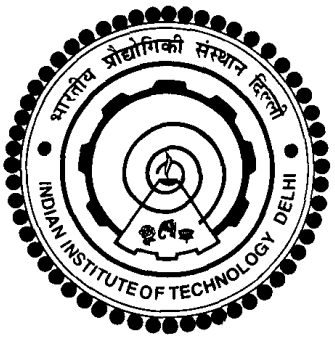
CHARACTERISING ONLINE VIDEO SHARING AND ITS DYNAMICS

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

By
Siddharth Mitra
2007MCS2102



Under the supervision of
Dr. Anirban Mahanti

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
July 2009

Certificate

This is to certify that the thesis titled **Characterising Online Video Sharing and its Dynamics** submitted by Siddharth Mitra, in partial fulfillment of the requirement of the degree of Master of Technology in Computer Science and Engineering, is a record of bona fide work carried out by him under my guidance and supervision at the Department of Computer Science and Engineering, Indian Institute of Technology, Delhi. The work presented in this thesis has not been submitted elsewhere, in part or in full, for the award of any other degree or diploma.

Dr. Anirban Mahanti and Dr. Kolin Paul
Department of Computer Science and Engineering
Indian Institute of Technology
Delhi

Abstract

Video sharing and publishing services that allow ordinary Web users to upload videos of their choice and watch video clips uploaded by others have recently become very popular. Our principal contribution in this paper is a comparison of the video sharing workload characteristics of five of the most popular video sharing services. This also includes more than a million YouTube videos that we traced weekly, for over a period of 8 months and provides us insight into the popularity dynamics on such online systems. Our traces contain meta-data of over 2 million videos which together have been viewed approximately 70 billion times. Using these traces, we study the similarities and differences in use of several Web 2.0 features such as ratings, comments, favorites, and propensity of uploading content. In general, we find that active contribution, such as video uploading and rating of videos, is much less prevalent than passive use and also that the uploaders in general are skewed with respect to the number of videos they upload. Similarly, we identify both invariants, and differences, in the video popularity distributions of the five services. We also identify differences in how popularity is measured that may be important in workload modeling.

Using our longitudinal YouTube data set, we study the evolution of user generated video files, and determine factors that affect popularity of videos. Our analyses shows that while video popularity does follow preferential attachment models, other factors also play important roles. For example, we present empirical evidence that suggests that videos uploaded by those with larger social networks may become relatively more popular. Similarly, our analyses suggests that age of the content also plays an important role. Our data set also allows us to present insights to factors that contribute toward unequal distribution of video views and content deletion.

Acknowledgements

This work has been a mix of ideas, advice, code, immense amounts of data and certain traces of frustrations that came with it. There have been numerous people who have contributed in one way or another over the course of my research.

I would like to express my sincere gratitude and respect to my advisor Dr. Anirban Mahanti, who instilled enough confidence in me right from the start. He is also responsible for sowing the seeds of many of the ideas presented in this work and also helping me bring them to fruition. Along with his guidance, I would also like to thank him for freedom he gave me to explore this topic on my own and ofcourse the trips to the nearest Barista.

I would also like to thank Dr. Derek Eager and Dr. Niklas Carlsson for their ideas, reviews, and revisions of my work over the past year. I am also thankful to Brian Gallaway at the University of Sasketchwan, for his help in setting up system resources for our data collection. Also, my fellow student's advice and company have been greatly appreciated.

Finally, I would like to make a special mention of my family, for their support, accepting my long and strange working hours in the right spirit, and also for the infinite cups of tea every day.

Contents

Certificate i

Abstract ii

Acknowledgements iii

List of Figures vii

List of Tables viii

Abbreviations ix

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Goals . . . 2
- 1.3 Methods . . . 3
- 1.4 Thesis Overview . . . 4

2 Background 5
- 2.1 Video Sharing . . . 5
  - 2.1.1 Dailymotion . . . 5
  - 2.1.2 Yahoo! Videos . . . 7
  - 2.1.3 Veoh . . . 7
  - 2.1.4 Metacafe . . . 8
  - 2.1.5 YouTube . . . 9
- 2.2 List of Questions . . . 11
  - 2.2.1 Static Characterisation and Invariants . . . 11
  - 2.2.2 Popularity Evolution . . . 11
  - 2.2.3 Social Network Effects and Content Deletion . . . 11
- 2.3 Data Sets . . . 12

3 Related Work 13
- 3.1 Power Law, Zipf and Pareto . . . 13
- 3.2 Traditional Media Workload and Online Access Patterns . . . 13
- 3.3 Web-based Video Sharing . . . 14

3.4 Miscellaneous 15

4 Tools **16**
4.1 Requirements 16
4.1.1 Python, Mechanize, urllib2 17
4.1.2 MySQL 17
4.1.3 SQLite 17
4.1.4 Matlab and R 17
4.1.5 Environment 18

5 Methodology and Data Collection **19**
5.1 Video and Uploader Attributes 19
5.2 Dailymotion Data Collection 21
5.3 Yahoo! Video Data Collection 22
5.4 Veoh Data Collection 23
5.5 Metacafe Data Collection 24
5.6 YouTube Data Collection 25
5.7 Summary of Data Sets 26
5.7.1 Static Characterisation 26
5.7.2 Temporal Characterisation 27
5.8 Limitations of Data Sets 28
5.9 Summary 29

6 Invariants in Video Sharing **31**
6.1 Video Duration 31
6.2 Ratings and Average Rating Score 32
6.3 Comments and Favourites 34
6.4 Video Uploads and Uploaders 35
6.5 Video Popularity 37
6.5.1 The 80-20 Rule for Video Popularity 37
6.5.2 Zipf and Power Law Analysis 41
6.6 Summary 45

7 Popularity Evolution **48**
7.1 High-level Characterisation and Validation 48
7.1.1 Activity Distribution 48
7.1.2 Age of Videos 50
7.1.3 Duration of Videos 50
7.1.4 Tagging of videos 51
7.1.5 Ratings and Comments 52
7.2 Longitudinal Characterisation of User Access Patterns 53
7.2.1 Overview 53
7.2.2 Hotset Dynamics 55
7.2.3 Viewing Rate Dynamics 58
7.2.4 Rich Get Richer Models 59
7.2.5 An Econometric Viewpoint 62
7.2.6 Unified Growth Model 65
7.3 Summary 66

8 Social Network Effects and Content Deletion 68
8.1 Impact of Social Networking 68
8.1.1 Who uploads videos? 69
8.1.2 Influence of Social Networks on Popularity of videos 70
8.1.3 Growth of Social Network 71
8.2 Content Deletion 72
8.3 Summary 74

9 Conclusions and Future Work 76
9.1 Thesis Summary 76
9.2 Results and Contributions 77
9.3 Future Work 79

List of Figures

6.1 Cumulative distribution of video duration. 32
6.2 Cumulative distribution of rating count. 33
6.3 Average rating score of videos. 34
6.4 Cumulative distribution of comments/favourites. 35
6.5 Skewness in the number of video uploads. 36
6.6 Distribution of uploads by uploaders. 36
6.7 Skewness in the total views popularity of videos. 38
6.8 Skewness in the average viewing rate over a two week period (Dailymotion), and the average viewing rate since upload. 39
6.9 Distribution and models of the total views popularity of videos. 40
6.10 Distribution and models for viewing rate popularity. 44

7.1 Initial file popularity. 49
7.2 Age distribution at seed time. 50
7.3 File durations (CDF). 51
7.4 Tags at seed time. 51
7.5 Aggregate Popularity Trends. 54
7.6 Hot set analysis. 55
7.7 Rank Changes. 57
7.8 CDF of immediate growth function. 58
7.9 Rich gets richer. 59
7.10 Weekly changes and correlations. 61
7.11 Welfare function by age. 63
7.12 Theil decomposition. 65
7.13 Growth model validation. 65

8.1 User Rank vs. Total Uploads. 69
8.2 New Uploads. 70
8.3 Total increase in views Vs users network. 71
8.4 Growth of friends network. 72
8.5 Deletion of videos. 73

List of Tables

5.1 Dailymotion table schema. 22
5.2 Yahoo! videos table schema. 23
5.3 Veoh table schema. 24
5.4 Metacafe table schema. 24
5.5 YouTube table schema. 26
5.6 High-level summary of four of our data sets. 27
5.7 High-level summary of data sets. 28

6.1 Models for the total views popularity distribution. 42

7.1 Inequality Indices (seed views). 62

Abbreviations

3GP	Third **G**eneration **P**artnersip Project
AAC	**A**dvanced **A**udio **C**oding
AVI	**A**udio **V**ideo **I**nterleave
BBC	**B**ritish **B**roadcasting **C**orporation
CDF	**C**umulative **D**istribution **F**unction
CCDF	**C**omplementary **C**umulative **D**istribution **F**unction
FLV	**FL**ash **V**ideo
HD	**H**igh **D**efinition
HTTP	**H**yper **T**ext **T**ransfer **P**rotocol
IPTV	**I**nternet **P**rotocol **T**ele**v**ision
JSON	**J**ava**s**cript **O**bject **N**otation
MKV	**M**atros**k**a **Video**
MPEG	**M**otion **P**ictures **E**xperts **G**roup
MP4	**M**otion **P**ictures Experts Group - **4**
MSN	**M**icro**s**oft **N**etwork
SQL	**S**tructured **Q**uery **L**anguage
URL	**U**niversal **R**esource **L**ocater
WMV	**W**indows **M**edia **V**ideo

Chapter 1

Introduction

As the popularity of the World Wide Web ("Web" or "WWW") continues to increase, we also witness the proliferation of channels for information dispersal. The current form of the Web has moved away from a static medium to a highly interactive one. No longer are the users mere consumers of information. The current highly interactive and dynamic form has often been termed 'Web 2.0'. With growth in technology, the people are seeking additional media for interaction other than textual. Thus, people are moving to exploring the full potential of the Web using multimedia. This new kind of usage brings new content on the Web in the form of videos and dynamic content, which is unlike the static traditional Web. One of the new popular destinations of users on the Web are online video sharing sites. Popular online video sharing services such as YouTube allow users to watch, upload, and share video content. With a huge selection of easily accessible videos, these services have quickly become extremely popular. The difference or similarities in characteristics, as well as their impact on the Internet are issues that need to be explored. We anticipate as users grow more accustomed to using online video sharing for their daily entertainment or informational dosage, these services will become an even more important part of the Internet.

In this chapter, we first discuss the motivation behind our work and also familiarise the users with some of the problems we considered and our goals. Then we talk about the methods we use to attain the aforementioned goals. Towards the end, we present a brief outline of the rest of the work.

1.1 Motivation

There are numerous video sharing services available today, which serve as some of the most popular destinations of user on the Web. Among the many video sharing services

that are available today, YouTube by far is the most popular, and its *Alexa* global Internet ranking (which is currently 3) illustrates this point. Furthermore, some recent report claim that YouTube traffic comprises nearly 20% of all HTTP traffic and around 10% of all Internet traffic [16], which is a highly significant percentage. Dailymotion, Metacafe and Veoh are not far behind with rankings 64, 127 and 176 [7] respectively. Thus, their usage and the higher bandwidth requirements for video than text, makes a significant fraction of the entire traffic on the Internet. Studying their characteristics, the user access patterns not only help us in designing better systems but also allow us to improve the existing ones. The interactive features on these services also provide a window into user behaviour, and social activities on the Web.

Video sharing services such as YouTube are inherently complex. These services host millions of videos, while several thousand are uploaded everyday, and several million are viewed daily. In addition to allowing users to upload and share videos with users around world, these sites typically facilitate tools and opportunities for social interaction among users that can help discover videos that the users are more likely to enjoy. Such social networking features include the ability for users to express friend or fan relationships that simplify staying up-to-date on the videos and likes of these users. Ratings and opinions are typically facilitated through the use of rating and comments services which allow users to discuss individual videos. By measuring viewing rates and taking advantage of the interactive features, these sites can also help users unveil videos they are more likely to appreciate. Users can also identify movies using internal search engines or lists that highlights popular movies of various categories that are sorted based on criteria such as "most recently uploaded", "most popular", or "most commented".

Studying popularity progression, its fluctuating nature and identifying the growth process behind it, is an important first step towards predicting popularity and designing efficient content distribution and recommendation systems. Studying these services, is also an important step in predicting workloads, understanding general dynamics of popularity, discerning user behaviour, his online relationships and their impact.

1.2 Goals

The primary goal of our work is to characterise the workload of online video sharing services like Dailymotion and YouTube, among others. In addition to studying the workload profile of these sites, we also seek to find invariant properties or trends across vastly different models of video sharing on the Internet. This is an important step towards building a broader understanding of this type of workload. Prior work in this

domain has only focused on YouTube, thus we also validate and build upon those results in our work.

Another important goal of our work is to understand the dynamics of how the popularity of videos evolve and change with time. In particular, we are interested in studying how addition of new videos impacts the popularity of existing videos, what factors influence the popularity of videos, what impact social networking has on future popularity of content, and how frequently and why videos are removed from these services. To capture changes and trends in the popularity, we collected weekly snapshots of more than a million YouTube videos. Our data collection spans more than 238 days and allow us to study how the popularity depends on factors such as the age of the video, the number of views it has gained since upload, its past popularity, how it has been rated thus far, the past success of the uploader of the file, as well as various other social aspects.

While there has been some recent works characterising various video sharing services [18, 20, 26, 55], this work differs from these in that it provides a longer-term study in which we capture some of the more fundamental popularity dynamics that affects how frequently different movies will be viewed. We believe that this work serves as a basis for understanding evolutionary traits on current systems and will help in building efficient media content distribution systems and modeling dynamic workloads for these systems. We believe this research will aid in developing models for these complex systems, and is complementary to prior work.

1.3 Methods

We surveyed prior literature on content and user characteristics, and traditional workload scenarios on the Internet, to enable us to understand our goals in terms of the existing works. This is essential as we compare our results to prior ones using similar data sets. Further, we acquire a set of five data sets from five different video sharing services. Four of those data sets, we use to understand the static characteristics of video sharing workloads. We also have multiple snapshots of YouTube, taken for more than 8 months regularly every week. This enables us to look at longitudinal dynamics of popularity and other attributes. We create our own tools to gather this data. Then, we use statistical analysis to research our data set, and present those using graphical representations where warranted. Since our data set contains well over two million videos from a wide variety of sources, we believe that our data set itself serves as an important contribution.

1.4 Thesis Overview

In the remainder of this work, we introduce some background information on the services and data we analyse. Chapter 2 presents a list of guiding research questions. Chapter 3 discusses related work. Chapter 4 describes some of the tools we use for our research, while Chapter 5 outlines our data collection process and strategy with a summary of our data sets. We then move on to static characterisation and a discussion on invariant trends we find across the different services in Chapter 6. Chapter 7 discusses the longitudinal aspects of popularity characteristics from our YouTube data set. Chapter 8 describes the uploader behaviour on such sites, the impact of social network of a user on popularity of content and also deletion of content. Finally, we summarize the results, our contributions and also present some directions for further research on the topics in this work.

Chapter 2

Background

In this chapter, we first introduce the reader to the services that we analyse in this work. We describe these services in detail along with their features, and also ways that a user can interact with these sites. Then, we introduce a list of questions we intend to answer here, enumerated from our goals in the previous chapter. This chapter concludes with a brief overview of the data sets we use for analysis.

2.1 Video Sharing

Video sharing services such as YouTube, Dailymotion allow users to share, view and discuss videos with other users. The lower barrier to entry on the current Web 2.0 sites have enabled millions of ordinary Web users to contribute back to multiple streams of information existing on the Internet. Some sites also employ pages with list of videos based on criteria such as 'most recent', 'most popular' or 'most commented' which might help a user in filtering the spectrum of noise and unveil good content. In this section we discuss some of the major video sharing services on the Web. We base our results on the empirical data obtained by crawling their Web sites.

2.1.1 Dailymotion

Dailymotion [24] is a leading video publishing and sharing Web service headquartered in Paris. This service was launched in March 2005, around the same time as YouTube was launched. As of this writing, the maximum allowed size and duration of uploaded videos are 150 MB and 20 minutes, respectively. Only users designated as "motion makers" are exceptions to the above-stated limits. Since February 18, 2008, the site supports video content that can play at 720p on an HD set, but the bit rate is significantly

less than that expected for HD quality. Dailymotion's `Alexa` [7] global Internet traffic rank, based on the number of visits to the site is 64 (May 2009), with more than 50% of its users coming from France, the United States, and Japan. According to the Businesswire [17], the site recorded 55 million unique visitors monthly in March, 2009 with over 1.2 billion page views [25] recorded in 2007, and 15,000 new videos are uploaded daily into Dailymotion's global network of 18 localized video entertainment sites. In March 2009, Dailymotion delivered over 975 million videos to users including curated content from premium and Motionmaker creative contributors. Dailymotion allows users to browse videos by searching tags, channels or user-created groups; the search system also introduces results based on things other users have searched for.

Dailymotion classifies videos into a number of channels, some of which are:

News & Politics - World events, pop culture, entertainment and celebrity gossip. Includes videos from NBC, Fox, TVGuide and citizen journalists.

Funny - Comedy shows from the WB, NBC and Comedy Central.

Film & TV - Shows and movies from television.

Music - Music videos, concerts and interviews from artists.

Auto-Moto - Videos on motorsport racing.

Arts - Art videos, animation, stop motion, cartoons and independent short films or videos.

Gaming - Videos on console and PC gaming.

Webcam & Vlogs - Webcam videos, opinions.

Travel - Travel videos on global destinations, vacation spots.

Sports & Extreme - Sports videos from professional, college and classic league sports.

Animals - Funny home videos of animals.

People & Family - Parenting tips and home movies.

Tech & Science - Science technology news and product reviews.

College - Videos on college experiences.

Life & Style - Design, style and do-it-yourself.

Latino - Latino culture, and some others

A user can also sign up to become a *MotionMaker*, and gets some additional features such as:

- Some popular or interesting videos are 'featured' more prominently on the site than others. Being a MotionMaker, gives one's videos an ability to be featured.
- A user can upload video files of up to 1GB size and of unlimited length.
- Ability to upload videos in a superior quality. A user can then receive HD coding for the videos that he uploads.

A typical user on the site also has the ability to comment, rate the videos and mark them as some of his favourites. He can form social networks on the site through Fan or Friend relationships.

2.1.2 Yahoo! Videos

Yahoo! video [52] is Yahoo's video publishing and sharing service. It was initially launched as a video searching site, and in June 2006 launched it as a video sharing service. According to the service's home page,the maximum allowed size of uploaded videos is 150 MB; no information regarding any upper limit on video duration is available. Yahoo! has an overall `Alexa` Internet traffic rank of 1; however, its various services are not individually ranked. We also could not find any published information on the number of page views, unique visitors, and uploads per month. The free service provides users with a means to search and play videos, save videos as their 'favorites', subscribe to channels, create playlists, and embed videos in Web pages and blog posts. The homepage contains editorially-featured videos that change daily and are skewed towards comedy, viral videos, talented users, odd stuff, animation, and premium entertainment content. Yahoo! Video accepts videos in WMV, ASF, QT, MOD, MOV, MPG, 3GP, 3GP2 or AVI formats and transcodes to a 700 kpbs bitrate. Video playback is in Flash and presented in a 16:9 aspect ratio by default. Like Dailymotion, it also classifies videos into 20 separate categories like Action, Animals, Art & Animation, Commercials among many others. Videos can also be filtered according to current popularity, while allowing viewers to explore videos that have been the most popular ever.

2.1.3 Veoh

Veoh [50] is a video publishing, sharing, and Internet television service based in San Diego, California. This service is geared towards content from major studios, independent production houses, and ordinary Web users. Veoh offers two different services. The

`veoh.com` site allows users to browse videos, create accounts, upload content, and preview videos within one's browser. The second service called VeohTV requires download of software that facilitates delivery of full-length, high-quality content, to desktops and portable devices. Veoh recommends upload of videos encoded for playback at 500 Kbps or higher. Videos of duration less than 20 minutes can be watched using a browser, while longer duration content requires use of the Veoh player application. As of October 2008, according to `Alexa`, a majority of the visits to this service were from the United States and Japan, and the service's global Internet traffic rank was 176.

The Veoh.com site, which is currently serves approximately 19 million viewers per month, allows viewers to watch streaming video from across the Web. It also allows access to its content via devices like phones and other portables. In a similar fashion to YouTube, Veoh.com offers a broad selection of content from network television and allows viewers to watch full episodes of television shows and full-length movies. Veoh.com hosts a range of programming, from user generated content to professional studio content. It also provides a software application, VeohTV Beta, that allows a "lean back" and remote controllable viewing of video content on the Web. In December 2008 Veoh transitioned from VeohTV to the Veoh Web Player. The Veoh Web Player enables users to watch full-length videos on Veoh.com from within their browsers itself. Users are also allowed to download content from Veoh.com and some other websites. It uses peer-to-peer technology on its player software application, and Adobe Flash-based streaming video on its website. It claims its use of peer-to-peer in the player application enables distribution of longer form video files at a much lower cost, with the expectation that the bandwidth costs will not rise in direct proportion to the number of users. Veoh systems leverages LAMP and Java technologies. Veoh also recommends videos, that a user might prefer, based on his behavior. As a user watches, rates, and downloads videos, the Veoh recommendation system 'learns' what interests the user and presents more video choices that meet similar criteria.

2.1.4 Metacafe

Metacafe [37] was founded in July 2003 in Israel, with the stated goal of promoting short videos that are specifically developed for entertaining the Internet audience. The company is headquartered in Palo Alto, California, with offices in Tel Aviv and New York. Metacafe is similar to other video viewing websites such as YouTube or Dailymotion, but with several differences. It includes duplication elimination, a different type of Adult content filtering, a community member reviewer panel, VideoRank, and Producer Rewards. It uses a VideoRank system to estimate the reaction of a views to videos in order to feature those more prominently to users. Additionally, it also pays video

creators for original work that has exceeded a certain threshold of both total views and VideoRank score through its Producer Rewards Program. The site features short-form videos in a variety of channels, including Animation, Comedy, Entertainment, How To, News and Events, People and Stories, Sports, Video Games and others. It also hosts original content posted by independent video creators, small to mid-sized production groups, and major media companies.

