Characterizing Popularity Dynamics of User-generated Videos: A Category-based Study of YouTube

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Shaiful Alam Chowdhury

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Abstract

Understanding the growth pattern of content popularity has become a subject of immense interest to Internet service providers, content makers and on-line advertisers. This understanding is also important for the sustainable development of content distribution systems. As an approach to comprehend the characteristics of this growth pattern, a significant amount of research has been done in analyzing the popularity growth patterns of YouTube videos. Unfortunately, no work has been done that intensively investigates the popularity patterns of YouTube videos based on video object category. In this thesis, an in-depth analysis of the popularity pattern of YouTube videos is performed, considering the categories of videos.

Metadata and request patterns were collected by employing category-specific YouTube crawlers. The request patterns were observed for a period of five months. Results confirm that the time varying popularity of different YouTube categories are conspicuously different, in spite of having sets of categories with very similar viewing patterns. In particular, News and Sports exhibit similar growth curves, as do Music and Film.

While for some categories views at early ages can be used to predict future popularity, for some others predicting future popularity is a challenging task and require more sophisticated techniques, e.g., time-series clustering. The outcomes of these analyses are instrumental towards designing a reliable workload generator, which can be further used to evaluate different caching policies for YouTube and similar sites. In this thesis, workload generators for four of the YouTube categories are developed. Performance of these workload generators suggest that a complete category-specific workload generator can be developed using time-series clustering. Patterns of users' interaction with YouTube videos are also analyzed from a dataset collected in a local network. This shows the possible ways of improving the performance of Peer-to-Peer video distribution technique along with a new video recommendation method.

Acknowledgements

I would like to take this opportunity to thank and express my gratitude to the people who helped me and made the successful completion of this thesis possible.

First and foremost, I would like to express my genuine gratitude and sincere appreciation to my supervisor Dr. Dwight Makaroff who helped me in every aspects of my life here in Saskatoon. With my very little idea about research, my supervisor guided me in the right directions from the beginning of my M.Sc. program. He helped me to understand everything in detail and always tried his best to answer my questions. I must say, I was very lucky to have him as my supervisor. I could not have imagined having a better guidance for my M.Sc.

Besides my supervisor, I would also like to thank the rest of the members of my thesis committee: Dr. Derek Eager, Dr. Chanchal Roy, and Dr. Seok-Bum Ko for their suggestions and insightful comments. Special thanks to Dr. Nathaniel Osgood and Dr. Christopher Dutchyn for their valuable suggestions on the algorithms I have used in my thesis. I am also grateful to Greg Oster for his help during the data collection period.

I am very thankful to my roommates and friends here in Saskatoon who provided unconditional support and encouragement for the successful completion of my thesis. Last, but not the least, I would like to express my sincere gratitude to my parents, my wife and other family members who were always there for me. I would have been lost without the support they showed me from Bangladesh.

Contents

List of Tables

List of Figures

List of Abbreviations

CATV	Cable TV Systems
CDF	Cumulative Distribution Function
CCDF	Complementary Cumulative Distribution Function
CDN	Content Delivery Network
P2P	Peer-to-Peer Networks
QoS	Quality of Service
UGC	User-generated Content
UCC	User-copied Content
VoD	Video-on-Demand

Chapter 1

Introduction

Multimedia services are achieving enormous popularity with advances in Communication and Networking technologies [50]. Over the past decade, Video-on-Demand (VoD) streaming applications, especially user-generated content sites like YouTube, have been considered one of the most popular classes of multimedia applications on the Internet [54]. Video-on-Demand, unlike traditional television, is a service which enables a client to watch a program/video according to her convenience [41, 50]. Because of this user-friendly feature, VoD has been expected to replace both traditional TV programming and the culture of DVD movie rentals, which has prompted new applications and service models [43].

A typical VoD system operates on a client-server architecture with different types of transport networks, e.g., cable TV systems (CATV) [41]. A client in a VoD system selects a video of choice, which is sent as a request to the video server. The server fetches the video file into its memory and starts delivering the requested video using a dedicated channel to the client. This enables the user to watch the video with VCR-like controls [49, 54]. In general, VoD services are classified into two subclasses: Traditional VoD services and user-generated video (UGC) services.

1.1 Traditional Video-on-Demand Services

In typical VoD systems, videos have been produced and distributed by a number of qualified professionals including production organizations and licensed broadcasters. Hong Kong introduced the first commercial VoD service in 1990. Unfortunately, that project was not economically successful because of the immature technology and lack of knowledge of PayTV dynamics and business models.[1] Nowadays, VoD services are categorized into three broad classes: free VoD, pay-per-view VoD and subscription VoD. In spite of the early failure of large scale pay-per-view VoD (Hong Kong, 1990), other two types of services have played important role to attract a large number of people to watch videos on their demands (Netflix[2] and HBO on demand[3] for examples). This increasing number of VoD users poses challenges to the service providers in maintaining a high level of quality of service(QoS) [50]. The number of concurrent channels that can be supported, however, is not unlimited.

[1] http://vod.szm.com/en.html last accessed: 22-08-2013
[2] http://www.netflix.com/ last accessed: 22-08-2013
[3] http://www.hboondemand.com/ last accessed: 22-08-2013

As a consequence, many recommendations have been offered to address this issue. Instead of following single channel per viewer, repeated broadcasting with appropriate modifications have been offered [6, 47]. The idea is that multiple users might be interested to watch a single very popular video at the same time. However, user interruptions (e.g., pause or forward) are not allowed with such kinds of video distributions. Moreover, the clients have to wait until the next broadcast time. Other techniques that were suggested are batching [4, 20, 51] and patching [24, 32]. Batching waits for a predefined number of requests and then starts multicasting among different customers. In case of patching, a client simultaneously listens to the multicast stream of the video, and collects the segments that were delivered before it joins the network. However, in these kinds of VoD services, identifying the most popular videos are not that difficult as the providers, in most cases, know in advance which videos are going to be popular [12, 57]. For example, it is a valid hypothesis to make that a movie that was in some Box-office hit list will attract large number of clients in sites like Netflix.

1.2 Web 2.0 and User-generated Content

Sites where people are allowed to get together and share personal views/content makes humanity a giant community. Web 2.0 is defined as the second generation of the World Wide Web, which focuses on user interaction and collaboration as well as sharing of user-generated contents. The conspicuous difference between Web 2.0 and prior web technologies is that, in Web 2.0, users can share their own content rather than being limited to passive viewers of content. Some of the main Web 2.0 services are Blogs, Wikis, and Multimedia sharing [48].

User-generated content (UGC) is a collection of objects on websites, such as video, text, pictures, etc., created by ordinary users rather than by marketers, administrators, or other professionals associated with a site. In UGC sites, contributions come mostly from amateurs, which makes it very difficult to predict future attraction of a particular content as content is often newly created at the time of submission to such sites [57]. The enormous freedom in uploading and watching content for free has motivated people to a great extent in producing their own materials. People do not have to go through mainstream media in order to propagate their content to the audience, which led more than 82 million people in the US to create on-line content in 2008. This number is expected to reach nearly 115 million by 2013.[4] This new rise is referred as "citizen journalism" and considered as a triumph of amateurism over professionalism.[5] The success of UGC has pushed other commercial organizations to invite people to post their own contents in the organizations' sites. CNN's iReport[6] is one of the most remarkable examples, a section of CNN website for people to upload their stories as well as pictures and videos, that achieved significant success through UGC.

In spite of not being paid for their content, people contribute to these sites for several possible reasons:

[4] http://mashable.com/2009/02/19/user-generated-content-growth/ last accessed: 22-08-2013
[5] http://www.guardian.co.uk/media/2012/jun/11/rise-of-citizen-journalism last accessed: 22-08-2013
[6] http://ireport.cnn.com/ last accessed: 22-08-2013

to connect with like-minded peers, self-expression including political campaigns, rewards and fame. UGC also plays a vital role in on-line marketing. Interestingly, 65% of users aged 18-24 consider evaluations and opinions shared on social networks in order to make a purchasing decision, indicating consumers value what other consumers say about a product lot more than the manufacturer's advertisement.[7] UGC communities can also be helpful for the manufacturers themselves. While traditional marketing research suffers from less or no customer feedback, in UGC sites people express their opinions without being asked. This information can be collected and analyzed to have better knowledge about customers.

Although UGC includes photos, text, graphics etc., notable numbers of videos are created and uploaded everyday in sites like YouTube. YouTube, with the motto "Broadcast Yourself," is the most popular user-generated video site. It has altered the way people watch video on the Internet with a huge number of video producers and consumers. Moreover, it seems becoming popular in YouTube has become vital for marketing services and products [15]. YouTube has been popular since its inception in 2005. It became the 4^{th} most accessed Internet site in 2007 [16]. Statistics from 2009 shows that 20 hours of videos were being uploaded in YouTube every minute [17]. As a consequence, YouTube was spending $1 million per day for its server bandwidth.

Recent statistics show that on average 72 hours of videos are uploaded in YouTube every minute and more than 4 billion hours of video are watched every month [42]. This estimates 3 million hours of video are uploaded each month. Besides, more than 800 million unique users are reported to visit YouTube each month.[8] Not surprisingly, according to Alexa,[9] YouTube is now the 3^{rd} most accessed Internet site, after Google and Facebook [42].

1.3 Thesis Motivation

The enormous popularity of YouTube—number of videos and users—created challenges in its video distribution performances and cost. YouTube is still the most bandwidth intensive service of today's Internet, and it accounts for 20-35% of the global Internet traffic [26, 35, 38]. YouTube performance has never been satisfactory compared to other measured sites [16]. In 2010, YouTube performance was worse than 58% of other surveyed sites [17]. Recent studies also support two central observations—increasing number of videos and users [22, 46] as well as dissatisfying experiences of users in watching YouTube videos [34] in terms of delivery speed. Alternate architectures of video distribution, rather than pure client/server model, were needed for YouTube and similar sites. A traditional client/server system is vulnerable to service bottlenecks due to inadequate resources, especially when the server is overwhelmed by the huge number of incoming requests. In fact, in order to improve the end-users experience, adaption of Content Delivery Network (CDN) and

[7] http://www.socialmediaexplorer.com/content-marketing-2/how-user-generated-content-powers-better-social-seo-results/ last accessed: 22-08-2013

[8] www.youtube.com/t/press_statistics last accessed: 22-08-2013

[9] http://www.alexa.com/topsites last accessed: 22-08-2013

multi-layer caching has been implemented by YouTube [3]. The idea is to replicate the most popular videos in the CDN.

Identifying the most popular videos in a given time, however, is much more challenging for YouTube than in traditional VoD systems. YouTube videos are not generally produced by professionals. The amount of content uploaded to YouTube in a 2 month period is more than all the content that can be aired for 60 years altogether by NBC, ABC and CBS, because of the freedom and flexibility in uploading almost unrestricted number of videos [42]. This phenomenon leads to a very asymmetrical viewing distribution among contents in YouTube and similar sites [14, 22, 39]. Careful characterization is needed to understand video popularity dynamics along with discovering the processes that dictate popularity over time. For example, identifying the most popular contents at their early ages in order to fetch them into the replica servers is considered as one of the most promising solutions in reducing bandwidth cost and improving user experience [12, 42, 48].

Much research has been done investigating request characteristics from both the client [27, 57] and server side [10, 14, 22, 25] in order to understand YouTube traffic pattern and thus improving service to YouTube users. However, none of this earlier work considered the types of video objects while analyzing the growth pattern of YouTube videos. This aggregate data may not tell the whole story. Szabo *et al.* [48] observed significant errors in the first week of videos popularity prediction while evaluating a proposed popularity prediction model. This can be because of the non-stationarity observed in videos early popularity [10]. However, Szabo *et al.* anticipated that different categories in YouTube might have different growth curves, and thus, category-based models could lead to more accurate predictions. In fact, Johnsen *et al.* [33] observed very short active life spans of News videos in News-on-demand VoD services, which is different than the findings of other studies based on YouTube without considering video types [10, 25].

This thesis starts with the hypothesis that the growth patterns of YouTube videos are related to the videos' categories. The category-specific methodology, proposed in this thesis, would be useful for UGC sites that have a single cache for the region of requests captured. YouTube operates on such a global scale that a single cache would not be sufficient. Rather, multiple regional caches satisfy regional demand patterns, which has been shown to be significant between different regions in the world [11]. If regional request data was available through the standard YouTube API, recommendations for multiple caches could be made in this thesis. However, Mitra *et al.* [39] observed similar viewing characteristics among different user-generated video distribution sites that are popular in different regions in the world. This intensifies the confidence that although YouTube videos are mostly local, popularity patterns of videos in different regions could be very similar.

The fundamental questions this thesis addresses are the following. Do YouTube video categories follow significantly different popularity dynamics so that category-specific characterization can be more beneficial for the network engineers and marketers? Are the earlier techniques, such as baseline model [48] or three-phase characterization [10], able to be applied for modeling category-specific growth patterns? Does the users interaction with videos depend on video categories (which videos attract the same user multiple times)?

The answers to these questions will enable more accurate workload prediction models for simulations and deployments of UGC servers.

1.4 Thesis Contributions

As previously mentioned, user-generated content delivery via the Internet is interesting but has challenges for efficient distributions. Understanding demand in more detail will help design infrastructure to meet that need.

In this thesis, *the time-varying global viewing patterns of different sets of YouTube videos from their introduction into the system are analyzed, considering the categories of videos (as defined by their uploaders). Characterization is done from those aspects of video popularity growth patterns that can be used for caching mechanisms and advertisement policies. Finally, the characterization enables the development of category-specific workload generators which can be combined to form the input for simulators and prototype systems. Local viewing patterns are analyzed briefly to give insight into further specific details that may influence local caching policies.*

Patterns of users' interaction with YouTube videos are investigated using a local dataset [57]. This helps to recommend the possible improvements in video distribution techniques under a local network (P2P, local caching and proxy caching). In addition to YouTube and other similar UGC sites (Dailymotion and Metacafe for examples), these observations can also be useful for category specific video on demand sites (e.g., sports.yahoo.com, news.yahoo.com, Netflix). A proper understanding of the most popular VoD system's workload will aid in design of new systems, capacity planning, and network management for similar types of systems. The findings of this thesis can be summarized as follows:

- Different categories in YouTube follow different viewing patterns—some categories enjoy high average views for only the first couple of days of their lifetimes, whereas some other categories enjoy a constant viewing pattern over time.
- Fractions of popular videos vary significantly according to the category.
- For some of the categories, the future popularity of videos can be predicted at their very early ages.
- Number of views of the popular videos for most of the categories follow a Zipf like distribution whereas views of the unpopular videos form a light tail portion and fits better with Weibull distribution.
- The uploading trend in YouTube is changing over time. People are now uploading more user-generated contents (UGC) compared to the earlier observations.
- Time-series clustering, such as K-Spectral algorithm, can be applied to model the growth patterns of YouTube categories.

- Repeated views of YouTube videos have a strong correlation with video categories, which could be used to design a better cache replacement policy than Least Recently Used (LRU) mechanism.
- YouTube users form separate communities based on their interests in YouTube videos. This community information is useful for effective video recommendation systems.

These observations can contribute to a better understanding of the popularity dynamics of YouTube videos, which can, in turn, be used to develop a more realistic workload generator for YouTube. This enhances the ability to develop and evaluate various caching mechanisms for YouTube and similar UGC sites. Moreover, successful prediction of videos' future popularity can be instrumental in appointing appropriate advertisement policies for YouTube and similar UGC sites.

1.5 Thesis Organization

The rest of this dissertation is organized as follows. Related work is presented in Chapter 2, where characterization of UGC from different aspects including user's interaction is analyzed. Chapter 3 presents the method and summary of data collection. Global viewing characteristics of different YouTube categories is presented in Chapter 4. Process of category-specific workload generation is presented in Chapter 5. Interaction of individual users with different categories is presented in Chapter 6. Chapter 7 concludes this dissertation along with some directions to future work.

Chapter 2

Related Work

It has been a challenging task to understand the growth patterns of on-line content. This becomes more difficult for UGC systems where lots of users interact with an enormous amount of content. YouTube, the most popular video site, attracted lots of researchers in this area of content popularity analysis [10, 11, 12]. It has been observed that most of the user-generated video sites (e.g., YouTube, Dailymotion, Metacafe etc.) follow similar characteristics including popularity dynamics and view distribution among videos [39]. This indicates that observations from careful characterization of any of these systems can be used for the others. This thesis concentrates on YouTube, as it is the most popular and thus may be helpful to indicate what problems may be faced by others in the future.

This chapter is divided into two sections. In the first section, earlier work on YouTube popularity characterization is explained. The factors that influence video popularity in YouTube are analyzed in detail. Uploaders behaviour, content aliasing, comment rating, and different external and internal referrers have been found to be related to videos popularity. In particular, more attention is given to understand popularity dynamics. The second section describes the characterization of other on-line content systems, such as Peer-to-Peer(P2P) file sharing, and News-on-demand. This helps to find the similarities/differences between YouTube and other sites.

