-layer characteristics of CLS. For infrastructure characterization, we will focus on host behaviour of CLS servers, load balancing schemes, power-law behaviour from the client and server ends, and ISP and geographic distribution of CLS servers. For network-layer characterization, we will focus on flow-level properties such as flow size, flow duration, and flow inter-arrival, and host-level properties such as transfer volume, host on-time, and flow concurrency. We will use these metrics and compare them with Web and P2P flows. This analysis will highlight details about the infrastructure of CLS (e.g., how files are hosted on CLS data centres and how they are retrieved) and present methods for distinguishing traffic flows from CLS.

Implications of CLS on Web 2.0 and the Internet: We will study the growth and evolution of these CLS over time. We will try to understand how usage and user behaviour evolves for these services. The consequences of CLS use such as ease of sharing (compared to other similar services), traffic demands, caching, and copyright issues will also be studied. This analysis will help us understand how content might be distributed in the future and how it could impact edge networks.

5. METHODOLOGY AND DATASETS

Figure 3 illustrates our eight-step methodology for measurement and analysis of the CLS ecosystem. Step 1 involves passive monitoring of campus traffic. Step 2 uses active measurements via crawling. In Step 3, we collect Web analytics

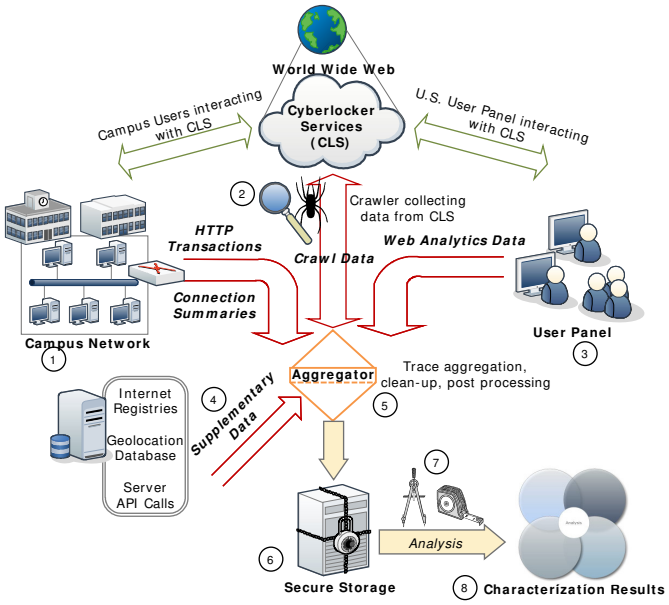

Figure 3: CLS measurement/analysis methodology

data followed by collection of supplementary data in Step 4. Data aggregation is performed in Step 5 and the data is securely stored during Step 6. We analyze the data in Step 7 and finally apply the results in Step 8. In total, four datsets - 2 local and 2 global - were collected.

Local data: We collected two datasets of the local CLS usage. The primary dataset used is a trace of HTTP transactions (later referred to as HTTP trace) collected over a one-year period (Jan-Dec 2009) from our university's 400 Mbps full-duplex link to the Internet. The data contains application-layer information such as HTTP headers (e.g., HTTP method, status code, Host header, etc.) and transport-layer information (e.g., transfer duration, bytes transferred, etc.). The HTTP traces were produced by a Bro[4] Network Intrusion Detection System script that summarized HTTP transactions (request-response pairs) on the university's Internet link in real time. User identifiable information such as IP addresses and cookies are not stored. This method allows for greater privacy for users, however, it does not allow us to perform some long-term analysis of user characteristics. We aggregated the HTTP transactions of interest from the data. These transactions were identified based on the HTTP `Host:` header field.

We also collected flow-level data (connection summaries) coinciding with the HTTP data collection period to augment the analysis. We used Bro to collect these summaries. Each connection summary contains information such as the source and destination IP addresses and port numbers, and the number of bytes transferred in each direction. We extracted the relevant CLS flows from the dataset using the IP addresses of the Cyberlocker sites in the HTTP data.

Global data: The first global dataset was collected by crawling a CLS search engine. This search engine indexes over 100 million publicly available CLS links. The search engine offered an API that allowed us to crawl several CLS files. Since CLS do not allow their hosted files to be searched,

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

the CLS search engine can only index files that are available in the public domain. The following information about the crawled files were collected: name, size, extension, URL, tags, ratings, and date added. The crawl was performed during March and July, 2010.

The second dataset comprised of CLS analytic statistics from `Compete.com`. `Compete` provides information collected from 2 million Internet users in the United States (1% of the total U.S. Internet population). This dataset includes normalized data from the entire U.S. Internet population containing information such as user count, page visits, page views, user attention and stay period, the share of Internet population reached by these services, and content referrers. This dataset allows us to understand the usage pattern of these services on a larger scale and contrast the statistics with that of the analysis done on the local-level data.

Supplementary data: Some additional data for RapidShare files were also collected using its API. The API allowed us to know the status of a file and its server location (`server id`). Internet Registries and Geolocation databases were queried to gather information about ISP, AS, organization, and geographic locality of CLS.