Metacafe, uses a review system where each uploaded video is reviewed by a pool of volunteers to determine its suitability to the site. It also offers an unique revenue sharing model. Like all video sharing services, revenue is earned from advertisers. Uploaders of videos that achieve more than a fixed number of views and an appropriate average rating become eligible for a share of the revenue earned from advertisements. According to the site, some content producers have earned in substantially from their videos. Metacafe's current `Alexa` Internet traffic rank is 127; close to 40% of the visitors to this site are from India and the United States.

2.1.5 YouTube

YouTube [53] is not only the most popular video sharing service but it is also one of the most popular sites on the Internet. `Alexa` places the site at number 3 on its rankings. YouTube is is owned by a company based in San Bruno, California, and uses Adobe Flash Video technology to display a wide variety of user-generated video content, including movie clips, TV clips, and music videos, as well as amateur content such as video blogging and short original videos. In addition to content posted by individuals, media corporations such as CBS, BBC and other organisations also offer some of their material on YouTube.

Unregistered users can watch the videos, while users that register on the site are permitted to upload an unlimited number of videos. Accounts of registered users are called "channels". YouTube also allows seperation of content into different categories like Comedy, Music etc. The site restricts content which is potentially offensive to registered users over the age of 18. Such offensive content includes videos containing defamation, pornography, copyright violations, and material encouraging criminal conduct is prohibited by YouTube's terms of service. Violating these terms of service, can also lead to suspension of accounts by the site. The content in such cases is also removed from the site. In this work we look at content deletion on YouTube in detail.

YouTube's video playback technology for Web users is based on the Adobe Flash Player. This technology allows the site to display videos with quality comparable to more established video playback technologies (such as Windows Media Player, QuickTime, and

RealPlayer) that generally require the user to download and install another Web browser plug-in just in order to view videos. Flash videos do require a plugin, but Adobe market research [6] shows that the Flash plug-in is installed on over 95% of the computers.

Videos uploaded to YouTube are limited to 1GB in file size or upto ten minutes in length. Prior to 2005, it was possible to upload longer duration videos, thus we occasionally see videos that are longer than ten minutes. This restriction was enabled in order to curb uploading of copyrighted content such as TV shows. The site allows uploads in most formats, including .WMV, .AVI, .MKV, .MOV, MPEG, .MP4, DivX, .FLV, and .OGG. It also supports 3GP, allowing videos to be uploaded directly from a mobile phone. Originally it offered videos in a single format, but since then it has moved on to three main formats, including a "mobile format". The standard quality format displays videos at a resolution of 320x240 using the Sorenson Spark codec, with mono MP3 audio. It also introduced "High quality" videos, introduced in March 2008, are shown at 480x360 with mono MP3 sound. It utilises the H.264 codec and stereo AAC audio, which can be accessed by adding &fmt=18 to the end of the video URL. 720p HD support was also added in November 2008. At the same time, the YouTube player was changed from a 4:3 aspect ratio to widescreen 16:9. 720p videos are shown in full 1280x720 and encoded with the H.264 codec, with stereo audio encoded with AAC. From February 2009, all new videos are shown using H.264 and AAC by default for both"Standard" and "High Quality" viewing, for the approximately the same file size.

YouTube has a numerous ways of ranking its videos, such as "most viewed", which is further split into four categories: today, this week, this month, and all time. The current rankings are:

- Featured
- Rising Videos
- Most Discussed
- Most Viewed
- Top Favorited
- Most Popular
- Most Responded
- Top Rated

2.2 List of Questions

After exploring the services from which we present our data material for analysis, we move on to some of the questions that we investigate in this work. These are enumerated from the goals we outlined in the previous chapter.

2.2.1 Static Characterisation and Invariants

- Invariant properties across different video sharing services
- Distribution of duration
- Usage of interactive features like rating, commenting, marking as favourite
- Video upload distribution
- Distribution of video popularity in terms of total views
- Distribution of viewing rates

2.2.2 Popularity Evolution

- Characterising changes in the ranks across weeks
- Impact of video attributes like age, comments on the popularity of a video
- Relationship between short term popularity and long term popularity
- Quantifying inequality in popularity of videos
- Specifying contributions to this inequality
- Defining a growth model that explains future views in terms of present attributes of videos

2.2.3 Social Network Effects and Content Deletion

- Distribution of upload frequency and identifying the growth patterns in video uploads
- Social network distribution on the site in terms of fans or friends
- Impact of social network on the popularity of videos
- Growth of a user's social network
- Distribution and properties of deleted videos

2.3 Data Sets

We have discussed the video sharing services that we look at in our work. We collect a data set from each of the five sites we outline above i.e. Dailymotion, Yahoo! videos, Veoh, Metacafe and YouTube. Our data sets contain well over 2 million videos, with approximately 70 billion views in total. In addition to we trace a million YouTube videos weekly for over a period of 8 months, during which we have witnessed over 14 billion new requests to videos. We present our data sets in more detail in subsequent chapters.

Chapter 3

Related Work

In this chapter, we introduce the reader to prior work on related topics and how it ties in with ours.

3.1 Power Law, Zipf and Pareto

Power law is a polynomial relationship that exhibits scale invariance, while Zipf's law is a power law that models frequency of use with its popularity rank. Zipf's law originates from George Kingsley Zipf who noticed that the distribution of words in a text follow a certain statistical pattern. It states that the frequency of an object is inversely proportional to its rank. Zipf's law has been applied to many areas in social sciences, specifically video-on-demand. Pareto's law or the '80-20 rule' is usually given in terms of the cumulative distribution function (CDF), i.e. the number of events larger than x is an inverse power of x. Clauset *et. al.* [22], Newman [41], Adamic and Huberman[4, 5, 33] discuss these in greater detail.

3.2 Traditional Media Workload and Online Access Patterns

Significant effort has been applied into understanding traditional media and Web server workloads with focus on video popularity and locality of access[1, 2, 8, 47]. Cheshire *et. al* and Veloso *et. al* [21, 49] compares aspects of live and stored video streams. Huang *et. al.*[32] have analysed popularity distribution and user behaviour on MSN video.

Golder *et al.*[29] have looked at temporal access and social patterns in Facebook, while Nilsen [42] examines analysis of news-on-demand characteristics and client access patterns.

3.3 Web-based Video Sharing

There have been numerous works investigating the phenomenon of video sharing in recent times. Most of them have focused mainly on YouTube for their observations and are based on either crawling [18, 20, 31] the site or collection of YouTube specific traffic at university networks [26, 27, 55]. Throughout this work, we refer to results on YouTube workload whenever appropriate.

Various properties of YouTube Flash video files have been examined. For example, the encoded bit rate of a large fraction of the videos was found to be between 300 and 400 Kbps as reported by Gill *et. al.* [26]. Duration of videos in YouTube as an aggregation of different normal distributions has been investigated by Cheng *et. al.* [20]. A typical video on YouTube is around 3 to 5 minutes long, illustrating that short videos are the norm, and on average between 8 to 10 MB in size [26, 55]. We validate some their results in our work.

Video popularity and file referencing behavior have also been studied in previous literature [18, 20, 26, 55]. For example, using meta-data on videos obtained by crawling YouTube's science and entertainment categories, Cha *et al.* [18] find that the Pareto principle applies to the total views since a video's upload. Gill *et al.* [26] and Zink *et al.* [55] find that the Pareto principle applies weakly to YouTube video accesses as seen at the gateway of their respective university networks. The applicability of Zipf's model to the number of video views (references) has also been considered. Crawling-based techniques show that the number of video views since upload follows Zipf-like behavior with cut off [18, 20], while edge-based analysis finds a Zipf model to be reasonable for video accesses [26].

The "fetch at most once" model [30] has been suggested as providing one possible explanation for deviation from Zipf-like behavior [18]. It has also been suggested [20] that recommendation systems and common crawling approaches (that start from a list of the most popular videos and follow links to related videos) may skew results towards popular content and weed out the unpopular content.

We believe that it is important to distinguish between popularity as measured by the total views since upload, and popularity as measured by the viewing rate; the former is considered in [18, 20], while [26, 55] implicitly consider the latter. Note that Zipf models

for Web and media file popularity concern popularity as measured by referencing rate (as determined by the number of references over the fixed period of a trace), and not the total number of references to a file since its creation (e.g., see [9, 15, 54]).

There has also been some effort towards understanding how users interact with YouTube. For example, it has been found that YouTube videos in the entertainment, comedy, and music categories are viewed the most [20, 26]. Halvey and Keane [31] explored the social dynamics in YouTube's video sharing service based on meta-data obtained by crawling. They found that most users do not form social networks and only a small number of users post comments, ratings, and use other interaction tools. However, it does seem those people who do use the available tools have a much greater tendency to form social connections. Similar observations were made in a Reuters news article but with reference to Google video, YouTube, and Flickr [10].

Almeida *et. al.*[14] have looked at a feature unique to YouTube namely, video responses.

3.4 Miscellaneous

Our work also delves into inequality measures from the field of Econometrics. The various coefficients like Gini [28, 44]have been applied to calculate, for example, the welfare function [44]. The Theil [23] index on the other hand has also been applied to varying fields such as wine marketing [51].

We use dissimilarity metrics similar to Cha *et. al.* [19], that have previously used Spearman [45] rank correlation coefficient to calculate dissimilarity for changes in ranks of channels in a deployed IPTV system.

Chapter 4

Tools

Since we are analysing content sharing on the Web, a bulk of our analysis rests on crawls of the corresponding sites over a period of time. We also require significant data analysis capabilities and tools for storing them. In this chapter, we outline details on our requirements, and the tools we choose for our work along with the environment within which we use these tools.

4.1 Requirements

The questions we posed in the preceding section require us to substantiate them with empirical evidence. Thus, to provide this empirical data we conduct site wide crawls on all the sites we described in Chapter 2. To enable us to crawl efficiently we require tools and libraries that allow us to send HTTP requests for Web pages and also extract a very specific set of data among the pages received. We also aimed at gathering a large representative sample, thus some parts of our crawl had to be conducted in parallel. For this purpose, we required significant network capabilities.

Another important requirement was the storage of data. The data storage options should allow us to store vast amounts of relational data and also perform operations like filter, sort easily. Thus, we chose to go with an open source relational database management system.

After our data collection was completed for a site, we required data analysis tool to allow us to explore relationships between attributes of videos. Our study ranges from correlation analysis to building regression models. Thus, we used data exploration tools with the required capabilities. We describe our exact tool set below.

4.1.1 Python, Mechanize, urllib2

The foremost reason for choosing Python [43] was our familiarity with most of its features. Since one of our main requirements was data collection, we knew that it also comes with an impressive number of libraries for sending resource requests over the Internet. The primary library we used was *urllib2* for sending HTTP GET requests. The ease of use was also a major factor in choosing this library. On some occasions, we also used the Mechanize [36] library for crawling. Mechanize comes with better cookie handling, observance of robots.txt and convenient parsing features.

4.1.2 MySQL

We decided to use MySQL [40] as our primary data storage option for numerous reasons. The first being that it has an open source version and is free to use. It also comes with a huge install base, with appropriate documentation. We also found impressive array of libraries for using MySQL with any language of our choice. Thus, to afford us the flexibility to use any data analysis tool with our main data storage, we concluded on using it. The version we used was MySQL 5.0.

4.1.3 SQLite

SQLite is a lightweight embedded database engine. During data collection by our crawlers, the attributes of videos were stored directly in SQLite [46] databases instead of text or binary files. We preferred this method, because it allowed us to easily perform operations and check status of our crawl in greater detail. We also did not have to create our own file formats to store data, which made the data highly portable and easy to use. After the crawls were done, we imported the data into MySQL tables for analysis.

4.1.4 Matlab and R

Matlab [35] is a numerical computing environment and programming language. It comes with a wide variety of toolboxes for performing statistical analysis. We used Matlab to for regression and correlation analysis. It also comes with graphing capabilities that allowed us to explore the data interactively.

R [48] is a language and environment for statistical computing and graphics. As much of our work consisted of statistical analysis of content and popularity, this was a natural choice for us. We used it in conjunction with Matlab for most of our analysis.

4.1.5 Environment

Our data collection requirements were quite significant. We required multiple crawls to be run in parallel. For this purpose, we used a set of 15 servers at the University of Sasketchwan, Canada for conducting all our Web crawls. The server environment were all GNU/Linux based. Our crawl tools and libraries were then set up on the server and replicated. The collected data after a finished crawl was then downloaded on our personal computers for analysis, which constituted our main work environment.

Chapter 5

Methodology and Data Collection

In order to answer our questions on invariant trends across different video sharing services on the Internet, we needed meta data for a large representative set of videos from various file sharing services. Towards meeting the objectives of this work, attributes of videos from five different video sharing services were collected. This was done by obtaining a snapshot of a random sample of videos from each of the five different sites mentioned in chapter 2. We use the snapshots from Dailymotion, Yahoo! videos, Veoh and MetaCafe described in this chapter to carry out a static characterisation of our different video services. To answer question on evolution of videos we use our data set from YouTube of over a million videos, which we tracked for over 8 months in total. We also use this data set to validate previous results and compare them against our own results obtained from the other data sets. Our crawl strategies and description of meta data for our five data sets are described in more detail in this chapter.

5.1 Video and Uploader Attributes

Each video on these sites has a number of attributes associated with it like the number of views, comments etc. These attributes describe the popularity of these videos, its age or even its content among other things. Each of these uploaded videos is also associated with its uploader. Thus, for some of the sites we were also able to capture the status of the uploader. We begin by describing all the different types of attributes of videos and their uploaders that are available on such sites, below:

Video Identifier (Vid) - Each video on these site has a identifier that allows us to identify that video uniquely on the site. For example, Dailymotion uses a combination of the title and a 6 digit alphanumeric characters at the beginning to

form the identifier, whereas Yahoo! videos uses an integer to determine the video. We describe formats separately with examples for each of these sites later in this chapter.

Source Identifier (Sid) - Username/handle of the user that uploaded the particular video.

Duration - The video length mentioned on the pages were in the format *mm:ss*, which were subsequently converted into seconds for processing.

Age - On some of the services like Dailymotion, the date of upload was mentioned on the video page. This enabled us to accurately determined the age of the video from the time of capture. On Yahoo! videos on the other hand, the ages of video are mentioned as number of days/weeks/months ago, which makes it impossible to accurately determine the age of a video at a smaller granularity. We converted the age description into the nearest number of days to give us an approximation of how old the video is.

Rating - The users of the sites are also allowed to rate the video based on how they liked it. Thus each video also had a rating attribute which could be a value out of a maximum of five stars.

RatingCount/Votes - This denotes the number of people that have rated a particular video. The term 'Rating Count' and Votes denote the same measure in our work.

Views - The sites also display the total number of requests that a video has got since upload. This allows us to get a sense of its popularity profile.

Comments - The users on these sites are also allowed to create discussions around videos. The users can comment and reply to each other. Thus the comments attribute denote the total number of comments received since upload.

Favourited - A user is allowed to mark a video as one of his favourites. Some of these sites also depict the number of people that marked the video as a favourite.

Tags - The content of the video can also be described by using common keywords. This attribute lists all the keywords that were used to describe a particular video. We use the term 'Tag Count' to denote the number of tags a video has.

Channel/Category - Each video on these sites is classified into different categories or channels like Entertainment, Music etc.

Videos - This is an uploader attribute. This denotes the number of videos the particular uploader of a video has uploaded since he joined the site.

Fans - On sites like Dailymotion and YouTube, you can construct a one way relationship where you follow someone else and their uploads. On YouTube, the channel subscription is one such relationship, where a user can subscribe to another users page and be update of all new content uploaded by him/her. We term all such uni-directional relationships as a 'Fan' relationship. This attribute then denotes the number of fans any particular uploader has.

Friends - One can even form bi-directional relationships on such sites. This attribute indicates the total number of friends any uploader has since he joined the site.

We listed some of the most common attributes were were interested. Due to site specific issues we did not get all of the above attributes from all sites. We do list the attributes we obtained from each site separately in their corresponding sections ahead. In further sections we discuss about how we acquired data from each of the sites.

5.2 Dailymotion Data Collection

Videos on Dailymotion are divided into many categories such as 'Music', 'Action', and 'Humor'. Videos within a category can be found under several sub-categories such as "most popular", "most recent", "most viewed", and "most commented". These sub-categories sort the videos of a certain category according to the criterion chosen for listing. We performed experiments to determine whether or not these sub-categories produce mutually exclusive video listings, and found that there are substantial overlaps. In particular, it appeared that "most recent" and "most popular" gave us access to a large fraction of the videos in the other sub-categories. For this work, we crawled the 'most recent' and 'most popular' listings under the 'music' category, which consisted of approximately 90,000 and 50,000 pages, respectively, with each page listing 14 videos. Dailymotion has since changed its site layout; currently, only a listing of 100 pages is accessible. For each of the sites analysed in this work, customised crawlers, written in Python, were designed to obtain meta-data associated with videos such as the number of views, number of ratings, and number of comments. Requests issued by the crawlers were spaced in time to limit overloading the services. The crawlers did not download any videos and did not contribute to the view counts of the video clips; only textual information was downloaded. This approach reduced the overall network bandwidth consumption of our crawls, and also limited the load placed on the services. Only publicly available information is retrieved from the services. The meta data by these crawlers were directly stored in embedded SQLite [46] databases, which was subsequently used for processing. Our crawler downloaded the HTML source code of each video page

and performed string matching operations on it using regular expressions to extract the meta-data and obtain per-video statistics. For each video, we collected the following information: video identifier, uploader identifier, number of views, time of upload, video duration, number of ratings, average rating, and number of comments. Table 5.1 depicts the database table schema we used for storing the meta data for each video. It also shows the attributes that we captured for Dailymotion.

TABLE 5.1: Dailymotion table schema.

Column	**Type**	**Example**
Vid	String	x3a956_bliss-cover_music
Sid	String	Julien1843
Duration	Integer	242
Age	Integer	172800 (minutes)
Rating	Float	3.0
RatingCount(Votes)	Integer	1
Views	Integer	105
Comments	Integer	2
Favourited	Integer	0
Category/Channel	String	music

We crawled Dailymotion twice, first on 8 March 2008, and again on 22 March 2008. Except for the analyses of video popularity, the aggregate from these two traces was used. For videos found in both crawls, meta-data from the latter crawl was used. Table 5.1 depicts the attributes that we captured as meta-data for each Dailymotion video. Each of these attribute served as a field in our SQLite table. In total we captured over a million unique videos, that had been requested over a billion times in all.

5.3 Yahoo! Video Data Collection

Yahoo! video lists 20 categories ranging from 'Action' to 'Travel'. Each category listing contains 100 pages, with each page consisting of 20 videos. For each video on a page, the following four pieces of information were visible on the browser: video identifier, uploader identifier, video duration, and the number of views. Initially, we tried obtaining these fields using a modified version of our Dailymotion crawler; however, we noticed that the attribute fields of the videos were missing from the fetched pages' HTML source. These missing fields were being set after the page fetch, through a client-side Javascript, which our simple page crawler failed to replicate.

To overcome the above problem, we crawled each category by using a customised Firefox [39] client that could automatically browse through pages by clicking on links and

that could automatically save the contents of the page to disk. Specifically, our customised Firefox used a macro recorder called AutoHotKey [11] to browse pages of interest, and a Firefox extension called AutoSave [12] to automatically save the fetched page to a file on disk which we subsequently processed to obtain the relevant fields.

Our automated Firefox client provided us with data on approximately 50,000 videos. Our experience with Yahoo! video is that not all videos available on the site are displayed under categories. By using Yahoo! video's search engine, we discovered additional videos. Our customised client was used to discover additional videos by searching for English words from an online dictionary [1], discovering in this process approximately an additional 50,000 video identifiers. Once we had the list of video identifiers, the respective video pages were fetched for further details such as the number of ratings, average rating, and the time of upload. We noted that the individual video also issued asynchronous requests were used to fetch some of the information of interest to us. Using `Ethereal`, we identified the URL pattern of these extra requests. We wrote a Python script to issue these requests. The data returned was in the JSON format, which we processed to obtain the necessary statistics.

Table 5.2 depicts the table schema and attributes captured for videos from Yahoo! videos. Our Yahoo! data set doesn't contain the number of people that marked a video as among their favourites. Initially we captured more than a 100,000 videos but due to missing fields we had to disregard some of the captured videos. Our data set still contains 99,207 videos, requested over 770 million times.

TABLE 5.2: Yahoo! videos table schema.

Column	**Type**	**Example**
Vid	Integer	88
Sid	Integer	42
Duration	Integer	51
Age	Integer	390 (days)
Rating	Float	3.2
RatingCount(Votes)	Integer	10
Views	Integer	17613
Category/Channel	String	music

We obtained the list of video identifiers over a period of 3 days from 13 to 15 March 2008. The additional data was obtained on 17 March 2008 using our Python script.

[1] http://personal.riverusers.com/ thegrendel/enable2k.zip

5.4 Veoh Data Collection

Veoh associates uploaded videos with channels. We automated download of all channel pages, and extracted the video identifiers of all videos listed on each page of those channels. The URL structure of the site is fixed, thus we were able to pull the individual video pages which contained their attributes using just their identifiers and channel name. Some of the channels were skipped in our crawls because of a family filter scheme. Furthermore, duplicates of videos found to be listed on multiple channels were pruned. Our Veoh data collection was initiated on 18 March 2008. Table 5.3 depicts the table schema and attributes captured for videos from Veoh. Our Veoh data set does not contain favourites, or comments. Our Veoh data set consists of over 250,000 videos.

Table 5.3: Veoh table schema.

