2011-01-13

Fresh Analysis of Streaming Media Stored on the Web

Rabin Karki
Worcester Polytechnic Institute

Fresh Analysis of Streaming Media Stored on the Web

by

Rabin Karki

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Computer Science

by

Jan 2011

APPROVED:

Professor Mark Claypool, Advisor

Associate Professor Emmanuel Agu, Thesis Reader

Associate Professor Michael A. Gennert, Head of Department

Abstract

With the steady increase in the bandwidth available to end users and Web sites hosting user generated content, there appears to be more multimedia content on the Web than ever before. Studies to quantify media stored on the Web done in 1997 and 2003 are now dated since the nature, size and number of streaming media objects on the Web have changed considerably. Although there have been more recent studies characterizing specific streaming media sites like YouTube, there are only a few studies that focus on characterizing the media stored on the Web as a whole. We build customized tools to crawl the Web, identify streaming media content and extract the characteristics of the streaming media found. We choose 16 different starting points and crawled 1.25 million Web pages from each starting point. Using the custom built tools, the media objects are identified and analyzed to determine attributes including media type, media length, codecs used for encoding, encoded bitrate, resolution, and aspect ratio. A little over half the media clips we encountered are video. MP3 and AAC are the most prevalent audio codecs whereas H.264 and FLV are the most common video codecs. The median size and encoded bitrates of stored media have increased since the last study. Information on the characteristics of stored multimedia and their trends over time can help system designers. The results can also be useful for empirical Internet measurements studies that attempt to mimic the behavior of streaming media traffic over the Internet.

Acknowledgments

I owe my deepest gratitude to Professor Mark Claypool. He has been a tireless and patient mentor, and colleague not only throughout the work of this thesis but also during my time at WPI. For everything, from giving personal and professional advice, helping me transition smoothly in the new environment to adapting to my odd meeting schedules, thank you.

I would also like to thank everyone at PEDS and CC research groups. Their comments, suggestions and insights have helped me a lot in shaping this thesis, and my knowledge. Also thanks to Professor Emmanuel Agu for going through the thesis and providing suggestions and comments.

The entire process has been the experience of a lifetime for me. I am sure I will remember and apply the things learnt from this experience for a long time to come.

Contents

1 Introduction **1**

2 Related work **5**

3 Methodology and design **9**

3.1 Starting points 9

3.2 Crawling and gathering data 11

3.3 Tools and extrication of media characteristics 12

4 Analysis **15**

4.1 Summary analysis 16

4.2 Audio 21

4.3 Video 24

4.4 Comparison with previous studies 31

4.5 Sampling issues 33

5 Conclusion **36**

6 Future work **39**

A Appendix **41**

List of Figures

4.1 Overlap percentage 16

4.2 URLs per domain name 16

4.3 Media URL counts per starting point 20

4.4 Last modified dates for media URLs according to the servers 21

4.5 Audio codecs 22

4.6 Audio/video lengths 23

4.7 Audio/video file sizes 24

4.8 Encoded audio bitrates 25

4.9 Video codecs 25

4.10 Media file sizes (FLV vs Non-FLV) 27

4.11 Encoded video bitrates 28

4.12 Video resolutions 29

4.13 Video aspect ratios 29

4.14 Number of formats in a YouTube URL 30

4.15 Overlap percentages 31

4.16 CCDF of video durations for different sample set sizes 34

4.17 CDF of video bitrates for different sample set sizes 35

List of Tables

4.1 Crawler starting points 18
4.2 Overlap percentages 18
4.3 Top 15 domains 19
4.4 Media clips analyzed 20
4.5 Top ccTLDs encountered in media URLs 21
4.6 Audio channels 22
4.7 Comparison of results 32

Chapter 1

Introduction

Internet access for the world population has continued to grow rapidly[1], as has broadband access[2], enabling growth in Web 2.0 Web sites and user contributed content. User generated content sharing in sites like YouTube, along with growing processing power and faster internet connections mean that the multimedia content on the Web is more accessible than ever before. Videos which are served from a single, central administration like news, sports and entertainment sites, whether it is free or paid access, are also contributing to the overall multimedia content available on the Web.

Multimedia streams require higher data rates and consume significantly more bandwidth than traditional Web objects. Traffic generated by multimedia tends to be bursty [9] and is more sensitive to delay. Streaming media also requires significantly more storage, thereby increasing the storage requirements of media servers and proxy caches. Multimedia streams traffic typically take longer to play than does downloading traditional Web objects. Therefore streaming media, as

[1]World Internet Usage Statistics News and World Population Stats. By Miniwatts Marketing group. http://www.internetworldstats.com/stats.htm

[2]Broadband subscribers per 100 inhabitants.
http://www.oecd.org/document/60/0,3343,en_2649_34225_39574076_1_1_1_1,00.html

compared to traditional Web workloads, present a number of new challenges to system designers. Information on the characteristics of stored multimedia can help system designers to minimize the workload of streaming media on the Internet by proper capacity planning of content delivery infrastructures. Such information can also be useful when selecting representative streaming media clips for empirical Internet measurement studies that attempt to mimic the behavior of commercial streaming media traffic over the Internet.

Unfortunately, there is little recent published work on specific characteristics of streaming media stored on the Web. While there have been studies to measure and analyze streaming media at the client side [5, 7] and at specific sites like YouTube [7, 4], there have been no recent studies of the general attributes of streaming media clips stored on Web servers. Studies done by Acharya and Smith [1] and Li et al. [8], with data gathered in 1997 and 2003 respectively, are dated given the fast changing nature of media today. In 1997, Acharya and Smith [1] studied video content stored on the Web by analyzing every video available via then popular Alta Vista search engine. They found that the Internet could not support real-time streaming given the encoded bitrates and last-mile connection capacities avaliable in 1997. In 2003, Li et al. found that 29% of the videos are encoded for modem bitrates [8]. We hypothesize that the percentage is significantly lower today, i.e. encoding rates are higher. They also found that nearly half of the video have resolutions 320x240 or less. We hypothesize that the number of videos having resolutions larger than this has increased. Similarly, the study showed online multimedia was dominated by three formats: Real Network's RealMedia, Microsoft's Windows Media and Apple's QuickTime. However in the last few years, Adobe (formerly Macromedia) Flash has quickly risen to become seemingly dominant format in online video. We hypothesize that newer codec types, including Flash video, for

both audio and video have emerged, replacing previous leaders. Moreover, recent data and analyses can enable longitudinal comparison of trend analysis across time.

While there is substantial audio and video content stored on peer-to-peer (p2p) file sharing systems [11], p2p content is typically not streamed. Typically p2p applications download the multimedia content completely before beginning playout. Thus, network traffic behavior for p2p file systems is similar to bulk file transfers rather than streaming media. As our study focuses on the characteristics of streaming media that is played out in real time, analysis of multimedia content stored on p2p systems is left outside the scope of our study.

For our work, Larbin[3] was customized to create a Web crawler and was launched from 16 different starting points, each instance crawling to at least 1.25 million URLs. Other tools were then built to identify whether a particular URL contained multimedia content. The content thus found was analyzed using tools described in Section 3.3 and media parameters were recorded that have been previously indicated as potentially impacting the perceived quality of media content streamed over the Web. Those parameters included size, encoding format, frame rates, resolution, duration and average encoded bit rate. in addition to this fresh analysis, the result of our study was compared longitudinally to study how characteristics have changed over time.

Results of our study show that most streaming media clips are relatively short and they have grown slighly longer since the last study done in 2003. The median file size of media clips has also increased. The majority of the audio clips are encoded at typical standard bitrates instead of being targeted for lower bandwidth connections. Resolution wise, the trend shows that videos having higher resolutions are increasing and high definition videos are also present on the Web today.

[3]Larbin: Multi-purpose Web crawler. http://larbin.sourceforge.net/index-eng.html

The rest of the thesis is organized as follows. Chapter 2 explores related work, and compares and contrasts it with our work. Chapter 3 discusses the methodology and design of our work. In this chapter, Section 3.1 presents the method used for selection of starting points for the study, Section 3.2 discusses the crawling process and how data was gathered, and Section 3.3 provides description of tools and process used for extracting characteristics from media content. Chapter 4 presents the detailed analysis of the data and results of our study. This chapter is divided in four sections. Section 4.1 discusses the summary of URLs gathered and statistical overview of data collected. Section 4.2 presents the information and attributes of the audio contents. Similarly, Section 4.3 provides the results of our study about video contents and Section 4.4 compares results our study with previous studies. Section 4.5 discusses sampling issues related to our study. Chapter 5 presents the conclusions of our study and Chapter 6 provides some possible future work extensions.

Chapter 2

Related work

In this chapter, we present work related to our study and discuss how they compare or contrast with our study. There have been a few studies in the past characterizing streaming media. Some other studies have been conducted to study streaming media workloads from the client and server perspectives. Also discussed are the tools that have been used in our study. Some tools were customized to meet our requirements and a tool called MediaProbe was created to probe the newer media formats.

There is little recent published work on specific characteristics of streaming media stored on the Web. More specifically, two studies similar to ours have been conducted previously in 1997 and 2003. In 1997, Acharya and Smith studied video content stored on the Web by analyzing every video available (over 57,000 AVI, QuickTime and MPEG files) via then popular Alta Vista search engine [1]. They found that the most common video technology in use at that time was QuickTime. They also found out that given the encoded bitrates of the videos, Internet bandwidth at that time was at least an order of magnitude too slow to support streaming playback of the video. However, the nature of streaming media has changed considerably since that time.