2.1 YouTube

2.1.1 YouTube Video Characteristics

A detailed investigation of the characteristics of YouTube videos and their request patterns was done by Cheng *et al.* in 2007 [16]. Information from approximately 2.6 million videos was collected by following the related video links of some popular videos in the YouTube population. The estimated number of uploaded videos in YouTube was around 42.5 million by that time. The main characteristics considered were video length, category, active life span and relationship to other videos.

The uploading rate of YouTube videos could be fitted with a power law curve, and out of 15 categories,[1] Music and Entertainment videos were found to be uploaded most frequently. In case of video lengths, almost

[1] http://support.google.com/youtube/bin/answer.py?hl=en&answer=94328 last accessed: 22-08-2013

98% of the videos were found to be less than 600 seconds in playback duration. This is due to the limit imposed by YouTube on video length in 2006, which is why videos for television shows and movies uploaded during that time were found in several segments. No correlation is found between video length and video popularity. In spite of having a heavy tailed portion in the popularity distribution curve of YouTube videos, distribution for the popular videos in YouTube follows Zipf distribution. This implies that popular videos of YouTube are as popular as Zipf's law predicts.

With respect to *active life spans* of videos, investigation suggests that most videos have been watched frequently only in a short span of time. These characteristics can be fitted well by a Pareto distribution, which indicates the low probability of watching a video after its active life span. Finally, the YouTube video network is found to be similar to the small-world network as the graph of related videos in YouTube exhibits similar characteristic path length and clustering coefficient to small-world networks.

Considering the small-world properties of YouTube network, this paper concludes that the Peer-to-Peer technique, with proper modifications, can be employed to improve the performance of YouTube and similar sites. Even in case of employing proxies close to the clients, approximately 80% hit-ratio can be achieved with only 8GByte of disk space, using prefix caching of related videos. That is, if a group of videos are significantly related to each other, then a user is likely to select another video from the same group after finishing the current one.

The first set of data collection was based, however, on some standard feeds provided by YouTube API that only return popular videos. Therefore, collection of information of videos by using related links of those videos has a high probability that the dataset contains information for popular YouTube videos only. Although this kind of dataset can be used to evaluate the caching policies, it is unlikely that appropriate understanding of video distribution sites can be gained by ignoring the characteristics of the dominating number of unpopular videos. For example, while fitting the viewing pattern of YouTube videos with Weibull and Gamma distributions, it is likely that the shape and scale parameters for both of the distributions might be changed if an unbiased data set is used. Moreover, for unbiased data set, the tail section of the distribution would be longer than the authors found.

2.1.2 Popularity Growth Pattern of YouTube Videos

To observe the time-varying popularity of YouTube videos (crucial for efficient object caching), Borghol *et al.* [10] collected information of 29,791 YouTube videos by using the Most Recent standard feed provided by the YouTube API. Their collection procedure was good enough to have an unbiased dataset; the Most Recent standard feed returns video information randomly of videos that were uploaded very recently, regardless of their number of views. Their investigation shows that most of the videos achieve their peak popularity within fewer than six weeks from their uploading time. Moreover, as an approach to investigate whether or not the current popularity of a video is an indicator of future popularity, Pearson's correlation coefficient was calculated between added views at consecutive snapshots computed on a weekly basis.

The correlation coefficient between snapshots four and five is found to be very weak. This coefficient becomes approximately 0.7 between the snapshots eight and nine, and increasing to 1 between snapshots 16 and 17. This is because of the significant rank shift among videos at their very early ages as depicted in Figure 2.1. This observation suggests that current popularity of an older video can reflect its immediate future popularity, but this is not the case for a very young video. Three-phase characterization was performed to further understand the growth pattern and thus developing a workload generator for YouTube. The authors placed YouTube videos into groups according to whether they were after, before or at peak in a particular week. Videos under the same group were found to follow very similar view distributions over time, which is week-invariant. As a result, the average viewing rate was found nearly constant over time for the videos pertaining to a specific group as in Figure 2.2. In spite of little variation in the curves, no particular trend is observed. These observations encouraged the authors to model the popularity dynamics of YouTube using only three fixed distributions with a known peak distribution. Acceptable accuracy was found comparing synthetic and empirical data.

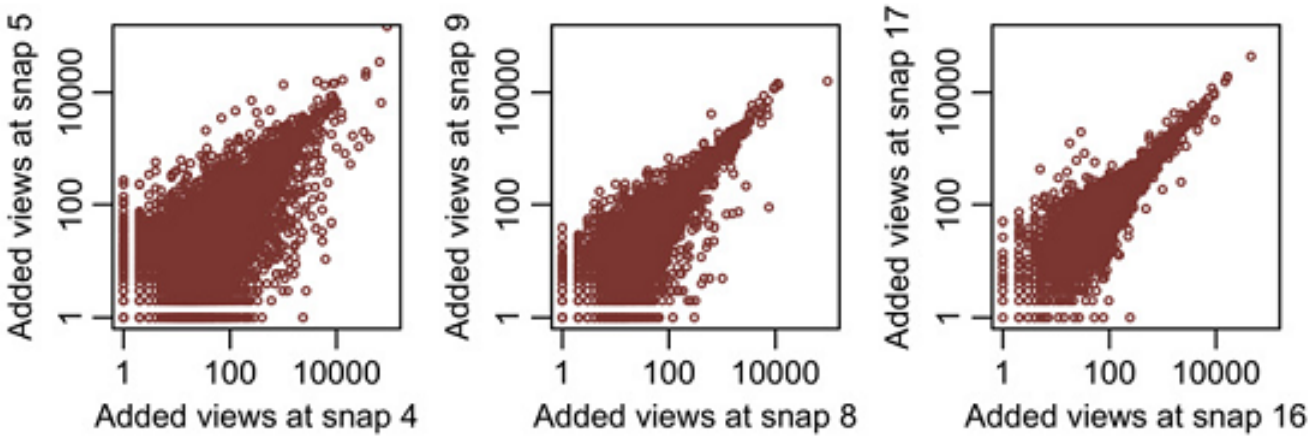


Figure 2.1: Number of Views at Two Adjacent Snapshots [10]

Unlike most of the earlier work, video information was collected randomly, which helps to characterize YouTube videos with the least possible known bias. However, it would have been better if the prediction of future popularity were conducted only for the popular videos. It is expected that viewing patterns of the unpopular videos follow a different distribution than the popular ones, as shown by Cheng *et al.* [16] even for their biased dataset.

Figueiredo *et al.* [25] applied a novel technique, Google charts, to collect the number of views over time for YouTube videos. Their results suggest that popular videos usually experience huge number of views on a single peak day or week. Then the time varying viewing patterns of popular videos, deleted videos and randomly selected videos were analyzed separately. They found that videos that were deleted because of copyright violation, tend to get most of their views much earlier in their (short) life times than other videos,

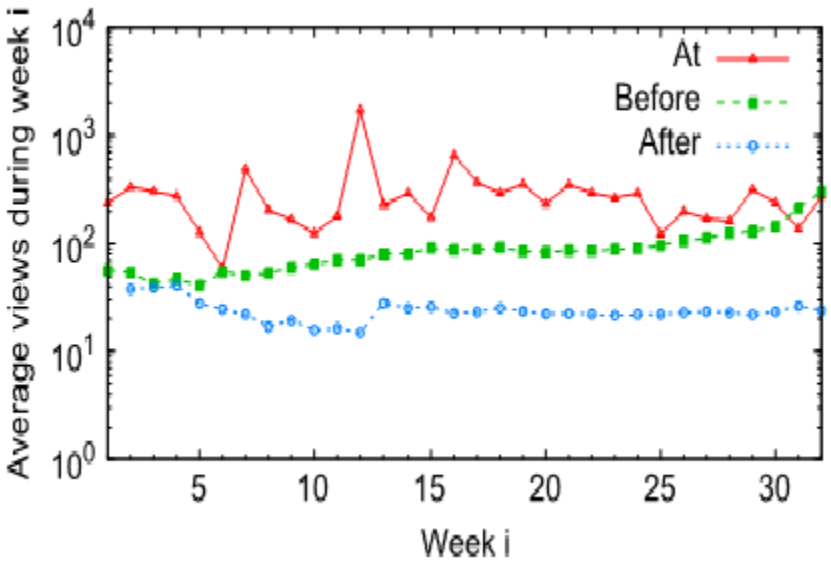


Figure 2.2: Viewing Rate for Videos At, Before and After Peak [10]

when lifetime is measured as the upload date until the time of data collection (April 2010). For instance, for half of the videos in the popular, deleted and random datasets, it takes at most 65%, 21% and 87%, respectively, of their lifetimes until they experience at least 90% of their total views. For 50% of the total views, it takes 26%, 5% and 43% respectively for the previously mentioned three datasets. In addition to popularity over time, they also investigated the impact of different types of referrers (internal and external), that can positively influence the views of a video. Out of all different referrers, Featured and Social referrers have significant impact on the number of views.

The attributes of the set of videos is not adequate to have a proper understanding of the dynamics of video popularity since the Google charts API provides at most 100 data points for each video, regardless of video age. This procedure limits the details of the viewing pattern; weekly or daily viewing patterns are unavailable for many videos.

It is another research issue to identify whether one referrer might influence the number of views from other referrers. For instance, a popular video may experience further popularity growth from Social referrer after being featured by YouTube. Similarly, it may first receive a large number of views from Social referrer; thus leading it to be featured by YouTube. Although the videos that violate copyright laws experience most of the views early in their lifetimes, the actual time until deletion is not mentioned. Therefore, the metrics on views over time for deleted videos are of questionable validity, as some of the videos could have actually only existed for a few days and been uploaded as early as 2007 (when the YouTomb[2] project, from which the videos were selected, was initiated), and the lifetime measured of greater than one year, would include time during which the video was unavailable for viewing.

[2]youtomb.mit.edu last accessed: 22-08-2013

The growth pattern of social videos—videos that enjoy most of their views from Internet sharing—was observed in detail by Broxton *et al.* [12]. Growth patterns were analyzed for 1.5 million YouTube videos. The videos were selected randomly in order to capture an unbiased dataset. Videos were then classified as social and non-social based on the fraction of social views. For example, views come from YouTube search or related video links were considered as non-social source whereas requests directed from external links were considered social. Search and related video links contribute more than any other referrers in YouTube's overall popularity. As the social videos receive very few requests through these two referrers, their active life times were found very short compared to the non-social videos. Short active life time was not found as an aberration even for the very popular social videos. This phenomenon suggests that identifying a social video at its very early age is very crucial for efficient caching decisions. The authors, using their observations on these videos, developed a ranking criteria for websites and blogs that indicates the ability of a site to make a video popular. This methodology, however, can be applied only for social videos that represent a very small fraction of YouTube population. Moreover, all the videos shared in a site (e.g., Facebook, Twitter) do not become that popular to be considered for caching, which makes the methodology more difficult to apply.

Future views of videos can be predicted if there is a good relationship between early and late popularities. Szabo *et al.* observed significant correlation between the long-term views of YouTube videos and their early views [48]. This observation led them to design a model that predicts, based on a linear function, future popularity of a video (views at *reference day*) when the present popularity (views at *indicator day*) is fed as the input to the model. A constant additive factor is needed to cope up with the noise. This additive factor varies based on the indicator and reference days. This model, in spite of being simple, shows acceptable performance, especially when the indicator day is more than a week of the uploading time. The drawback of this model is that it considers two videos to have same number of views at a reference time if their indicator times' views are same. In reality, two videos can have (e.g., News and Music) very different future popularity even though they enjoyed similar early views. In a word, this model uses the preferential attachment or *rich-get-richer* phenomenon. Moreover, the measurement period (30 days only) and the number of videos (7000) are too low to be very confident about the findings.

Motivated by these drawbacks, Pinto *et al.* offered two extended models to predict future popularity of YouTube videos more accurately [42]. The first one was multivariate linear regression model (ML), which considers not only the early views but also the historical growth patterns of a particular video. The basic idea is that the number of views of a video for a period of time are not uniformly distributed. As a consequence, different weight values are assigned to the measured days to design the future prediction model. This leads two different videos to have different numbers of views in a reference day in spite of their similar number of views up to an indicator given that they had different growth curves. The second proposed model used a training set of known growth patterns of different videos. A video is compared with the training set to find another video that has similar growth pattern with the video of interest. Then the future views are assigned according to the known video's popularity. These two models were found to outperform the prediction model

by Szabo *et al* [48]. However, these models are more complicated and can not be applied when the history of viewing pattern is not available. Furthermore, these two models are computationally expensive when the history being considered is very long.

Crane *et al.* [19] considered the endogenous and exogenous effects on the growth curves of YouTube videos. Videos placed on the front page by YouTube authority were considered to gain popularity by endogenous effects. On the contrary, exogenous effect was attributed to the videos that were in the list of "most-viewed-today" as these videos mostly become popular from users' interests. Unsurprisingly, the videos that become popular by exogenous effect were observed to retain their popularity much longer than the endogenous ones. Further, based on these popularity evolution patterns, YouTube videos were classified into four different groups. A significant portion of the videos that shows no noticeable peak were referred as *Memoryless*. Growth pattern of this class can be designed using a stochastic process. Videos with significant peaks, that enjoy very symmetrical popularity before and after peak were referred as *Viral*. On the contrary, *Quality* videos suddenly become popular and retain their popularity after the peak at least for a significant amount of time. *Junk* videos, however, observe sharp rise and sharp decay in the popularity curve.

2.1.3 Content Aliasing in YouTube

Video popularity can be affected negatively by their clones—a group of videos with very similar contents. This section focuses on content aliasing in YouTube along with a clone detection process and impact of clones in video popularity. A video can become unpopular when another video with the same content is uploaded by a more popular uploader or the late uploaded video offers better quality to the users [9].

Pedro *et al.* [40] investigated content duplication and overlap in YouTube. In order to detect duplicate scenes among different videos, content-based copy detection tools (CBCR) has been used. Sets of graphs were formed such that the edges in a graph represent highly related videos in YouTube. The components of the fingerprint-based CBCR can be described as having three steps: fingerprint generation module, reference content database and search module. In the fingerprint generation module, all videos are transformed into a sequence of points in the fingerprint feature space. The reference content database is a database of known fingerprints that can be developed using supervised training. Finally, in the search module step, fingerprints for all incoming video streams are compared with the reference content database. In order to evaluate the effectiveness of the CBCR technique, a pilot experiment was conducted. For this experiment, a reference content database was formed by collecting randomly 2000 music videos based on private contributors and TV programme producers. Duplicate videos were eliminated manually in order to ensure that the database does not have redundancy. From the set of 2000 music videos, 200 videos were selected randomly. Artist and song names of these 200 videos were used as keywords in order to search YouTube. Although YouTube API's keyword-based search provides up to 1000 videos' information, only the top 10 of those were selected in order to have a more relevant list of videos. After manual screening, almost 550 near-duplicate videos were found; the videos that are not completely identical but have significantly similar number of screens

are defined as near-duplicate. From this pilot project, approximately 90% accuracy has been found for the fingerprint-based CBCR.

As the final step to investigate the content redundancy in YouTube, 703 queries were collected by using top 10 gaining weekly queries provided by Google Zeitgeist.[3] After filtering, 579 queries were used to collect 28,216 videos information. Although this dataset was biased towards more popular videos, it does not hamper the analysis as it is expected that relatively popular videos suffer more from content aliasing. Results suggest that, from the collected dataset, almost 16% of the YouTube videos suffer from content duplication along with significant amount of overlapping among videos. Figure 2.3 shows how the popular videos suffer more from content duplication than the less popular ones; the X-axis shows the order of the search results (from the top 30 search results) and the Y-axis is the distribution of duplicate videos.

The authors of this paper claim that this video duplication happens mainly for two reasons. Firstly, many users re-upload popular content in order to increase their popularity as a uploader. Secondly, many users upload different versions of a video with the subtitle in their own language, which is referred as multilingualism.