Column	**Type**	**Example**
Vid	String	v1467483zJSHGRYR
Sid	String	jason200
Duration	Integer	1409
Age	Integer	172800 (minutes)
Views	Integer	5246
Rating	Float	4
RatingCount(Votes)	Integer	13
Category/Channel	String	–familyepisodestowatch

5.5 Metacafe Data Collection

Similar to Veoh, Metacafe also associates uploaded videos with channels. The users are allowed to create their own channels. Thus, channel is a more prominent feature in Metacafe. So we used a similar strategy and automated the download of all channel pages, and extracted the meta-data of videos listed on each page. Some of the channels were skipped here as well because of a family filter scheme. Furthermore, duplicates of videos found to be listed on multiple channels were removed. Our Metacafe data collections were initiated on 2 April 2008, respectively. Table 5.4 depicts the table schema and attributes captured for videos from Metacafe. Metcafe did not show favourites, or rating count on their pages, thus those attributes are not in our data set. We managed to capture 239,000 approximately. The details of all our data sets is summarized in Section 5.7.

Table 5.4: Metacafe table schema.

Column	Type	Example
Vid	Integer	967769
Sid	String	Free+Minds+TV
Duration	Integer	1765
Age	Integer	113760 (minutes)
Rating	Float	2.06
Views	Integer	633
Comments	Integer	0
Category/Channel	String	–familyepisodestowatch

5.6 YouTube Data Collection

The preceding sections, provided details on four of our five datasets. Most of these previous data sets were based on a single snapshot, except Dailymotion. To answer questions on evolution of video popularity we needed multiple snapshots of a data set. Thus we gathered an initial set of YouTube videos which we tracked for multiple weeks. This enabled us to watch the growth profile of videos at the granularity of a week. The results from this data set are provided in Chapter 7 seperately, where we also compare it to our other data sets.

Data was collected from YouTube in two partially overlapping phases. In the inital first phase, we identified a set of approximately 1 million videos and collected metadata (e.g., number of views, ratings, and comments made thus far) related to these videos. We refer to this initial set of information as our *seed*. In the second phase, we collected metadata for the videos in the seed on a weekly basis for a period of 34 consecutive weeks. Both phases relied on scripts developed using YouTube's API .

When collecting our initial seed, our aimed to gather a data set that had the following properties:

1. The data set should be representative of the videos of various age groups found on the site; i.e, it should contain samples of old as well as new videos.

2. The recent videos collected in our seed should capture the upload trends across an entire week.

3. The data set should be sufficiently large, yet manageable to allow capture of weekly snapshots without significant load on the servers and our measurement infrastructure. The seeding process is discussed next.

The YouTube API provides a call that returns details on 100 recently uploaded videos. Using this API, we collected metadata on approximately 70,000 videos. The API also

allows retrieval based on keyword searches. To expand the set of videos to be monitored, with a representative set of old videos, we performed keyword searches using words chosen randomly from the same dictionary we used above. As search results for some words return a very large number of videos, we added only the first 500 videos returned by each API call, so as to not skew our data sets towards popular words. The initial seeding phase was from 27 July 2008 through to 2 August 2008. Table 5.5 provides the table schema that we used for storing the data set and also lists all the attributes we used for analysis.

TABLE 5.5: YouTube table schema.

Column	**Type**	**Example**
Vid	String	Gs8oCZk3RJs
Sid	String	brapetur
Duration	Integer	364
Age	Integer	10194 (minutes)
Rating	Float	NULL
RatingCount(Votes)	Integer	0
Views	Integer	68
Comments	Integer	0
Favourited	Integer	0
Category/Channel	String	Entertainment
Tags	String	brapetur,cabrera,grupo,jeanette,jueves
Tag Count	Integer	5
Videos	Integer	44
Fans	Integer	3
Friends	Integer	0

In the second phase, we collected metadata once per week. Using the timestamp at which the meta data for a video was first captured, we ensured that subsequent snapshots are exactly one week apart. For example, if a video was identified on Tuesday evening of the seeding phase, then each weekly measurement for this file was performed as close to the same time as possible, on Tuesday evenings, in the following weeks. This form of staggering allowed us to track a large number of files without exceeding rate limitations, offload our own resource usage, and at the same time offset any diurnal request patterns that takes place within a week. Our data set contains an additional 34 snapshots, taken once per week, from 3 August 2008 to 29 March 2009. We summarize the data set at a high level in the next section.

5.7 Summary of Data Sets

Our entire analysis, rests on empirical analysis based on 5 data sets from different sites. We had multiple snapshots for YouTube, thus we use that data set for our growth and evolution results. The 4 other data sets we use to find invariants across video sharing

services. We also perform the same set of experiments for YouTube videos seperately in Section 7.

5.7.1 Static Characterisation

TABLE 5.6: High-level summary of four of our data sets.

Item	Dailymotion	Yahoo!	Veoh	Metacafe
Category	Music	All	All	All
Total videos	1,194,186	99,207	269,531	239,250
Total views	1,794,790,877	770,066,629	587,729,318	3,075,778,864
Median views	210	884	283	408
Maximum views	2,895,396	4,051,080	2,387,554	9,747,625
Videos with no/one views	615/1,386	938/246	1,779/1,908	2,193/2,274
Total votes	4,525,481	1,340,713	1,240,094	—
Median votes	1	0	1	—
Maximum votes	3,814	10,535	502	—
Videos with no/one votes	427,695/256,602	54,232/12,406	115,692/41,116	—
Total uploaders	199,108	31,560	20,874	29,256
Uploaders with only one upload	93,533	21,037	7305	12,770
Average duration (minutes)	3.88	4.76	17.38	2.44
Median duration (minutes)	3.65	2.67	13.4	1.68

Table 5.6 summarises our static characterisation data sets. For these sites we only had only one snapshot each, except Dailymotion, for which we had two. We used aggregate of both snapshots for our analysis in Section 6. Our data sets contain meta-data on approximately 1.8 million video clips. Together, these video clips have received close to 6 billion views. From the table, it appears that videos are rated much less frequently than they are viewed, that typically videos are significantly shorter than the typical full-length movie or television show, and that a small number of uploaders account for a large fraction of the video uploads.

We note that Veoh serves significantly longer duration content than the other sites. Interestingly, as we will show later in this paper, Veoh appears similar to the other services, with respect to many other metrics, suggesting that there may be "invariants" that are not specific to services with YouTube-length videos, but that also may be applicable to services with longer content. Another noticeable difference between the workloads is the fraction of uploaders that upload content more than once. With Veoh 65% of the uploaders are multi-timers, while with Yahoo only 33% uploads more than once. With Dailymotion and Metacafe the corresponding percentages are 53% and 56%, respectively.

5.7.2 Temporal Characterisation

For our YouTube evolution data set, we tracked approximately 1 million video files of various ages. Table 5.7 below summarises the data set. Overall, during our measurement period, the videos we tracked received an additional 23 billion views and 18 million comments. We focus on this data set seperately in Chapter 7, which analyses the evolution and growth of videos. We also compare some of our basic static characterisation results for the previous data sets to validate ours and other's claims.

TABLE 5.7: High-level summary of data sets.

Item	**YouTube**
Start date	Jul. 27, 2008
End date	Mar. 29, 2009
Weekly snapshot	35
Videos	1,165,381
Uploaders	682,345
Views (start)	40,095,735,434
Views (end)	64,059,865,546
Comments (start)	87,511,314
Votes (start)	106,711,240
Favourited (start)	169,374,035
Tags (start)	11,966,431

5.8 Limitations of Data Sets

Collection of meta-data of videos via crawling has some limitations that can potentially impact the conclusions drawn from analyses of such data. In this section, we discuss some of these issues, and outline how we tried to address them so as to help the reader interpret the results in the context of these limitations.

How a site is crawled can also have an impact on the resulting analyses. In particular, biases may be introduced during data collection. A typical approach is to download pages that list videos. Often, a video sharing service offers multiple video listings, each under a different category (such as most recent videos, all time most popular, most popular this week, all time most rated, most rated this week, etc.). One strategy is to crawl all such pages and prune duplicate information. We applied this approach to crawl Dailymotion and Yahoo! video. This approach does not, however, guarantee obtaining a good snapshot of the content on the site because older video clips that have not received a threshold number of views may not be listed. In addition, services may limit the number of videos listed under a category. For example, Yahoo! video limits the video listings per category to 100 pages. One can argue that content not easily accessible to

the crawlers is also not readily visible to the users of the site, and thus the information gleaned by the crawler is representative enough. This motivated us to augment our data by searching for videos using words from an online dictionary, to get a random sample.

Another potential cause for concern is the continually evolving nature of video sharing workloads. New videos are added every day, and videos get new views, ratings, and comments. This makes the task of workload characterisation much more difficult than with static content. A single snapshot may provide representative data on the general characteristics of the service at the time of data collection. To understand video popularity, a highly non-stationary attribute, in particular, we obtained multiple YouTube snapshots and also internally validated most of our claims with our two snapshots from Dailymotion.

Most evolutionary crawl strategies require a pool of items to be collected and then tracking their progress over time. By fixing such an initial pool of items, or videos in our case, we fail to capture the variations in new videos entering the system. Thus, it becomes difficult to capture the seasonal effects. For example, during elections one might expect a lot of videos related to politics to be uploaded and also viewed. Thus our seed set, might end up capturing or even missing out on such trends. Thus, we have tried to keep our YouTube data set large enough and also containing a mix of both older and recently uploaded videos over a week.

Media sharing services often employ various mechanisms for the exposure of their content. The might also label some of the content as 'featured', thus making them more accessible. Search criteria based on 'ratings','views' etc. also add more to the exposure of some videos. In fact, given the size of these sites, it is unclear if it is feasible to track when and how often any single file is featured in these listings. Thus, our crawl, for example, might not represent some older videos with less than a threshold views. These would not feature highly on search results or appear on any listings. From our analysis we observe that our data set contain a significant number of videos in each age group. The fewer number of videos more than two years old, is to be expected since these sites are fairly new. To further test for representativeness based on views, we carried out Monte-Carlo simulations, for samples of various sizes taken from our data set. For each such random sample we calculated its mean and median. We carried out such random subsampling for sample sizes as low as 1000 videos, and observed that the sample means and medians converge to the aggregate values for the entire set. Hence, we remark that all our data set do exhibit representativeness to an extent for the entire population of these sites.

5.9 Summary

In this chapter, we first discussed the attributes of videos and uploaders that we were primarily interested in. We also discussed the methods, and strategies that we used to collect those attributes for a large sample of videos. We also described a high level summary of our data sets, along with some of the limitations that we believe should be kept in mind while analysing our results. We believe that the characterisation of our extensive data sets ranging from five different sources for over two million videos, provide an important contribution to the field. Some of the topics we discuss in further chapters have not been analysed in prior works relating to video sharing in general.

Chapter 6

Invariants in Video Sharing

Previously we saw our data collection strategy and a summary of our data sets. In this chapter, we analyse the questions from Chapter 2 and present our characterisation results, with particular emphasis on the invariants and differences of four different services. Firstly, video duration is studied in Section 6.1. Sections 6.2 and 6.3 discuss how users are interacting with the different video sharing services via ratings and comments, respectively. Section 6.4 analyses characteristics of video uploaders. Detailed analyses of video popularity are presented in Section 6.5. Finally, Section 6.6 summarises our measurement results and identifies characteristics that may be considered to be invariant across video sharing services.

6.1 Video Duration

This section studies the duration of the videos found in our data sets. One problem with duration data, common across all data sets, was that a few video pages reported erroneous video durations. For example, in our Metacafe data set, we found one video for which a duration of 120 days was reported! A manual check of this video showed that it was, in fact, only a few minutes long. Similar issues were reported in the YouTube video duration analysis carried out by Gill *et al.* [26]. To minimize the impact of erroneously reported video durations on our analysis, when computing average video durations (reported in Table 5.6), videos whose reported duration was longer than 6 hours were ignored. Note that videos with reported durations less than this threshold accounted for 99.99% of all videos in each of our data sets.

Figure 6.1 presents the cumulative distribution of video duration for four of our data sets. We draw several inferences from this figure and the results in Table 5.6. In general,

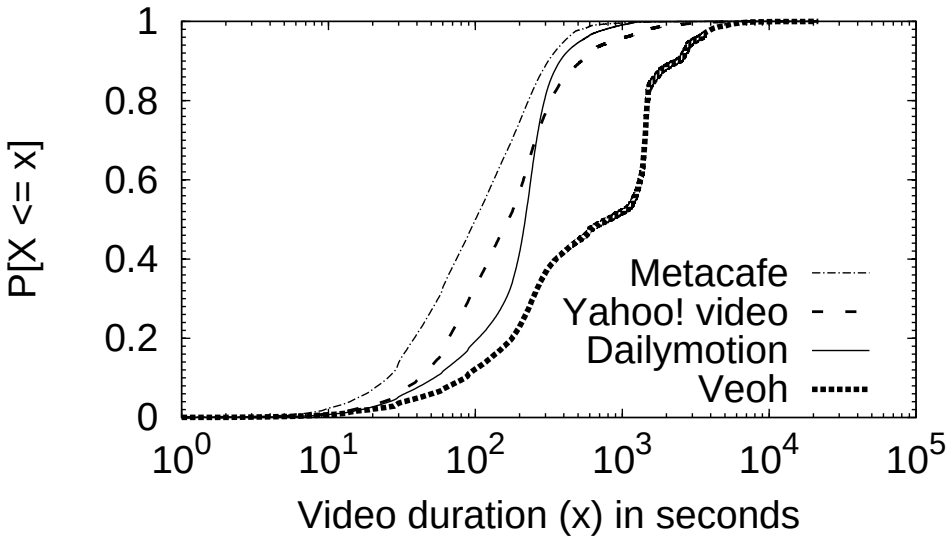


FIGURE 6.1: Cumulative distribution of video duration.

these Web-based video sharing systems are concerned with short duration videos. A typical video is between 2 and 4 minutes long for all but the Veoh service, which hosts some longer duration content from major production houses. We also find that only a very small fraction of the videos are very short; for example, less than 1 minute long. Metacafe, which is focused towards shorter-duration videos, is an exception as approximately 32% of the videos from this service are less than 1 minute long.

Note that the services we considered, with the exception of Veoh, either place explicit limits on video duration, or implicitly impose limits on duration by limiting the size of the files that can be uploaded. Only privileged users of the service are allowed to upload longer duration videos or larger video files. Therefore, not surprisingly, 98% or more of the videos in the Yahoo! video, Metacafe, and Dailymotion data sets are shorter than 20 minutes. However, Veoh does not place any limits and does host content from major studios and independent production houses. Veoh also facilitates streaming of videos longer than 20 minutes using a peer-to-peer player. We find a substantial number of videos that are longer than 20 minutes in Veoh, with approximately 30% of the videos being between 20 and 30 minutes long (the typical length of a television program). Approximately 98% of the videos in the Veoh data set are shorter than 1 hour.

6.2 Ratings and Average Rating Score

The four services that we consider allow users to assign videos an integer rating between 0 and 5, with 0 indicating low quality or satisfaction and 5 indicating high quality or satisfaction. From the services considered, we were able to collect both the rating

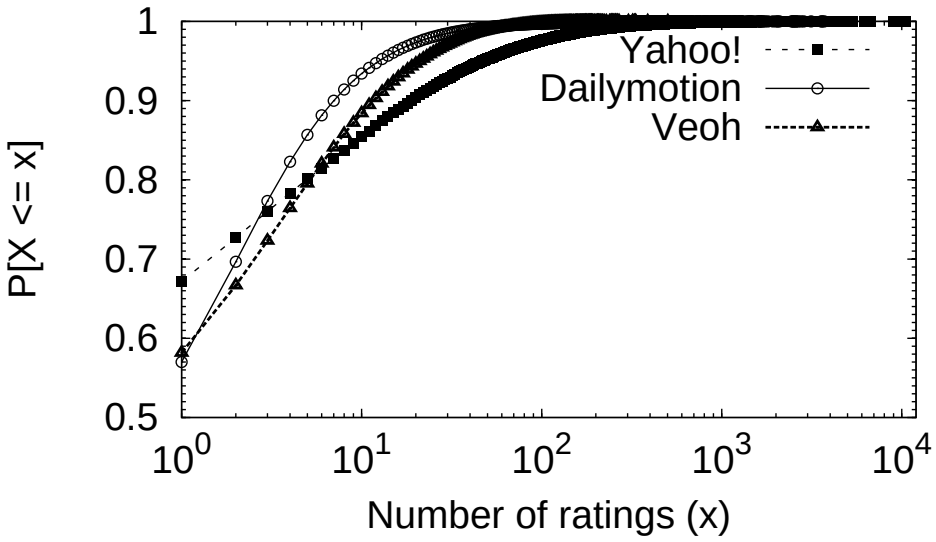


FIGURE 6.2: Cumulative distribution of rating count.

count(or votes) and average rating score from all but the Metacafe service (from which we were only able to obtain the average rating score associated with each video).

Table 5.6 tells us that videos are not rated as often as they are watched. For example, there are more than 1 billion views to videos in our Dailymotion trace, but these videos have been rated only 4 million times. Results for Yahoo! video and Veoh are qualitatively similar.

Figure 6.2 shows the cumulative distribution of the number of times videos have been rated. This figure reaffirms observations made above regarding the paucity of ratings. For example, from the figure we find that 90% of Yahoo!'s videos were rated 20 or fewer times; for Dailymotion and Veoh, the corresponding numbers of ratings are 8 and 12, respectively. Only a small fraction of views translate into ratings, and only a small fraction of the videos receive substantial (e.g., 50 or more) ratings. The number of ratings and views to a video can be expected to exhibit a strong positive linear correlation. Pearson's product-moment correlation between the number of views and ratings is 0.68, 0.66, and 0.24 for Dailymotion, Yahoo!, and Veoh, respectively. Clearly, the more a particular video is watched, the higher the expected number of ratings, however, the correlation is stronger for Dailymotion and Yahoo! then it is for Veoh.

The ability to rate videos is one of the features that enables visitors to these video sharing sites to express their degree of liking of the videos that they watch. The average of the rating scores can, therefore, be used as a metric to evaluate how satisfied users are with the sites' content. A histogram of the average rating score for the videos in each of our data sets is shown in Figure 6.3. The "NR" column represents the fraction of videos that were not rated. Recall that we did not have rating counts in our Metacafe data set.

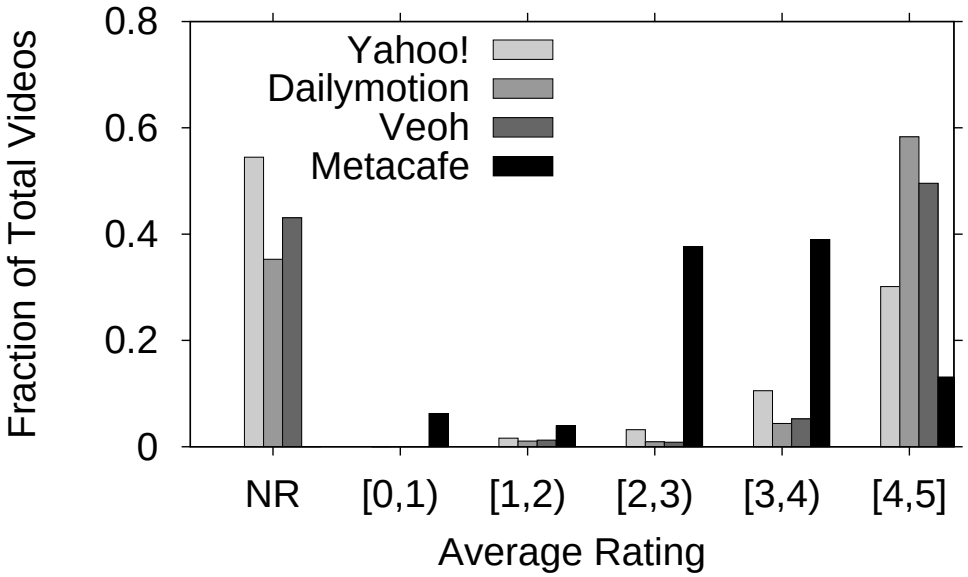


FIGURE 6.3: Average rating score of videos.

Therefore, we were not able to distinguish between videos that have not been rated so far and videos that have been given a score of zero by the rater(s). In our Metacafe data set, 14,956 videos had an average rating of fewer than one out of which 14,944 had an average rating of zero; the latter 14,944 videos are included in the "$[0, 1)$" column but these could possibly be videos that have never been rated (and thus should have been part of the "NR" column). Overall, our results may indicate that people tend to rate videos that they enjoyed watching; for Dailymotion, Yahoo! video, and Veoh, among the videos that have received ratings, we find that a majority have an average rating of 4 or higher. For Metacafe, we find that a majority of the videos have an average rating of 3 or higher, again possibly pointing towards the propensity of people rating videos that they enjoyed watching.

6.3 Comments and Favourites

Commenting on and bookmarking videos as favourites are two other features available on many video sharing services. Both the number of comments and the number of times a video has been marked as a favourite provide some indication of the level of interest a particular video has generated, with more comments/favourites indicating higher interest. These features, along with the ability to rate videos, are key Web 2.0 features offered by video sharing services.