In 2003, Li et al. did an extensive study to characterize streaming media stored on the Web [8]. They crawled 17 million pages, yielding nearly 30,000 audio and video clips. By comparing the results with work in past studies, they found that the volume of streaming media stored on the Web has increased by more than 600% over the previous five years. They also showed that streaming audio and video content were dominated by properietary streaming products, specifically Real Media and Windows Media. This study is closest to our work, but while they found only 30,000 media clips, every minute 24 hours of video is uploaded to YouTube[1] alone today[2], which indicates the need for a fresh study. In Chapter 4, we compare the results of our study with those of Li et al. study, whenever possible, to discuss trends and changes.

However, there have been more recent studies characterizing the video content of specific sites, mostly YouTube. Cha et al. studied popularity distribution, popularity evolution and content duplication of user generated content videos [4]. Zink et al. gathered a large trace of YouTube traffic and investigated the caching problem [13]. Gill et al. presented a YouTube traffic analysis which tracked YouTube transactions in a campus network, and focused on deriving video access from the network edge perspective [7]. Duarte et al. showed that there is a relationship between geography and the social network features available in YouTube [6].

Other studies [5, 12] have analyzed streaming media workloads. At the client side, Chesire et al. [5] collected traces of RTSP client activity originating from a large organization, compared media workload charactestics to traditional Web-object workloads in terms of bandwidth utilization, server/object popularity and sharing patterns and looked at the effectiveness of performance optimizations on streaming-media workloads. At the server side, Veloso et al. [12] characterized live

[1] http//www.youtube.com/
[2] http://www.youtube.com/t/fact_sheet

streaming media content delivered over the Internet. They found that nature of interactions between users and objects is fundamentally different for live versus stored objects.

To individually access each of the media content objects in order to collect the audio and video clips information, Li et al. used customized tools built from commercial application SDKs, open source programs and custom built components. They used MPlayer[3] to analyze Apple QuickTime, MediaTracker[4] for recording video information of Windows Media and RealTracer[5] for providing information on Real Media. Although there are many new media formats which are not recognized by these tools anymore, we used these tools to the extent to which they were useful.

A whitepaper by Bergman [2] noted that a tremendous amount of content on the Web is dynamic, pages it called the "deep Web". It estimated that deep Web is 550 times larger than the surface Web. Raghavan et al. [10] classify the dynamism of Web pages in three types: temporal, client-based and input dynamism. They proposed a prototype for hidden Web crawling called HiWE. While our crawler methods do not target the hidden Web per se, we do take into account the fact that the Web today is too large and unreachable by a standard crawler in many cases for a comprehensive data collection as done by Acharya and Smith [1].

For our study, we required a crawler that would be able to crawl and gather data according to our needs. To the best of our knowledge, there are no crawlers built specifically for multimedia content. ht://Dig[6] is tailored for indexing and searching a domain or Intranet, which was not our intent. We did not deem Tubekit[7], which can crawl YouTube based on a set of seed queries, general enough for our use. We

[3]http://www.mplayerhq.hu
[4]http://perform.wpi.edu/real-tracer/#mediatracker
[5]http://perform.wpi.edu/real-tracer/#realtracer
[6]http://www.htdig.org/
[7]TubeKit - A YouTube crawling toolkit. http://www.tubekit.org/

had sought to use OpenWebSpider[8] in the early stages of our work, which is an open source Web spider ported in C# from Java, but due to stability issues and the speed of crawling, we decided against it. We modified Larbin[9] according to our requirements and used it for crawling.

[8]OpenWebSpider - The open source web spider and search engine. http://www.openwebspider.org/

[9]Larbin: Multi-purpose Web crawler. http://larbin.sourceforge.net/index-eng.html

Chapter 3

Methodology and design

The following methodology was used to collect data on the characteristics of streaming media stored on the Web.

- A strategy was devised for selecting the starting points for the crawler such that the data gathered would contain representative sample of stored media on the Web (see Section 3.1).
- Larbin[1] was customized to crawl the Web from specified starting points, and collect and store URLs (see Section 3.2).
- Methods were identified and tools were developed to identify media content and extract media characteristics by reading packet headers (see Section 3.3).

3.1 Starting points

While selecting starting points for the crawler, the strategy was to pick Web pages that are popular, being likely to be accessed by well-connected users so that the

[1]Larbin: Multi-purpose Web crawler. http://larbin.sourceforge.net/index-eng.html

data gathered would contain the representative sample of stored media on the Web. The starting points were chosen from six different countries (USA, UK, China, France, South Korea and Japan) to make the sample geographically diverse. The geographical separation was also intended to help reduce the overlaps in search space amont the individual crawls. Using the report by Nielsen[2] and About.com ratings[3], some of the most popular video sites were chosen so that we could sample from sites with videos being watched by users. To diversify the samples, some popular news and sports sites were also included as starting points.

Apart from videos, we believe that podcast services also contribute to the amount of multimedia content on the Web. The following popular podcast directories were chosen as additional starting points: Podcastdirecoty[4] and Podcast pickle[5]. For direct comparison with streaming media content found in [8], a few starting points used in that study were taken directly from that paper. The final list of starting points in alphabetical order is given in Table 4.1. In the table, the first column is the Web page address, the second column is the name of the page and the third column is the geographical location of that particular Web page.

Starting from these 16 distinct starting pages, each of the crawler instances gathered URLs until at least a 1.25 million URLs had been reached. Before starting the analysis, exactly 1.25 million URLs crawled were taken from each crawling instance. Although the crawler records unique URLs within a single crawl, the output from the different crawls may overlap and include the same URLs on multiple crawls. Chapter 4 discusses the amount of overlaps between different crawls.

[2]Nielsen - Ratings and rankings. Aug 2009 report.
http://en-us.nielsen.com/rankings/insights/rankings/internet

[3]About.com - top ten most popular video sites.
http://websearch.about.com/od/imagesearch/tp/popularvideosites.htm. Retrieved on October 2, 2009.

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

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

3.2 Crawling and gathering data

A crawler was used to traverse the Web and gather data for analysis. Writing an entirely new Web crawler was out of project scope and timeline. So instead it was decided to make use of the open source alternatives. Different options were considered for this purpose. For example, ht://Dig[6] is tailored for indexing and searching a domain or Intranet, which was not our intent. Tubekit[7], which can crawl YouTube based on a set of seed queries, was not deemed general enough for our use. OpenWebSpider[8] was proposed, which is an open source Web spider ported in C# from Java, but due to stability issues and the speed of crawling, ultimately not selected. Finally, Larbin[9] was modified to suit our requirements and used for the crawling. Starting from a specified starting point, the Larbin crawler recursively traverses the embedded URLs and stores them in files until it is stopped.

Larbin is an open source Web crawler available under GPL. It is intended to fetch a large number of Web pages very quickly. The program is multithreaded but prefers using select instead of a lot of threads (for efficiency purposes). The advantage of Larbin over other crawlers is that it is much faster when getting files over many sites because it opens a lot of connections at the same time and is easily customizable. However, Larbin does not index the pages that it fetches. It works in Linux and uses standard libraries and adns[10] which is provided with the distribution. For each run, Larbin can be configured to have multiple connections open at the same time, each one working independently for faster URL fetching. For our study, each run of Larbin crawl was configured to use 5 parallel connections. It uses combined

[6]http://www.htdig.org/

[7]TubeKit - A YouTube crawling toolkit. http://www.tubekit.org/

[8]OpenWebSpider - The open source web spider and search engine. http://www.openwebspider.org/

[9]Larbin: Multi-purpose Web crawler. http://larbin.sourceforge.net/index-eng.html

[10]http://www.chiark.greenend.org.uk/~ian/adns/

breadth-first and depth-first algorithm and during a single run, it maintains a data structure of previously crawled URLs to make sure that it does not crawl the same URL again.

Be default, Larbin logs URLs that begin with the prefix HTTP only. We customized Larbin to store the URLs that begin with other protocols too. This includes different multimedia protocols like RTSP, MMS and podcasts protocols like feed and ITPC.

3.3 Tools and extrication of media characteristics

Once crawling and gathering URLs was finished, the next step was to use specialized tools to go through all the URLs gathered and identify if they were streaming media URLs or whether they had streaming media embedded in their pages. URLs gathered from a general Web crawler cannot always give us a direct link to actual media files, as more and more media is played via embedded players (e.g. the YouTube video player). It was not possible to get to all the streaming media content stored on the Web as was done in a previous study [1] not only because of the sheer size of the Web today, but also because a large portion of multimedia content is dynamically generated [2], paid or private. It is relatively easy to detect media which uses specific streaming protocols like RTSP, MMS or objects with specific extensions like avi, mp3, wmv, asx, rm etc. However with embedded players, the URL alone does not identify streaming media content. Thus, we had to manually extract the direct link to media files for sites like like YouTube and Dailymotion[11].