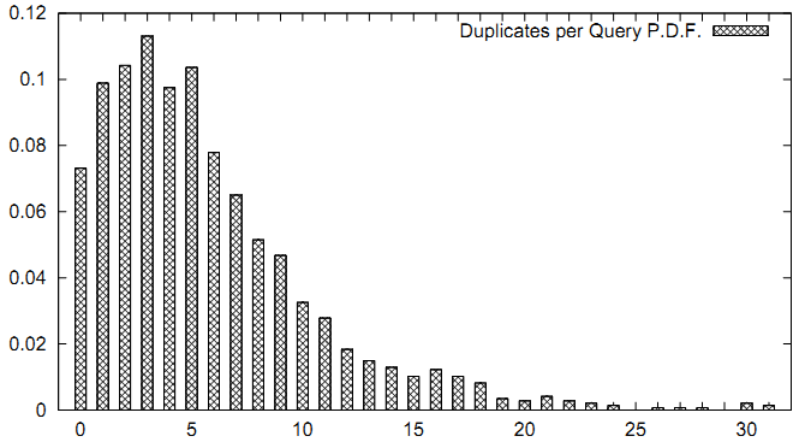


Figure 2.3: Distribution of Duplicates among Different Queries (Top 30 Search Results [40])

However, some cases of duplication (e.g. videos with a common descendant) were not considered, although common ancestor of different videos was considered during the detection process. Most importantly, impact of content aliasing on the original videos—considering the first uploaded video as the original one—are not presented, which might be very crucial for the on-line marketers. For instance, Cha *et al.* [14] show that total view counts from different copies of a video can be more than two orders of magnitude than that of the original version. Unfortunately, the dataset is not rich enough to estimate the actual amount of content duplication.

The impact of content aliasing on videos' popularity, along with other factors, was investigated in depth by Borghol *et al.* [9]. Although a video's popularity primarily depends on the content subject/quality, other

[3]http://www.google.com/zeitgeist/ last accessed: 22-08-2013

factors such as uploader's number of subscribers and previous success as well as the number of associated keywords with the video can play very important role; keywords usually help a video to become more visible to the users through YouTube's search referral. Moreover, for older videos, popularity history effects current number of views of a video, i.e., *rich-get-richer* phenomenon. Importance of these factors on video popularity can be well understood by analyzing the popularity of clones in YouTube. The authors defined a clone set as a collection of videos that have very similar content and duration. It has been found that videos, in spite of having similar content, enjoy significantly differing views because of the difference among the uploader's social network size as well as the number of keywords associated with the videos. In general, the earliest uploader of the similar content accounts for most of the views. However, a video can become significantly more popular, even in case of late uploading, than all other videos in its clone-family just because of its uploader's popularity. This observation can be instrumental for making caching or advertisement policies for very young videos when the popularity prediction is impractical.

The dataset size in this research is, however, very small compared to the YouTube's population. Only 48 clone sets consisting of 17 to 94 videos were investigated. No indication is given about what percentage of videos in YouTube suffer from content aliasing.

2.1.4 Playback Quality Concerns/Potential Solutions

Dissatisfying experience of YouTube users in watching videos was illustrated by Khemmarat *et al.* [34]. The authors also proposed and evaluated a promising solution to improve latency in delivering videos to the end-users. At first, an experiment was conducted to evaluate user experience in watching YouTube videos—the number and duration of pauses during video playback. The information of pause frequency was collected automatically by examining video download traces. 12 volunteers from 12 different environments (different network access technologies) were asked to use the Wireshark network protocol analyzer[4] to capture YouTube traffic. A model was developed to estimate the number of pauses in playback. From the sample dataset, it was found that 10 out of 12 environments contained playbacks with pauses, and 41 of 117 playbacks contained pauses, which represents approximately 35% of the total playbacks. This observation demonstrates that YouTube users experience noisy playbacks, possibly more significantly for higher quality videos. This problem can be intolerable in resource-poor network environments as high definition videos become increasingly popular in YouTube. The authors suggest that prefetching can be applied to solve this problem. Two different kinds of prefetching agents (PA) were considered: PF-Client and PF-Proxy. PF-Client is dedicated only for one client and is located at the client whereas PF-Proxy is located at proxy server and serves for all the client under the same proxy. All the YouTube requests from a client are directed to the PA. The PA serves the client with the prefix of the video if the video is available in the local server, and starts retrieving the remaining part of the video from the YouTube server. If the prefix is not found locally, the PA retrieves the whole video from YouTube and sends it to the client. Two different referrers were used to select videos for

[4]www.wireshark.org last accessed: 22-08-2013

prefetching: YouTube search results list and related video lists, as these two lists were found as the two most frequently used referrers. Although these two referrers return up to 25 videos' titles in the list, it was quite challenging to estimate the actual number of videos that need to be prefetched for optimal performance. The top N videos were selected for prefetching where the value of N was varied in different parts of the experiment.

Figure 2.4 shows the hit ratio of different prefetching techniques against different values of N. For example, SR-N/PF-Client represents the hit ratio of the PF-client agent that prefetches the top N videos when the search referrer is used. PF-Proxy—one agent for all the local clients—outperforms all other techniques given that videos are selected for prefetching from the related video list provided by YouTube. Moreover, only the top 15 videos are enough to store so that 75% hit ratio can be obtained. Although this part of examination was conducted for an infinite cache size, similar results were found for reasonable amount of cache size. Interestingly, this paper found that a combination of caching and prefetching can increase the hit ratio by 5-20% as compared to the prefetch only mode, depicted in Figure 2.5.

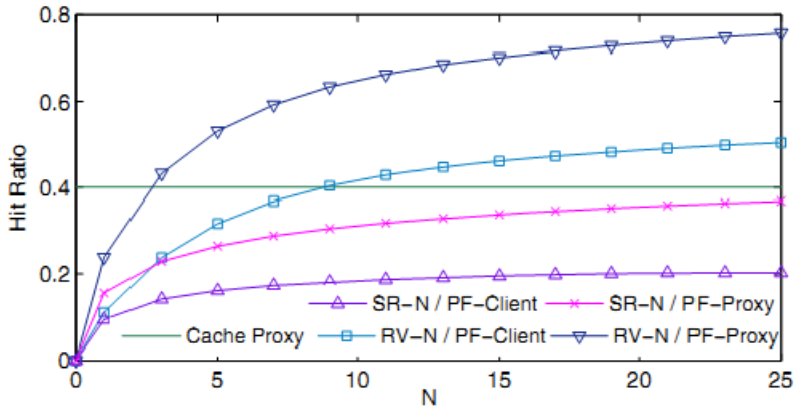


Figure 2.4: Performance of Different Prefetching Techniques [34]

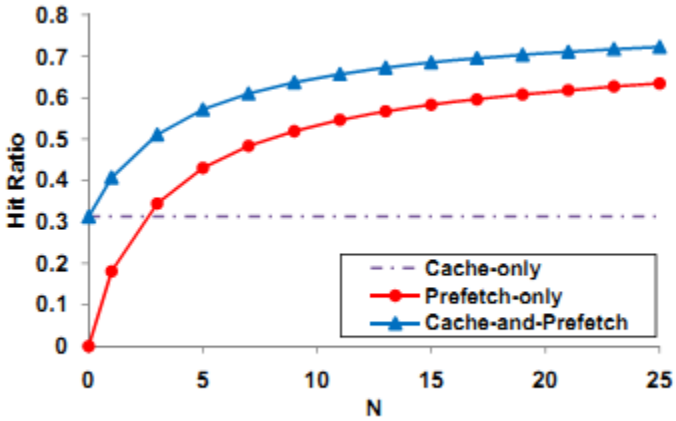


Figure 2.5: Improvement of Hit Ratio after Combining Prefetching and Caching [34]

However, the amount of data required to resume from pausing had to be estimated as the actual amount used by YouTube was unknown. Moreover, it would have been better if the performance of prefetching was examined by using other referrers like Most Viewed, and Top Rated. Besides, the performance of prefetching was not compared to other potential techniques like batching; batching improves playback quality as well as reduces the network bandwidth required via multicasting [45].

2.1.5 YouTube video Traffic Under Campus Networks

From an extensive review of the prior literatures, no earlier work is found that investigates how individual users interact with YouTube categories. This analysis is important for designing video distribution techniques in a local network. Zink *et al.* [56] examined You Tube traffic between YouTube servers and University of Massachusetts campus network. Three different periods were used for this measurement in 2007 and 2008. The result shows that only approximately 25% of all requested videos were requested more than once. Surprisingly, no correlation was found between local and global popularity of YouTube videos, indicating YouTube video requests have a different popularity characteristic based on the region of viewing. Three different content delivery techniques were examined: proxy caching, client-based local caching and P2P-based distribution. In local caching, a video is stored in user's own memory for future requests, so that repeated requests can be satisfied without any interaction with the YouTube server. This has the benefit of reducing bandwidth consumption and latency. Local caching, however, can not take the advantage of fetching the same video from another user in the same subnetwork. Thus, a client has to forward its request to the YouTube server in spite of a local copy existing at one of its neighbours. A potential solution to this problem of local caching is to employ P2P for video sharing. In P2P, a user first checks if the requested video is stored in its own memory. For a negative response, this video is searched throughout the scope of the network to find if any peer contains that video. Otherwise, the request is forwarded to the main server. The P2P technique, because of the scalability it offers, has been considered as the most promising solution for YouTube like sites [16]. This technique, however, also has its own drawbacks. Peers holding copy of a requested video have to be on-line at the appropriate time to satisfy the requests.

The proxy caching mechanism—one cache for all users—outperforms all other techniques and exhibits an effective low-cost solution. The size of the proxy cache was varied between 100 MB and 150 GB, and when the cache size changes from 100 MB to 1 GB, the performance increases 10%. Finally, maximum performance was found when the cache size was 100 GB. P2P-based caching shows worse performance than the client-based caching architecture. This is because of the peers unavailability when another peer looks for a video contained by the unavailable peer. In their later work [57], using two more datasets, significant improvement in P2P distribution was observed when multiple copies of a video were stored among different peers in the network. The results of their simulation illustrate that, compared to the other content delivery method enhancements, client-server architecture for example, caching is more effective to decrease network traffic and latency. The longest measurement period, considering both of the studies, was only 14 days. This

limits the opportunity to evaluate the aforementioned distribution techniques for a significantly long period of time. The number of videos that were requested multiple times, increased substantially (more than 35%) for a measurement period of 14 days.

A similar experiment was conducted by Gill *et al.* [27] by collecting the traffic information of YouTube videos in University of Calgary campus network. They investigated file properties, usage patterns, and transfer behaviours of YouTube videos along with the social networking aspects. While analyzing the global usage patterns, the authors found that 52.3% of the videos in the all-time popular category were between 3 and 5 minutes in duration. This suggests that, on the contrary to other studies claiming no relation between video length and popularity [1, 57], short videos observe higher long-term popularity than longer videos. However, this does not necessarily indicate that short videos are always popular. Interestingly, 73% and 5% of the videos that were requested on Campus were at least one month and one year old respectively, suggesting campus users are not exclusively attracted to the new videos. This may also indicate that these users are usually interested for videos propagating through social networks like Facebook. Music, Entertainment, Comedy, and Sports were found as the most popular categories on Campus, among the 12 YouTube video categories during that time. Surprisingly, only 3% of the watched videos were News & Politics, although this category was found to be very popular in global scale on a daily basis. Unlike the global request pattern observed in other works [1, 17], YouTube video frequency distribution was found to follow Zipf distribution in a local network. Zipf-distribution in video popularity suggests that appropriate caching decisions not only can improve the end user experience, but also reduce network bandwidth requirements.

These two results can be biased, however, by the measurement locations which appropriately restrict the context of the studies and the solutions that are proposed. For instance, finding video requests in YouTube similar to Zipf distribution is controversial when the analysis is performed on a global scale. For the purposes of this thesis, global access patterns are essential.

In the Least Recently Used (LRU) mechanism, followed by Zink *et al.*, a client stores a video in its own cache, only if that video is not stored anywhere in the network. If the cache space is full, it replaces a video that has been used least recently with the new one, without evaluating if the user might be interested to the new video again in future. This might decrease the cache hit rate if the user, in future, requests the replaced video multiple times, and it is no longer contained by any other on-line peers.

2.1.6 Regional Popularity of YouTube

Brodersen *et al.* [11] investigated the relationship between locality and popularity of YouTube videos. The number of daily views for more than 20 million videos were collected. Including official states and minor territories, this paper considered 250 different regions for the analyses. There were about 40% of YouTube videos that enjoy at least 80% of their views in a single region. This evidence indicates that many YouTube videos tend to become popular in a locally confined area, rather than in a globally wide region. Different categories were found to exhibit different patterns of global and local popularity. Therefore, it would seem

that many factors contribute to enable a video to attract viewers from all over the world, as this occurs rarely. Likewise, strong correlation is found between the location of a video's uploader and its regional popularity. For instance, because of similar interests, videos uploaded from the USA exhibit similar popularity in UK, Mexico, and Canada. On the contrary, videos uploaded in Japan and Brazil enjoy on average 90% of their views in their uploading region only.

The impact of social sharing on YouTube videos popularity is investigated as well. Although the amount of social sharing experienced by YouTube videos is different for videos with different number of lifetime views, very surprisingly, the impact of social sharing is found to be significant for unpopular videos, while for popular videos social sharing is less prominent. On average, a video tends to become popular and to peak in its own focus location (where a video has most number of views in its lifetime), and only then this video becomes popular in other regions.

Regional popularity characteristics of YouTube videos, concentrating more on Latin America, was also investigated [23]. It has been indicated that geography influences popularity of YouTube videos. YouTube was the 4^{th} most accessed site in Ecuador and Venezuela, which was 5^{th} in Mexico and Chile and 6^{th} in and Argentina and Brazil. In terms of number of uploaded and watched videos, Brazil, Mexico and Argentina led Latin America. However, compared to other regions in the world, such as the USA, the number of videos that are uploaded or watched by Latin American users is very low. Even these users were found to be reluctant to use social features offered by YouTube. The authors claimed that Latin American users were constrained by the lack of broad-brand infrastructure. The videos uploaded in Latin America were usually less popular than other videos (e.g., videos uploaded from the US), and 76% of their views had come from only 10% of the uploaders. The regional popularity (number of views) of YouTube videos also deviates from a Zipf distribution because of the long tail. However, number of comments and responses were found to show Zipf behaviour in all the considered regions. Interestingly, the duration of videos uploaded in different regions were very similar.

These findings can be instrumental for the design of local caching mechanisms and advertisement policies of YouTube and similar content distribution sites. As different regions exhibit different popularity characteristics, similar caching strategy might not yield optimal results in different parts of the world.

2.1.7 YouTube Workload Analysis and Generation

Abhari *et al.* [1] designed a workload generator for YouTube, and then evaluated the performance of proxy caching with two different datasets. The first dataset (popular dataset) is collected by using the standard feed Most Viewed in a day and Most Viewed in a week provided by the YouTube API. On the other hand, for the second dataset (regular dataset), the Most Discussed, Most Viewed, Recently Featured, and Top Rated standard feeds were used first. Data collection was continued by following the related links of the first two datasets and thus ensuring a significantly large video population. This paper then characterized the properties of YouTube videos. The similar crawling approach provided results similar to Cheng *et al.* [16] in

terms of distribution of video lengths and correlation between length and popularity. Likewise, popularity of YouTube videos was found to fit with a heavy-tailed Weibull distribution. The amount of time that a video remains in the most popular video list is also examined. The short active life span of the popular videos is confirmed by observing the daily Most-Viewed list provided by the YouTube API.

Based on these observations, two different workload generators were developed: server workload generator and client session generator. The server workload generator simulates the files available on the YouTube server, whereas the client session generator simulates user accessing the server by selecting a video from available videos. The client session generator was designed in a way that videos with the larger value of view counts are more likely to be selected by the client. A Poisson distribution was used to generate subsequent requests from a client. The performance of proxy caching was measured according to the request patterns generated by the workload generator. Figure 2.6(a) and 2.6(b) depict the performance of proxy caching considering infinite and finite cache size respectively. When the cache size was considered finite, Least Recently Used (LRU) technique was applied for video replacement. Figure 2.6(a) shows that with a larger number of requests from the clients, both daily and weekly traces have a higher hit ratio, and a better hit ratio is found for longer traces (weekly) than shorter traces (daily). On the other hand, the hit ratios achieved by proxy caching and LRU policy are in the range of 12% to 90% for different cache sizes.

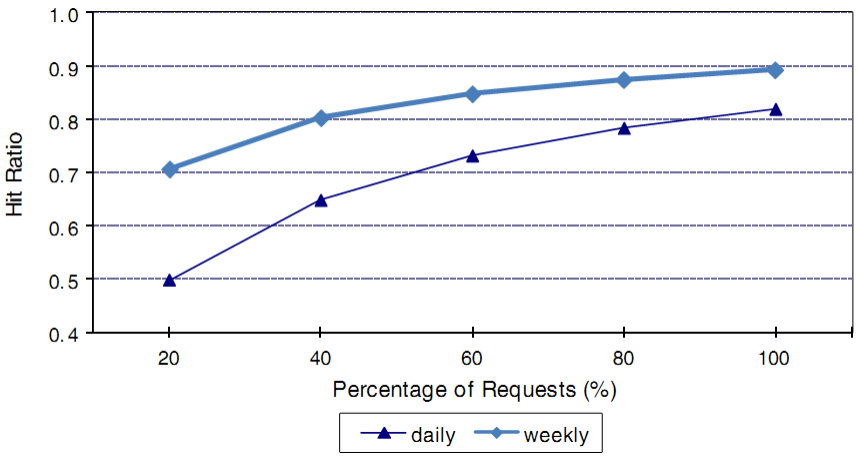
Datasets of this paper are biased to the popular videos and suffer from similar problems as observed by Cheng *et al.* [16]. Moreover, while generating subsequent requests from a client by the workload generator, it was considered that a client does not send a new request without watching the earlier requested video completely, which is not the general case for YouTube videos. Likewise, the category of video object was not considered at all in this paper, which might be crucial to understand the actual growth pattern of YouTube videos.