Unfortunately, we were able to obtain the number of comments for each video from only Dailymotion and Metacafe, and the number of favourites assigned to each video from only Dailymotion. For lack of space, we do not present the cumulative distribution

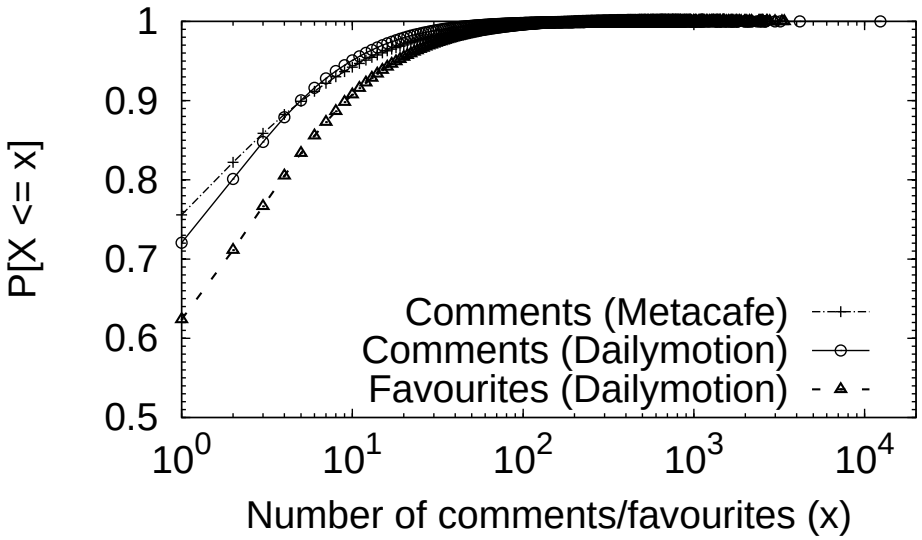


FIGURE 6.4: Cumulative distribution of comments/favourites.

of comments and favourites among the videos but present some comments based on our analyses. In general, the number of comments (favourites) exhibits high variability with many videos receiving a small number of comments (favourites) and a handful of videos receiving many comments (favourites). In particular, a total of 624,885 videos, approximately 57% of the videos in the Dailymotion data set, have never been commented upon, and 95% of the videos have received 10 or less comments; however, the maximum number of comments observed for a video is 12,377. Qualitatively similar observations can be made for comments in the Metacafe data set. The favourites feature is also sparsely used. Approximately 47% of Dailymotion videos have never been bookmarked as a favourite by any user, while 16% of the videos have been bookmarked as a favourite exactly once, and approximately 85% of the videos have been bookmarked as a favourite 10 or less times. Nevertheless, the video bookmarked as a favourite the most was bookmarked by 3,338 users.

6.4 Video Uploads and Uploaders

Another important characteristic of video sharing is how frequently people publish or upload new videos. To upload videos, services typically require that an account be created, and videos can be uploaded only when the creator of the account is logged on to the system. Using the uploader identifier, we analyse the characteristics of uploaders.

Figure 6.6 shows the cumulative distribution of the number of uploads per uploader. Here we find that a significant number of uploaders uploaded only one video. In the Yahoo! video data set, approximately 67% of the uploaders uploaded only once, whereas

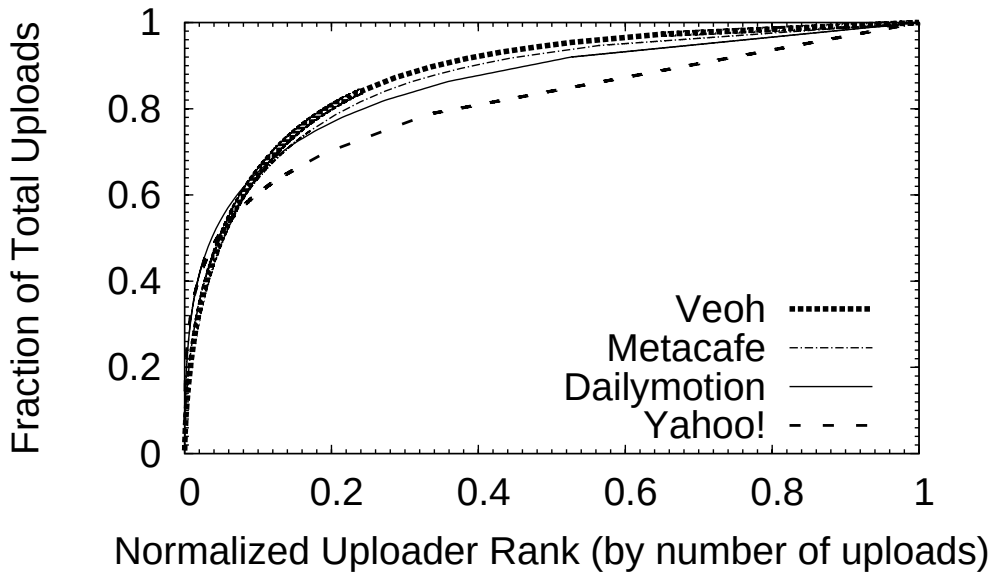


FIGURE 6.5: Skewness in the number of video uploads.

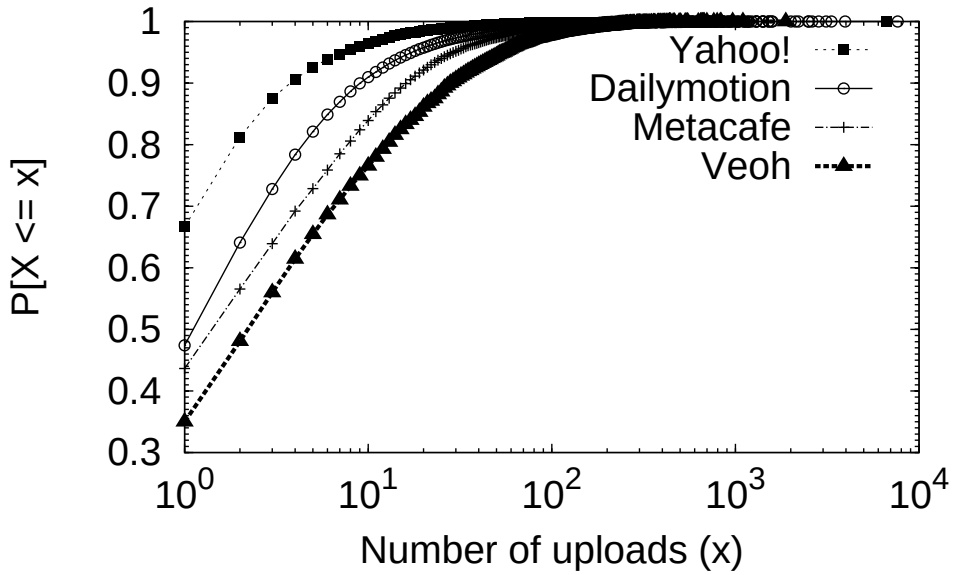


FIGURE 6.6: Distribution of uploads by uploaders.

for Dailymotion, Metacafe, and Veoh the corresponding percentages are approximately 47%, 44%, and 35%, respectively. In general, from this figure we observe that most, approximately 95% or more, of the uploaders uploaded less than 50 videos.

We also analysed whether or not the Pareto principle (cf. Section 6.5) applies to the distribution of the number of videos uploaded by each unique uploader. Our analysis suggests that the Pareto principle largely applies, with the top 20% of the uploaders accounting for close to 75-80% of the total videos in each data set.

We manually analysed the top 100 uploaders in each data set. For Metacafe and Veoh,

most of the top uploaders appeared to be independent production houses. In the Yahoo! data set, the top contributors appeared to be Yahoo! applications such as Yahoo! music, news, and health; in fact, Yahoo! music is listed as the uploader of 191,263 videos in our data set. However, our analysis did not discover any such trend from the Dailymotion workload.

6.5 Video Popularity

This section presents results concerning video popularity distributions as observed for our data sets. We distinguish two quite different measures of popularity, with differing applications and significance: the total number of views to videos since they were uploaded, referred to here as the *total views popularity*, and the rate with which videos accumulate new views, referred to here as the *viewing rate popularity*.

6.5.1 The 80-20 Rule for Video Popularity

Often we are interested in understanding how skewed the references are to the most popular videos, because presence of such skewness can have immediate positive implications with respect to the potential effectiveness of content management strategies such as caching. When discussing skewness of distributions, the "80-20 rule", also known as the Pareto principle, is often considered as it is found to be applicable in many diverse contexts. This rule, in its orginal context of wealth distribution, states that 20% of the wealthiest people account for 80% of the total wealth of the population [41]. Applicability of this rule, and the potential skewness of references, has previously been discussed in the context of references to Web servers and proxies [9, 34], on-demand streaming systems [54], and more recently in the context of video views in YouTube [18, 26, 55].

Figure 6.7 shows the cumulative distribution of the *total views popularity*; i.e., the total number of views to a video since it was first uploaded, as measured at the time of our crawl. With respect to this *life-time* metric, we find that the Pareto principle generally holds for video views; 20% of the most popular videos accounted for approximately 85% or more of the total views, in the four data sets we analysed. In general, our observations are quantitatively similar to those from a similar analysis on YouTube videos [18]. However, we note that the Metacafe service appears to exhibit significantly more skew than the other services we considered.

The total views popularity distribution is useful for understanding service features such as "all time" most popular listings, but does not provide an accurate picture of the

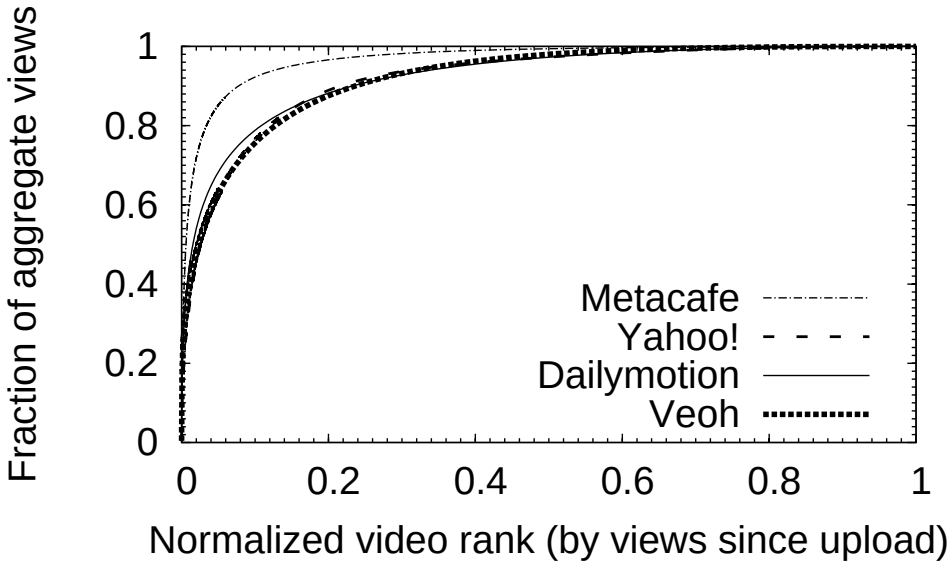


FIGURE 6.7: Skewness in the total views popularity of videos.

distribution of the rates at which videos are viewed. The latter is very important when attempting to model the video reference process, and in understanding the potential of different content distribution and caching architectures. For example, with the total views popularity metric, an older video with many views in the past may appear to be more popular than a recently uploaded video (and, erroneously, a better caching candidate) simply because the newer video has not been available for enough time to acquire more views.

We note that the viewing rate is highly non-stationary. To measure the *viewing rate popularity*, i.e., the rate with which videos accumulate new views, we resort to measuring the *average* rate over some particular time period. One approach to obtaining such a measure for a site is to crawl the site multiple times. With two crawls, the (average) viewing rate popularity of a video can be obtained as the *increase* in the number of total views between the two crawls, divided by the time between the measurements. In the absence of at least two crawls, another measure of (average) viewing rate popularity can be obtained using the *average viewing rate since upload*, which we define as the number of views received since a video was uploaded divided by the current age of the video at the time of the crawl. This latter measure removes, to some extent, the age bias in the total views popularity measure.

Figure 6.8 shows the cumulative distribution for the viewing rate popularity of the videos. The two measures of the viewing rate popularity described above are used in this figure: the average viewing rate over a two week period, specifically the time span separating our two crawls of Dailymotion, and the average viewing rate since upload for each of the services. Our results show that, with respect to the average viewing rate

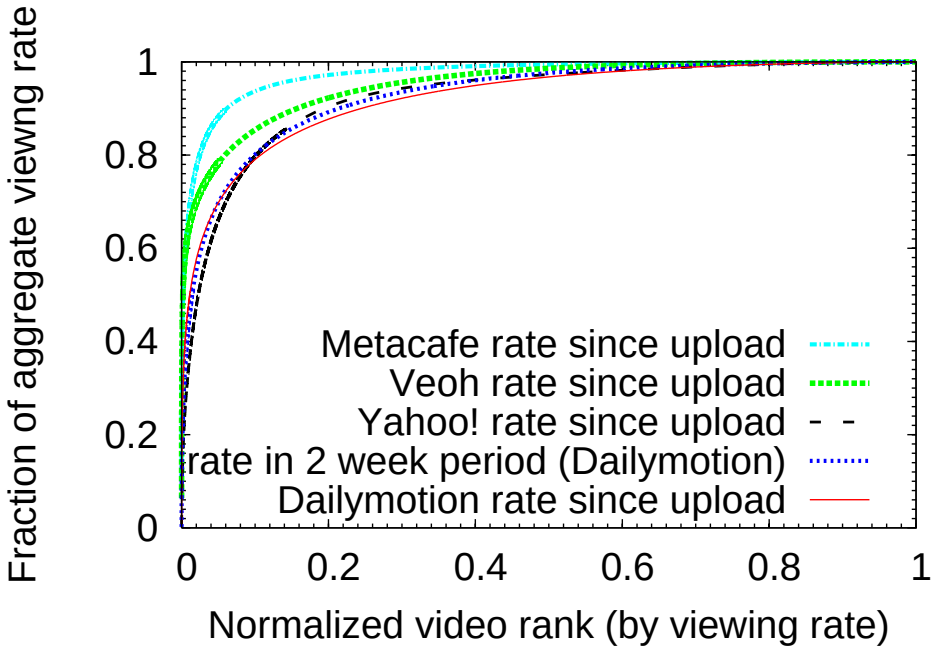
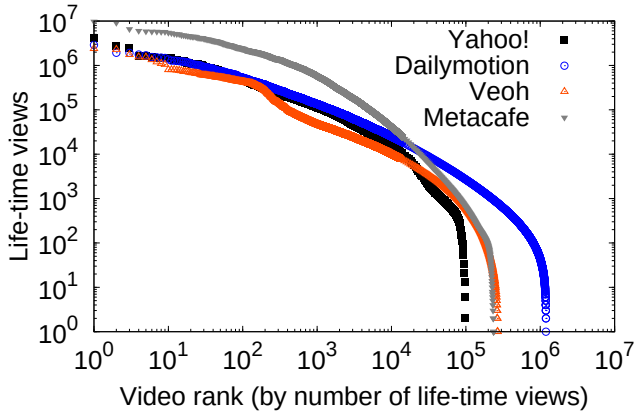


FIGURE 6.8: Skewness in the average viewing rate over a two week period (Dailymotion), and the average viewing rate since upload.

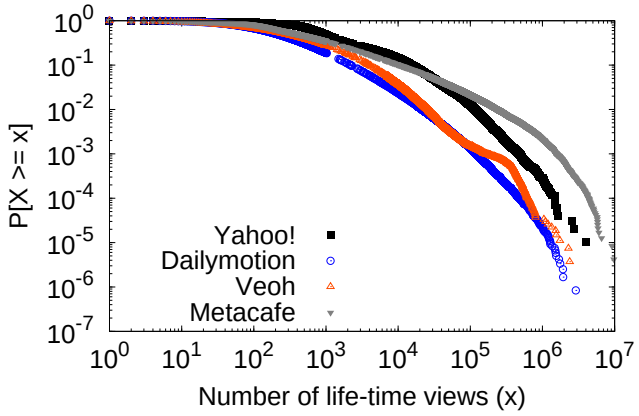
since upload, videos exhibit skewness with 20% of the most popular videosby viewing rate accounting for 80% or more of the total viewing rate. Similar to results for the total views popularity metric, the videos in the Metacafe data set exhibit more skewness than videos in other data sets. We also note that with respect to viewing rate popularity, and unlike results for total views popularity, Yahoo! videos exhibit more skewness than Dailymotion and Veoh videos.

The results in Figure 6.8 also show that in terms of both measures of viewing rate popularity, videos in the Dailymotion data set exhibit similar skewness properties. For the Dailymotion data, we find that 20% of the most popular videos account for close to 88% of the total viewing rate. Interestingly, these results are similar to what we observed for the total views popularity for videos, with the average viewing rate popularity measures indicating only a slightly increased skewness in video popularity. The most popular videos according to viewing rate popularity, however, may be quite different than with total views popularity. In fact, plots of average popularity as a function of age for the four data sets show that popularity as measured by the average viewing rate since upload is generally lower for older videos (particularly for Dailymotion, Metacafe, and Veoh), while the total views popularity is generally higher for older videos. Other important differences between the total views and viewing rate popularity measures are discussed in the next section.

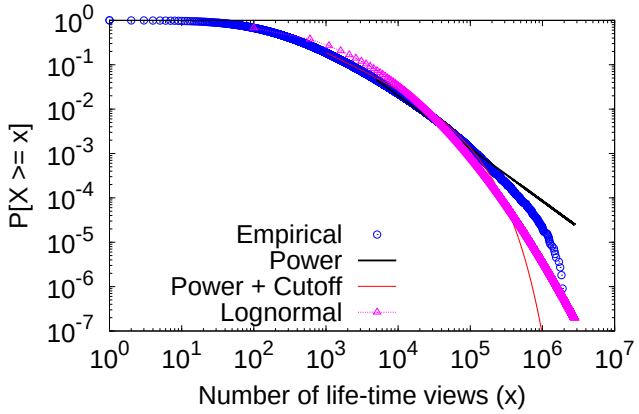
When comparing popularity distributions for user generated content with those for traditional Web and media workloads, it should be noted that the measure of popularity used in the latter context is *number of accesses to (Web or media) files over the fixed*



(a) Zipf plot



(b) CCDF plot



(c) Models (Dailymotion)

FIGURE 6.9: Distribution and models of the total views popularity of videos.

time period of the trace (see, for example, [3, 8, 9, 34, 54]), which in our context corresponds to the *viewing rate popularity*, and not to the total views popularity of the videos. Comparing our observations regarding viewing rate popularity with previous work on Web and media servers, we find that the popularities of videos on video sharing and publishing sites appear to be *more skewed* than object popularities in these other

domains. In the traditional Web domain, for example, it has been reported that typically the top 20% of the most visited Web pages account for approximately 70% of all visits to a Web server [9, 34].

6.5.2 Zipf and Power Law Analysis

Power laws can often be successfully used to describe phenomena in which "large" events are uncommon while "small" events occur frequently. A random variable X is said to follow a power law if $P[X \geq x]$ is approximately $Cx^{\alpha-1}$, where both C and α are constants; the parameter α is referred to as the exponent, shape, or scaling parameter of the distribution. A characteristic feature of power law distributions is the presence of a straight line on a complementary cumulative distribution function (CCDF) plot over several orders of magnitude when a logarithmic scale is used on both axes. In the literature on traditional Web and media workloads, for example, power laws have been found to apply to reference counts seen at Web proxies [15, 34], and media servers [8]. The presence of power law behavior in these reference streams has important implications for the design of caching systems, which may store only a relatively small number of the most popular objects (i.e., Web or media files) in the cache with the goal of improving response times and saving bandwidth. This section considers the issue of whether or not power laws can be used to describe video popularity in the four measured video sharing services. Zipf's law, an alternative characterisation of power law behavior [5, 41], is frequently used in the literature on traditional Web workloads. Zipf's law states that if objects are ranked in order of their frequency of occurrence, with the most frequently occurring object assigned rank one, the second most frequently occurring object assigned rank two, and so on, then the number of occurrences y relates to the rank of the object r as $y \sim r^{-\theta}$, where θ is the exponent of the Zipf distribution. A rank-frequency plot is often used to study Zipf-like behavior, with the presence of an approximate straight line when a logarithmic scale is used on both axes indicating the likely presence of this behavior.

While the rank-frequency plot shows most clearly the distribution of the "lukewarm" and "cold" objects, the CCDF plot shows most clearly the distribution of the "hot" and "lukewarm" objects. We use both types of plots here.

Figure 6.9 shows rank-frequency and CCDF plots for the total number of views since a video was uploaded (i.e., the *total views popularity*) for each of our data sets, along with best fit models for Dailymotion's total views popularity distribution. A number of inferences can be drawn from Figure 6.9(a). We observe that the total views popularity

TABLE 6.1: Models for the total views popularity distribution.

Data set	x_{min}	α	Candidate models
Dailymotion	1000	1.72	Power + cutoff, lognormal
Yahoo! video	10000	2.25	Power
Veoh	1000	1.76	Power + cutoff, lognormal
Metacafe	100	1.43	Power

appears to be Zipf-like for a substantial range of video ranks for all data sets, with a pronounced exponential cut off for the least popular videos.

The presence of an exponential cut off suggests that there are not that many videos that are hugely unpopular; that is, we do not observe a *long tail* of unpopular videos that would be necessary for the distribution of the total views popularity to exhibit Zipf-like behavior for the coldest videos. While sampling bias could be a contributing factor to the exponential cut off, we note that it is an invariant across all services. For each of the four data sets, the CCDF plot for the total views popularity, shown in Figure 6.9(b), has a right tail that spans four to five orders of magnitude. This is indicative of high variability in the total views achieved by the videos. The presence of skewness (i.e., a small number of videos account for a large fraction of the aggregate total views) combined with the long right tail indicates that the total views popularity distribution is heavy-tailed. Visual inspection of the graph suggests that the total views popularity distribution may have power law behavior over a portion of its range. For Metacafe, for example, power law behavior appears to exist for life-time views in excess of 100, with a drop-off for the hottest videos.

Figure 6.9(c) shows the best fit power law (key: "Power"), power law with exponential cut off (key: "Power + Cutoff"), and lognormal (key: "lognormal") distributions, for total views popularity as measured for the Dailymotion data set. It is often difficult to distinguish among the mathematical distributions that we consider in this graph, with respect to their goodness-of-fit to measured data. For example, the lognormal distribution can also exhibit a near straight line in the right tail of the CCDF plot when there is high variance in the distribution [38, 41]. From the best fit curves, it appears that no fit is qualitatively significantly better than the other fits. For example, while the middle region of the curve (e.g., life-time views between 1000 and 100,000) appears to be best modelled by a power law, the left-most region and part of the middle region (e.g., life-time views between 10 and 10,000) appears to be better modelled by a lognormal distribution or by a power law with cut off. In general, however, the total views popularity appears to be heavy-tailed. Qualitatively similar results hold for the other data sets as well.