To individually access each of the media content objects to collect the audio and video clips information, Li et al. used customized tools built from commercial

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

application SDKs, open source programs and custom built components. They used MPlayer[12] to analyze Apple QuickTime, MediaTracker[13] for gathering video information on Windows Media and RealTracer[14] for measuring the information on Real Media. Although there are many new media format which are not recognized by these tools anymore (e.g. Flash), we used these tools to the extent they were useful. We also built our own tool to analyze the newer media formats which were not recognized by these tools using FFprobe[15]. FFprobe is a simple multimedia streams analyzer with a command-line interface based on the FFmpeg[16] project libraries. The mime types, modified dates, filesizes etc. were obtained by sending the HTTP request to the media URLs served by HTTP servers and recording the response obtained from the servers.

We developed a tool called *MediaProbe* to get the headers from streaming media, extract the information contained in the headers and record them. MediaProbe works in multiple stages:

- URLs from the files, containing URLs list generated by Larbin are read into memory.
- If those URLs contain streaming media, URLs are added to a linked list. Whether a URL contains streaming media or not is determined by comparing it with regular expressions of known media URL patterns.
- An HTTP request is then sent to those URLs and the Web page is downloaded. This step is done only if the URL is a Web page.
- The text of the page downloaded is parsed to see if it actually contains the

[12]http://www.mplayerhq.hu
[13]http://perform.wpi.edu/real-tracer/#mediatracker
[14]http://perform.wpi.edu/real-tracer/#realtracer
[15]http://sourceforge.net/projects/ffprobe/
[16]http://www.ffmpeg.org/

streaming media. If it does, the direct link to the streaming media is extracted.

- The header of streaming media is then downloaded (we downloaded the initial 50 KBytes), and stored it in a temporary file.

- Lastly, FFprobe[17] is executed on that temporary file. The information generated by FFprobe is recorded, along with the mime type and last modified date of the media reported by the Web server, Web page URL in which the media was found, and direct link to the media.

[17] http://sourceforge.net/projects/ffprobe/

Chapter 4

Analysis

We started our crawler from 16 distinct starting points listed in Table 4.1. The crawling was done from Worcester Polytechnic Instute (WPI) between 10 December, 2009 and 24 January, 2010. In addition to storing and crawling URLs that began with prefix HTTP, the crawler was also configured to store URLs that began with other protocols. Each crawler run continued until it fetched at least 1.25 million URLs. For those crawls with more than 1.25 million URLs, only the first 1.25 million URLs were used in analysis.

Analysis of data thus collected is done in the following four sections. Section 4.1 presents a summary of URLs gathered and their clustering. Section 4.2 presents an analysis of the attributes of the streaming audio. Section 4.3 presents an analysis of the attributes of the streaming video. Section 4.4 provides a comparison of the results of our studies with previous study showing how the characteristics of multimedia content stored on the Web has changed. Section 4.5 discusses sampling issues related to this study.

4.1 Summary analysis

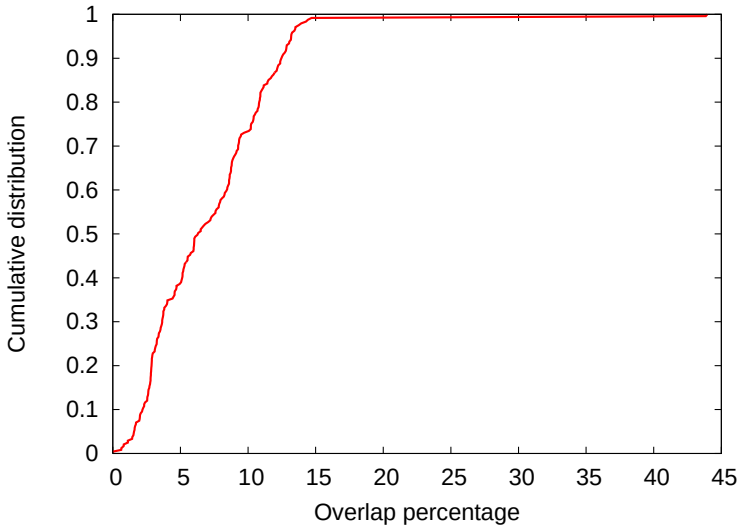

Figure 4.1: Overlap percentage

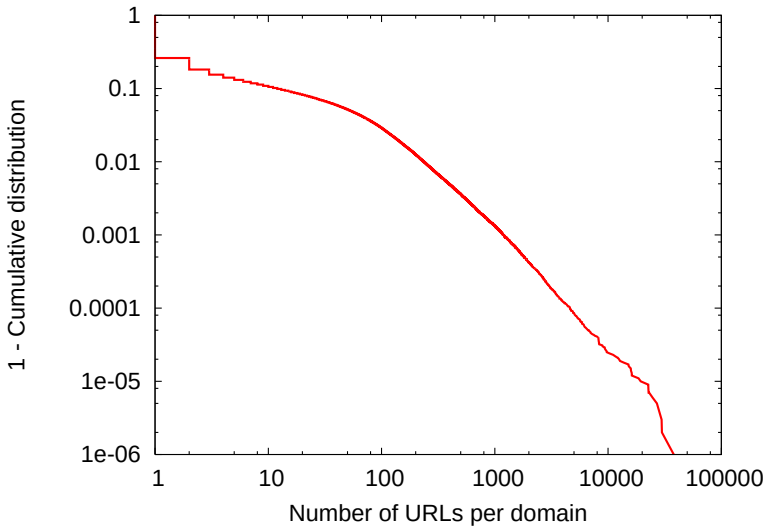

Figure 4.2: URLs per domain name

Despite running the crawler from different starting points, many of the crawled URls overlap those from other starting points. The overlap ratio from domain A to

domain B is calculated using the following equation.

$$ratio(A->B) = overlap(A,B)/sizeof(A) \tag{4.1}$$

Therefore, the overlap ratio from A to B might be different from the overlap ratio from B to A. This overlap is shown in Table 4.2. The starting points are given in the first row and first column of this table. Each cell gives the percentage of URLs that overlapped between those two starting points. The cumulative distribution function (CDF) of the same data is given in Figure 4.1. The horizontal axis in this figure represents the overlap percentage between any two starting points and the vertical axis is the cumulative distribution of those overlap percentages. The figure shows that the overlap between any two starting points is below 15%, except for the overlap between bbc.com and veoh.com, which is significantly higher at 43.9%. Overall, we crawled 15.32 million unique URLs.

The complementary cumulative distribution function (CCDF) of the number of URLs found per domain is given in Figure 4.2. The x axis in this figure is in logscale and shows the number of URLs per domain name. The y axis is also in log scale and shows the complementary cumulative distribution. The URLs in this figure came from 1,070,591 different distinct Web servers. Although there are a few "heavy hitters" that contribute a lot of URLs, more than 55% of the domains contribute only one URL. The top 15 domains that contributed the most URLs and the number of URLs they contributed are given in Table 4.3. Twitter, Facebook and Flickr occupy the top three spots.

Figure 4.3 depicts the URLs that were identified as containing streaming media, based on either the protocols or extentions, for each starting point. The x axis contains the name of each starting point sorted in increasing count for total media

Table 4.1: Crawler starting points

Starting URL	Starting Page	Country
bbc.com	BBC Homepage	UK
cbsnews.com	CBS News	US
chinatimes.com	Chinatimes	China
cnn.com	CNN	US
dailymotion.com	Dailymotion Video	France
empas.com	SK Communications	S. Korea
espn.com	ESPN Sports	US
free.fr	Free ISP	France
hulu.com	Hulu	US
ntt.co.jp	NTT Corp	Japan
podcastdir.com	Podcast Connect	US
podcastpickle.com	Podcastpickle	US
veoh.com	Veoh Network	US
video.yahoo.com	Yahoo Video	US
vids.myspace.com	Myspace Video	US
youtube.com	YouTube	US

Table 4.2: Overlap percentages

	bbc	cbs	ct	cnn	dm	emp	espn	fr	hulu	ntt	pcd	pcp	veoh	yah	mys	yt
bbc		8.6	2.3	13.5	12.2	2.8	12.0	8.8	12.5	3.7	6.5	13.0	43.9	9.1	9.3	5.1
cbs	8.8		2.8	10.9	8.7	2.9	11.8	10.8	11.5	2.9	4.6	9.3	8.4	9.2	7.3	3.3
ct	2.4	2.9		5.3	1.1	9.1	5.7	1.6	2.2	1.7	0.8	1.7	2.0	3.1	2.9	0.6
cnn	13.8	10.9	5.2		10.9	4.7	14.4	10.2	10.4	4.0	5.6	11.1	10.9	8.1	10.5	3.7
dm	12.4	8.6	1.1	10.9		2.3	11.9	10.2	13.2	2.6	6.0	12.8	11.7	10.1	12.5	6.5
emp	2.8	2.8	8.7	4.6	2.3		6.7	2.5	2.8	3.9	2.0	2.7	2.6	3.5	5.4	1.4
espn	12.4	2.0	5.8	14.7	12.2	7.1		10.8	13.5	6.3	8.5	14.3	13.1	10.7	11.2	7.2
fr	9.0	10.8	1.5	10.2	10.2	2.6	10.6		9.5	3.6	6.0	8.8	8.0	8.2	7.7	3.7
hulu	12.7	11.4	2.2	10.4	13.2	2.9	13.2	9.5		3.1	5.2	12.9	13.4	9.3	8.9	5.3
ntt	3.7	2.9	1.7	4.0	2.6	4.0	6.1	3.6	3.1		3.3	3.7	3.0	2.8	3.8	1.6
pcd	6.6	4.6	0.8	5.5	6.0	2.1	8.3	6.0	5.2	3.3		6.0	6.0	5.3	4.7	2.8
pcp	13.2	9.3	1.6	11.0	12.8	2.8	13.9	8.7	12.8	3.7	6.0		12.2	8.5	7.9	5.2
veoh	44.0	8.3	2.0	10.7	11.5	2.6	12.6	7.8	13.3	2.9	6.0	12.1		9.4	8.7	4.5
yah	8.6	8.5	2.9	7.5	9.3	3.4	9.7	7.6	8.6	2.6	4.9	7.9	8.8		10.4	3.2
mys	9.3	7.2	2.8	10.4	12.4	5.5	10.8	7.6	8.9	3.8	4.7	7.8	8.7	11.2		3.5
yt	5.1	3.2	0.6	3.6	6.3	1.4	6.9	3.5	5.2	1.6	2.7	5.1	4.4	3.4	3.4	
bbc:bbc.com, cbs:cbsnews.com, ct:chinatimes.com, cnn:cnn.com, dm:dailymotion.com, emp:empas.com, espn:espn.com, fr:free.fr, hulu:hulu.com, ntt:ntt.co.jp, pcd:podcastdir.com, pcp:podcastpickle.com, veoh:veoh.com, yah:video.yahoo.com, mys:vids.myspace.com, yt:youtube.com																