2.1.8 Predicting Comment Ratings in YouTube

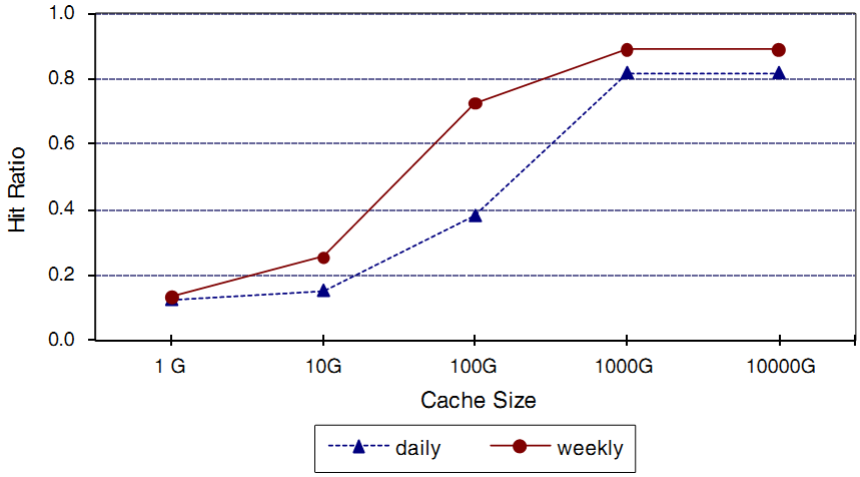
Popularity can also be related to the number/rating of comments experienced by a video. Predicting comment ratings is useful to predict a video's future popularity as Chatzopoulou *et al.* [15] found strong correlation among YouTube videos' total views, number of comments, number of ratings and number of favorites.

The predictability of comment ratings was investigated by Siersdorfer *et al.* [46]. More than 6 million comments on 67,000 YouTube videos were collected to analyze the dependency between comment ratings and sentiment expressed in a comment. To calculate both positive and negative sensitivity, the publicly available SentiWordNet thesaurus was used. In SentiWordNet, a word is represented by three sentivalues called positive, negative, and neutral. The sentivalues are in the range of [0, 1] and sum to 1 for each triple. For example, a triple (0.875, 0.0, 0.125) represent a good word in SentiWordNet whereas (0.25, 0.375, 0.375) usually represents a bad word. Sentivalue for a comment is calculated by computing the averages of positive, negative and neutral values for each word.

The distribution of ratings is asymmetric for positive and negative ratings in YouTube, and in fact, the



(a) Infinite Cache Size



(b) Finite Cache Size

Figure 2.6: Performance of Proxy Caching [1]

data indicates that YouTube users tend to cast more positive votes than negative. Interestingly, 50% of the comments are considered to be neutral. Category dependencies of ratings are also investigated, and it is found that due to the impartial nature of the Science videos, they present a majority of neutral comments. Politics videos are found to have significantly more negatively rated comments compared to other categories, while Music videos enjoy more positively rated comments than all other categories.

Different categories of YouTube tend to attract different kinds of users and produce more or less discussion as a function of the controversy of the topics. Further analysis found that rating of a comment can be predicted to some extent. The findings are useful for promoting interesting comments even in the absence of community feedback. In other words, automatically predicted comment ratings can be helpful as a supplementary ranking criteria for search results. However, this work would have been much better if the results were verified by using another tool besides SentiWordNet. The sample size of 67,000 randomly selected videos is not enough

to draw a conclusion on this complex issue. Many more videos' information could have been collected using the YouTube API.

2.1.9 YouTube Uploaders

Ding *et al.* [22] examined the uploaders' behaviours in YouTube extensively. Data analysis shows that the number of videos uploaded by the users follows a Zipf-like distribution and it shows that this uploading rate follows the 80-20 rule [2], which means 80% of the videos are uploaded by only 20% of the uploaders. Not surprisingly, numbers of subscribers for the uploaders also follows a Zipf like distribution.

Approximately 31% of the videos are uploaded by the users in the USA. After analyzing the YouTube social network, results suggest that the social users not only upload more videos but also their videos are watched more than the non-social users. This paper also demonstrates that male users upload more videos than female users. Finally, a comparison between user-copied content (UCC) and UGC videos was made. This analysis shows that most of the popular uploaders usually upload more UCC videos, although their UGC videos have more views than their UGC videos. Some of the findings of this paper are very important. For instance, the most popular 20% of the uploaders attract approximately 97% of the total views. Moreover, 20% of the uploaders only upload to a single category, and more than 85% of the uploaders upload more than 50% of their videos only to their three top categories. These findings can be very useful in order to predict the future popularity of videos at their very early age.

This paper also has some mentionable drawbacks. Most of the results presented in this paper are based on estimation. For instance, identification of UGC and UCC was done by examining a sample of the videos, and then conclusions were drawn for the whole dataset. It would be worthwhile research to identify if a video is UGC or UCC by a methodological approach, which would be able to give more accurate results. Moreover, the crawling approach is not presented clearly. For example, a seed user was selected first to collect its uploaded videos, and then all the related videos were crawled to capture their uploaders. This process was repeated many times with a new seed. Unfortunately, how the seed was selected is not mentioned. The BFS approach to crawl theYouTube social network can be biased to capture only the information of high degree users. However, findings of this paper address some issues that need further investigation and thus exposing an open research issue in this area.

2.1.10 User Categories in YouTube

In April 2006, YouTube announced the Director program in response to a video length limitation that was imposed to prevent copyright violations. A user could apply for a Director account after proving himself/herself as a legitimate creator of his/her uploaded content, which allowed to upload videos longer than 10 minutes. Then Musicians and Comedians accounts were introduced for publishing performer information and schedule of show dates. In 2008, Guru and Reporter accounts were added. Guru is for those who like to post videos that teach skills and how to do something, whereas Reporter was for the people who like to share news and

events occurring around them. Finally, Non-profit and Politician accounts were introduced.

Biel *et al.* [7] analyzed these user categories of YouTube along with their uploading rate, viewing rate and social aspects. In YouTube, a user is labeled as standard user when he/she first registers. Then a user can change his label to be Director, Comedian, Musician, Guru, or Reporter after applying for any of these profiles. Statistics suggest that most of the users (almost 90%) in YouTube do not belong to any of these special categories and continue their watching or sharing as standard users. In spite of their lower numbers, the special users contribute more than the standard users (uploading, watching and subscribing). This indicates that only the active users are interested about these categories; many users are still not aware of these categories. This paper also shows that, in YouTube, male users dominate female users in terms of uploading and watching videos. Female users are more social in YouTube than male users, however, as they have more subscribers, subscriptions, and they also favorite more videos than male users. Interestingly, although Politicians and Reporters upload many videos, their number of watched videos is very low, which indicates that these people are more interested in releasing their work or spreading their messages, rather than exploring people's interests. This paper is a good example to show how YouTube can be used to reveal the attitudes and behaviours of different kinds of people. This study, however, could not capture the statistics about anonymous users as it is not required to login to watch a video in YouTube.

2.1.11 YouTube and other Video Delivery Sites

Mitra *et al.* analyzed the growth patterns of four popular user-generated video sites including Metacafe, Dailymotion, Veoh and Yahoo! video [39]. While most of the earlier studies concentrated only on YouTube, this paper focused on finding similarities among the sites that offer similar contents and features to the users. From the aforementioned four sites, metadata for 1.8 million videos was collected that collectively consumed 6 billion views. The authors found that, like YouTube, users are mostly interested in watching videos rather than commenting, bookmarking or using other social features. The Pareto principle [2] was found common in uploading trend as well as popularity distribution; 20% of the uploaders upload 80% of the videos and 80% of the total views come from 20% of the uploaded videos. All the considered sites exhibit Zipf's law in video request frequency for the popular videos only. Similar to YouTube, a large section of the videos form the tail in total view distribution and thus fitting with an exponential cutoff. The authors suggested that these observations can lead to proper distribution mechanism for all user-generated video sites.

Cha *et al.* [14] show that publishing characteristics of YouTube are significantly different than non-UGC sites. As of June 9^{th}, 2008, the largest on-line movie data-base IMDb[5] carried only 1,039,447 movies and TV episodes, whereas approximately 65,000 videos were being uploaded daily in YouTube. These statistics imply that it takes only 15 days for YouTube to produce the same number of videos as listed in all of IMDb. While comparing the video publishers between YouTube and non-UGC sites, based on Lovefilm, in YouTube

[5]www.imdb.com last accessed: 22-08-2013

there are some publishers who post more than 1000 new videos over a few years; it usually takes more than 50 years for a single producer to produce 100 movies in the film industry.

YouTube videos are found to be shorter than non-UGC videos by two orders of magnitude, although the length of YouTube videos varies according to the category. In order to compare the viewing rates, information was collected from Netflix and Yahoo! Movies whereas views of Science & Technology videos were collected from YouTube. For Netflix movies, numbers of customer ratings were used to estimate the number of views since view information is not provided by Netflix. The scale of consumers per video is very different for YouTube and non-UGC. The views distribution of YouTube spans more than 6 orders of magnitude, while the number of ratings per movie in Netflix and Yahoo! Movies span only 4 orders of magnitude, illustrating the natural diversity of in YouTube uploader and consumer population.

Unlike the non-UGC sites, the authors observe that most of the popular Science & Technology videos in YouTube have incoming links from external sites. Surprisingly, in spite of that enormous number of incoming links, the authors observed that only 3% of the total views comes from these external sites. The view statistics suggest that video popularity in YouTube follows a power-law distribution with an exponential cutoff. These findings suggest that video delivery techniques in YouTube like sites should be designed more carefully and traditional approaches might not be well suited for the UGC systems.

2.2 Characterization of Other On-line Contents

2.2.1 Peer-to-Peer File Sharing

Before the introduction of YouTube and similar video sites (e.g., Dailymotion), Peer-to-Peer (P2P) file sharing was dominating the Internet traffic. For example, it was reported that P2P file sharing accounted for 43% of the traffic in a study of the University of Washington network [44]. This motivated Gummadi *et al.* to characterize the traffic pattern of Kazaa, a popular P2P file sharing system that time, under the same university network [30]. 20 terabytes of Kazaa traffic was collected over a period of 200 days in order to investigate the long-term pattern of user interaction with the Kazaa files. Kazaa objects suffered from *fetch-at-most-once* behaviour; the users were not interested to download the same file more than once in most cases. This is because of the immutability of the Kazaa objects, unlike other popular sites like CNN; multimedia objects do not change over time which is not the case in Web pages. The popularity of the Kazaa system was rather governed by the introduction of new objects and users. The *fetch-at-most-once* behaviour led Kazaa objects to deviate from showing Zipf popularity in their rank distribution. This is because of the behaviour of the most popular Kazaa files which is significantly flatter than Zipf's law predicts. The flatter section alone was responsible for the 50% of the Kazaa traffic. The main difference between Kazaa and YouTube like systems is that life time of objects in YouTube are not limited by the number of new users joining the network everyday—a video in YouTube can be watched by the same users more than one time unlike *fetch-at-most-once* behaviour observed in Kazaa objects.

Zhang *et al.* analyzed the Bit-torrent system more elaborately with the collection of 4.6 million torrents and 38,000 trackers over a period of nine-months [53]. In Bit-torrent systems, a file is divided into smaller pieces, which can be downloaded in parallel. Five of the most popular torrent discovery sites, including Pirate Bay, were used for crawling the data. The authors found that Bit-torrent is dominated by the tail section of contents; 82% of the files had no more than 10 peers although more than 10,000 active peers were found for the most popular torrent file. Uploaders' contributions in Bit-torrent are very similar to YouTube. Most of the files were uploaded by a small fraction of the users. In Pirate Bay, for example, a large number of the users contribute very little in content uploading. At each discovery site, however, approximately 100 users uploaded more than 1000 torrent files individually. In contrast to YouTube, US was at 15^{th} position in popularity rank of Bit-torrent among different regions in the World. UAE, Singapore and Canda are the first three countries that contributed mostly to Bit-torrent traffic. Similar to the YouTube popularity dynamics [14], the authors observed that young torrents are more popular than the older ones. Popularity of different file types was also analyzed (e.g., Music, Movie, TV shows, Software, Games, and Books). Although Movies, Music and TV shows contain most of the popular torrent files, the popularity of other categories (Books, Games and Software) are also noticeable. This indicates the diversity of content propagated by Bit-torrent systems.

In recent work, Carlsson *et al.* characterized the download patterns of Bit-torrent files from both global and local perspectives [13]. Data was collected for a period of 48-weeks at the University of Calgary along with the collection of global view of content popularity. A small fraction of the trackers at the university were found to be responsible for a significant traffic load, and thus following a Zipf-like distribution. Nonetheless, the tail section in the tracker distribution also contributes noticeably in traffic generation although load per tracker is very low for them. The users on campus downloaded significantly more larger files proportionately than the global users. This is because of the many global users that suffered from limited access bandwidth making them unwilling to download very large files. On average, more than 70% of the downloads of a file took place at the university before the global peak of a file. This suggest that university users are the early viewers/downloaders of on-line content, and this phenomenon can be used to predict global content popularity. Although this observation was common for all the categories including Games and TV shows, university users were found as the late followers of Music. Unlike YouTube, Bit-torrent files (specially the popular ones) reach peak popularity much later than their uploading/discover time. Likewise, many of the popular videos can retain their position in the hot set for a much longer time, 1 week to 5 weeks, unlike YouTube videos [1].

Some of the findings of this work, however, do not complement the earlier findings. For example, late popularity of torrent files found in this research contradicts the findings of Zhang *et al.* [53]. Early flash crowd phenomenon in Bit-torrent-like systems was also observed by Guo *et al.* [31].

2.2.2 Peer-based Personal Video Recorder System

Guebert *et al.* [29] envisioned a peer-based shared Personal Video Recorder (PVR) system where videos are stored/watched using the peers resources instead of traditional client-server architecture. The authors considered an extension of a PVR that can store videos broadcast by the content-providers so that users can watch video through time-shifting. Individual user's behaviour was simulated in order to form Zipf-like pattern in video popularity, assuming the aggregate video popularity follows the Zipf law. In other words, the authors analyzed how Zipf-like aggregate behaviour originate from individual peer-assigned object utilities. Besides storing the most popular videos, the peers also allocate spaces for those videos that otherwise could disappear completely from the entire network. Although Zipf-law states that the most popular video is requested twice more (considering α =1) than the second most popular one, the authors suggested that this is not because of individual user requesting most popular video twice as much as the second one. Rather, the number of peers that request the most popular video could be twice than the number of peers interested for the second-most popular video. The parameters for the simulation were number of channels, number of users, number of videos and a decay factor that controls video popularity dynamics. The authors considered constant decay factor, assuming a video would be at its peak popularity at the day of its publication, and the popularity will be faded over time according to the assigned decay factor. It was challenging to find an accurate decay value that generates a Zipf-like distribution in the overall video request pattern, with the addition of new users and videos in each time unit. A very high decay value can lead to a situation where many of the popular videos observe similar number of requests, and thus producing a flatter head section than the Zipf-law predicts. On the other hand, a very small decay value causes the popular videos published earlier to consume most of the requests, whereas newly published videos get very little chance to become globally popular.

The consideration of constant popularity decay is, however, impractical for most of the video sharing sites. For example, the effect of sharing through social networks (e.g, Facebook, Twitter) has been found to be significant in altering the growth curve of on-line content [12, 25, 52].

2.2.3 News-on-demand

The proliferation of on-line newspapers have gained lots of users, which in turn faded the popularity of paper editions of News. This motivated Johnsen *et al.* to analyze the server load and user behaviour in News-on-demand sites [33]. The authors used *Verdens Gang*, the most popular news-site in Norway at that time. Data was collected for a period of 2-year (with 4.6 million users and 3,500 different files) in order to investigate popularity dynamics and peak distribution along with the access patterns. The data set contains 1000 audio files while the rest of the files are videos. The audio files were mostly dominated by hit music. This is why the audio files' sizes, unlike video files, were very similar to each other. The authors found that access to the site varied according to the time and date. For example, on weekends, the amount of traffic was very low

compared to weekdays. In particular, the server observed significant load around noon each weekday.