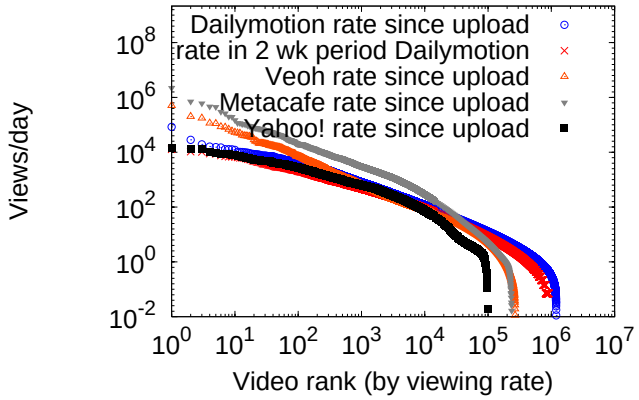
Using the likelihood ratio test [22], we compared power law fits with power law plus exponential cut off and lognormal fits. Note that this test only tells us which of the competing distributions is a better model for the data but does not tell us whether or not the winner is a good model for the underlying empirical distribution. For the Yahoo! video and Metacafe empirical video popularity distributions, the comparison indicates that a pure power law is a better model, while for Dailymotion and Veoh we find that both power law with exponential cut off and lognormal distributions provide better fits. Table 6.1 presents the power law exponent α along with the minimum value of x for which power law or power law with cut off behavior holds for the data sets; the exponent ranges between 1.43 and 2.25, and is consistent with prior observations for YouTube's science and entertainment videos [18].

Figure 6.10(a) shows a rank-frequency plot of the *average viewing rate since upload* for each of the services, as well as the *average viewing rate over a two week period* for the Dailymotion data set. Figures 6.10(b) and (c) show the corresponding CCDF plots along with the corresponding best fit models for both the Dailymotion measurements. Plots of best fit distributions for the other sites are omitted owing to space constraints. Several inferences can be drawn from the rank-frequency plots in Figure 6.10(a).

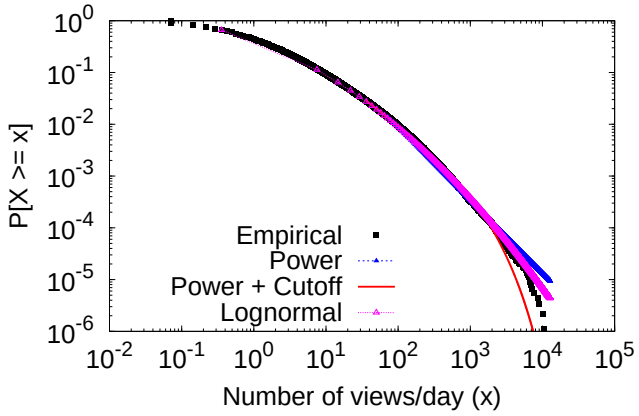
First, Zipf-like behavior is considerably more apparent in these plots than in those for the total views popularity in Figure 6.9(a). In particular, although the rank-frequency plots in Figure 6.10(a) do show a drop for the least popular videos, this drop is not as pronounced as those in Figure 6.9(a), indicating that, with respect to viewing rate, there are more unpopular videos than one would conclude from data on total views popularity.

Second, although the rank-frequency plot for the average viewing rate since upload is quite similar to that for the average viewing rate over a two week period, the latter plot shows somewhat less of a drop for the least popular videos, than does the former plot. We conjecture that rank-frequency plots for average viewing rate over an even shorter time scale such as a day, may show even more purely Zipf-like behavior than the two week average viewing rate plots. This conjecture is a topic of future work.

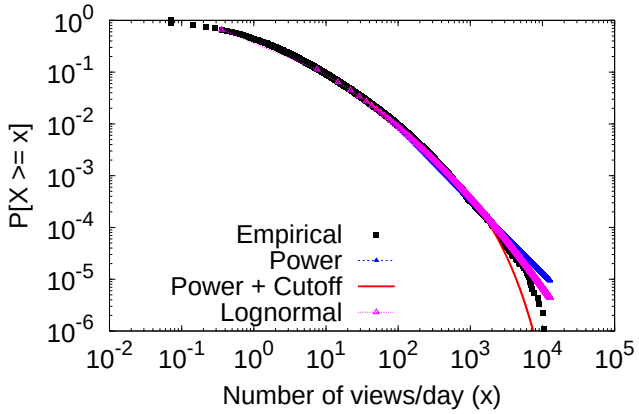
Third, while there are very few, approximately 0.01%, Dailymotion videos (for example) with only one view since they were uploaded, close to 10% of the videos had only a single view within the two week period between our crawls. The percentage of these so called one-timers, i.e., videos that have been viewed only once in a trace period, is comparable to that observed in the context of traditional Web and media servers. In the latter context, studies have reported between 15 and 75% of the total referenced objects to be one-timers [9, 34]. Some caution is necessary here. In the Web context, the percentage of one-timers are found to be roughly independent of the time duration



(a) Zipf plot



(b) Two-week average rate CCDF



(c) Average rate since upload CCDF

FIGURE 6.10: Distribution and models for viewing rate popularity.

of the trace period. Whether or not the percentage of one-timer in the video sharing context exhibits similar time invariance needs to be investigated.

Figures 6.10(b) and (c) present the empirical CCDF of the average viewing rate over a two week period and average viewing rate since upload, respectively, for the Dailymotion data set, along with the best fits for power law, power law with exponential cut off, and

lognormal distributions. These graphs can be used to explore the differences in these two average viewing rate metrics with respect to the most popular videos. Plots of best fit distributions for the average viewing rate since upload are omitted for the other services owing to space constraints. More formally, using the likelihood ratio test, we find that a power law ($\alpha = 2$) is the best candidate when modelling the distribution of the average viewing rate over a two week period for the Dailymotion data set (as it beats both a power law with cut off and lognormal fitting). A similar analysis on the distribution of the average viewing rate since upload for this data set suggests that power law ($\alpha = 1.93$) with cut off is the best candidate for that metric; similar analysis on Metacafe also suggests power law ($\alpha = 1.46$) with cut off, while for Yahoo! and Veoh it appears that both lognormal and power law with cut off are good contenders. Our analyses shows that power law and related variant distributions might be useful for modeling the tail of the viewing rate popularity distribution. In addition, we also find that for most of the tail, the two metrics for viewing rate exhibit similar behavior although they diverge for the values on the extreme right (i.e., at the most popular end) of the distribution.

6.6 Summary

The preceding sections presented multi-dimensional analyses of video sharing workloads. Our analyses found several characteristics that appear to be common across video sharing services including the YouTube service studied in prior work [18, 26, 31, 55]. We note that the video sharing services considered here span a wide range, including Veoh which serve longer content and Metacafe that uses a revenue sharing model to give uploaders incentives. Characteristics that appear to be *invariants* are summarized below:

1. Users are primarily interested in watching videos; social interaction tools for rating, bookmarking, and commenting on videos are infrequently used. Similar observations hold for YouTube [31].

2. The number of uploaders to a service are an order of magnitude smaller than the number of uploaded videos. This again reiterates that users are primarily interested in viewing videos. We found that the Pareto rule applies, with the top 20% of the uploaders contributing between 75-80% of the total videos.

3. The typical video available from these services is of short duration. This observation, based on data from the Dailymotion, Veoh, Yahoo! video, and Metacafe services, is similar to that made for YouTube [26, 55]. Some services, such as

Veoh, however, are starting to feature longer duration videos, using peer-to-peer technology to address the problem of efficiently distributing such videos.

4. Both the total number of views to videos, and the rate with which new views occur, follow the Pareto rule, with 20% of the most popular videos accounting for 80% or more of the views. Similar observations regarding the number of views since upload [18] and the viewing rate [26] have been made for YouTube.

5. The total views popularity distribution is heavy-tailed and may be modelled as power law or power law with exponential cut off, with power law exponent between 1.4 and 2.5. However, we note that neither a power law (or variants), nor a lognormal distribution, appear to fit the entire distribution well. These results are consistent with those reported for YouTube's videos in the science and entertainment categories [18]. Similarly, the total views to videos may be modelled as Zipf with cut off.

6. The average viewing rate since upload is generally more Zipf-like than the total views popularity distribution. (While we have two snapshots only for the Dailymotion data set, we note that the average viewing rate for that two week period is even more Zipf-like than the average viewing rate since upload.) The viewing rate popularity distribution of videos can be modelled as power law with cut off, with power law exponent between 1.4 and 2.

7. In contrast to traditional Web and media server workloads, there are not many "one-timers", when this term is defined for this context according to views since upload. When considering views over a fixed-length trace period, however (such as the two week period between our snapshots of Dailymotion), we expect that the proportion of "one-timers" may become more comparable to that with traditional workloads.

In the ensuing discussion, we provide some concluding remarks regarding the above observations.

Our first set of remarks concern the limited use of social interactivity tools. Given the sparsity with which these tools are used, it may initially appear unlikely that these would be useful in discovering new content and in designing video recommendation systems. However, as shown in Section 5.1, ratings can be used to discover popular content, and in on-going work we have found that the current rating count may be used to gauge the future popularity of videos. In addition, while our results show that use of social interactivity features are not pervasive, we believe that these features are probably important to those that use them, and thus might play a role in retaining the clientele.

Over time, it is possible that the use of these features will increase. Investigating this question remains a direction for future work.

Our next set of comments is about the popularity characteristics of the videos found on video sharing services. We believe that the very nature of user generated videos, and the sociological behavior of video sharers, precludes the possibility of a large percentage of the video clips having few (e.g., less than 5) or no total views since upload. Typically, video sharing Web sites have a link to videos that have been recently uploaded. A typical surfer of such sites may navigate through this listing and perhaps view some videos that are thought to be of possible interest. Another potential cause of few one-timer videos may be publicity by the creator or uploader of a video. These sites are often venues for sharing interesting videos among friends and family members. A common practice of uploaders is to send friends and family the link of the uploaded video (or post a link to the newly uploaded video in one's own social network profile page or blog space), and people in the social network of the uploader may be expected to oblige. Finally, it is possible that the uploaders themselves view their uploaded videos.

While all uploaded videos may get some views over their life-time, there can be many more videos that are viewed only once (or, of course, not at all) over a fixed duration period of days or weeks. Our analysis of viewing rate popularity found that over a two week period approximately 10% of the videos are requested only once. We also found that the viewing rate popularity distribution is more Zipf-like than the total views distribution. We further look at the evolution of popularity in the next chapter.

Chapter 7

Popularity Evolution

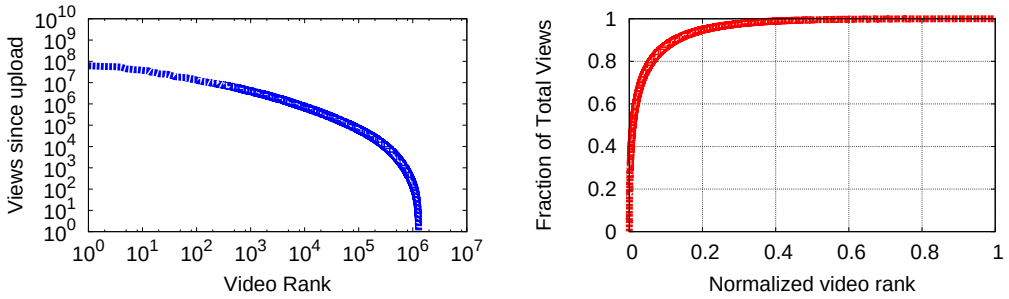
In this chapter, we outline the evolution of popularity from our YouTube data set. We first provide a static characterisation of the data set based on the *seed* snapshot in Section 7.1. We then move on to the longitudinal characterisation of requests in Section 7.2. An overview and discussion on hot-set dynamics are discussed in Sections 7.2.1 and 7.2.2, respectively. We then look at viewing rate dynamics and rich get richer models in Section 7.2.3 and Section 7.2.4. We also provide a novel approach of looking at popularity skewness and contribution to general inequality we saw in the previous chapter, by applying measures from the field of Econometrics in Section 7.2.5. Section 7.2.6 sums up our observations and provide a growth model that describes the evolution of popularity. We finally summarize our observations in Section 7.3.

7.1 High-level Characterisation and Validation

To validate our static characterisation claims in the previous chapter we next provide a high-level characterisation of the YouTube set captured during the seeding period, separately in this section. The data is also validated and compared with previous characterisation studies that have analysed YouTube or similar systems. In addition, we comment upon more general properties associated with this data set that may be helpful when interpreting our results in the later sections.

7.1.1 Activity Distribution

Focusing on video popularity, we note that previous studies [18, 20, 26, 55] have found that various Zipf-like models can be used to capture the number of video views. For



(a) Life-time views (at *seed*)

(b) Skewness (at *seed*)

FIGURE 7.1: Initial file popularity.

example, crawling-based studies have found that the number of views since upload follows Zipf-like behavior with cut off [18, 20], while edge-based Campus-network analysis of the access frequency to different YouTube files roughly follow a Zipf distribution [26]. We previously noted that these differences, in part, may be due to differences in how popularity is measured. Using measurements from four diverse set of video-sharing sites, we find that popularity as measured by the number of views within a fixed, relatively short, time period exhibits Zipf-like behavior, whereas popularity when measured by the total number of views to videos since their upload exhibits similar behavior but with cut off.

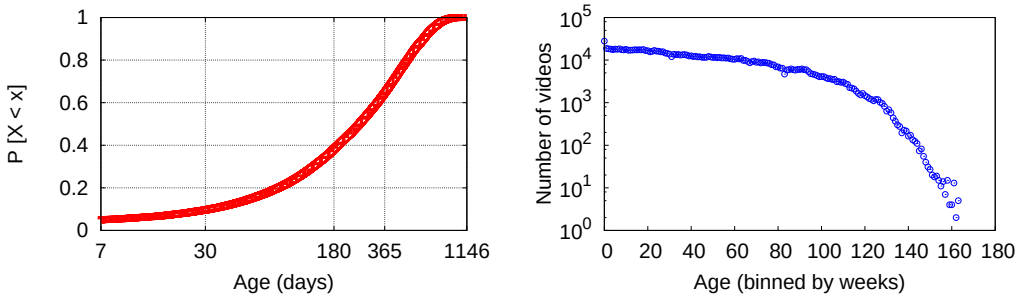
To validate the presence of Zipf like characteristics in our data set as against, these existing data sets, Figure 7.1(a) shows a rank-frequency plot of the total number of views up until the seeding time. In other words it shows shows us distribution of long-term popularity. The rank frequency plot emphasizes the number of views of the most popular files. As observed, the curve show straight-line characteristics for atleast five orders of magnitude. The plot for short-term popularity or views gathered by videos within the first week also shows a similar presence of a straight line. We note that straight-line characteristics, when a logarithmic scale is used on both axes, suggests that Zipf-like models may be appropriate to model the distribution. Our results are consistent with previous studies. For the long-term popularity, we observe a sharp (exponential) cut off for less popular files. For the number of views during a single week, the Zipf model, without cut off, applies.

Similar to previous studies [18, 26, 55], the data captured at seed time suggests that the skewness of views follow the Pareto principle (or the "80-20" rule), which states that that at least 80% of the views are contributed by only 20% of the files. As Figure 7.1(b) shows, the distribution of views is highly skewed, with 10% of videos in our data sets contributing around 80% of the total views since upload. While not shown here, we note that the Pareto principle also applies to the number of views during the first week

(*week1*). In fact, the principle holds for the number of views during any other week of our measurements.

7.1.2 Age of Videos

Fig. 7.2(a) shows the cumulative distribution function (CDF) of the number of files of different ages in our data sets. Here, the age is measured as the time since upload until the time it was first captured. The oldest YouTube and Dailymotion video found in our data were 38 and 44 months old, respectively. It is interesting to note that roughly 50% of the files are less than eight months old. The median age of the YouTube videos was found to be 256 days.



(a) CDF (all videos)

(b) Searched by keywords

FIGURE 7.2: Age distribution at seed time.

While our YouTube data sets contain a mix of very recently uploaded videos and videos that are significantly older, we note that the fraction of new videos (for example) may be skewed due to a significant fraction of files being identified through the list of recently uploaded files. We therefore believe that the upload frequency (over time) is better judged using the age distribution of the files obtained through our keyword searches. Fig. 7.2(b) shows the distribution of age of only those videos that were identified through searches by keywords. With the y-axis on log-scale, we note that this figure suggests that the upload frequency have steadily increased exponentially over the past two years with some smaller weekly variations. Furthermore, we note that there is a pronounced knee in the plot. It appear that files were uploaded at an exponentially increasing rate before this knee too; however, at that time the relative increase in the upload rate was significantly greater than in the last two years.

7.1.3 Duration of Videos

In contrast to traditional media servers, which typically serves high-quality long-duration videos, video sharing sites primarily serves short user generated videos. Fig. 7.3 shows

the cumulative distribution function (CDF) of the file duration for each site. We note that a sizeable portion (80%) of the YouTube has a duration of less than 400 seconds. A closer look at the frequency distribution function shows three peaks. The first peak occurs within one minute and contains around 20% of the videos. The second peak appears between 3-4 minutes and mainly contains music videos. Finally, the third significant peak appears at roughly 10 minutes. Closer inspection of this peak reveals that these files typically are part of longer videos that have been partitioned into smaller pieces to circumvent size limitations imposed by these sites. These files include (parts of) episodes from various television series. Finally, we note that the average duration of YouTube videos were found to be 281 seconds. Our observations coincide with previous findings on durations [20, 26].

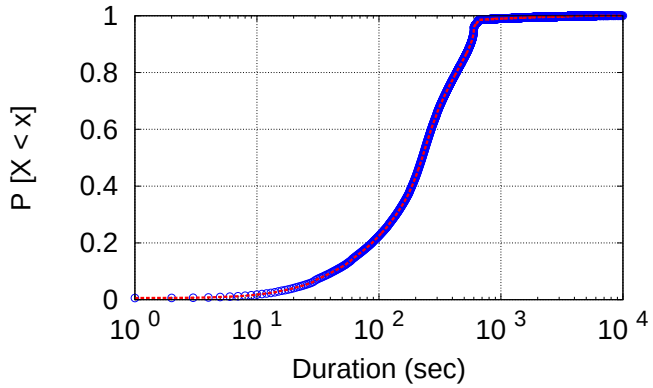
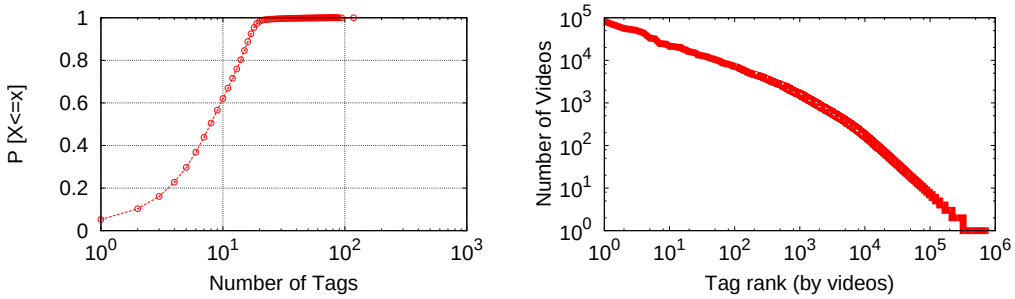


FIGURE 7.3: File durations (CDF).

Having noted that these sites typically impose file size restrictions, it should also be noted that YouTube, at the time of our seeding measurements, allowed users to upload up to ten files at a time, of a total size of 1GB. Clearly, the majority of files are much smaller than these limits. However, it should be noted that the large files still can consume much resources. YouTube now also supports uploading higher definition videos. Clearly, designing effective delivery mechanisms for these systems will be an important issue if these sites will serve more larger files, and/or the existing portion of large files start to see higher request loads than currently is the case.

7.1.4 Tagging of videos

We note that user generated content sites often allow their users to add 'tags' or descriptive words to objects. By adding tags to uploaded files, an uploader can add semantic values to a file and make it easier to find (using key words search, for example). The tag meta data was missing from our previous four data sets. Fig. 7.4(a) shows the cumulative distribution of tags found on each video in our data set. Approximately 80% of the videos were found to have ten or fewer tags. While we only report values for the initial



(a) Tags per file (CDF)

(b) Files per tag (Zipf)

FIGURE 7.4: Tags at seed time.

seed, it should be noted that 99.92% of the videos did not have any new tags added to them during the remainder of the measurement period.

Table 5.7 shows that our initial snapshot of YouTube has over 11 million tags. Removing duplicates, these numbers reduces to 790,289 distinct tags. Figure 7.4(b) shows a rank-frequency plot of the different tags. The videos are ranked based on the number of files they appear in. The presence of Zipf-like behavior is evident from the graph. The ten most frequently occurring tags contains some common English terms (e.g., 'the', which is the most common tag word on both YouTube and Dailymotion as observed) and some more descriptive words such as 'music', 'video', 'live', 'humor' and 'rock'. We further note that there was a big overlap between the sites, as well as between the ten most frequent tags and the ten tags that had the most views.

7.1.5 Ratings and Comments

A typical user on such video sharing services can interact with videos in a variety of ways. A user can comment on the video, he can also bookmark a video as a favourite, and vote on the same as well. As noted in Table 5.7, we note that users are primarily interested in watching videos; social interaction tools for rating/voting, and commenting on videos are infrequently used. Previous studies [31], along with our own observations in the previous chapter report similar findings.

We have found it interesting, however, that there is a high correlation between the amount of ratings and comments with the popularity of files, similar to our previous data sets. For example, we note that the Pearson's correlation coefficient for the number of comments with the total number of views since upload is 0.56. Thus it is evident that there is a high positive correlation between the two factors. Some of the videos also contain ratings provided by the user. The meta data captured in our crawl also has the number of people that have rated the videos. The value for the coefficient is , between rating (or votes) and seed views is even higher at 0.73.

The above observations indicate that the number of comments, votes and other interactions can provide some indication of the level of interest a particular video has generated, with more comments/favourites indicating higher interest. In addition to allowing social interactions among users, these Web 2.0 features can be used to provide supporting evidence with regards to videos that are more likely to receive future views.