objects and the y axis contains the count of URLs. The media protocols we used to identify the URLs were Multimedia Messaging Service (mms), Real Time Streaming Protocol (rtsp:), iTunes protocol (itpc:), Real Media Protocol (pnm:) and Feeds

Table 4.3: Top 15 domains

No. of URLs	Domain name
87566	twitter.com
38039	www.facebook.com
29905	www.flickr.com
29684	www.fontsearchengine.com
28134	www.amazon.com
26970	digg.com
24667	www.myspace.com
23011	del.icio.us
22758	ad.doubleclick.net
22675	www.blogger.com
22654	www.youtube.com
19406	www.stumbleupon.com
18575	technorati.com
16223	www.nytimes.com
16165	en.wikipedia.org

(feed:). Out of these protocols, itpc and feed are basically plain text files that are used to provide links and description to other (not necessarily streaming media) content. Similarly, to identify streaming media content, the following common extensions were used for identification: .3g2, .3gp, .aac, .aif, .asf, .asx, .avi, .caf, .divx, .f4v, .flac, .flv, .hdmov, .m3u, .m4a, .m4b, .m4r, .m4v, .mid, .midi, .mkv, .mov, .mp3, .mp4, .mpa, .mpeg, .mpg, .ogg, .ogm, .ra, .ram, .rm, .rmvb, .vob, .vpm, .wav, .wma, .wmv, .xvid. We identified 59,115 media files using this method. This is relatively small number compared to the number of URLs crawled. This figure does not include streaming media content that is embedded in the pages inside Flash and other players. Most modern sites like YouTube host their media content insided embedded players and are not accessible via direct links.

We recorded the last modified date as reported by the Web servers while gathering the media information. Subtracting the last modified date from the current date provides the age of the streaming media files. Although the crawling was done around Jan 2010, we gathered the properties of the media in May 2010. The latest

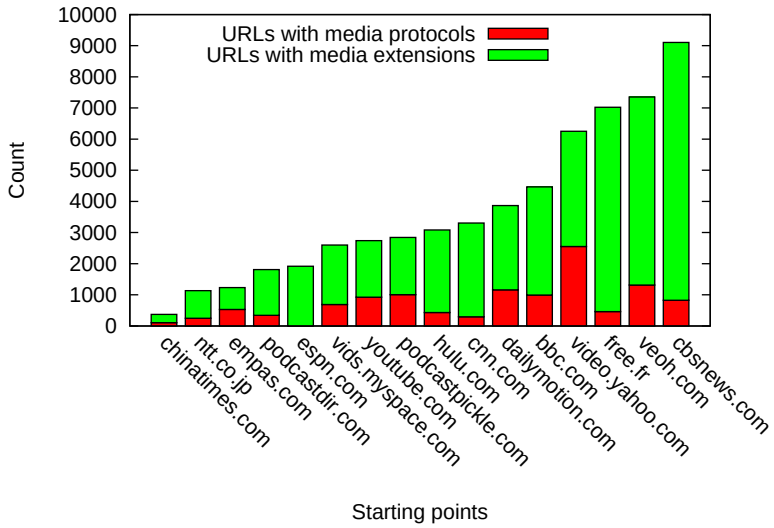

Figure 4.3: Media URL counts per starting point

date obtained was May 20, 2010, 09:10:08 AM. Figure 4.4 shows the cumulative distritubion function (CDF) of last modified date of the media files since May 20, 2010. In this figure, we can see that half of the content is less than 10 months old. However, some of the content is more than a decade old. The oldest streaming media clip we encountered was 170 months (i.e. more than 14 years) old.

Table 4.4 shows the number of media clips analyzed. Out of the clips analyzed, 48.5% were audio and 51.5% were video.

Table 4.4: Media clips analyzed

Audio	29274	48.49%
Video	31095	51.51%

We extracted Country code top-level domain (ccTLD) from the media URLs. Table 4.5 shows the ten most frequently encountered ccTLD. uk was encountered the most number of times. tv was second. Although tv is the ccTLD of Tuvalu, it is used for the television and entertainment industry purposes.

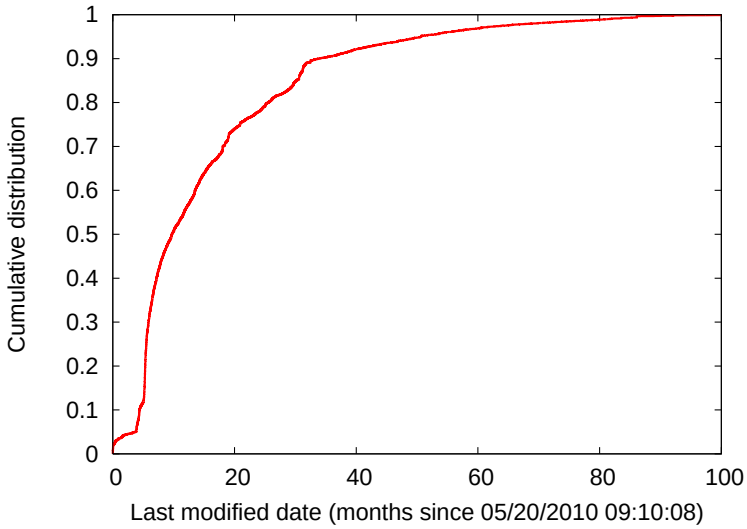

Figure 4.4: Last modified dates for media URLs according to the servers

Table 4.5: Top ccTLDs encountered in media URLs

Number of media URLs	ccTLD	Country
1155	uk	United Kingdom
1125	tv	Tuvalu (used by TV/entertainment)
520	jp	Japan
417	de	Germany
408	fm	Federated States of Micronesia(used by radio stations)
356	ca	Canada
251	cz	Czech Republic
233	us	USA
224	nl	Netherlands
201	cn	China

4.2 Audio

This section analyzes the various characteristics of the streaming audio content. Table 4.6 shows the number of audio clips having stereo and mono channels. The count in this table includes the audio embedded in the video clips as well as the audio only clips.

Table 4.6: Audio channels

Mono	18927	39%
Stereo	29231	61%

Figure 4.5 presents the histogram of major audio codecs. The x axis in this figure shows the major audio codecs encountered sorted from most to the least and y axis shows their count. The figure also includes the percentage of audio clips having that particular codec above each bar. A significant amount of audio content, more than half, is mp3. Second most prevalent is the Advanced Audio Codec (AAC). AAC was designed to be the successor of the MP3 format and generally achieves better sound quality than MP3 at similar bit rates [3]. 5% of the audio was encoded using cook codec, which is a lossy audio compression codec developed by RealNetworks. Windows Media Audio V2 codec follows. In total, there were 23 different types of audio codecs found.

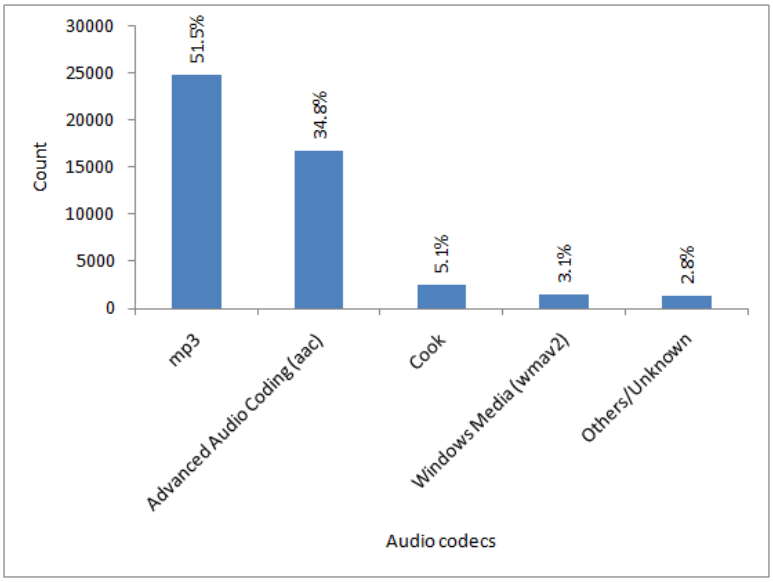

Figure 4.5: Audio codecs

Figure 4.6 graphs the CCDF of the stored audio and video length distributions.