In case of popularity distribution, a Zipf distribution was found to exhibit very good match for a very short measurement period. However, for the entire measurement period (2-years), News-on-demand shows total deviation from Zipf's law. The authors claimed that this is because of the popularity development of items over time. If different news items become the top news of a site everyday, then popularity of these news items can not be described by Zipf's law, especially for a very long measurement period; many popular items observed similar number of requests and thus leading to produce a much flatter head section in the distribution. The time-to-peak distribution and active life-span of popular News items were also analyzed. It has been observed that, not surprisingly, most of the items reached their peak popularity at the same day they were published. Some of the items were found to reach peak popularity later and these were mostly the Music audio files rather than news. News items were also found to become unpopular shortly after their peak day. This illustrates the fact that readers are attracted by only very recent News, with very few exceptions. In approximately 15% of the accesses, users were found not to start the videos from the beginning. Similarly, 34.8% of the users stop playback after watching only at most one third of the stream. This observations pose more challenges to the network administrator as lots of bandwidth can be wasted when the full file was loaded, but a user watched only a part of it. However, many users watched the same segments of some videos again and again. This suggests that segment caching such as adaptive segmentation can be much more effective than simple prefix caching.

In recent work, Gonzlez-Aparicio *et al.* characterized six most popular Spanish News sites including News videos posted on those sites [28]. Access distribution among different topics was analyzed for a period of 9-months. Along with the traditional Zipf, Mandelbrot and Stretched distributions, Box-Cox transformation has also been offered as a new model for on-line content access pattern. Popularity dynamics were observed with a per-day granularity as News videos can be more sensitive to time. In fact, the authors found that in most of the news-sites, new videos achieved 80% of their views at the very first day of their life times, which complements the earlier observation [33]. This indicates the challenges in caching mechanisms because of the required time to push the popular videos into caches. Similar to YouTube videos, most of the traffic was generated for a small fraction of the videos; this can be best explained by Pareto principle. In general, local-news videos were observed to attract more users than others like Sports, Economy, National, and International. While fitting data with known distributions, the authors found that there is no distribution that shows acceptable performance for all the six considered sites. Box-Cox distribution, that was not used in any of the earlier research for access frequency distributions, were found to outperform the well known Zipf, Mandelbrot (a slightly modified version of Zipf) and Stretched distributions in some cases. The authors suggested that applying Kolmogorov-Smirnov test is a better choice than the Chi-square test for model validation, as it is more restricted in accepting a model.

2.3 Summary

This chapter analyzed the earlier works, emphasizing mainly on understanding on-line contents' popularity dynamics. The driving forces behind the success of on-line contents in gaining users' attraction were discussed. Zipf behaviour in viewing distribution was found common in most of the studies. However, in UGC systems, Zipf's law can be used to model only the popular contents except for the measurements performed in local/confined networks. In global scale, UGC systems exhibit a very long tail because of the enormous number of unpopular contents. However, total deviation from Zipf's law can be observed for a very long measurement period, as the contents uploaded later can have sufficient time to be a member in the most popular list. Lots of contents with similar popularity create flatter head section than the linear decay suggested by Zipf's law.

Different factors, regardless of the content's subject/quality, were found influential in driving video popularity. These includes uploaders' previous history as well as techniques in uploading contents, such as associating large number of related keywords. The procedure and location of data collection can produce complete different results in contents' popularity. For example, Borghol *et al.* [10] found that keyword-based search using YouTube API returns more popular videos than the *most-recent* standard feed. Similarly, crawling approaches that start with the most popular videos and follow related video links produce dataset dominated by the popular videos and extirpate unpopular videos [39].

Although few of the studies raised the issue that contents' type can be used towards better understanding of contents' popularity, none of them had a detailed investigation to evaluate the hypothesis. This thesis leverages the best practices in the previous literature to investigate the popularity over time considering video categories in YouTube.

Chapter 3

Data Collection

As observed from the related work, the procedure of data collection is very important in order to measure the dynamics of video popularity. No earlier work is found that measured the daily views of different categories of YouTube videos from the very first day of their uploading time. Characterizing growth pattern using videos of different ages might not produce accurate results. The popularity distribution for the popular videos, observed in earlier works, can be different if videos of similar ages were considered; many videos uploaded later did not have sufficient time to become popular. Although Borghol *et al.* [10] collected videos that were uploaded in the same week, impact of video category on popularity was ignored. Moreover, considering per-week granularity may not be the proper way to have a broad understanding about the popularity in a site like YouTube. For some of the categories (e.g. News), the first week since uploading deserves more investigation, rather than considering the whole week's views as a single point of measurement [28].

3.1 Tools of Data Collection

Introduction of Google charts to obtain video view statistics in YouTube makes data collection easier than before. This chart is not available for many of the videos, however, and it is not clear which videos are associated with this tool and which are not. Returning views at 100 different points, regardless of video age, makes the task of comparing growth patterns an arduous task. In order to overcome these drawbacks and to get more insightful results, data collection was accomplished using the python client library of YouTube API.[1]

YouTube API provides several standard feeds for retrieving video_id. These are *Most Viewed*, *Most Linked*, *Most Responded*, *Most Discussed*, *Top Rated*, *Recently Featured*, and *Most Recent*. None of these standard feeds, except *Most Recent*, returns video lists of similar ages. More importantly, these standard feeds are biased towards the popular videos- thus restricting their applications in appropriate understanding of YouTube like sites where unpopular videos are dominant in number. The *Most Recent* standard feed, however, provides data set that were uploaded very recently without being biased towards popular/unpopular videos.

Multiple crawlers were deployed to obtain data used in this thesis. Based on the advantages and limitations of this data, comparison will be performed with earlier datasets used by other researchers. Since the crawler

[1] https://developers.google.com/youtube/1.0/developers_guide_python last accessed: 22-08-2013

obtained information from the API, the location of the execution of the crawlers is irrelevant.

3.2 Data Crawlers

3.2.1 Most Recent Crawlers

To obtain the video dataset for which the request analysis is carried out, 15 different crawlers were deployed on March 3^{rd}, 2012, collecting video IDs for 15 different categories.[2] For example, to collect the video IDs of music videos, the *Most Recent* standard feed restricted to videos in the Music category was used. All the crawlers collected video information for exactly 24 hours so that the age of any video in the dataset did not exceed 24 hours. This time limit was imposed to make sure that the views of all the videos can be collected from the very first day of their lifetimes.

The *Most Recent* standard feed provides video information randomly, increasing the confidence that the dataset is not biased based on popularity. A similar procedure was followed on March 4^{th}, 2012 in order to enlarge the dataset. After two complete days of data collection, a total of 71,208 videos' information was collected. Depending on the server load, the YouTube API returns only up to 100 videos' information for each category every one or two hours, which limits the size of the dataset. Considering that 72 hours of video is uploaded every minute, this translates into 600,000 ten-minute videos per day. Thus, an unbiased sample of up to 12% of the videos uploaded on a particular day is collected here.

3.2.2 Video View Collection Crawlers

Video view collection using two separate crawlers was started from March 4^{th}, 2012 and March 5^{th}, 2012. This process was continued for 149 consecutive days (approximately 5 months). The crawlers were programmed to switch to sleep mode until the next day after completing the collection of views for all the videos, thus ensuring a 24-hour difference between each subsequent view collection. The age of a video was calculated at the first day of view collection for that particular video, and then the number of views was normalized according to the video's actual age, by assuming that viewing rate is constant throughout the first day. For example, if a video age is less than 24 hours while collecting the first day views, a fraction of views from the next day's views was added as a compensation to make of a total 24 hours for the first day. Subsequent collections are offset by the same amount each day for those particular videos assuming viewing rate is constant throughout the day. This assumption will not influence the results as the granularity of views is one day.

A very small number of videos in the dataset were older than 24 hours at the time of the first view collection, and the opposite rule was applied for those videos. It is important to mention that, due to network connection failures, views for some of the videos in days 20 and 58 of the measurement period were

[2]http://support.google.com/youtube/bin/answer.py?hl=en&answer=94328 last accessed: 22-08-2013

not captured. Fortunately, those days are not that important for most of the videos, as most of the significant events occur at the very early age of a video. That is why in this thesis, after normalization, 147 day's views out of the first 149 days of each video's lifetime are analyzed.

Although the view collection was started with 71,208 unique video IDs, after 149 days of data collection the number of videos in the dataset fell to 47,711 (an average deletion rate of 33%). Although some of the videos were deleted by the uploaders themselves, by manually inspecting the data set, a large percentage of the sampled deleted videos had violated copyright issues and were forcibly removed. Table 3.1 shows the summary of the dataset. Different categories suffer differently from copyright infringement issues. Howto, Film, Entertainment and Tech videos suffer most frequently from copyright issues. Intensive analysis of these different deletion rates is left as future work. However, it was observed that deletion rates for all the categories decrease significantly over time and after a certain period of time, extremely few deletions take place.

In general, view collection in this thesis differs from Borghol *et al.* in three significant aspects: 1) size of the dataset is approximately 2.4 times, 2) the granularity of the measurements is much higher (per-day instead of per-week), and 3) the explicit consideration of video category by retrieving information of approximately equal numbers of videos from each category allows a more comprehensive comparison.

3.2.3 Uploading Rate Crawlers

In order to measure the current uploading rate of different categories, another crawler was developed that collected video information with the category names provided by the YouTube API's *Most Recent* standard feed. This standard feed returns a sample of videos that were uploaded very recently without any bias towards any specific category. The crawler collected video information over a period of 5 months, starting from February 2^{nd}, 2012 and collected approximately 365,000 unique videos' information with their categories. This allows the estimation of the short-term current uploading rates. While not an accurate representation of the entire content of YouTube, it does give some insight into the current uploading activity. The result from the current dataset is compared to the uploading activity for a previous data set from 2007 to determine if any obvious changes have occurred, or if trends can be identified.

3.2.4 Category Identifier Crawler

For some previous datasets that did not identify the category (e.g., dataset collected by Zink *et al.* [57] and Borghol *et al.* [10]), crawlers were deployed to obtain the missing information. A slight limitation of comparison with older datasets is that older datasets suffer from a deletion problem, which is also common in the current dataset after 149 days. The comparison between datasets collected in this manner becomes more unreliable given that popular videos suffer more from deletion rate than unpopular videos [25].

Table 3.1: Categories and Number of Videos

Category	**Number of videos (Day 1)**	**Number of videos (day 149)**	**Deleted videos Pct**
Howto	4773	1772	62.87
Film	4654	2346	49.59
Ent.	4991	2528	49.34
Tech	4942	2682	45.73
Games	4711	2966	37.04
People	4310	2730	36.65
Autos	4714	3245	31.16
Comedy	4744	3467	26.91
News	4623	3432	25.76
Travel	4918	3698	24.80
Sports	4812	3733	22.42
Music	4774	3477	21.93
Nonprofit	4624	3691	20.17
Education	4710	3801	19.29
Animals	4908	4143	15.58
Total	**71208**	**47711**	**33.00**

Chapter 4

Global Request Characteristics

Designing an efficient and scalable framework for content distribution is essential for sustainable development of large scale UGC systems. This requires an accurate understanding of the users' interaction patterns with the contents. This is helpful for both single-server and distributed systems [33]. Workload characterization of requests is considered as the primary tool towards this understanding. In systems where content popularity changes over time, detecting peak days is crucial for efficient distribution. This is more important if the peak days observe access frequency far more than the other days. The server can be overloaded unexpectedly by an enormous number of requests for specific content. This can be overcome by distributing the popular content in the CDN at the appropriate time, requiring the knowledge of content growth patterns. This chapter is dedicated to explore these issues. Examining popularity dynamics in different categories will help understand which categories are sensitive to time and if there is any common trend to reach peak popularity. It is also vital to classify the categories that have a much higher viewing rate in peak days than at other times.

Another important issue is to observe if the view distribution in a category is unevenly distributed. An uneven distribution of popularity indicates that significant success in cache hit ratio can be reached by only storing a small number of videos. While the evaluation of more traditional 80-20 rule—80% of the total views come only from 20% of the video population—can be misleading if there are even very few outliers, calculating the percent of popular videos in a category gives more insight in this regard.

These observations are helpful for making appropriate caching decisions and appointing proper advertisement policies. The current uploading rate of categories is also considered, which is necessary to design a complete request generator for YouTube. Finally, the applicability of a power law (e.g., Zipf) distribution in category-specific analysis is performed. A zipf distribution in content access frequency, similar to the 80-20 rule, indicates that caching would be beneficial in reducing network bandwidth. Besides, this part of analysis can be used as input for synthetic work load generation, as described in the next chapter.

4.1 Time-to-peak Distribution for Categories

Inspired by Borghol *et al.* [10], *time-to-peak* is defined as the day in which a video experienced the most views during its lifetime. Videos with at least 100 views are considered for this analysis; a video with a very small

number of views might contribute unfairly to the understanding of the actual growth pattern of a category. Figure 4.1 and 4.2 show the cumulative distribution functions (CDF), as in equation 4.1, of time-to-peak for the videos from different categories. It is clearly seen that the time to reach peak popularity is not the same for all the categories.

$$F_X(x) = P(X \leq x) \tag{4.1}$$

News and Sports categories follow a similar distribution and the time to reach peak popularity for these two categories is the shortest. Approximately 85% of News and Sports videos above the 100 view threshold reach peak popularity within only the first 4-5 days of their lifetimes. As well, in every category, between 48% and 67% of the videos experience their peak viewing on Day 1. Another group of categories (Music, Film, Howto, Tech and Education) follow a similar pattern to each other and many videos in these five categories reach peak popularity much later in their lifetime.

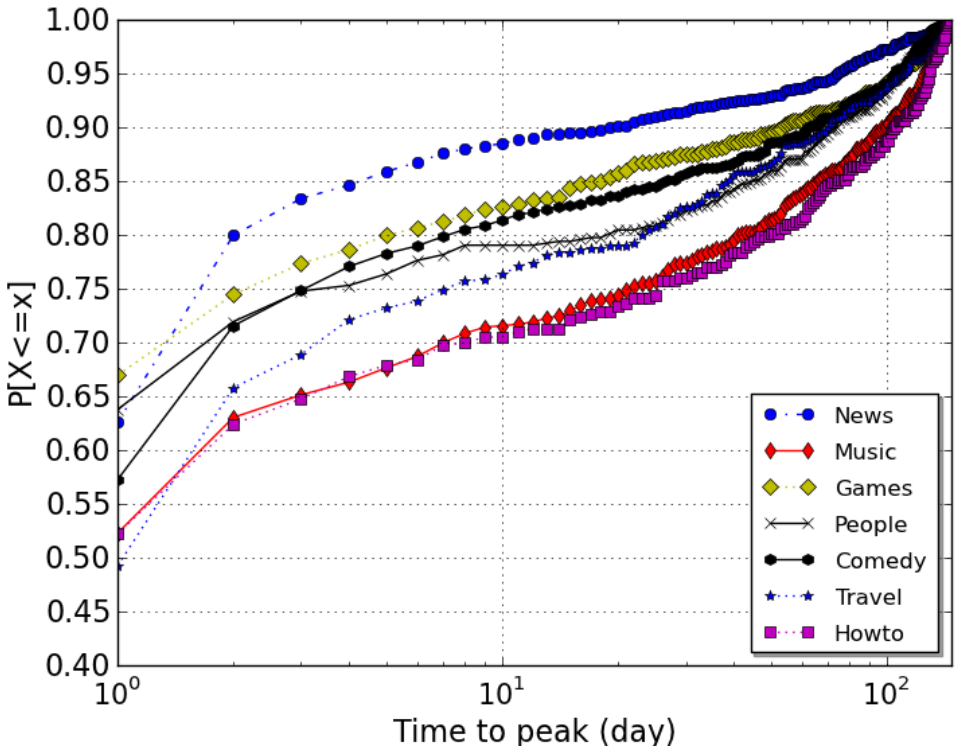


Figure 4.1: CDF of Time-to-peak (Selected Categories)

All other categories follow similar distributions to each other, and peak distributions of these categories lie within the aforementioned two groups of categories. Very similar results were observed for all the categories except Howto after considering videos with more than 1000 views; more popular Howto videos were observed to reach peak popularity faster than the less popular ones.