7.2 Longitudinal Characterisation of User Access Patterns

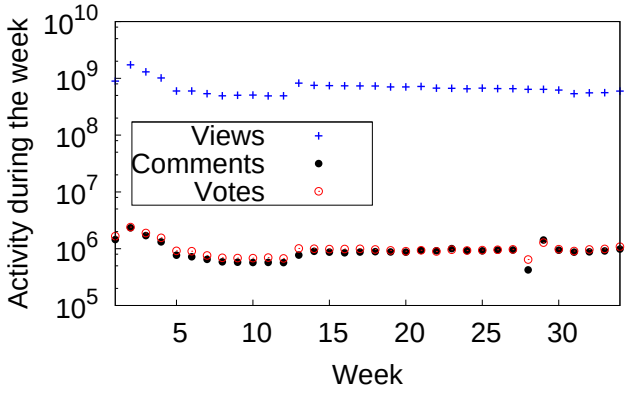
Understanding how users access the content of popular video sharing sites is an important step towards understanding the demands and systems requirements that these sites face. We saw in the previous section how our YouTube data set also had similar static characteristics in terms of video attributes, as compared to other video sharing sites. In this section we analyse trends and changes in the file popularity, as well as identify factors that impact the success of these videos. We also look at disparity in long and short-term popularity using various econometric inequality measures and their evolution over time.

7.2.1 Overview

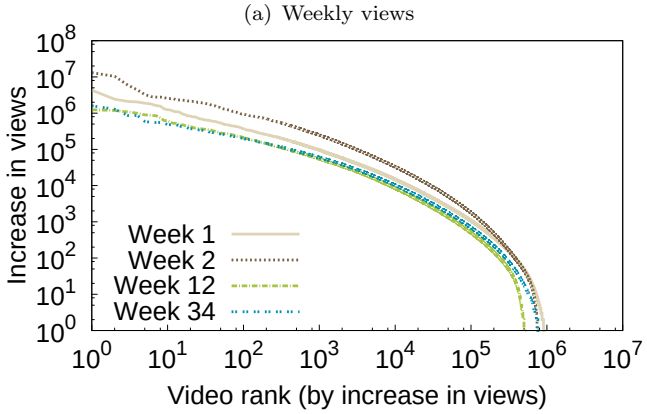
As a first step of understanding how the demand for content on these sites changes with time, we first analyse how the access pattern of the aggregate set of files (collected at seeding time) changes with time. Fig. 7.5(a) shows the number of views, comments, favourited and votes that the videos on YouTube get each week till the end of our crawl period.

We note that with the exception of the first week, the activity on YouTube slowly declines for the subsequent weeks. This observation holds true for both the set of newly uploaded videos and the set of videos including all videos. Thus, it suggests that the aggregate popularity of a fixed set of YouTube files decreases relatively slowly as the popularity shifts to a newer set of files. While it is clear that the aggregate popularity of a set of files only decrease at moderate rates, there may still be significant differences between individual videos. As a first step towards capturing how the file popularity of a set of files changes with time, Fig. 7.5(b) shows a rank frequency plot for the complete data set. Each of the curves, represent the number of views a video gets in the corresponding week. It is interesting to note that the distribution remains fairly steady across the weeks. Also, the weekly views accrued for the second week is higher for the most popular files, similar to what we observed in Fig. 7.5(a). We believe that this peak in popularity during the second week of our crawl is due to a transient variation

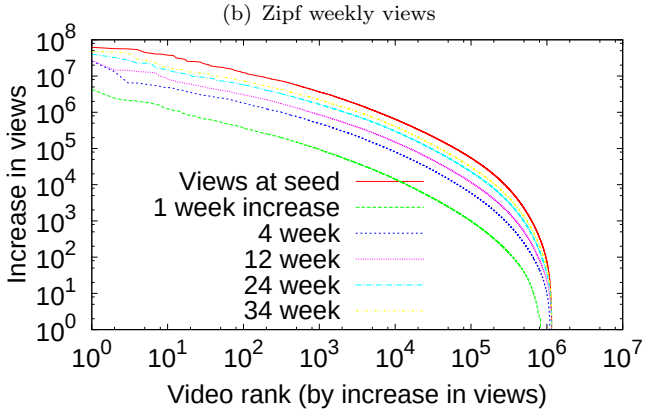
on YouTube. The daily reach of YouTube, had dropped around the middle of June and July 2008 according to figures on *Alexa*. Hence, we infer that an reasonable explanation for this could be due to an unusually low usage of the site during our seeding and first week measurements.



(a) Weekly views



(b) Zipf weekly views



(c) Zipf new views since since seed

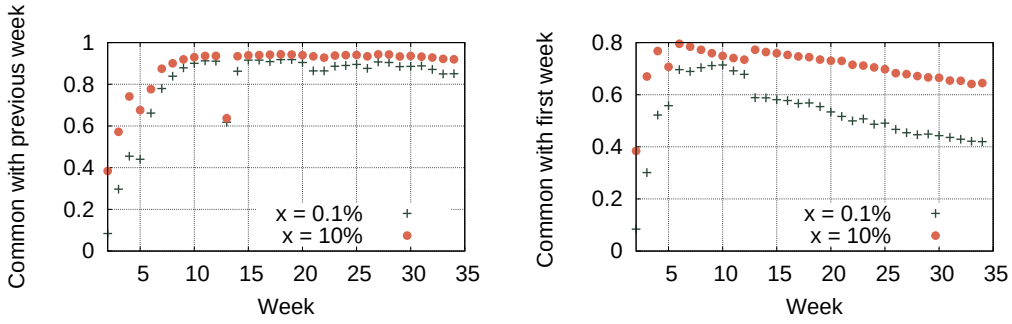
FIGURE 7.5: Aggregate Popularity Trends.

Furthermore, for the one-week old files, the shift in the rank-frequency distribution is somewhat greater, suggesting that it has a shorter period during which newly uploaded files can gain momentum. For example, while there were roughly $O(10^5)$ new files that where viewed during the first week, only $O(10^4)$ of these videos where viewed during the twelfth week. In general, we note that the life of a typical video might be very short

lived. The sheer volume of new videos uploaded every day, as well as the diverse set of avenues of information propagation (through blogs and syndication feeds, for example) add to this ephemeral nature, in which user attention quickly shift away to new object of interest.

We also wanted to observe the progression of shorter term distribution, and how it ties in with long-term. Thus, we plotted the distribution of total views (at seed) on a rank-frequency plot in Fig. 7.5(c). The figure also shows the distribution of new views gained in the first week, first four weeks, first twelve weeks and progressively longer intervals. It is evident from the same that with each passing week the shorter term popularity also start to achieve the same profile as the long term distribution. The concentration plot for new views gained every week show high degree of skewness with 10% of videos in any week accounting for around 90% of the views in that week, suggesting that the difference between popular and not so popular files is greater. They are also similar for longer intervals of first one week, first two weeks and so on, with 10% of videos in any interval again accounting for around 90% of the new views. This distribution also remains fairly constant across all weeks during our measurement. From our analysis, we infer that Pareto's rule applies to any interval long or short and is inherent to such content.

7.2.2 Hotset Dynamics



(a) Common with previous week

(b) Common with first week

FIGURE 7.6: Hot set analysis.

Having observed that the popularity remains relatively skewed, we now turn our attention to the relative popularity of files as measured by the number of views to a file in each week of our measurement period. Of special interest is the rate of change in the set of most popular files (called the *hotset*, in the following); i.e., how often do videos enter and leave this hotstet.

For simplicity, we define the *hotset* as the x% most popular files (i.e., the x% of files with the most views in the last week). This definition is motivated by a cache, for example, that only can store a smaller fraction of the files. Fig. 7.6(a) shows the fraction of files that remains the same between consecutive weeks. Fig. 7.6(b) shows the fraction of files in the hotset that also were in the set during the first week. There is a distinct transient behavior in the YouTube hotset. In particular, there is a huge change in the hotset, between week one and two (as observed by the sharp decrease in commonality from the first to the second week). In fact, after roughly six-to-seven weeks the commonality between hot sets of any two adjacent weeks hotsets stabilizes around 90%. In other words, the large change in the first few weeks quickly settles and the weekly changes stabilize at approximately 10% per week, at the end of the crawl. However, we note that the commonality with the first week in Fig. 7.6(b) steadily increases as the popularity of many of the files that gained in popularity from the first to the second week cools. In other words, by the end week 13, the hotset only differ from the original hotset by 20% to 30%. The figures provide us a glimpse into the natural selection process behind the most popular videos. The recently uploaded videos cause sort of a ripple effect and disrupt the existing hot set. Their effect subsides after a few days as expected, but also some videos in the original hot set steadily become stale and move out of the hot set and newer videos take their place. Thus we see a slower but steady decline after week 13 in Fig 7.6(b). In these figures, we have used $x = 10\%$ and $x = 0.1\%$. For example, these thresholds correspond to 789 and 83,349 views during week 1, respectively.

The set of common files clearly consist of files that are long-term popular, while the set of files that entered and left the hotset, have a much more short-lived popularity profile. To gain a better understanding of the files entering and leaving the hotset, we looked at the age distribution of hotset, and its changes with time. We note that much fewer of the recently uploaded files enter the hotset on YouTube. This might be(in part) due to us only capturing a small fraction of the total number of newly uploaded files on YouTube (as YouTube faces much more uploads per day, than what is available through the API). A factor that may cause the transient behaviour, with videos of short lived popularity, likely is caused by files being 'featured' in various YouTube listings. Although, of the set of files entering this set, only a smaller fraction is able to sustain their popularity for a longer period. We also analysed the total number of files that currently are not in the hotset, but at some other point during our measurement period have been in the hotset. The fraction of such files stabilize even faster. We observed, that by the end last week of the measurement period, for x=10%, there were 12% of total files that had once been in the hot set but were not present in the last week. The corresponding figure for x=0.1% was 0.18% of the total files.

Overall, the age distribution of the hotset is linear on a log-log scale. This suggests that the age of the files in the hotset follows a Pareto distribution. In fact, if we assume that the time each of these files have been in the hotset is proportional to their age, this observation would suggest that the time files stay in the hotset is Pareto distributed. With a Pareto distributed life-time in the hotset, a file that have been in the hotset for a year is as likely to remain in the hotset for another year, as a file that have been in the hotset for a month is to remain in the hotset for another month.

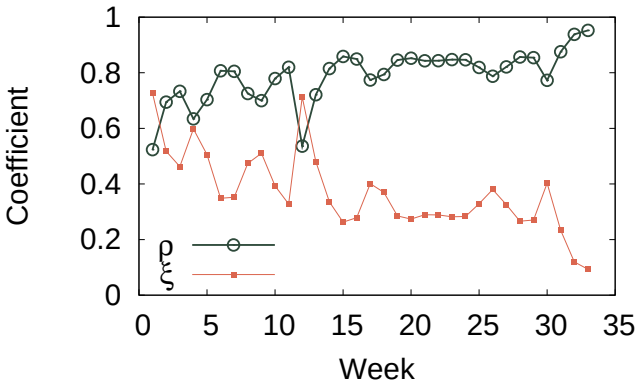


FIGURE 7.7: Rank Changes.

Our next question is how frequently the ranks of the videos change across weeks. In other words, we look at the level of dissimilarity in the ordinality of all videos. To infer this variation in ranks we use a dissimilarity coefficient, $\xi = 1 - \rho^2$, where ρ is the Spearman's correlation coefficient. Spearman's correlation coefficient compares two sets by using their rank information. the coefficient ranges from -1 to 1, with 1 denoting perfect agreement between the two sets of rankings and -1 for perfect disagreement between the ranks. A value of 0 denotes almost no correlation. Whereas, the value of ξ , varying between 0 and 1, measure the proportional variance in degree ranks unexplained by the initial set of ranks. The ρ coefficient is calculated as:

$$\rho = \frac{1 - 6\sum d_i^2}{n(n^2 - 1)} \tag{7.1}$$

where d_i is the difference in ranks across two adjacent weeks and n is the total number of videos.

To measure this variance, we gave each video a rank for each week, based on that corresponding increase in its requests. For each week, in Fig. 7.7, we took each video and compared its present rank with that of the next week to derive ρ. Since a dissimilarity greater than 0.l is considered significant, Fig. 7.7 shows that there are considerable rank changes in all, and even within the hot set. There were large variations in the rank

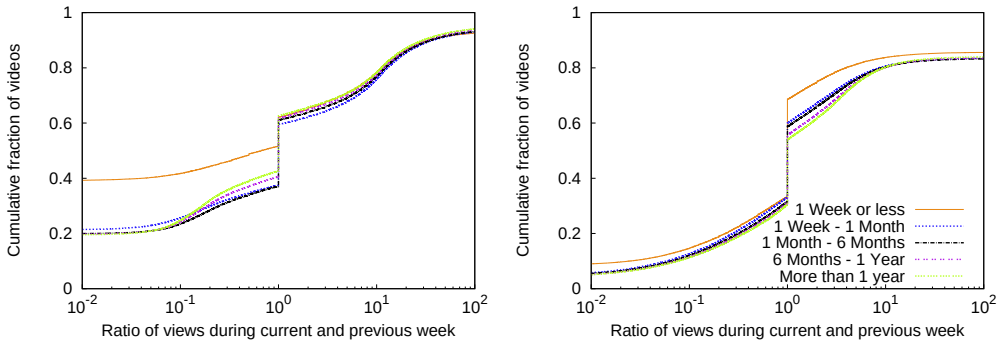
within the first couple of weeks, but after a few more weeks the variance again stabilized and fluctuated around a fairly lower value. It is evident that not only does the set of top videos change, but there is considerable upheavel within them as well.

7.2.3 Viewing Rate Dynamics

So far we have looked at popularity progression at an aggregate level, and the profile of the videos that were at the popular end of the spectrum. In this section, we will look at the evolution process at a more detailed level, the relationship of views in adjacent weeks and the impact of age on it.

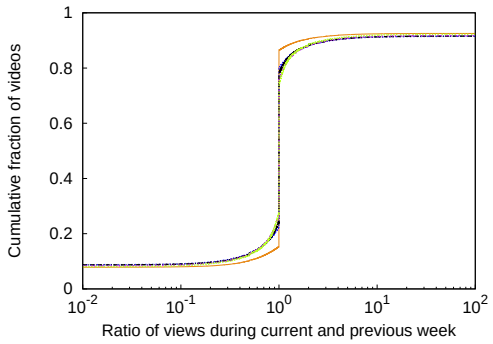
To relate the number of views in consective weeks for videos of the same age class, we define an *immediate growth function* as, the number of views during a week $W + 1$ divided by the number of views during week W. We plot the CDF of the function for data set in Fig. 7.8 for three different weeks. We note that the above definition can result in a ratio of zero or infinity. Thus we replace the occurrence of zero increase in views with a sufficiently small value between 0 and 1 while calculating the ratio. The reason behind doing the same was simply to enable us to figure out the fraction of videos that get less, more or same number of views, respectively, in the next week with respect to its current weekly views. Thus, the ratio is equal to 1, whenever the two adjacent week's views are the same. On the other hand the values are very large or very small if one of the weekly views is zero. With the x-axis on log scale, the frequency of occurrences of these extremes can be (roughly) identified by the left-most and right-most point in the figure.

From Fig. 7.8, it is evident that the behaviour of recent videos is different from the other age groups. In general, slighly more than half of the videos aged less than a week old get a smaller number of requests than in the previous week. This process is akin to a "natural selection" of sorts. It serves to illustrate the fact that among the numerous videos uploaded every single day not all of them become highly popular but half of those videos are already on a declining trend. We also note that approximately 40% of the new videos do not get any views at all. With each succeeding week, the behaviour of new videos starts to resemble the older set of videos. A point to note in this graph is that in the second weeks around 30% of the videos get the same number of views they got the previous week. Another distinct trend that we see is that for our fixed pool of videos and with each passing week, the number of videos with the ratio equal to one increases steadily from 30% of the videos, to approximately 80% in week 12. This means a lot of videos get the same number of requests that they got in the previous week and indicating that the short term views are a reasonable predictor for the future views.



(a) YouTube (week 2)

(b) YouTube (week 3)



(c) YouTube (week 12)

FIGURE 7.8: CDF of immediate growth function.

We will look at the correlation between adjacent weeks views in more detail in the next section.

7.2.4 Rich Get Richer Models

Barabasi and Albert [13], propose a model of growth in complex network, commonly know as the *preferential attachment* model, which states that any new nodes about to attach itself to another node in a network is much more likely to link itself to a node with more nodes attached already. Not surprisingly, this model has also been referred to as the rich-get-richer phenomenon. In this section, we provide empirical evidence that this model also applies to the growth of video popularity. We also quantify how much "richer" the rich become.

Consider first the impact of the total number of views since the video was uploaded in determining the files future popularity. Fig. 7.9 shows number of views gained during a week, as a function of the total number of views that a file had at the beginning of that week. Here, average values are presented with the original number of views binned

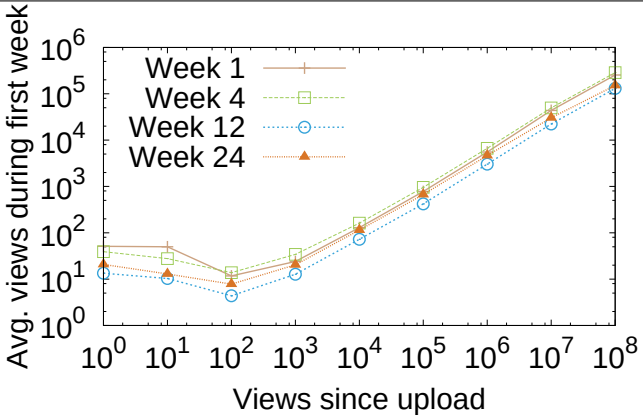


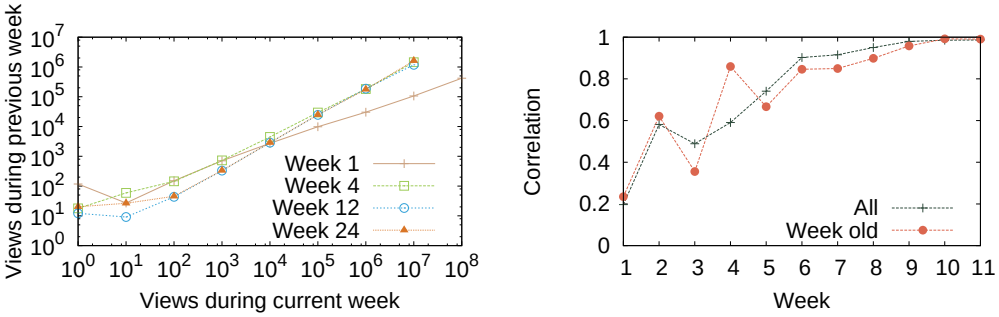
FIGURE 7.9: Rich gets richer.

logarithmically. We note that the curves show an increasing trend after a video achieves between roughly 100 views, indicating that a video needs that many views before it enters a state where it is more likely to gain views, just based on the fact that it already have been viewed. The correlation between seed views and delta views across weeks is around 0.55. We call files with sufficient number of views to benefit from this correlation to be members of the "rich" club, and note that this set includes files from a wide spectrum of popularities.

As noted in previous sections, some files remain popular for a long time period. A natural predictor for future popularity is therefore the current popularity. With weekly snapshots, we compare how the popularity of individual files changes from week to week. Of special interest is the correlation between consecutive weeks; as this can be used to predict a file's popularity in the following week (similar to what was discussed for the hotset analysis in Section 7.2.2). We are also interested in determining how long the current popularity is a reasonable predictor of future popularity. Fig. 7.10(a) shows the popularity in future weeks as a function of the number of views in the preceding weeks. We use logarithmic binning when calculating average weekly viewings for files of different popularity.

Clearly, there is a strong correlation between the number of views during consecutive weeks. Fig. 7.10(b) quantifies this correlation between consecutive weeks. We also know that popularity of a file at week one provide less and less insight to the popularity of a file at later weeks. This is due a decline correlation between the first weeks views and requests within any later weeks. We note that the correlations are fairly similar to the commonality observed with the hotset analysis. The correlations between views during adjacent weeks increases from 0.2 to 0.983 in the last week. In general, the correlations with the views during the first week are significantly smaller than between adjacent weeks. It increases from 0.2 between *week1* and *week2* to more than 0.75 between *week1*

and *week12*. These observations are consistent with our observations with the transient region in the hotset analysis.



(a) Past week

(b) Weekly correlation

FIGURE 7.10: Weekly changes and correlations.

Overall, the current popularity (as measured by the number of views in the last week, for example) appears to be a significantly better indicator of future views than the total number of views gained thus far. At this point it should also be noted that we also have considered using the *average viewing rate since upload* as a potential metric to predict future popularity. This metric has been considered previously by us in the last chapter, and is motivated by the fact that more than two snapshots are not always available and the current popularity therefore may be difficult to predict. This metric has the advantage that it removes the age dependency from the total number of views gained thus far, without requiring multiple snapshots. Also, average views per day show a declining trend with age i.e. older videos on average show a smaller value for requests per day that they received. The correlation between viewing rate at seed and delta views is less that 0.1 for most weeks. As our analysis suggests, this metric does not improve much over using the total number of views since upload, and is significantly worse than using last week's views. The problem with this metric is that it does not take into consideration when the cumulated views took place. For example, with this metric two different month-old videos, with the same number of views, would be weighted the same even if one may have had a steady rate of views while the other had a burst of views the first day and did not have any views since then.

Before further quantifying the growth of videos, it should be noted that comments, ratings, and other interactive features also may provide a good indication of future popularity. Similar to life-time views and views in the last week we have noted strong correlations/relationship between the original number of comments, ratings, etc. and the number of views in the following week. The Pearson's coefficient between increase in comments to the increase in views in the same week is 0.5 on average. Similarly, the value between new votes and weekly view gain is 0.7 on average.

TABLE 7.1: Inequality Indices (seed views).