The x axis in the figure is duration of clips in minutes and the y axis is the complementary cumulative distribution of the clip durations. Both the axes are in logscale. The plot shows that most stored audio clips are relatively short. The median length of audio clips is about 4 and a half minutes. While about 10% of the audio clips are 60 minutes or longer in length, only 0.5% of the audio clips are 2 hours (120 minutes) or longer. The longest clip found was 251 minutes long.

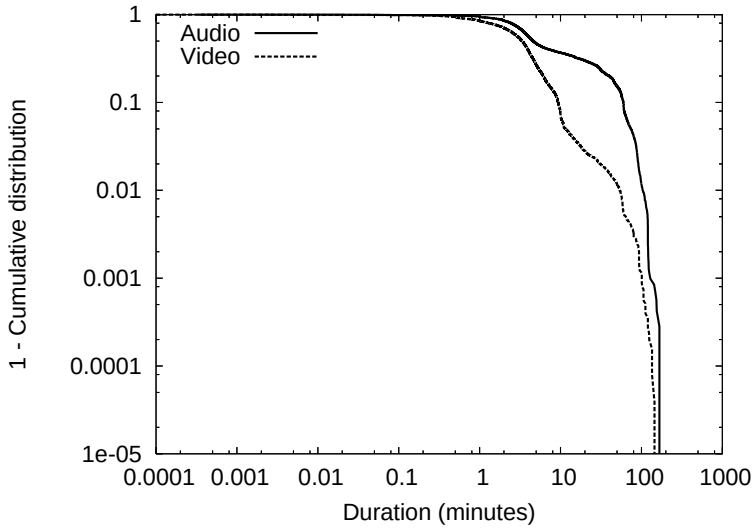

Figure 4.6: Audio/video lengths

The CCDF of the filesize in MBytes of stored audio and video is given in Figure 4.7. The axes in the figure are in logscale. The x axis shows the filesize in MBytes and the y axis is the complementary cumulative distribution. The median audio clip filesize is 6.5 MBytes, which is about as long as an average 4-5 minute song. The maximum size of an audio clip is about 1 GBytes which was an hour and 43 minute long ogg file.

Figure 4.8 provides CDF for encoded bitrates for audio in Kbits per second. The x axis is the bitrate in Kbps and the y axis is the cumulative distribution. The vertical gridlines represent the most commonly used audio bitrates. The median

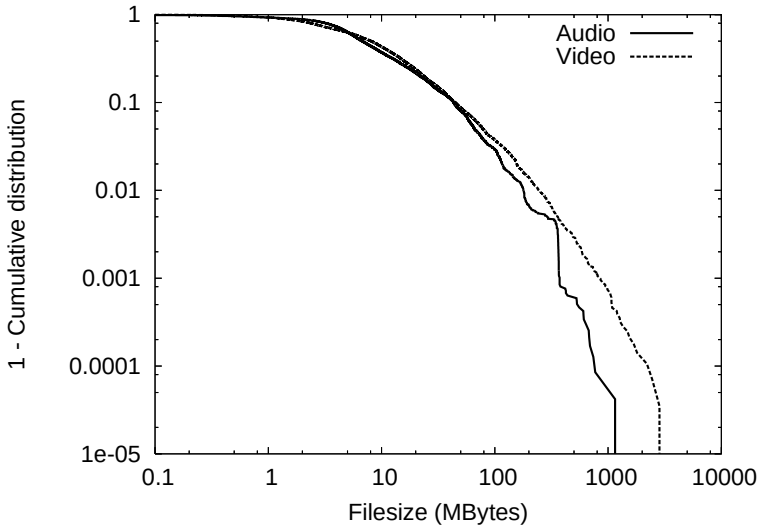

Figure 4.7: Audio/video file sizes

bitrate is 128 Kbits/sec, which is the typical encoding rate for an mp3 file. Distinct steps can be seen at 32 Kbps, 64 Kbps, 128 Kbps, 192 Kbps and so on, all standard bitrates at which audio clips are encoded. As we will see in Section 4.4, this graph also shows that quality of the audio stored on the Web has significantly increased since the study in 2003 [8]. This is likely because the bandwidth available is sufficient to support greater encoding rates for audio streaming and the bitrates do not need to be as limited due to bandwidth.

4.3 Video

Video content, like audio, can have wide range of characteristics in terms of codecs, encoding bitrates, length and resolution depending upon the applications, available bandwidth, visual quality etc. This section analyzes the characteristics of the gathered streaming video content stored on the Web.

Figure 4.9 depicts the major video codecs that we encountered. These are the

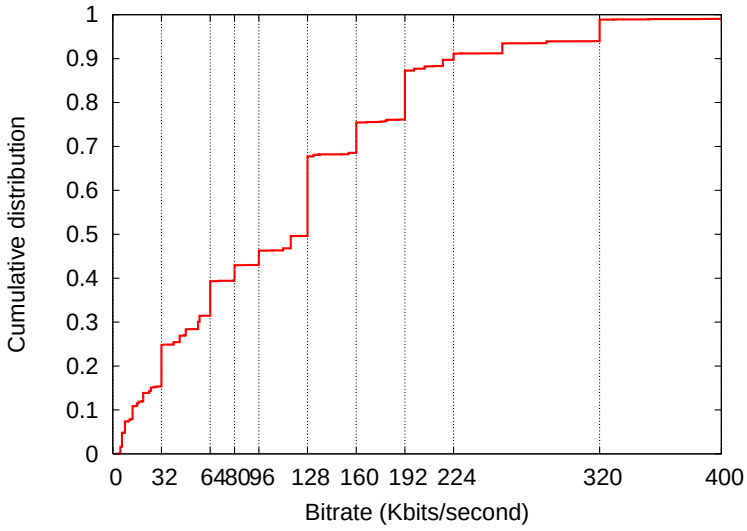

Figure 4.8: Encoded audio bitrates

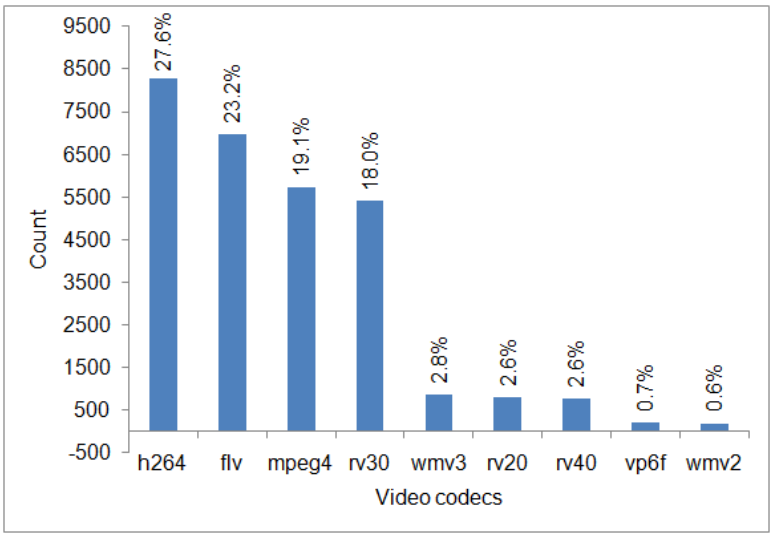

Figure 4.9: Video codecs

codecs reported by FFProbe which used the streaming video headers to obtain that information. The x axis shows the major video codecs encountered, sorted from most to the least and y axis shows their count. The figure also includes the percentage of video clips having that particular codec above each bar. Advanced Video Coding (H.264) makes up about 28% of the video codecs found. H.264 is used in wide array

of applications as players for Blu-ray Discs, videos from YouTube and the iTunes Store, Web software like Adobe Flash Player and Microsoft Silverlight, broadcast services for DVB and SBTVD, direct-broadcast satellite television services, cable television services, and real-time videoconferencing. In second place for about a quarter of the videos is the Flash Video (flv) codec. Flash Video is a container file format used to deliver video over the Internet using Adobe Flash Player versions 6 to 10. Flash Video content may also be embedded within Shockwave Flash (SWF) files. Notable users of the Flash Video format include YouTube, Hulu, Google Video, Yahoo! Video, Metacafe and many news providers. mpeg4 was used to encode $1/5^{th}$ of the video files, which then followed by various versions of Real Video and Windows Media Video codecs. Real Video 8 (rv30) and Real Video 9 (rv40) are the most popular proprietary Real Networks codecs, while Windows Media Video 9 (wmv3) and Windows Media Video 8 (wmv2) are the leading Microsoft Video codecs. In total, 36 different types of video codecs were encountered.

Figure 4.6 presents the CCDF of stored audio and video lengths, and shows video clips are a little shorter than audio clips. However there are a lot more videos having lengths between 1 to 10 minutes than there are audio clips in that length range. The median length of video clips is 3.2 minutes. While 10% of the audio content is an hour or longer, only 0.5% of the video clips are longer than an hour. The maximum length of video is 165 minutes. There were 9 videos having zero length and a video with negative length. These could be live videos, but we did not further investigate into it.