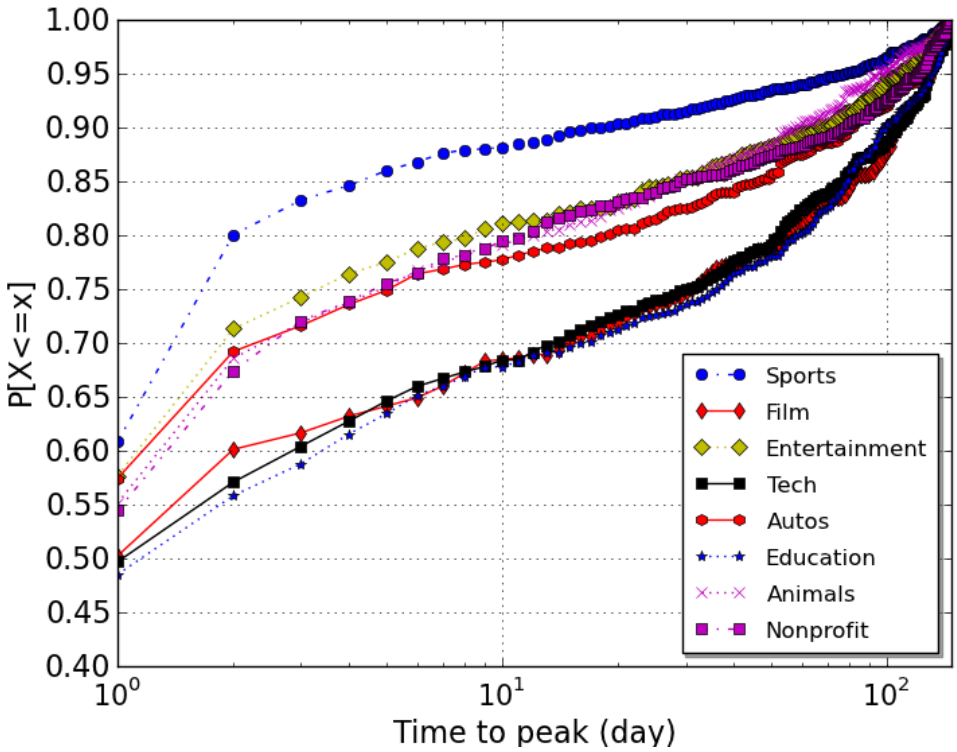


Figure 4.2: CDF of Time-to-peak (Remaining Categories)

4.2 Significance of Time-to-peak for Categories

It is important to understand if the peak day differs significantly from other days of a video's lifetime in order to determine if the previous statistic is helpful. A very high peak, compared to other days, indicates the urgency of employing advance resource allocation policy as well as the necessity of developing mechanisms to predict the peak time. On the contrary, for categories with almost constant viewing rates, the decision for resource optimization can be made anytime around the peak period. Even allocating appropriate resources on the peak day can still be beneficial for the upcoming days, in case the peak day was not successfully predicted; this is certainly not true for categories with very short active life spans.

One of the possible ways to observe the distribution of views among different days for a video is to calculate the view entropy as in equation 4.2: $E(k)$ is the entropy of video k and P_d is the fraction of views on day d.

$$E(k) = \sum_{d=1}^{N} P_d log_2(\frac{1}{P_d}) \tag{4.2}$$

This approach was followed by Google researchers [11] to observe the distribution of views among several regions in the world. In this case, higher entropy indicates longer active life span. However, this is not an accurate mechanism to achieve the goal of this thesis; equation 4.2 can not consider the days with no views,

and for most of the videos in the collected dataset, many days with zero views were observed. For this reason, the plan is to observe the complementary cumulative distribution function (CCDF) of x, as in equation 4.3, for different categories using equation 4.4, where $view(i)$ is the views at day i and $view(peak)$ is the number of views on the peak day.

$$\bar{F}_X(x) = 1 - P(X \leq x) \tag{4.3}$$

$$x = max(i) : view(i) \geq 50\% \times view(peak) \;\; \& \;\; i > peak \tag{4.4}$$

Again, only the videos with more than 100 views are considered for this analysis. The data in Figures 4.3 and 4.4 show that the peak day is a significantly unique point in the lifetime of videos for the categories that experience a faster growth pattern, e.g., News and Sports.

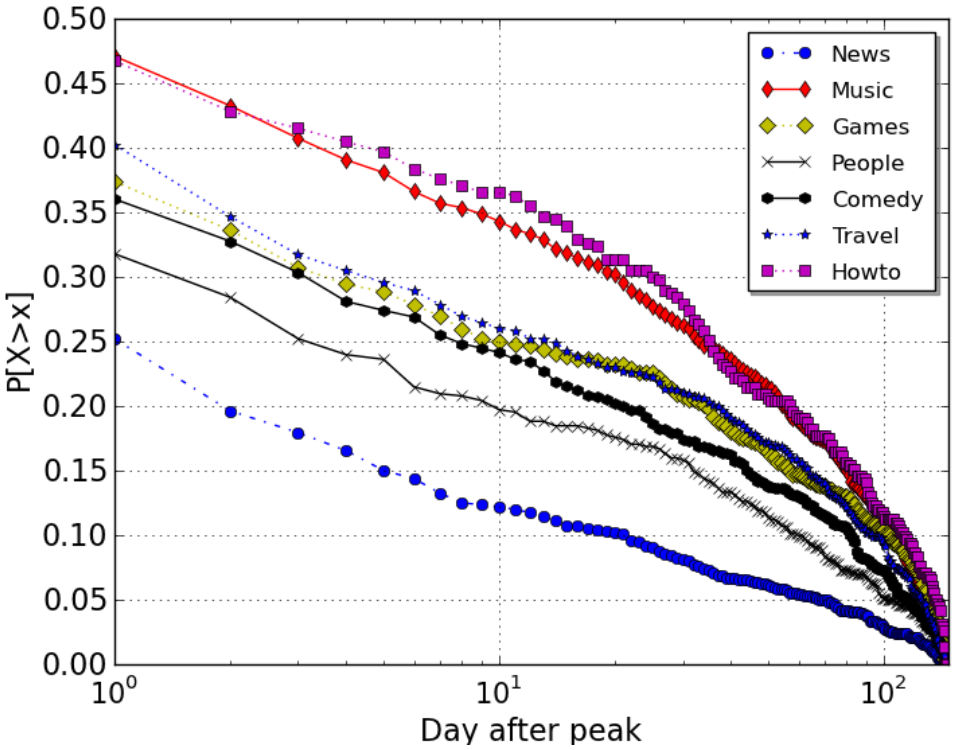


Figure 4.3: CCDF of Time-after-peak (Selected Categories)

These categories experience a short time of popularity, and decline to a lower, constant rate of viewing until the end of the data collection period. On the other hand, many of the Music, Film, Howto, Education and Tech videos that reach peak popularity comparatively later do not experience a sudden drop in their popularity. In other words, time to reach peak popularity is proportional to the active lifespan of a video. In particular, Figures 4.3 and 4.4 depict that more than 75% of the News and Sports videos never experience

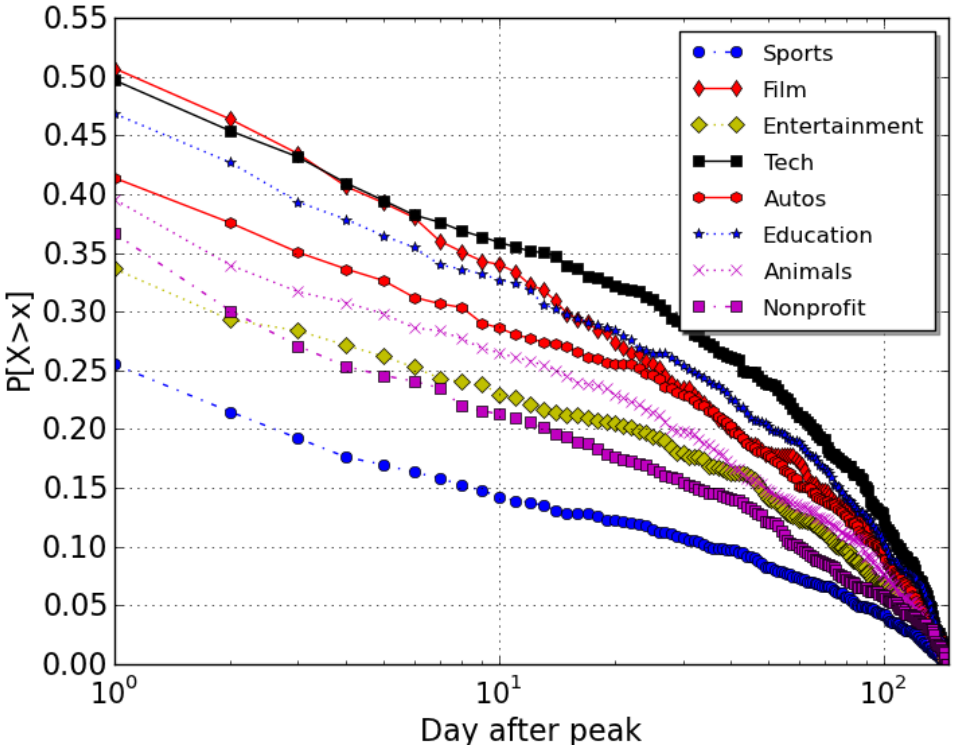


Figure 4.4: CCDF of Time-after-peak (Remaining Categories)

half of their peak days' views in their life time after the peak days; this drops to less than 50% for Film and Tech videos. The stability of Film, Music, Howto, Education and Tech videos, however, suggests that a longer measurement period would increase the difference between these categories and News/Sports. The results also suggest that more than 35% of the Howto and Tech videos can obtain a day in which a video experiences 50% of its peak days' views even after 10 days of the peak time. These findings complement the earlier observation that indicates the bulk of requests for News-on-demand systems in the very first day of videos' lifetimes followed by very sharp decay [33].

4.3 Relative Popularity Over Time

In order to to prioritize categories for time-dependent caching policies, it is necessary to know if the categories that reach peak popularity faster than others also experience differing numbers of views over time. Figure 4.5 and 4.6 depict the 95^{th} percentile of views of all the categories over time. The 95^{th} percentile instead of the mean is chosen to remove the potential effect of a single, large outlier that would obscure the intended insight as shown later in this section. This shows which categories have a minimum percentage of popular videos (5%) during every day of the data collection and the relative popularity of the categories for those popular videos.

As well, these graphs illustrate how viewing patterns of different categories change throughout the early part of their lifetimes. Although analysis of the dataset collected by Borghol *et al.* [10] shows that the views of the Music category exceed all other categories within their 8-month measurement period, it is clear from the current dataset that popular News, and Sports videos enjoy a significantly higher viewing rate than any other types of videos videos for the first couple of days since publication. Figures 4.5 and 4.6 also suggest that almost all categories have at least 5% of their videos which experience a high initial viewing rate; the difference is that, after these few peak days, views for most of the categories become very low, except Music and to a lesser extent, Film and Tech videos. The results indicate the variations in active life spans of different categories. For instance, in spite of having higher viewing rates of News and Sports videos than Music and Film for the first couple of days, the long-term dominance of Music and Film videos over News and Sports is observed.

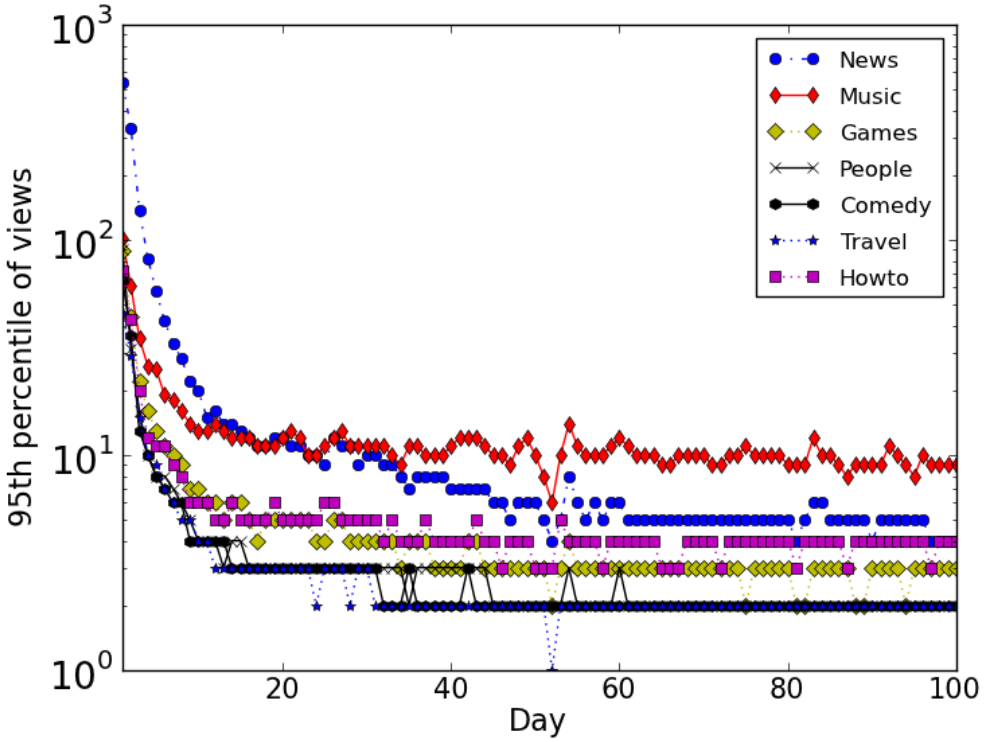


Figure 4.5: 95^{th} Percentile of Views Per Day (Selected Categories)

Although similar results can be observed from the average views per day as depicted in Figure 4.7, this statistic can be misleading sometimes because of the high variance of views; one extremely popular video can substantially increase the average views for a group of videos.

In the current dataset, the higher average views of Sports videos than News for the first few days is because of a single enormously popular Sports video. This video is the most popular video in the entire dataset and the number of views of this video is almost 24 times that of the second most popular Sports

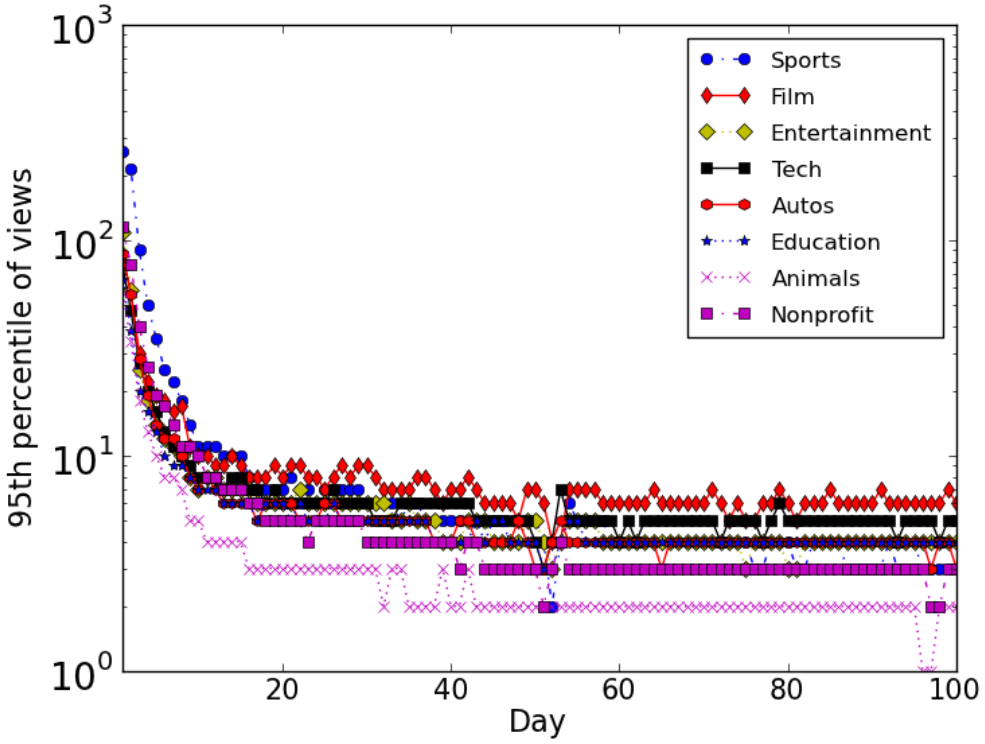


Figure 4.6: 95^{th} Percentile of Views Per Day (Selected Categories)

video. Nevertheless, this figure still supports the central observation; significantly higher viewing rate of News and Sports videos than Music and Film at their very early ages.

The viewing rates of different categories for significantly older videos are also analyzed by collecting category names and total views of the videos from the dataset of Borghol *et al.* [10], for 14 consecutive days starting from February 6^{th}, 2012. These videos are old enough to show some insight into the long-term viewing patterns of different categories. It is interesting to note that 19,860 videos still exist on YouTube after almost 4 years (a deletion rate of 33.4%, almost identical to the deletion rate after only 5 months in our dataset). The 95^{th} percentile of views is computed in each category for each day. The median of these 14 measurements is graphed in Figure 4.8 to show the comparative viewing rates for the most popular 5% of the videos in each category. This is to illustrate the importance of considering time dependent popularity of videos rather than total views after a long period of time.