6. PRELIMINARY RESULTS

We present preliminary results from analysis of the one-year long HTTP trace. We analyzed over 500 GB of compressed HTTP logs. These logs contained over 5 billion HTTP transactions representing over 60 TB of Web traffic. We identified over 13 million transactions attributable to CLS, which represented about 4% of the total Web traffic volume. Note that this is a conservative estimate since the traces contained several (large) transactions with incomplete byte counts. This was caused because the network monitor was unable to cope with the network load sometimes and missed some packets [8]. We focus on the top five CLS (generating over 60% of the CLS traffic volume) in the campus network. These CLS were RapidShare (RS), Megaupload (MU), Hotfile (HF), MediaFire (MF), and zSHARE (ZS).

Usage: Figure 4 shows the number of users per day for the top five CLS. The majority of the traffic volume of RapidShare and Megaupload is due to premium downloads. As a newly established service, Hotfile shows growth in its usage. Megaupload usage is stationary, while RapidShare shows some decline. We found that users preferred to download during weekends. They also had a proclivity towards using similar user agents for the downloads. We observe that some services have significant number of premium users, while others are mostly dependant on free users.

File size: Figure 5 shows the distribution of the size of the files downloaded. In the RapidShare curve we observe two steep increases at the 100 MB and 200 MB mark. These values represent the old and new file size limits, respectively. The new limit has been in effect since 2009, however, it appears that most of the files downloaded followed the older limit. About 1% of the files were larger than 200 MB, indicating that these files were uploaded by premium users and intended for premium downloaders. We also notice a steep increase at the 100 MB mark for Megaupload. Megaupload has a much higher upload size limit than 100 MB, however, the many files that are 100 MB in size indicate that uploaders tend to use a convenient file size that is common across several CLS. Megaupload also has a few files that are larger than 1 GB, which are specifically for premium users. All

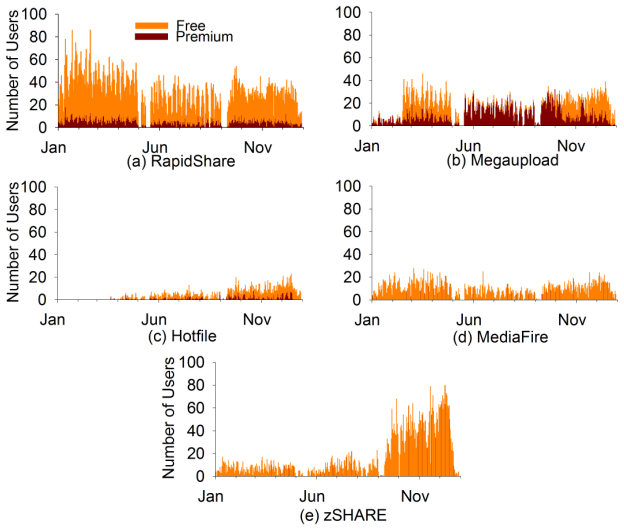

Figure 4: Number of CLS users

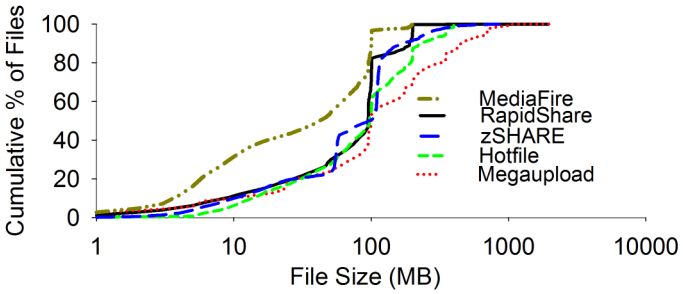

Figure 5: CDF of file size

files in Hotfile were less than 400 MB, which is the upload size limit for free users. MediaFire has a much smaller median file size than all other services indicating it is mostly used for uploading smaller files.

Although CLS are offered on the Web platform, their file size characteristics are different from traditional Web workloads. In the traditional Web, most of the transferred objects are small (85% of the objects are smaller than 10 KB) [7]. When considering a file size distribution of all Web traffic, CLS file downloads are concentrated in the tail of the distribution. The smaller files contributed by CLS traffic are due to images, text, and scripts associated with the Web site. Earlier studies [4] have shown that Web traffic introduced many mice flows, but only a few elephant flows. Increasing use of CLS can change this mix of mice and elephant flows. CLS appear to replicate the P2P phenomena on the Web (i.e., many mice flows and many elephant flows).

Content type: Table 1 shows the byte count distribution of various CLS content types. In terms of content count, majority of the content downloaded from RapidShare were Archives. Furthermore, this content category represents a larger share (86%) of the total RapidShare traffic volume. Video accounted for 22% in content count, while it was responsible for 12% of the RapidShare traffic volume. Megaupload content was dominated by Video both in terms of count (57%) and traffic volume (63%). About 67% of the Hotfile content downloaded were Archives. About 34% of

Table 1: Content type (%byte count)

Category	RS	MU	HF	MF	ZS
Archives	86.3	33.9	63.8	46.8	5.7
Video	11.9	62.7	29.0	40.5	91.3
Documents	0.4	1.5	0.5	0.3	0.1
Application	0.2	0.2	0.0	0.1	0.0
Audio	0.3	0.4	2.2	4.6	2.8
Others	0.9	1.3	4.5	7.8	0.1

MediaFire content was audio and it accounted for 5% of the traffic volume owing to small size of MP3 files. The majority of the zSHARE content was flash video; it allows its users to stream the video content instead of downloading it.