Figure 4.7 presents the CCDF of the filesize in MBytes of stored audio and video. Both the axes in this figure are in logscale. The x axis shows the filesize in MBytes and the y axis is the complementary cumulative distribution. In terms of space, although the video clips have distribution similar to audio clips, they are larger in size.

The median file size of video clips is around 8 MBytes, which is slightly larger than the 6.5 MBytes median for the audio clips. While the maximum size of video clips is nearly 3 GBytes, only 0.08% of them larger than 1 GBytes. The largest video clip we encountered was http://multimedia.cvut.cz:80/media/video/20091019_MKKandidatiRektor.wmv, a 2 hour and 4 minutes long Window Media video.

Figure 4.9 shows that Flash Video (flv) is one of the most prevalent video codecs used. So we compare the filesize of Flash videos to the sizes of other video formats. The CCDF of this result is presented in Figure 4.10. The x axis shows the filesize in KBytes and the y axis is the complementary cumulative distribution. The solid line shows the distribution of filesizes of Flash videos whereas the dotted line shows the distribution of filesizes of other videos. The graph shows that Flash videos have generally smaller filesizes than those of other types of media.

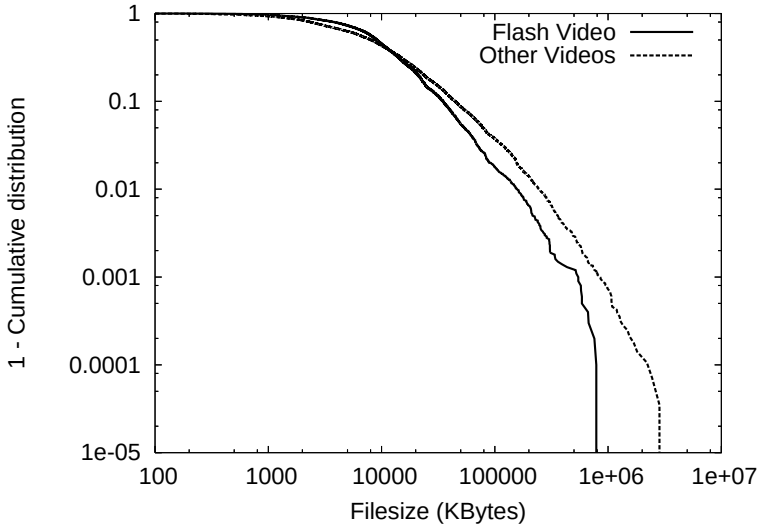

Figure 4.10: Media file sizes (FLV vs Non-FLV)

Figure 4.11 shows the CDF of encoded bitrates of video files. The x axis is the bitrate in Mbits/second and the y axis is the cumulative distribution. The median encoded bitrate of the video is around 0.3 Mbps. While the video bitrates

are significantly higher than those for audio, and higher than videos in previous studies as mentioned in Section 4.4, the encoded rates are still significantly lower than studio quality videos (3-6 Mbps) and HDTV quality videos (35-34 Mbps).

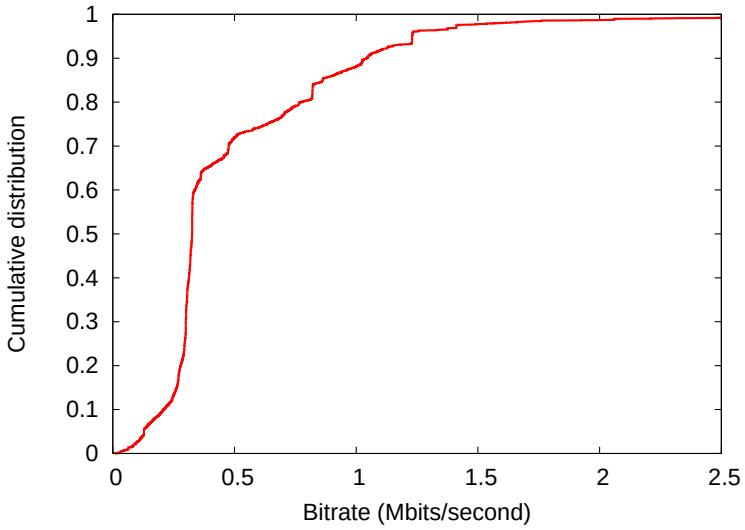

Figure 4.11: Encoded video bitrates

Figure 4.12 presents the CDF of the video clip resolutions. The x axis is squared resolution (i.e. pixel width multiplied by pixel height) and the y axis is the cumulative distribution. The dotted vertical lines indicate the most commonly used video resolutions like 320 x 240, 640 x 480 and 1024 x 768, and the steps in the distributions correspond to these typical resolutions. While significant amount of videos have resolution of 320 x 240, there are videos which have High Definition resolution (720p, 1080p i.e. with resolutions 1280 x 720 or 1920 x 1080).

Figure 4.13 presents the CDF of video aspect ratios. The x axis shows the aspect ratio and the y axis shows the cumulative distribution. The aspect ratio was obtained by dividing frame width by frame height. 4/3 is the most prevalent aspect ratio, while a significant number of videos have the aspect ratio of 16/9. Aspect ratios encountered ranged from 0.5 to 16.

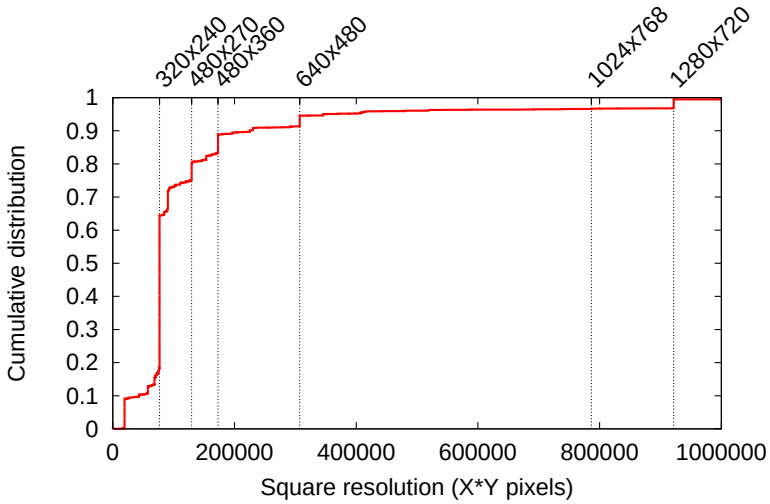

Figure 4.12: Video resolutions

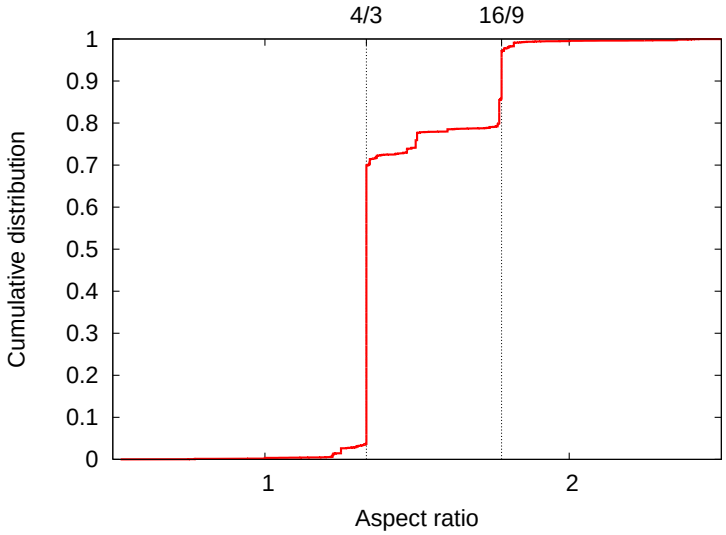

Figure 4.13: Video aspect ratios

For many videos, YouTube stores multiple copies of the same video in different formats (codecs and resolutions). The appropriate video is served according to the device's playback support and bandwidth available. While probing for YouTube URLs, we analyzed whether that particular URL contained the same video in mul-

tiple formats. More specifically, we looked for the following six formats in each YouTube URL: normal Flash Video, Hi-Resolution Flash Video, MP4 file, 720p video, 1080p video and 3gp video. So, each YouTube URL could contain from 1 to 6 formats. The CDF of this data is shown in Figure 4.14. The x axis is the number of formats contained in each YouTube URL and the y axis is the cumulative distribution. The graph shows that more than 65% URLs have only one video. About 20% of the URLs have 3 video formats.

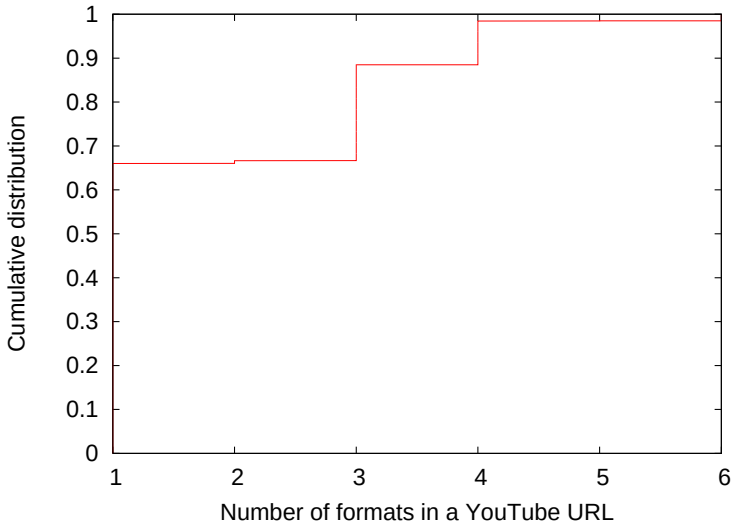

Figure 4.14: Number of formats in a YouTube URL

4.4 Comparison with previous studies

This section compares the results of our study with the results from previous similar studies, primarily with data gathered in 2003 [8]. Figure 4.15 shows the CDF of URLs overlap percentage between any two starting points. The x axis is percentage of URLs that overlap between two starting points and y axis is cumulative distribution. The solid line is the CDF of previous study and the dotted line is the CDF from our study. In the previous study, more than 80% of the crawls had 15% or less overlap, while in our study 90% of the crawls have about 15% or less overlapping URLs.