This analysis confirms the long-term domination of Music and Film videos over others, and the fact that News videos become extremely unpopular after more than 3 years. However, inconsistent results are also found for some of the categories. Although earlier analysis shows the superiority of Tech videos over Autos and Education videos, different results are found for this specific dataset. This may be because of the significantly higher deletion rate of Tech videos compared to Autos and Education, as shown in Table 3.1; Figueiredo *et al.* [25] found that popular videos suffer from deletion due to copyright violation more than

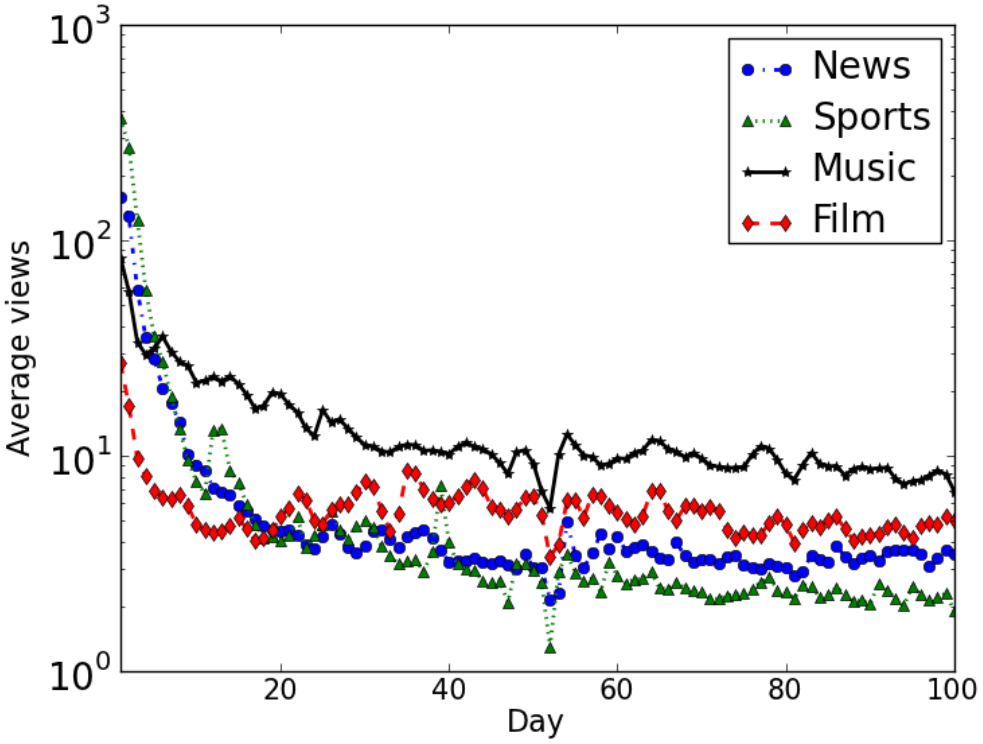


Figure 4.7: Time Varying Average Daily Views (Selected Categories)

the unpopular videos. Moreover, the videos remaining from the dataset collected by Borghol *et al.* [10] do not contain similar numbers of videos in each category because of their crawling approach, as shown in Table 4.1, which might affect the results.

There is an order of magnitude difference between the number of videos in the largest and the smallest category and deletion rate alone cannot account for this difference. With fewer videos in a particular category, the influence of a small number of outliers could dominate the statistic. It is interesting to note that the People and Comedy categories have a large number of remaining videos, but they are uniformly unpopular after 4 years.

More information about the deletion rate over time is needed to draw conclusions about the viewing patterns over long periods of time. One conjecture may be that the deletion of videos under the People category favours deletion of popular videos and the ones that remain after a number of years are only the unpopular videos. Additionally, if popular Film videos are deleted, then the remaining Film videos would be somewhat unpopular, but the data shows that film videos retain a high popularity. A longitudinal study is required that retains category information for deleted videos to make any accurate explanations for deletion rates.

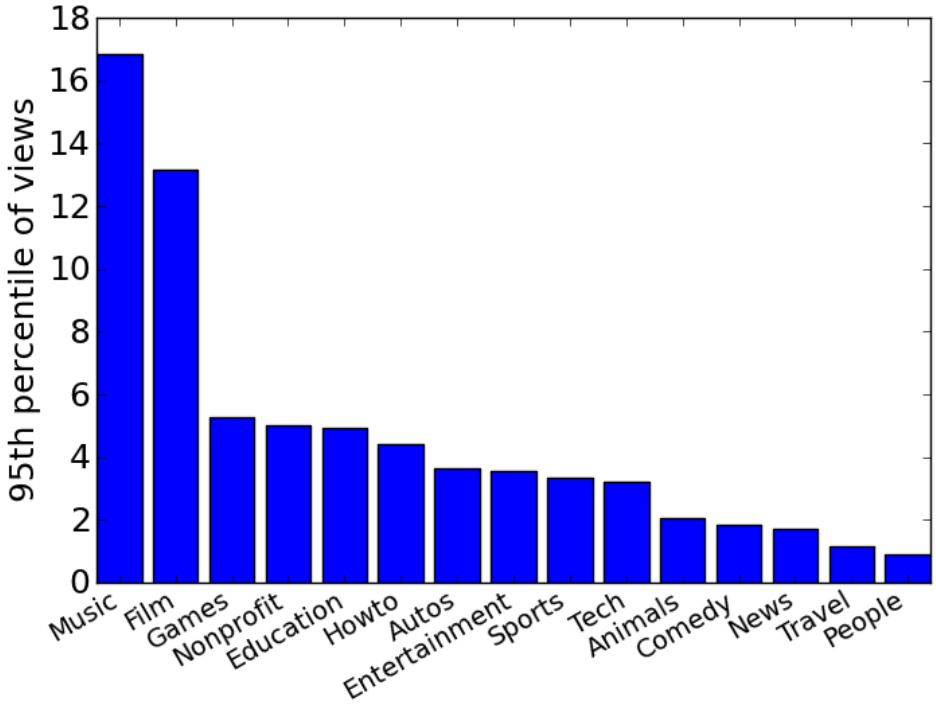


Figure 4.8: Viewing Rate of Old Videos

Table 4.1: Videos/Category (Borghol *et al.*)

Category	No. of videos
Music	4079
Entertainment	3471
People	2640
Comedy	2531
Sports	1469
Film	932
Travel	885
Autos	765
Animals	679
Games	655
Education	601
News	406
Howto	310
Tech	307
Nonprofit	110

4.4 Fractions of Popular/Unpopular Videos

One of the common findings in the earlier literature is that popularity is very unevenly distributed among on-line content. This observation is more prominent in UGC systems as an enormous amount of content comes from non-professional contributors. While it is very challenging to say which content is going to be popular, a category-based study can be helpful to identify the popular/unpopular videos to some extent. Although all the videos in a category are not popular/unpopular, it is not surprising to observe that some of the categories are more biased to popular videos. As an approach of this estimation, the percentages of videos with different views of the YouTube categories are shown in Table 4.2. The categories are ordered based on the percentage of moderately popular videos (10001 to 100000 views).

Table 4.2: Percent of Popular Videos

Category	**≤10 views**		**11 to 100**		**101 to 1000**		**1001 to 10000**		**10001 to 100000**		**> 100000**	
	Pct	Num	Pct	Num	Pct	Num	Pct	Num	Pct	Num	Pct	Num
News	18.85	647	39.57	1358	31.61	1085	8.42	289	1.4	48	0.15	5
Music	10.44	363	48.72	1694	32.87	1143	6.38	222	1.29	45	0.29	10
Film	23.06	541	49.53	1162	20.84	489	5.46	128	1.07	25	0.04	1
Sports	20.79	776	46.0	1717	26.12	975	5.97	223	1.04	39	0.08	3
Entertainment	27.77	702	46.88	1185	20.61	521	3.88	98	0.75	19	0.12	3
Autos	30.57	992	41.45	1345	23.17	752	4.07	132	0.68	22	0.06	2
Tech	22.56	605	47.28	1268	24.61	660	4.85	130	0.63	17	0.07	2
Games	27.51	816	49.36	1464	19.08	566	3.44	102	0.51	15	0.1	3
Nonprofit	24.11	890	48.04	1773	23.49	867	3.85	142	0.46	17	0.05	2
Howto	43.79	776	34.59	613	17.04	302	4.01	71	0.45	8	0.11	2
People	29.52	806	49.93	1363	17.69	483	2.42	66	0.4	11	0.04	1
Education	24.73	940	48.83	1856	21.7	825	4.34	165	0.37	14	0.03	1
Comedy	32.33	1121	51.08	1771	14.08	488	2.08	72	0.35	12	0.09	3
Animals	25.59	1060	56.48	2340	15.52	643	2.05	85	0.34	14	0.02	1
Travel	33.75	1248	48.89	1808	15.44	571	1.76	65	0.14	5	0.03	1

The data shows that only approximately 10% of the Music videos enjoy *fewer* than 10 views in the long run; this is much higher for categories like Howto, People, Autos, Comedy, and Travel. For Howto videos, it is more than 43%. News, Music, Film and Sports are the categories that contain most of the popular videos in the dataset, with over 25 videos having more than 10000 views. In spite of containing many very unpopular videos, the fraction of extremely popular Entertainment videos (0.12) are even more than Sports (0.08), and Film (0.04) videos. A similar observation is found for Games videos. On the contrary, the most unpopular videos are contained in the Travel category, followed by Animals, Comedy, Education and People

categories.

Although the number of Music videos dominates the number of News videos in case of views of the unpopular as well as the extremely popular videos, the number of moderately popular news videos supersedes the number of moderately popular Music videos, i.e., with 1000 to 10000 views. This phenomenon, however, might not be found for a significantly long measurement period, as Music videos enjoy a significant number of views for long period of time. It is noticeable that only 0.40% of the People videos experienced more than 10,000 views, in spite of the highest uploading rate (shown later) of this category. This suggests that although uploaders are currently uploading high rate of UGC videos, users are still not noticeably attracted to the UGC videos compared to the UCC (user-copied content) ones. This is further confirmed by recalling Figure 4.8, where old videos from the People category are uniformly unpopular.

Another measure that is analyzed is the CCDF of total views over the measurement period. There are a substantial number of videos in certain categories that had at most 1 view in their first 5 months of viewing. This can skew the popularity measures for some categories. In particular, the HowTo category had 17% of videos with at most 1 view and Autos had 12.6% with at most 1 view. Furthermore, there were 8.1% of the HowTo videos with 0 views. This indicates that there may be a section of completely unpopular videos that get published, which remain in the YouTube universe, but only take up space on a centralized server, and have not ever been downloaded to a proxy server or cache. Figure 4.9 shows the CCDF of the total views for a selected number of categories. This description shows the different behaviour in terms of viewing rates for the unpopular videos between the categories; the outliers (extremely popular videos) are not shown in order to observe the behaviour of unpopular videos more clearly. Music has very few videos with fewer than 20 views, but HowTo has over 60% of the videos with less than 20 views.

4.5 Current Uploading Rate of Categories

This section analyzes the activity of YouTube users in video uploading. In a complete workload generator that accepts new content in each time unit, a video must be assigned to a category. This needs the knowledge of current uploading trend in YouTube. Figure 4.10 shows the current uploading trend of YouTube videos using the dataset described in 3.2.3.

Cheng *et al.* [16] determined that Music was in the top position in number of uploaded videos followed by Entertainment, Comedy, Sports and Film. After manually examining some of the videos in these categories, it is found that these categories are dominated by user-copied content (UCC) rather than user- generated content (UGC). Outcomes of their analysis clearly shows that most of the videos in YouTube were UCC, although YouTube is considered to be a UGC site.

After collecting a random subset of uploaded videos for a period of approximately 5 months in 2012, the new dataset shows that the uploading trend in YouTube has changed over time. Based on the collected sample, an increased uploading rate of UGC videos is observed. Samples from the People category contain

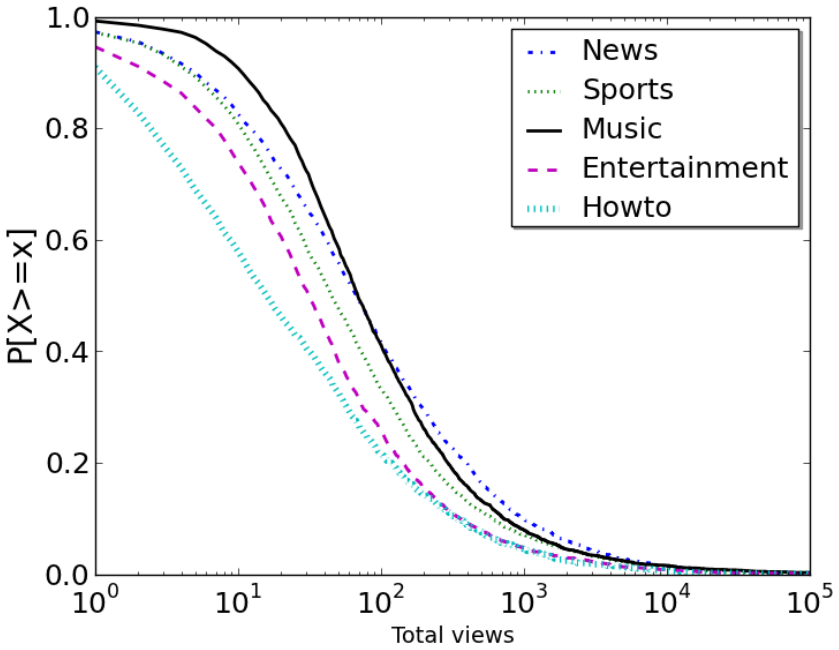


Figure 4.9: Selected CCDF of Total Views

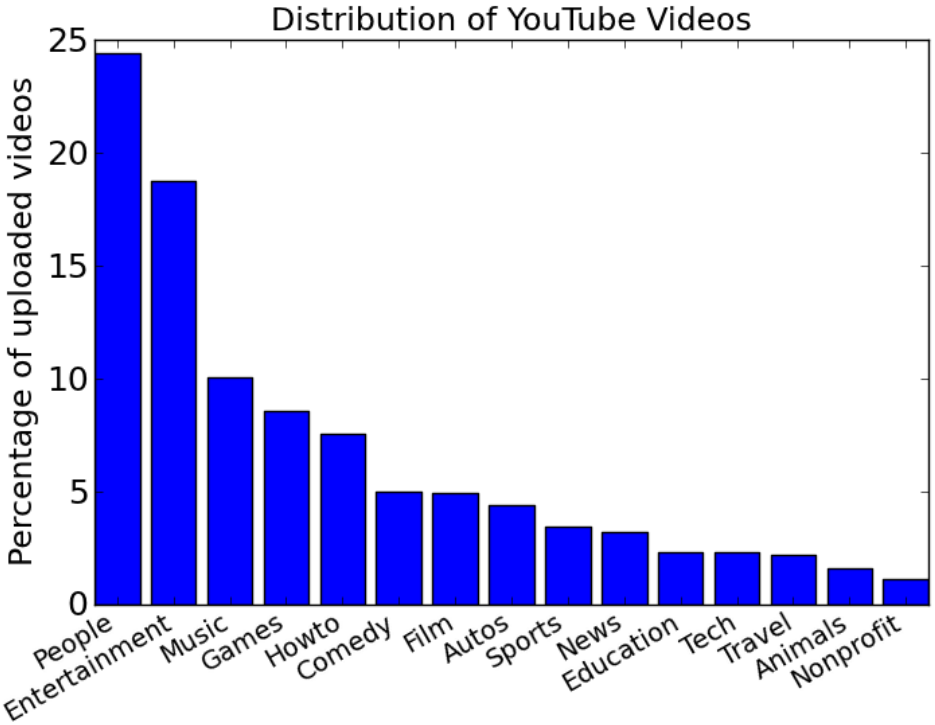


Figure 4.10: Category Uploading Rate

comparatively more UGC objects than other categories. The People category is now at the top position with approximately 24% of all the new videos; People was at the 6^{th} position in 2007 [16]—only 8% of all the videos.

This frequent uploading rate of People category videos poses more challenges to the system designer, as this category is clearly dominated by very unpopular videos compared to Music, News etc. Although significant amount of storage is needed to store them, in most cases, these videos are not requested by many users in YouTube-like sites.

4.6 Category Popularity Distributions

Zipf's law states that if objects are ranked according to their access frequencies, with the most accessed content as rank one, the second most popular content as rank two, and so on, then the frequency of occurrence is related to the rank (i) of a content according to the relation in equation 4.5 (N is the total number of content and a is the shape parameter). The observance of a straight line by plotting the access frequencies of content against their ordered ranks in a log-log scale indicates the correspondence with Zipf's law. Direct statistical measurements, however, are needed to confirm whether Zipf's law holds.