Index	**YouTube**	**Dailymotion**
Symmetric	95.5%	95.7%
Atkinson	93.3%	90.0%
Gini	91.4%	87.4%
Europe	95.5%	93.3%
Hoover	77.5%	72.2%
Symmetric redundancy	3.107	3.139
Theil redundancy	2.702	2.307
Welfare(using Gini)	2994.862	294.401
Welfare(using Atkinson)	2345.627	233.202
Welfare(using symm.)	1563.838	101.539

However, a problem with these metrics is that they are used sparsely. Therefore, they typically only provide good predictions for videos that are highly popular. For files that recieve many views, the additional feedback these ratings provide is not necessarily useful for short term popularity. We believe, however, that these metrics can be usefull when estimating a movies long-term popularity. For example, two files with equal current popularity may recieve highly distinct ratings and interactivity through comments. Chances are that the movie with more positive ratings will see a higher life time.

7.2.5 An Econometric Viewpoint

In this section, we will look at various econometric inequalities as a measure of heterogeneity in both long and short term popularity. We will also look at welfare functions of different age groups of videos on YouTube, along with their contributions to the total inequality across several weeks.

The field of Economics regularly deals with various disparities in the distribution of income and assets of countries or even individuals. Inequality measures, are one such tool to quantify this heterogeneity or disparity in wealth and income. There are various indexes for measuring this inequality. Some of the common measures include Gini, Hoover and Theil coefficients. In such indexes, perfect equality is usally denoted by the value 0 for the index. In the case of Gini [28], and Hoover a value of 1 indicates perfect inequality, where one single individual has all the income or wealth. These measures are also sometimes indicated using the percentage notations. Theil index on the other hand is not constrained between the values 0 and 1. A value of 1 for Theil indicates that the distributional entropy is similar to a system with an 82:18 distribution, which is very close to the Pareto distribution. A higher value signified even higher degrees of skewness.

To measure inequality in wealth, we calculate some popular indexes by considering the total views since upload as our wealth indicator. The difference being that, although wealth of an individual can decrease with time, in our case the total views obtained can only increase. The values of the some inequality measures are shown in Table 7.1. We consider our YouTube data set with the Dailymotion data set along as comparison. The values were calculated using the views at seed time. As is evident from the values and our previous discussions, there is presence of a high degree of skewness in popularity measured on total views gained since upload. Amartya Sen[44] proposed a welfare function(W) based on the Gini coefficient as shown in equation 7.2, which measures the average per capita income of a randomly selected individual from the entire population. The table also shows the value of the welfare function based on the wealth (instead of income) accumulated by videos. We also observe that the values of these measures remain similar across our entire collection period of 8 months. Thus total inequality in wealth is likely to remain constant on such sites.

$$W = \overline{Income}.(1 - GiniCoeff) \tag{7.2}$$

A more appropriate measure of disparity in current popularity would be to look at weekly income of these videos. We consider the views gained during a week as an indicator of its income. We observed a very high degree of overall inequality in the income of videos. The Gini coefficient calculated for weekly income remained above 0.9 for all the weeks that we observed, thus indicating that very few videos get a major share of all views within a week.

To measure the differences in inequality between videos of various age groups, we also calculated the Gini index for weekly increase in views for all the videos in five age groups. We observed that the inequality within each group are similar. The index varies from 0.95 to 0.92 among the different ages, with the younger videos on YouTube exhibiting a slightly greater inequality than other age groups. These values remain the same across all the different weeks.

Fig. 7.11 shows the value of the welfare function for new views across the first twelve weeks, i.e. it indicates the number new views in a week that a random video selected from a certain age group would be expected to have. We infer from the higher values of the function for older videos on YouTube, that they consistently get a larger share of the new views. The trend on the tail end of the curves continue till the end of the collection period, and have not been shown for brevity. The Gini index, although popular, has a disadvantage that it cannot be decomposed. Thus, the individual contributions to the

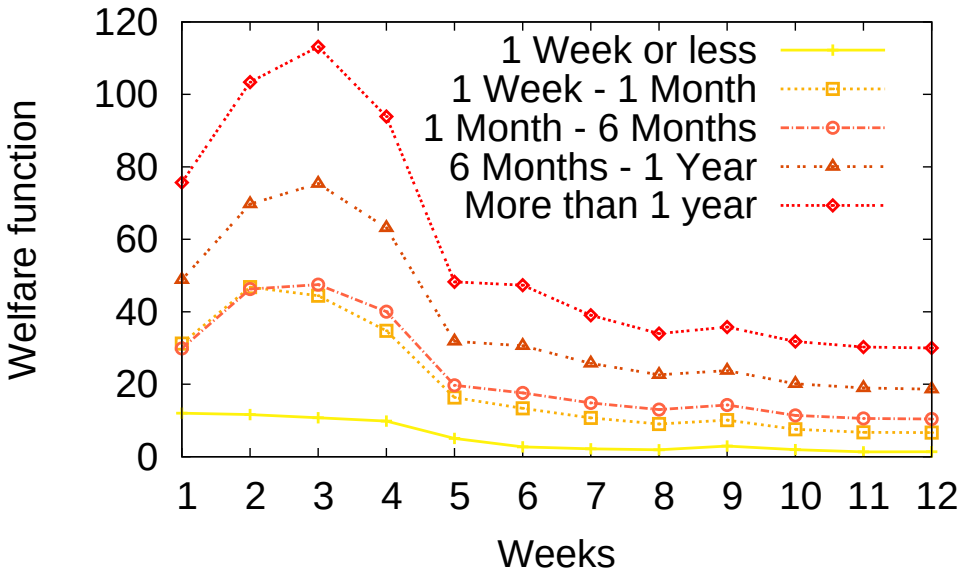


FIGURE 7.11: Welfare function by age.

total inequality cannot be determined through it. So we look at another index in detail called the Theil index.

The Theil index produces values that are unconstrained but always larger than 0. A value of 1 for Theil indicates that the distributional entropy is similar to a system with an 82:18 distribution (i.e., Pareto rule applies). Higher value signifies even higher degrees of skewness, and values lower than 1 indicate less skewness in the distribution. One key advantage of the Theil index is that it is decomposable, that is, the individual contribution to inequality of various groups (e.g., by age in our case) can be computed. Theil index is typically written in the following form:

$$T_{total} = \sum_{1}^{m} T_{within_{grp}} + T_{between} \tag{7.3}$$

The Theil index (T) can be decomposed into the contributions to inequality by each age group. Also, the weighted contributions of inequality 'within' each group and the heterogeneity 'between' each age group, can simply be summed to derive the total inequality. This is shown in equation(7.3), where 'm' is the number of groups. In order to calculate each groups contribution to T_{within} and $T_{between}$, we use equation(7.4) and (7.5) respectively. I_{ind}, I_{grp} and I_{total} refers to income of the individual, group and the total population, respectively. Similarly, P_{grp} and P_{total} refer to the group and total populations.

$$T_{within_{grp}} = \frac{P_{grp}}{P_{total}} . \sum_{1}^{n} \frac{I_{ind}}{I_{grp}} \; . \; ln \left[\frac{I_{ind}/I_{grp}}{1/P_{grp}} \right] \tag{7.4}$$

$$T_{between} = \sum_{grp=1}^{m} \frac{I_{grp}}{I_{total}} \; . \; ln \left[\frac{I_{grp}/I_{total}}{P_{grp}/P_{total}} \right] \tag{7.5}$$

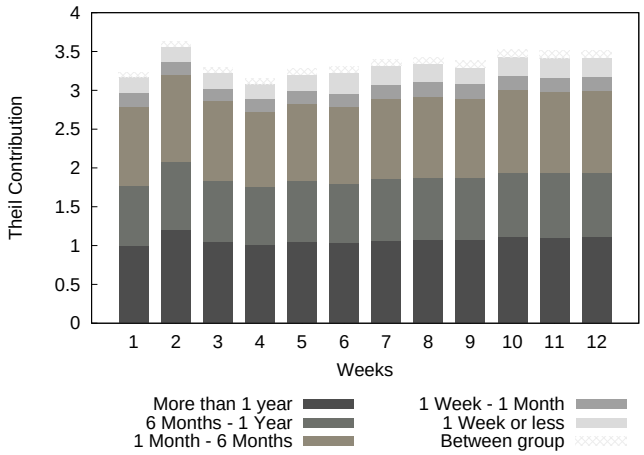


FIGURE 7.12: Theil decomposition.

Using the above equations, we calculated the individual Theil contributions of videos in different age groups. The results are shown in Figure 7.12. Note that the individual contributions can be negative but the weighted summations are always positive. The key take-home from this analysis is that groups containing older videos gets a larger share of view (increases) than than younger videos when considering the distribution of older and younger videos. This analyses thus tells us that popular older videos contribute significantly toward the skewness observed in views to videos.

7.2.6 Unified Growth Model

So far we have seen that future popularity is directly proportional to the total views since upload and also the views gained in a week. We also know that age provides a significant impact on it. Next, we try to combine and quantify these separate relationships between these factors into a growth model.

From our analysis, we observed that there was a distinct relation between age of a video μ and the ratio of a week's views to the total views at the start of the week. The nature of this relationship is depicted in equation 7.6.

$$\sqrt[x]{\frac{v_x}{v_0}} - 1 \propto \mu^{\gamma}. \tag{7.6}$$

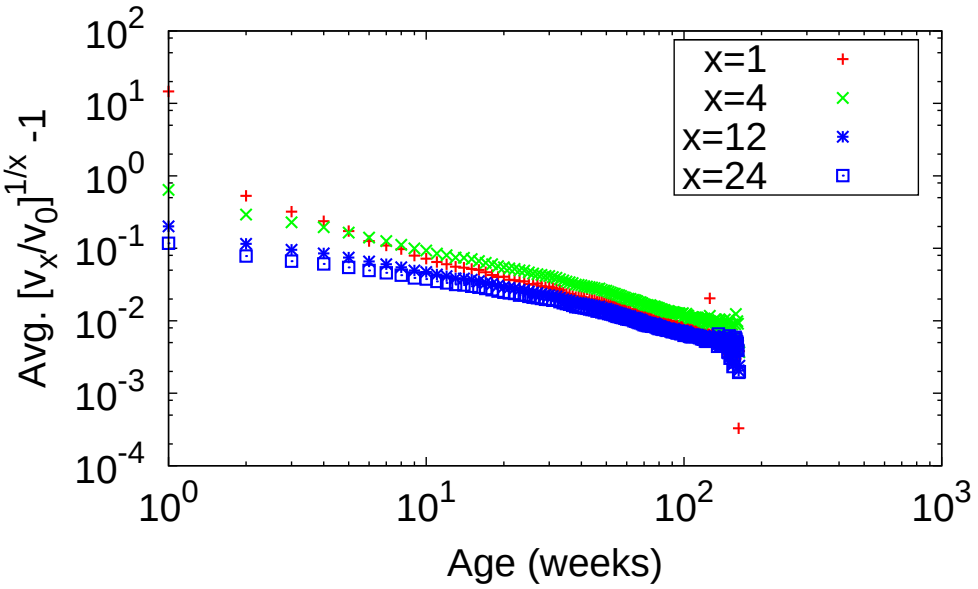


FIGURE 7.13: Growth model validation.

where, v_0 is the total number of views at the start, whereas v_x is its total views after x number of weeks. Here, $\sqrt[x]{\frac{v_x}{v_0}}$ is the average rate with which the total number of views increases for the video and gamma is a modeling parameter. Fig. 7.13 helps evaluate the accuracy of this model. In this figure, we plot the age of the video in weeks at the time of capture on the x axis. The y axis shows the average value of function $\sqrt[x]{\frac{v_x}{v_0}} - 1$ for all videos with the corresponding ages. We observe that the relationship between age and the function is linear on a log scale up to over 100 week old files, suggesting that our model is very accurate for this range. The drop off at the end is perhaps due to lack of sufficient old videos. Referring to Section 7.2(a), we note that there are very few files older than 100 weekson YouTube. Fitting a power function, we determine $\gamma = 1.1$ for $x = 1$, with a coefficient of determination (R^2) of 90% . The graph also shows the relationship for other values of x.

7.3 Summary

This chapter presented a characterisation of the evolutionary aspects of online video sharing. We looked at our YouTube data set, tracing over a million videos for 34 weeks. Some of our key observations are:

1. We validated some of the prior characterisation work on YouTube. We also compared YouTube with four other video sharing services and found that the popularity on YouTube also has Zipf-like characteristics and follows the Pareto rule.

2. Frequency of tag keywords on YouTube also follow Zipf like pattern.

3. We have also characterised the rank changes between the videos in terms of popularity. There is considerable change in the ranks in the initial few weeks, which later stablise.

4. We found significant impact of age on the rate of growth of a video. This relationship is further enumerated in our growth model.

5. We also show that these sites exhibit a preferential attachement pattern, i.e. they exhibit a 'rich get richer' phenomenon with a video with higher initial views gaining more views.

6. The views of a video has a significant correlation with the previous weeks gain in views and this correlation grows with each passing week.

7. We also present an Econometric view of the inequality on these sites. A major portion of this inequality is contributed by the older videos.

8. Finally, we aggregated our observations into a popularity growth model, in terms of present views and its age.

Chapter 8

Social Network Effects and Content Deletion

In previous chapters we have looked at static characterisation of video popularity and other attributes. We also looked at the evolution of videos in great detail in Chapter 7. Besides the characteristics of the video itself, there are a number of other factors that affect the growth of a video on such online video sharing sites. The influence of the uploader may also propel a video to more popularity. Content deletion on such systems is one topic that remains largely untouched in previous literature. Content deletion is also an important facet in the entire evolutionary process on these systems. In this chapter, we characterise these factors and outline some of their effects.

8.1 Impact of Social Networking

Video sharing sites provide certain features to their users, to allow them to build a social network. These features have not been studied in great detail in the context of video sharing, in prior works. On YouTube, a user may add other users as friends, which might reflect a real world relationship, a like-mindedness or common interests. Any user that is interested in the content posted by another user could also opt to become a fan. While a friend is a bi-directional relationship, a fan represents a uni-directional link between users. YouTube also provides data on the number of videos a user has posted. These videos also include videos that were not captured in our data sets. Thus, they provide a better window to the uploading profile of the random users collected in our crawl. In this section, we attempt to understand the upload behaviour of users along with the characteristics of these social networking features. We talk about the distribution of fans, friends relationship for all the uploaders that we encountered in our crawl. We also

look at how this network of a user evolve and its effect on popularity of videos posted by him.

8.1.1 Who uploads videos?

Another essential attribute of video sharing is how frequently people upload or publish new videos. The pool of uploaders in our data set have uploaded around 17.5 million videos in total. To discern the distribution of uploads by users, Fig. 8.1 depicts this rank-frequency plot. The X axis denotes the rank of the users based on the number of videos they have uploaded since they joined. In other words, the user with rank 1 is the user who has uploaded the most videos. We observe the presence of Zipf-like characteristics, from the straight line on a log-log scale for over 5 orders of magnitude. We also noted similar results for our previous data sets. Skewness in uploads for the previous data sets has already been discussed in Section 6.4.

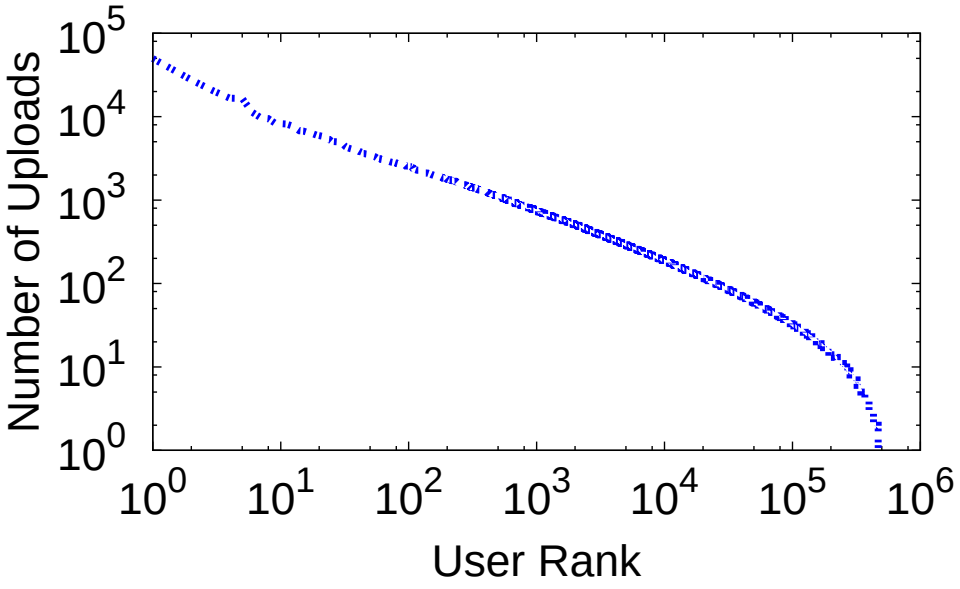


FIGURE 8.1: User Rank vs. Total Uploads.

Our incremental snapshots also allow us to capture the growth trend in video uploads over a significant period of time. Fig. 8.2 illustrates the relation between initial number of posted videos to the average number of new videos uploaded across the first twelve weeks of the crawl. Both axes of the graph utilize a logarithmic scale, with the initial number of videos binned logarithmically at scale 10. The graph provides evidence to the fact that uploading behaviour sustains over a period of time. While a smaller number of users account for a larger share of the uploads, its also means that, a user who posts a larger number of videos is likely to continue posting a higher number of videos in

relation to other users with low uploading profiles. We observe similar results from our two snapshots of Dailymotion.

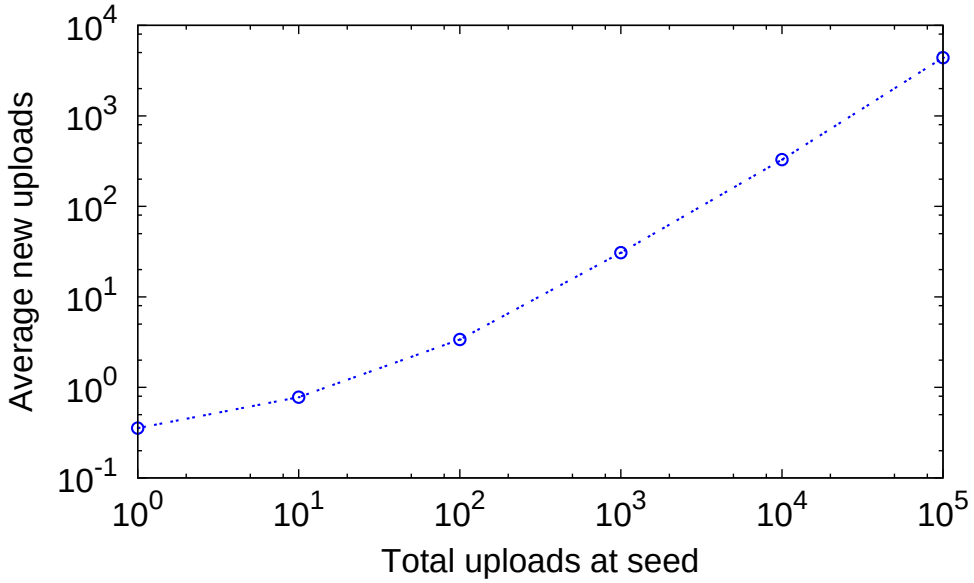


FIGURE 8.2: New Uploads.

8.1.2 Influence of Social Networks on Popularity of videos

The previous chapter has already discussed the fact that the rise in popularity of a video is directly proportional to its original number of views. Also, the increase in views in a week have a high correlation with last weeks increase in views. In our attempt to further understand the reasons behind the rise in popularity of videos, we look at the impact of a users network on the popularity of videos posted by him.

Our data set also contains records of the number of friends, fans for approximately 1 million users that we came across during our crawl. We found that of all the users captured in our set each had atleast 1 fan, while the percentage of users with no friends on YouTube was found to be around 29%. We analysed the relationship between rank of the users, based on his number of friends, with the actual number of his friends. The curve, although not shown here for the sake of brevity, shows us that YouTube fans and friends relations also follow Zipf-like characteristics for alteast five orders of magnitude. The presence of a cutoff at the tail end of the graph signifies the lack of the expected number of users with a smaller number of friends or fans. The missing fraction of tail end users might be due to a crawling artifact, perhaps accounted by all the users that did not post any videos and failed to show up in our crawl for videos.

The above graphs tell us the distribution of the network of a user. Next, we'll look at its effect on the popularity of the uploaded videos. Fig. 8.3 shows the relation between the number of people in a users network and views recieved on videos posted by him. The x axis depicts the number of fans or friends a user has, with logarithmically binning at scale 10. The y axis is the average increase in views observed during the entire crawl period. It clearly shows a monotonically increasing relation between the two i.e. videos posted by a user with a larger network is likely to get more views on average than a user with a smaller network. The reasons for the same could be manifold. The user could either actively seek to spread the content among his network of friends and thus facilitating the videos increasing views. On the other hand, the frequency with which a user posts interesting content could also lead to him having a larger network of followers and friends. The same also holds true for only a week old videos i.e. the vastness of a user's network leads to higher views even for the recently uploaded videos. This hints towards the fact that the network may play an important role in the initial number of views that a video gets after being uploaded.

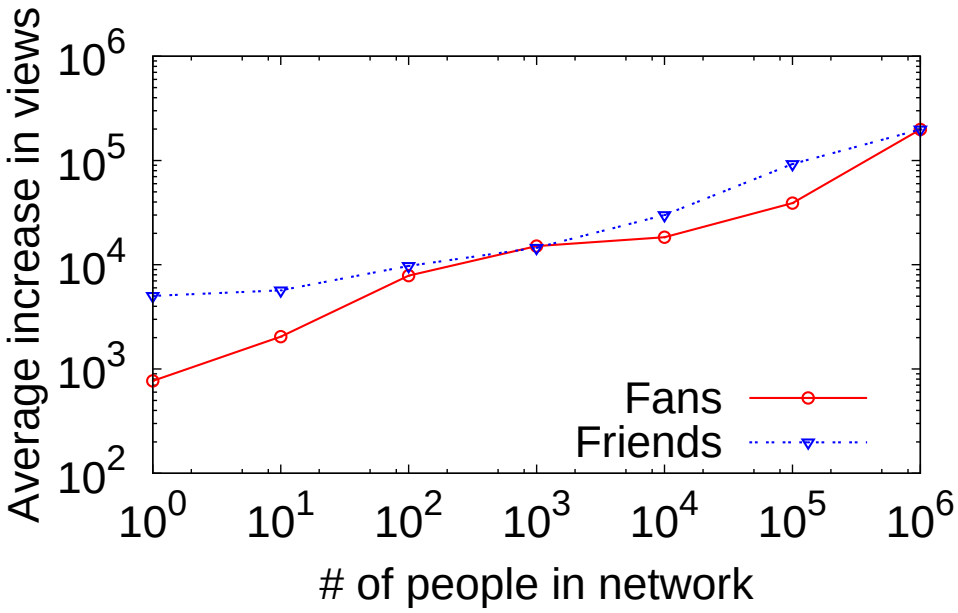


FIGURE 8.3: Total increase in views Vs users network.