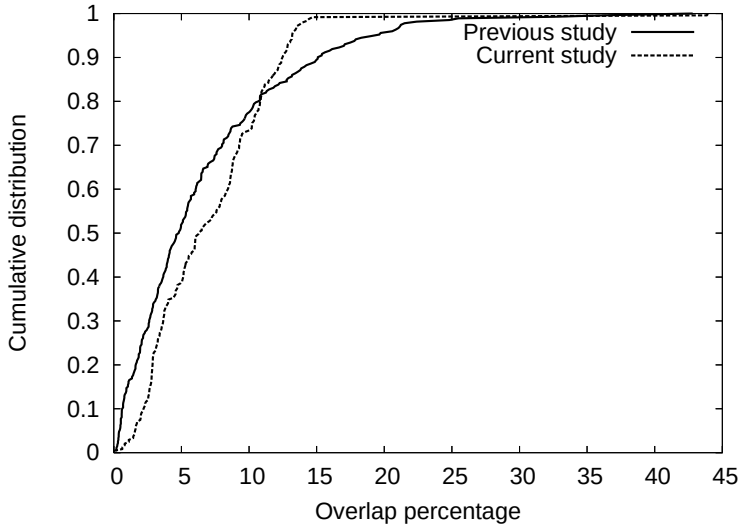

Figure 4.15: Overlap percentages

Table 4.7 shows the comparison of main results of the previous studies done in 1997, 2003 with our study.

In 2003, Li et al. [8] showed that the median video and audio clips durations were about 2 minutes and 4 minutes respectively. Our study shows median video

Table 4.7: Comparison of results

Characteristic	Study in 1997	Study in 2003	Our study in 2010
Median audio clip duration	NA	2 minutes	3.2 minutes
Median video clip duration	15 seconds	4 minutes	4.5 minutes
Audio clip encoded at less than 40 Kbps	NA	90%	About 20%
Median video clip size	1.1 MB	NA	8 MB
Videos targeted for broadband (768 Kbps) or higher	50%	1%	More than 20%
Videos with resolutions greater than or equal to 640x480	NA	Less than 1%	More than 10%
Prominent audio types	NA	Real Audio, MP3	MP3, AAC
Prominent video types	QuickTime, AVI	Real Video, WMV	H264, Flash Video
Videos with aspect ratio 1.33	74%	70%	70%

and audio clip durations to now be 3.2 minutes and 4.5 minutes respectively. While videos are still relatively shorter than audio in length, both are longer now than they were in 2003. Moreover, videos seem to have grown relatively longer in length, possibly due to the decreased price in disk storage and the increase in available bandwidth.

Our study finds the median encoded bitrates of video clips to be around 300 Kbps. In 2003, the median encoded bitrate was 200 Kbps [8]. The median has shifted higher, but the significant difference is between the percentage of videos that are targeted for broadband connections (768 Kbps) or higher. While Li et al. [8] found that only 1% of the videos are targeted for those higher bandwidths, our study shows that more than 20% of the videos are targeted for those connections. Interestingly, Acharya and Smith [1] found in 1997 that 52% of the videos had encoding rate of about 700 Kbps or higher. They found that video creators seldom dropped below 500 Kbps when creating video, below which the quality became unacceptable. But with advent of better encoding techniques, it has been possible to create higher

quality videos with lower encoding rates. Similarly for audio, the median encoded bitrate has increased from around 20 Kbps to 128 Kbps. In 2003, 90% of audio clips were encoded at less than 40 Kbps, but now only around 20% of the audio clips have encoded bitrates less than 40 Kbps.

Video resolutions have significantly increased since 2003 [8]. In 2003, less than 1% of the videos had resolution higher than or equal to 640 x 480, while now more than 10% videos have resolutions higher than or equal to 640 x 480. With more Websites adding support for High Definition (HD) videos, one can expect the video resolutions to increase in the future.

There are significantly more videos that have widescreen (16:9) aspect ratio today.

4.5 Sampling issues

While collecting data for large scale measurement studies on the Web, there are issues related to the number of samples compared to the size of the overall population. In 1997, Acharya and Smith were able to locate and download all the videos found on the Web [1], but today that is practically impossible to do due to the size of the Internet and the time limit of our study.

This section discusses the issues related to sampling and the data gathering approach used in obtaining 20 million URLs with the crawler. To make sure that this set of URLs is an adequate sampling of the Web, the strategy was to compare the results of our study with smaller sub-samples of our data. More specifically, this section analyzes whether a different number of URLs would affect the overall distribution shapes.

Figure 4.16 is the CCDF of video durations for different sample set sizes. The

x axis is duration in minutes and the y axis shows the complementary cumulative distribution. The solid line is plot for the entire sample i.e. 20 million URLs, the dashed line is for half the sample size, 10 million URLs, and the dotted line is for 5 million URLs. This figure shows that even for smaller sample sizes, the distribution remains visually similar. This suggests that crawling beyond our sample size of 20 million URLs is unlikely to change the results.

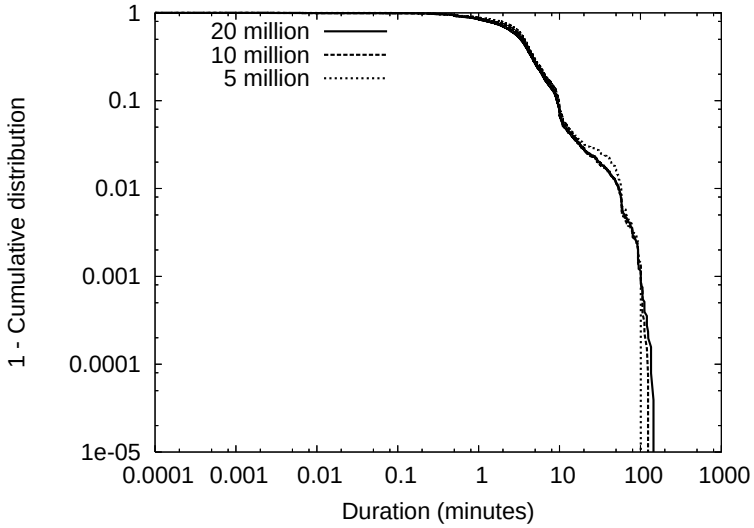

Figure 4.16: CCDF of video durations for different sample set sizes

Figure 4.16 is the CDF of video bitrates for different sample set sizes. The x axis is the bitrate in Mbits/second and the y axis is the cumulative distribution. The solid line is a plot for the entire sample, i.e. 20 million URLs, the dashed line is for half the sample size, i.e. 10 million URLs, and the dotted line is for 5 million URLs. Again, this figure shows that for smaller sample sizes, the distribution remains visually similar. This also suggests that crawling beyond our sample size of 20 million URLs is unlikely to change the results and our sample size provides adequate sampling of the Web for the study.

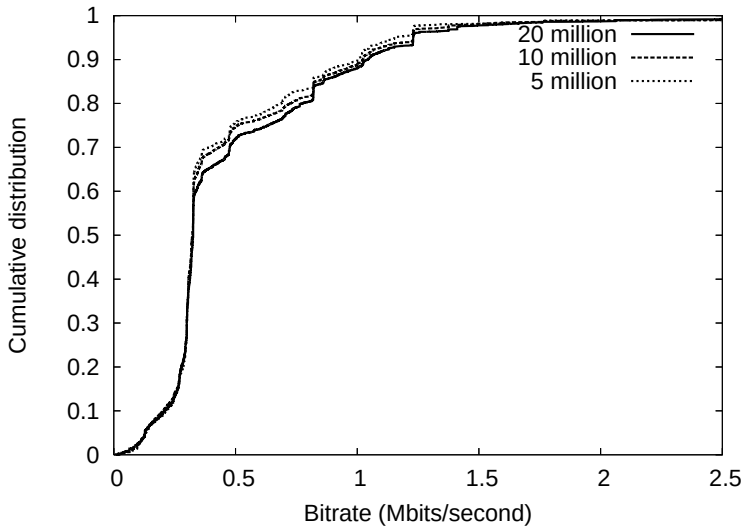

Figure 4.17: CDF of video bitrates for different sample set sizes

Chapter 5

Conclusion

The influx of user generated content onto the Internet brought about by Web 2.0, along with the increase in Internet access and bandwidth means that multimedia content stored on the Web continues to grow. Streaming media generally requires higher data rates and consumes significantly more bandwidth than do text-based Web pages. In addition, the quality of service required for streaming media tends to be different than for traditional Web pages, with streaming media being more sensitive to variation in bandwidth than text-based pages. The lack of adequate data on the current state of streaming media avalailable on the Web makes it difficult for system engineers and people working on optimizing streaming media performance to predict the short term and long term impact of increasing streaming media traffic. Moreover, network models are desinged on the assumption based on the result of studies conducted several years ago, making their validity suspect. Significant changes in user access patterns and capabilities and changes in characteristics of streaming media makes it unlikely the previous characterizations accurately represent current audio and video content on the Web today.