The prevalence of archived files is not surprising. It provides an easy method to split large media files and upload to the CLS. Additionally, archiving allows the user to provide a password for the content. Users can then share the password with their intended audience only.

7. CONCLUSIONS

We proposed a comprehensive multi-level characterization of the CLS ecosystem. We presented a methodology for collecting CLS traces from multiple viewpoints. We also presented results from preliminary analysis of a year-long HTTP trace collected from a large edge network. The results highlighted the similarities and differences in the usage and content characteristics of CLS. The results of this research will help in better understanding the evolution of the new Web, provisioning future ISP networks, and designing better content delivery systems.

8. ACKNOWLEDGEMENTS

The author thanks Carey Williamson, Niklas Carlsson, and Martin Arlitt for their guidance during this research. Financial support for the research is provided by Alberta Innovates Technology Futures and Natural Sciences and Engineering Research Council of Canada.

9. REFERENCES

[1] D. Antoniades, E. Markatos, and C. Dovrolis. One-click Hosting Services: A File-sharing Hideout. In *IMC*, 2009.

[2] M. Arlitt, R. Friedrich, and T. Jin. Workload Characterization of a Web Proxy in a Cable Modem Environment. *Performance Evaluation Review*, 1999.

[3] M. Arlitt and C. Williamson. Understanding Web Server Configuration Issues. *Software Practice and Experience*, 2004.

[4] N. Basher, A. Mahanti, A. Mahanti, C. Williamson, and M. Arlitt. A Comparative Analysis of Web and Peer-to-Peer Traffic. In *WWW*, 2008.

[5] 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 *IMC*, 2007.

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

[7] P. Gill, M. Arlitt, N. Carlsson, A. Mahanti, and C. Williamson. Organizational Use of Web-based Services. Technical report, University of Calgary, 2009.

[8] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube Traffic Characterization: A View from the Edge. In *IMC*, 2007.

[9] A. Greenberg. The "Mega" Sites: Bigger Than Facebook (Forbes Magazine). `http://tinyurl.com/ForbesMag-Article`, Nov 2009.

[10] 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 *SOSP*, 2003.

[11] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang. Measurement, Analysis, and Modeling of BitTorrent-like Systems. In *IMC*, 2005.

[12] C. Labovitz, S. Johnson, D. McPherson, J. Oberheide, and F. Jahanian. Internet Inter-domain Traff. In *SIGCOMM*, 2010.

[13] A. Legout, N. Liogkas, E. Kohler, and L. Zhang. Clustering and Sharing Incentives in BitTorrent Sys. In *SIGMETRICS*, 2007.

[14] A. Mahanti, C. Williamson, and D. Eager. Traffic Analysis of a Web Proxy Caching Hierarchy. *IEEE Network*, 2000.

[15] G. Maier, A. Feldmann, V. Paxson, and M. Allman. On Dominant Characteristics of Residential Broadband Internet Traffic. In *IMC*, 2009.

[16] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *IMC*, 2007.

[17] G. Neglia, G. Reina, H. Zhang, D. Towsley, A. Venkataramani, and J. Danaher. Availability in BitTorrent Systems. In *INFOCOM*, 2007.

[18] I. Poese, B. Frank, B. Ager, G. Smaragdakis, and A. Feldmann. Improving Content Delivery Using Provider-aided Distance Information. In *IMC*, 2010.

[19] J. Pouwelse, P. Garbacki, D. Epema, and H. Sips. The Bittorrent P2P File-sharing System. In *IPTPS*, 2005.

[20] Sandvine. 2009 Global Broadband Phenomena. Technical report, Oct. 2009.

[21] Sandvine. Fall 2010 Global Broadband Phenomena. Technical report, Oct. 2010.

[22] F. Schneider, S. Agarwal, T. Alpcan, and A. Feldmann. The New Web: Characterizing AJAX Traffic. In *Proc. PAM*, 2008.

[23] H. Schulze and K. Mochalski. IPOQUE Internet Studies (2006, 2007, 2008/2009). Technical report, 2009.

[24] R. Singel. Peer-to-Peer Passe, Report Finds (Wired Magazine). `http://tinyurl.com/P2PDying`, Oct 2009.

[25] C. Zhang, P. Dhungel, D. Wu, and K. W. Ross. Unraveling the BitTorrent Ecosystem. *IEEE TPDS*, 2010.

[26] M. Zink, K. Suh, Y. Gu, and J. Kurose. Characteristics of YouTube Network Traffic at a Campus network - Measurements, Models, and Implications. *Comp. Netw.*, 2009.