8.1.3 Growth of Social Network

We have provided evidence of the 'rich get richer phenomenon' pertaining to views, in the previous chapter. We have also seen how video popularity and users network have a very high correlation. Thus it becomes worthwhile to investigate how the users network grows which may indirectly affect the popularity of his videos. Our analysis of the uploaders has led us to the conclusion that preferential attachment might apply to the

social network on YouTube as well, due to presence of a correlation between network of the uploader and popularity of their content. Fig. 8.4 shows the relation between initial number of friends (binned logarithmically at scale 10) and the average increase in friends observed during the entire crawl. A user with a higher number of friends is likely to further this divide across weeks. The corresponding scatter plot for the fans relationship has been omitted for the sake of brevity, which showed a similar trend. 47% of the users on YouTube, showed no increase in the number of friends overall. The corresponding fraction for the number of fans was 17% respectively. We also noticed similar patterns from our snapshots of Dailymotion.

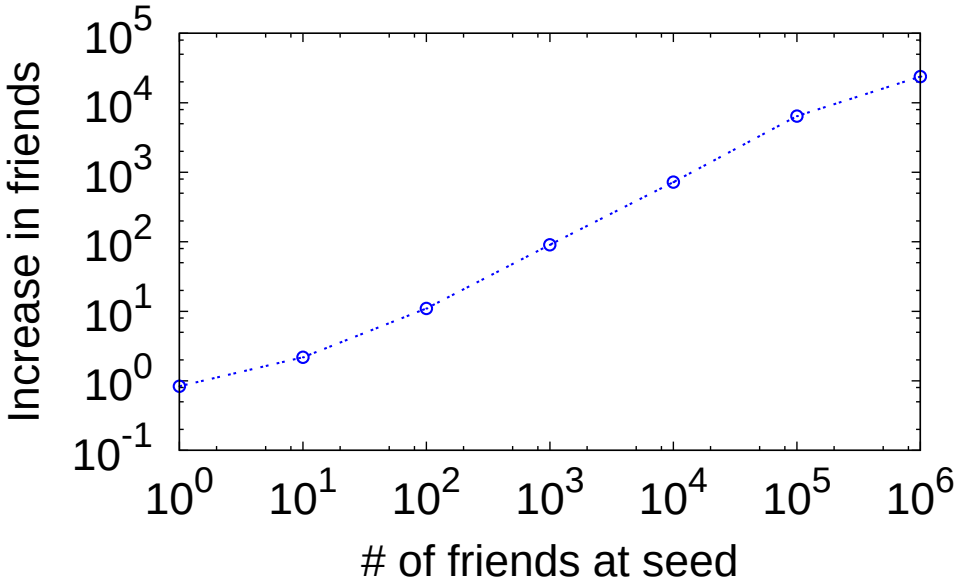
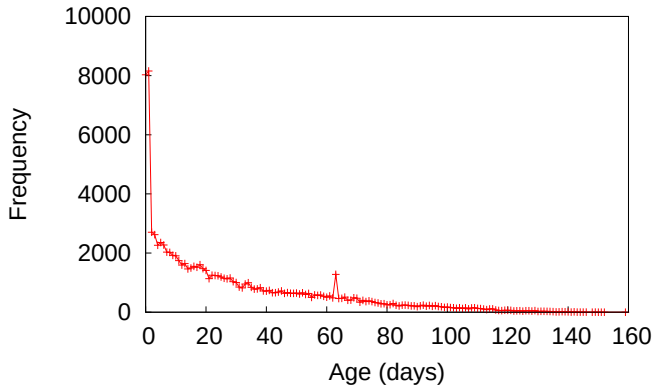


FIGURE 8.4: Growth of friends network.

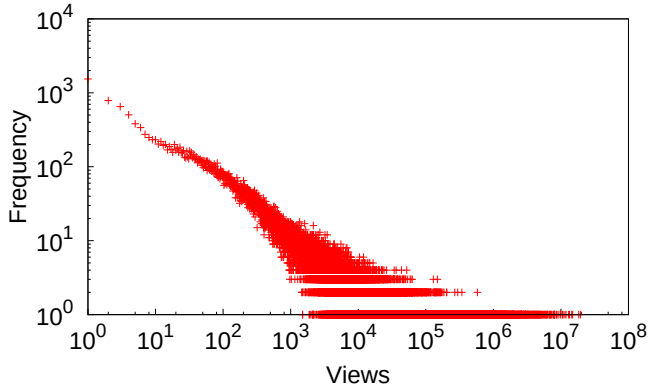
8.2 Content Deletion

In previous sections we have traced the growth of popularity of videos. In particular, we showed how upload behaviour of a user is persistent across weeks. A user with higher number of uploads in a week is likely to maintain the same upload profile. This allows us a glimpse into the addition of content to the system. From the aggregate analysis in Chapter 7 we have also seen that a set of videos continue to receive requests much after hitting their peak, thereby giving them an effectively infinite lifespan. Hence, the only way for a content to truly 'die' on such systems is for it to be deleted.

The number of videos that we collected during our seeding process was initially larger. We noticed that retrieving those content in further snapshots resulted in HTTP 404 errors, which were logged by our crawler. We had deleted those videos from all analyses



(a) Distribution of Age



(b) Distribution of views

FIGURE 8.5: Deletion of videos.

in the preceding sections. In our weekly crawls we noticed a total of 100,178 videos being deleted from YouTube across the 34 weeks of our crawl. Breach of copyright laws could be a reason for the removal of the content by the content provider. An additional reason for the removal could be deletion by the uploader himself. On analysing the number of videos deleted across weeks, it is evident that a significant number of videos got deleted during the first twelve weeks of the crawl period. The 404 error structure returned by the YouTube API, provided us a 'reason' field for the error. Analysis of the log files revealed that there were mainly three types of errors. The first being 'video not found' indicating that the content was either removed by YouTube or by the uploader himself. The other two reasons were due to suspension and closing of the uploader's account. A user is allowed to delete his own account from the system. The site may also suspend a users account for violation of terms of usage. Out of the all the videos deleted, we observed about 78% of them were not found, as in they were deleted by the users or by YouTube. Approximately 4% of these videos, were were due to the uploader closing their account and remaining such videos were due to users with suspended accounts. We remark that the content published by these user accounts were also deleted along.

Further, we investigate the characteristics of these deleted videos. Fig.8.5(a) shows the

distribution of age for the same. The X axis shows the number of deleted videos in each age group. While videos of all age groups had been deleted, the majority of the videos that were removed had been very recently uploaded, i.e. less than or equal to a week old. It becomes apparent that quite a few number of videos get removed in the very first week of their upload. The reason could be manifold. It might be due to novice users trying out the upload process and then deleting the content. Since we also had the initial meta data on all videos captured initially, it allowed us to manually investigate the uploader's profile. We notice that some users had uploaded another video with the same title as the one that was removed, but which had a different video identifier and upload time. We expect that such users modified those videos and re-uploaded them after they were satisfied with the content.

To analyse further reasons for deletion, we also depict the popularity distribution using views, of those videos. Fig.8.5(b) shows the same. The graph for shows pareto behaviour, with a large number of videos with 1 view and very few videos with higher views getting deleted. We believe that a lack of expected number of views in the first few days could have also prompted deletion of very young videos.

8.3 Summary

This chapter considered the impact of social network of an uploader on the videos uploaded by him on YouTube. We discussed content deletion on YouTube as well. The main observeations in this chapter are:

1. The uploaders on video sharing sites show Zipf-like characteristics in terms of total number of uploads.
2. The uploading freequency of users persists over time as evident from YouTube and also Dailymotion.
3. The uploaders network profile has a significant correlation with the number of views that a video posted by him get on YouTube.
4. The degree of a user in the number of fans or friends also exhibits the presence of Zipf's law.
5. Growth of social network of users also follow preferential attachment, i.e the user with a larger circle of relationships is more likely to gain more fans or friends than a user with a smaller number.

6. A significant amount of videos on YouTube get deleted, and a majority of those are very young. A major share of the videos getting deleted are those that have smaller number of views.

Chapter 9

Conclusions and Future Work

In this chapter, we summarize the work we have done and present our major contributions. Towards the end we also outline ideas for future work within the topic of this work.

9.1 Thesis Summary

Chapter 1 presented the motivation behind the research. Two of our major goals were: to find invariants across different video sharing services in terms of popularity and other video attributes; and to study the longitudinal characteristics of popularity.

Chapter 2 presented an overview of the different services we analyse in our research. We outline some of the main questions that we investigated, followed by a brief look at our data sets.

Chapter 3 listed all the prior work that was related to ours. We enumerated several workload characteristics studies from traditional media, and also work on video sharing done before.

Chapter 4 briefly described our work environment and the tools we used to carry out data collection, and statistical analysis.

Chapter 5 lists the crawl strategies, data collection methodology we used for obtaining our data. We also give a high level summary of all our data sets. The study used data sets obtained from five different sites. This empirical data served as the basis for all of our analysis.

The remaining chapters present our main results. Chapter 6 investigated the key invariants that we found across different video sharing sites. It also presented our static

characterisation results based on four of our data sets. We list suitable models that can be used to explain the empirical distribution of popularity found on these sites. Chapter 7 uses our 34 week YouTube data, and compares it to our other data sets. We also use the weekly snapshots to gain insight into the evolutionary aspects of popularity and growth of videos. It concludes with a modelling approach that presents a growth model which explains the future views in terms of its present popularity and age. Chapter 8 examined the popularity from the viewpoint of the uploader. We investigate the impact of social network of a user on the content posted by him. It also presents a look at content deletion on YouTube, which has not been discussed much in prior literature.

9.2 Results and Contributions

In this section, we discuss the results we have found from the analysis of the questions in Chapter 2. Some of our major contributions, as described in Chapters 6, 7, and 8 are:

- We present and analyse traces of *five* video sharing sites, with our Youtube data set traced for over 8 months. Our data set contains more than 2 million videos in total with approximately 70 billion views in total.
- Users are primarily interested in watching videos. Social interaction tools for rating, bookmarking, and commenting on videos are infrequently used. We identified several key invariants across the different video sharing services.
- The number of uploaders to a service are an order of magnitude smaller than the number of uploaded videos. This again reiterates that users are primarily interested in viewing videos. We found that the Pareto rule applies, with the top 20% of the uploaders contributing between 75-80% of the total videos.
- The typical video available from these services is of short duration. This observation, based on data from the Dailymotion, Veoh, Yahoo! video, and Metacafe services, is similar to that made for YouTube [26, 55]. Some services, such as Veoh, however, are starting to feature longer duration videos, using peer-to-peer technology to address the problem of efficiently distributing such videos.
- Both the total number of views to videos, and the rate with which new views occur, follow the Pareto rule, with 20% of the most popular videos accounting for 80% or more of the views. Similar observations regarding the number of views since upload [18] and the viewing rate [26] have been made for YouTube.

- The total views popularity distribution is heavy-tailed and may be modelled as power law or power law with exponential cut off, with power law exponent between 1.4 and 2.5. However, we note that neither a power law (or variants), nor a lognormal distribution, appear to fit the entire distribution well. These results are consistent with those reported for YouTube's videos in the science and entertainment categories [18]. Similarly, the total views to videos may be modelled as Zipf with cut off.

- The average viewing rate since upload is generally more Zipf-like than the total views popularity distribution. The viewing rate popularity distribution of videos can be modelled as power law with cut off, with power law exponent between 1.4 and 2.

- In contrast to traditional Web and media server workloads, there are not many "one-timers", when this term is defined for this context according to views since upload. When considering views over a fixed-length trace period, however (such as the two week period between our snapshots of Dailymotion), we expect that the proportion of "one-timers" may become more comparable to that with traditional workloads.

- We validated some of the prior characterisation work on YouTube [18]. We also compared YouTube with four other video sharing services and found that the popularity on YouTube also has Zipf-like characteristics and follows the Pareto rule.

- Frequency of tag keywords on YouTube also follow Zipf like pattern.

- We have also characterised the rank changes between the videos in terms of popularity. There is considerable change in the ranks in the initial few weeks, which later stablise.

- We found significant impact of age on the rate of growth of a video. This relationship is further quantified in the growth model we present in Section 7.2.6.

- We show that videos on these sites also follow a *preferential attachment* pattern i.e, they exhibit a "rich get richer" phenomenon. We provide empirical evidence that videos with a higher number of initial views gain more views with a higher probability than videos with lower initial views.

- We find that new requests to a video in a week have a significantly high correlation with the previous weeks gain in views and this correlation grows with each passing week.

- We also illustrate the presence of a significant effect of activities such as commenting, rating on the gain in views in succeeding weeks.
- We also present an Econometric view of the inequality on these sites. We define inequality in terms of Gini, Theil and some other indices. A major portion of this inequality was found to be contributed by the older videos.
- Finally, we aggregated our longitudinal observations into a popularity growth model, which explains future growth in terms of present views and its age.
- We analyse social network characteristics of the video uploaders in our data set. We observe that friend and fan relationships on these sites exhibit Zipf-like properties with a cut-off at the tail end.
- Our study also shows that network growth of users on YouTube and Dailymotion also follow a "rich get richer" pattern.
- We illustrate that the number of people in the uploaders network has a noteworthy correlation with the number of requests to the videos posted by him. The users with a larger network of friends have been observed to have significantly higher views on their videos.
- Finally, we look at content deletion and characterise the properties of deleted videos on both Dailymotion and YouTube.

9.3 Future Work

Many avenues remain for future work. Our ongoing work is concerned with developing models for video reference streams, and studying novel content distribution mechanisms for user generated content. Other open problems include determining general characteristics of user generated content sharing services, developing efficient storage management solutions for large scale user generated content sharing services, and understanding the impact of geographic diversity.

We also seek to further our research along the lines of prediction of future popularity at an individual level and also comparison of various prediction models. Some of our interests also include studying behaviour of users on such services, predicting their intent and classifying them, as well as a comparative study of other similar services.

Bibliography

[1] Acharya, S., and Smith, B. An Experiment to Characterize Videos Stored on the Web. In *Proc. SPIE/ACM MMCN* (San Jose, USA, January 1998).

[2] Acharya, S., Smith, B., and Parnes, P. Characterizing User Access to Videos On The World Wide Web. In *Proc. SPIE* (2000).

[3] Acharya, S., Smith, B., and Parnes, P. Characterizing User Access to Videos on the World Wide Web. In *Proc. SPIE/ACM MMCN* (San Jose, USA, January 2000).

[4] Adamic, L., and Huberman, B. Zipf's law and the Internet. *Glottometrics*, 3 (2002), 143–150.

[5] Adamic, L. A. Zipf, power-laws, and pareto - a ranking tutorial.

[6] Adobe. Adobe Flash Player Version Penetration. `http://www.adobe.com/products/player_census/flashplayer/version_penetration.html`, March 2007.

[7] Alexa. http://www.alexa.com.

[8] Almeida, J., Krueger, J., Eager, D., and Vernon, M. Analysis of Educational Media Server Workloads. In *Proc. ACM NOSSDAV* (Port Jefferson, USA, June 2001).

[9] Arlitt, M., and Williamson, C. Internet Web Servers: Workload Characterization and Performance Implications. *IEEE/ACM Trans. on Networking 5*, 5 (October 1997), 631–645.

[10] Auchard, E. Participation on Web 2.0 Sites Remains Weak, April 2007.

[11] AutoHotKey. http://swik.net/autohotkey.

[12] AutoSave. https://addons.mozilla.org/en-us/firefox/addon/2704.

[13] Barabási, A., and Albert, R. Emergence of Scaling in Random Networks. *Science 286* (October 1999), 509–512.

[14] Benevenuto, F., Duarte, F., Rodrigues, T., Almeida, V., Almeida, J., and Ross, K. Characterizing Video Responses in Social Networks. Arxiv, 2008.

[15] Breslau, L., Cao, P., Fan, L., Phillips, G., and Shenker, S. Web Caching and Zipf-like Distributions: Evidence and Implications. In *Proc. IEEE INFOCOM* (New York, USA, March 1999).

[16] BusinessWire. Ellacoya Data Shows Web Traffic Overtakes Peer-to-Peer (P2P) as Largest Percentage of Bandwidth on the Network. `http://www.businesswire.com/news/google/20070618005912/en`, June 2007.

[17] BusinessWire. Roland Hamilton Joins Dailymotion US as SVP of Sales as Company Continues Rapid Growth. `http://eon.businesswire.com/portal/site/eon/permalink/?ndmViewId=news_view&newsId`
2007.

[18] Cha, M., Kwak, H., Rodriguez, P., Ahn, Y., and Moon, S. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Internet Measurement Conference* (San Deigo, USA, October 2007).

[19] Cha, M., Rodriguez, P., Crowcroft, J., Moon, S., and Amatriain, X. Watching television over an ip network. In *IMC '08: Proceedings of the 8th ACM SIGCOMM conference on Internet measurement* (New York, NY, USA, 2008), ACM, pp. 71–84.

[20] Cheng, X., Dale, C., and Liu, J. Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study. In *In Proc. IWQoS* (Enschede, Netherlands, June 2008).

[21] Cheshire, M., Wolman, A., Voelker, G., and H.M.Levy. Measurement and Analysis of a Streaming-Media Workload. In *in Proc. USITS* (2001).

[22] Clauset, A., Shalizi, C., and Newman, M. Power-law distributions in empirical data. http://www.santafe.edu/ aaronc/powerlaws/, 2007.

[23] Conceicao, P. N., and Ferreira, P. M. The Young Person's Guide to the Theil Index: Suggesting Intuitive Interpretations and Exploring Analytical Applications. *SSRN eLibrary* (2000).

[24] Dailymotion. http://www.dailymotion.com.

[25] Dailymotion. Fact Sheet. `http://www.dailymotion.com/press/fact_sheet_july2007.pdf`, 2007.

[26] Gill, P., Arlitt, M., Li, Z., and Mahanti, A. Youtube traffic characterization. In *Internet Measurement Conference* (San Deigo, USA, October 2007).

[27] Gill, P., Arlitt, M., Li, Z., and Mahanti, A. Characterizing youtube user sessions. In *Proc ACM/SPIE MMCN* (San Hose, USA, January 2008).

[28] Gini, C. Variability and Mutability. , 1912.

[29] Golder, S., Wilkinson, D., and Huberman, B. Rythms of social interaction: Messaging within a massive online network. In *Proc. of Intl. Conf. on Communities and Technologies* (2007).

[30] Gummandi, K., Dunn, R., Saroiu, S., Gribble, S., Levy, H., and Zahorjan, J. Measurement, Modeling and Analysis of a Peer-to-Peer File-Sharing Workload. In *Proc. ACM SOSP* (Bolton Landing, USA, October 2003).

[31] Halvey, M., and Keane, M. Exploring Social Dynamics in Online Media Sharing. In *Proc. of WWW* (Banff, Canada, May 2007), pp. 1273–1274.

[32] Huang, C., Li, J., and Ross, K. Can Internet Video-On-Demand be profitable? In *SIGCOMM* (2007).

[33] Huberman, B., and Adamic, L. Power-Law Distribution of the World-Wide Web. *Science 287* (March 2000), 2115.

[34] Mahanti, A., Williamson, C., and Eager, D. Traffic Analysis of a Web Proxy Caching Hierarchy. *IEEE Network 14*, 3 (May/June 2000), 16–23.

[35] Matlab by The Mathworks. http://www.mathworks.com/.

[36] Mechanize. http://wwwsearch.sourceforge.net/mechanize/.

[37] Metacafe. http://www.metacafe.com.

[38] Mitzenmacher, M. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics 1*, 2 (2004), 226–251.

[39] Mozilla Firefox. http://www.mozilla.com/firefox/.

[40] MySQL, Relational database management system. http://www.mysql.com/.

[41] Newman, M. Power laws, pareto distributions and zipf's law. *Contemporary Physics 46* (2005).

[42] Nilsen, E. Analysis of news-on-demand characteristics and client access patterns. Master's thesis, University of Oslo, Norway, April 2005.

[43] Python, Programming Language. http://www.python.org.

[44] Sen, A. *On Economic Inequality*. Clarendon Press, Oxford, 1997.

[45] Spearman, C. The proof and measurement of association between two things. Amer. J. Psychol., 1904.

[46] SQLite - Self contained database engine. http://www.sqlite.org/.

[47] Tang, W., Fu, Y., Cherkasova, L., and Vahdat, A. Long Term Streaming Media Server Workload Analysis and Modelling. HP Labs, Tech Rep., 2003.

[48] The R project for statistical computing. http://www.r-project.org/.

[49] Veloso, E., Almeida, V., Meira, W., Bestavros, A., and Jin, S. A Hierarchical Characterization of a Live Streaming Media Workload. In *in Proc. SIGCOMM* (2002).

[50] Veoh. http://www.veoh.com.

[51] Wilson, D., and Lockshin, L. Using the Thiel Index Coefficient to Analyse Variation in Wine Consumption Habits. .

[52] Yahoo! Videos. http://video.yahoo.com.

[53] YouTube. http://www.youtube.com.

[54] Yu, H., Zheng, D., Zhao, B., and Zheng, W. Understanding User Behavior in Large-Scale Video-on-Demand Systems. *SIGOPS Oper. Syst. Rev. 40*, 4 (2006), 333–344.

[55] Zink, M., Suh, K., and Kurose, J. Watch Global, Cache Local: YouTube Network Traffic at a Campus Network - Measurements and Implications. In *Proc. of SPIE/ACM MMCN* (January 2008).