The goal of this study is to provide a fresh analysis and current snapshot of

streaming media stored on the Web. Using a custom-built Web crawler, we crawled 20 million Web pages from 16 globally distributed starting points served from 1,070,591 different distinct Web servers. Media analysis tools were used to identify and analyze the streaming media content on those pages. Then initial segments of each clip were downloaded to extract media characteristics stored in the headers.

Audio clips and video clips are available in almost the same number. Most streaming media clips are relatively short. Audio clips have median length of 4.5 minutes but almost 10% of audio is 1 hour or longer in length. The median length of video clips is just over 3 minutes. Although the duration seen in our study is longer than that seen in 2003 which was about 3 minutes, the difference between median clip length is not significant. Median filesizes of audio clips is about 6.5 MB and that of video clips is about 8 MB. While in 2003, the majority of audio encoded bitrates were still targeted to be acceptable for modem connections, currently that is not the case. The majority of audio is encoded at the typical standard bitrates instead of encoding them for lower bandwidths. In 2003, 29% of the videos were encoded for modem bitrates (56K), while today less than 1% of the videos are targeted for modem bitrates (56K). Over half of the videos have a resolution higher than 320 x 240, while in 2003 only less than 20% of the videos had a resolution higher than 320 x 240. In 2003, videos having high definition resolution were almost non-existent, while today more than 35% of the videos have higher than 320 x 240 resolution and videos having high definition (720p and 1080p) are prevalent.

In 2003, online multimedia was dominated by three media formats - Real Network's RealMedia, Microsoft's Windows Media and Apple's QuickTime. Now these three have been replaced by newer media types. More than 80% of audio clips are encoded using either MPEG Layer 3 (mp3) or Advanced Audio Codec (AAC). Similarly, H.264 and Flash Video (FLV) are the most prevalent video codecs available

on the Web accounting for the 27% and 23% of video found.

Chapter 6

Future work

This chapter presents possible future work as an extension to our study.

As mentioned in the introduction, there is a considerable audio and video content available on peer-to-peer file systems. This type of content is typically downloaded to the local system before playing it out. One extension to our work could be to create tools to crawl peer-to-peer file sharing systems and analyze the multimedia content found there.

While the average encoded bitrates gathered from headers provides information about the possible network impact, the actual streaming rates experienced by end users during playback are likely to be different. Today many streaming technologies can take advantages of multiple target bitrates stored in a single media object. Determining bitrate levels for stored multimedia clips, if available, could be another valuable extension.

Our study does not include media content that is not publicly available – paid content, and the content in social networking sites, such as Facebook. More users are uploading videos to social networking sites. Paid content in sites like Netflix and Hulu Plus is also increasing. Finding effective methods to gather information about

freely inaccessible media content and studying the nature of that content would also add value.

Although generally a user has to explicitly press the play button to start streaming and play media on Web pages, many Web sites start playing streaming media automatically as soon as the page loads. This might be a usability decision on the part of Web designer, but studying whether the content providers use bandwidth available or geography for autoplay decisions would also be an extentsion to our study.

We identified the country from which the media URL originated, if available, using country code top level domain. Another extension to our work would be to study the correlation between the characteristics of streaming media and geography.

Appendix A

Appendix

The following multimedia clips were the most duplicated ones in our data, i.e. these URLs were reached from the most number of crawling instances. These URLs were visited manually to record information like title, number of views etc., if available. The second column is the URL and first column indicates the number of times that particular URL was encountered.

```
11 www.youtube.com:80/watch?v=2lXh2n0aPyw
        (Piano stairs - TheFunTheory.com - Rolighetsteorin.se. 12,248,071 views)
11 cstvpodcast.cstv.com.edgesuite.net:80/nba/110609_nba_final.mp3
10 cstvpodcast.cstv.com.edgesuite.net:80/fantasyplaybook/120209_fantasyplaybook_final.mp3
10 cstvpodcast.cstv.com.edgesuite.net:80/fantasypicknroll/120309_fantasypicknroll_final.mp3
 9 www.youtube.com:80/watch?v=ysIzPF3BfpQ
        (The Muppets: Ringing of the Bells. 1,391,305 views )
 9 www.youtube.com:80/watch?v=txqiwrbYGrs
        (David After Dentist. 61,540,146 views)
 9 www.youtube.com:80/watch?v=Qb9jY8yAxgs
        (Edward Sharpe and the Magnetic Zeros on Letterman. 573,732 views)
 9 www.youtube.com:80/watch?v=oHg5SJYRHA0
        (RickRoll'D. 35,837,988 views)
 9 www.youtube.com:80/watch?v=MvSeL_LfdbA
        (Ignite: Molly Wright Steenson - A Series of Tubes, EP 7. 23,381 views)
 9 www.youtube.com:80/watch?v=HhgfzOzPmH4
        (Google Goggles. 1,399,755 views )
```

The following 23 types of audio codecs were encountered, sorted from most to the least:

```
mp3
Advanced Audio Coding (aac)
Cook
Windows Media (wmav2)
Others/Unknown
```

```
nellymoser
pcm_s16le
Ogg Vorbis (vorbis)
mp2
QDesign Music Codec (qdm2)
pcm_s16be
Adaptive Transform Acoustic Coding 3 (atrac3)
pcm_u8
Adaptive diff. pulse-code modulation Interactive Multimedia Association (adpcm_ima_qt)
pcm_s24le
mp1
Adaptive Multirate Audio (samr)
Free Lossless Audio Codec (flac)
adpcm_ima_wav
PCM mu-law (pcm_mulaw)
Apple Lossless Audio Codec (alac)
adpcm_swf
pcm_s24be
real_144
```

The following 36 types of video codecs were encountered, sorted from most to the least:

```
h264
flv
mpeg4
rv30
wmv3
rv20
rv40
vp6f
wmv2
mpeg1video
Apple Quicktime (Sorenson Video) (svq3)
theora
h263
wmv1
SMPTE 421M (vc1)
msmpeg4
mjpeg
mpeg2video
svq1
rv10
MSS2
Apple QuickTime RLE (qtrle)
Others/Unknown
png
msmpeg4v2
cinepak
WVP2
GoToMeetingCodec (G2M3)
qdraw
Apple Graphics (smc)
```

```
indeo3
rgbb
rpza
dvvideo
Indeo v5 (IV50)
MSS1
msvideo1
```

Bibliography

[1] S. Acharya and B. Smith. An Experiment to Characterize Videos Stored on the Web. In *Proceedings of the ACM/SPIE Multimedia Computing and Networking (MMCN)*, pages 166 – 178, San Jose, CA, Jan. 1998.

[2] M. Bergman. The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing*, 7(1):07–01, 2001.

[3] K. Brandenburg. MP3 and AAC Explained. In *AES 17th International Conference on High-Quality Audio Coding*, Florence, Italy, Sep 1999.

[4] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 1–14, San Diego, California, USA, 2007.

[5] M. Chesire, A. Wolman, G. Voelker, and H. Levy. Measurement and Analysis of a Streaming Media Workload. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems (USITS)*, pages 1–12, San Francisco, CA, USA, Mar. 2001.

[6] F. Duarte, F. Benevenuto, V. Almeida, and J. Almeida. Geographical Characterization of YouTube: a Latin American View. In *Web Conference, 2007. LA-WEB 2007. Latin American*, pages 13–21. IEEE, 2007.

[7] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube Traffic Characterization: A View From the Edge. In *Proceedings of the ACM Internet Measurement Conference (IMC)*, San Diego, CA, USA, Oct. 2007.

[8] M. Li, M. Claypool, R. Kinicki, and J. Nichols. Characteristics of Streaming Media Stored on the Web. *ACM Transactions on Internet Technology (TOIT)*, 5(4):601–626, Nov 2005.

[9] A. Mena and J. Heidemann. An Empirical Study of Real Audio Traffic. In *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 1, pages 101–110. IEEE, 2002.

[10] S. Raghavan and H. Garcia-Molina. Crawling the Hidden Web. In *Proceedings of the 27th International Conference on Very Large Data Bases*, pages 129–138, Sep 2001.

[11] S. Saroiu, P. Gummadi, S. Gribble, et al. A Measurement Study of peer-to-peer File Sharing Systems. In *Proceedings of Multimedia Computing and Networking*, volume 2002, page 152, San Jose, California, USA, Jan 2002.

[12] E. Veloso, V. Almeida, W. Meira, A. Bestavros, and S. Jin. A Hierarchical Characterization of a Live Streaming Media Workload. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurment*, pages 117–130. ACM, 2002.

[13] M. Zink, K. Suh, Y. Gu, and J. Kurose. Watch Global, Cache Local: YouTube Network Traces at a Campus Network - Measurements and Implications. In *Proceedings of the SPIE/ACM Multimedia Computing and Networking (MMCN)*, 2008.