The similarity with Zipf-like distribution for individual categories is analyzed here. Figures 4.11 to 4.14 show the rank-frequency distribution for all of the categories. Cheng *et al.* [16] and Abhari *et al.* [1, 16] found that although requests for popular YouTube videos follow a Zipf-like distribution, overall the distribution fits better with a Weibull distribution because of the tail section, which matches with the large number of very unpopular videos in YouTube. However, after considering the types of video objects, only News videos are found to follow a Weibull distribution for the entire range of videos (and first 80% with better accuracy); for all the other categories, request distributions of only the popular videos follow a Zipf-like distributions and the light tail sections of these categories can be fitted to a Weibull distribution, as can be seen with the high goodness of fit statistic (R^2). The shape and scale parameters were determined by applying Maximum Likelihood Estimation (MLE) approach.

A Zipf-like distribution for contents popularity indicates that caching the ranked one content would improve the hit ratio linearly than caching the second most popular content. Fortunately it can be seen that relative frequencies of the popular video requests in YouTube for most of these categories are as Zipf's law predicts. This suggests that, even in a category-specific caching environment, proxy caching is a potential candidate for reducing bandwidth cost of YouTube and can improve end-users' experiences in watching YouTube videos.

$$P(i, a, N) = \frac{\frac{1}{i^a}}{\sum_{n=1}^{N}(\frac{1}{n^a})} \tag{4.5}$$

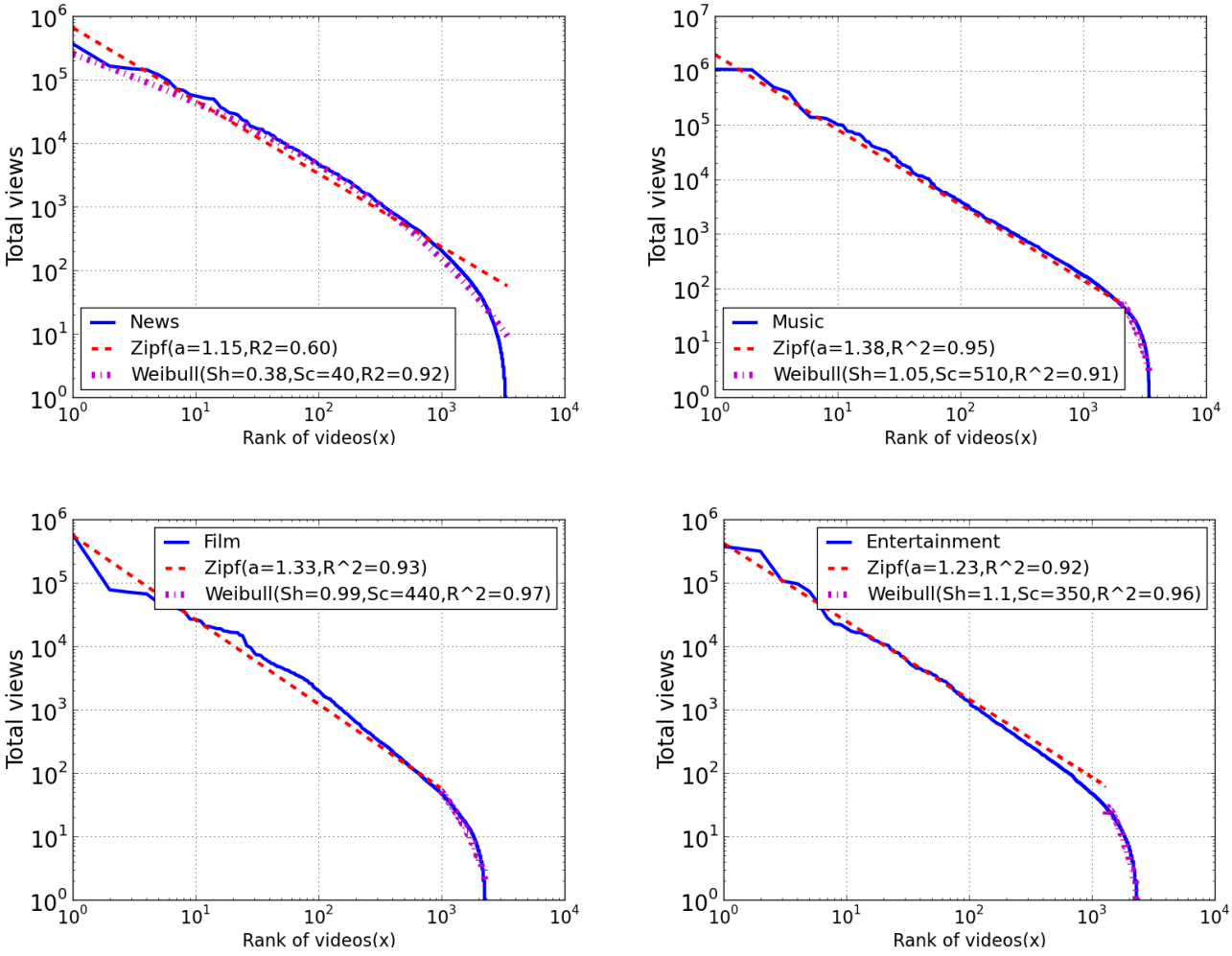


Figure 4.11: Number of Views Against Rank (Selected Categories)

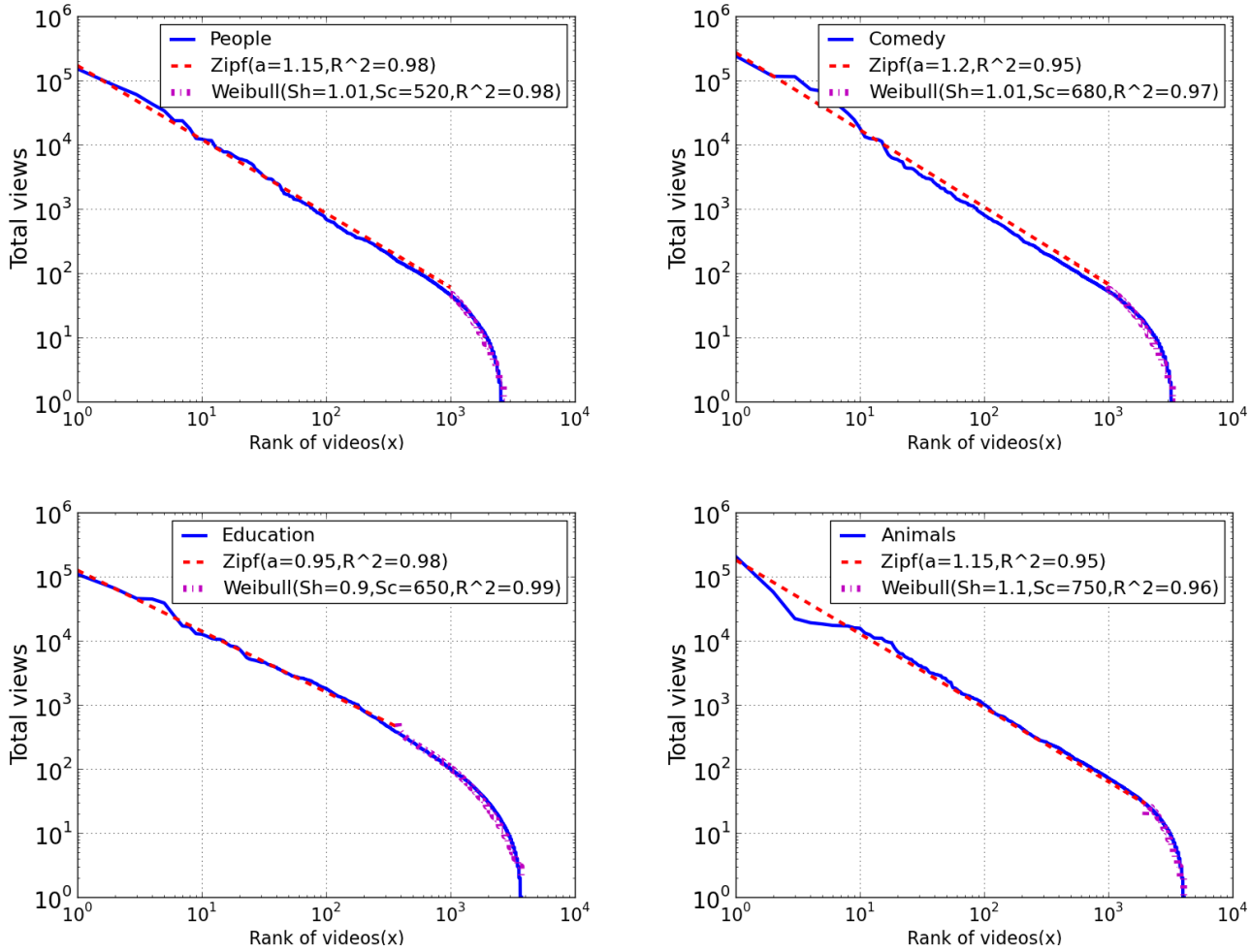


Figure 4.12: Number of Views Against Rank (Selected Categories)

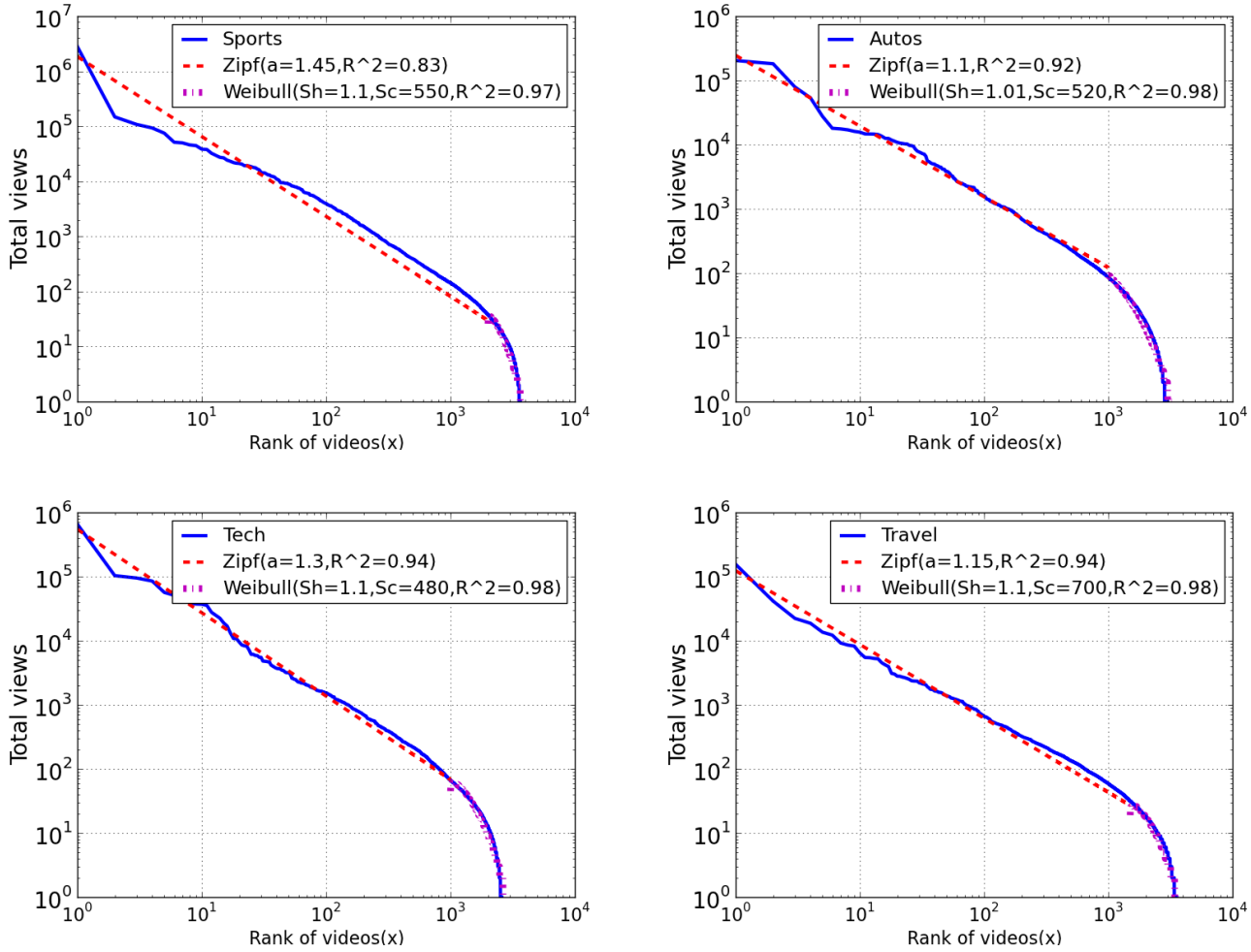


Figure 4.13: Number of Views Against Rank (Selected Categories)

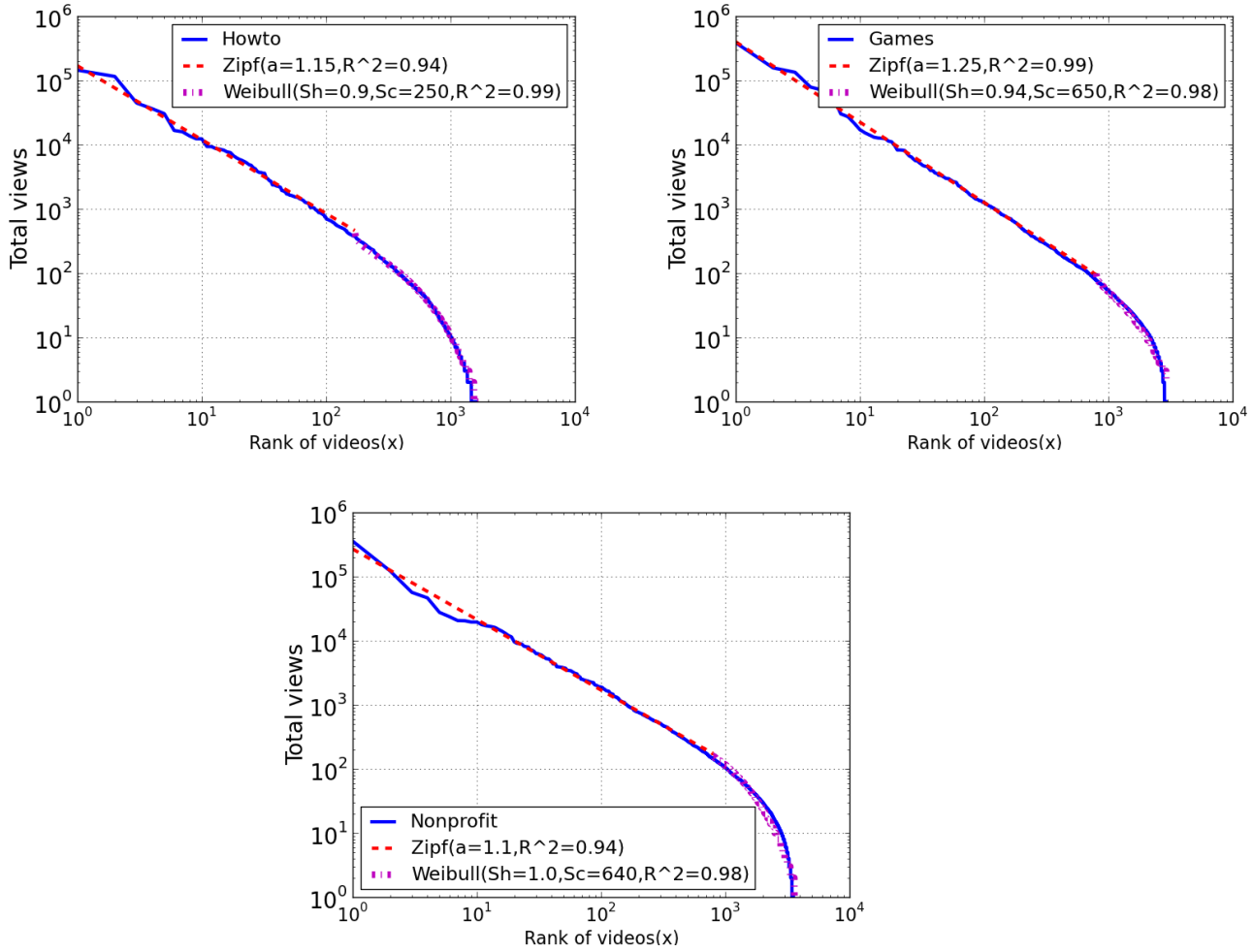


Figure 4.14: Number of Views Against Rank (Remaining Categories)

4.7 Summary

The results of this chapter confirm that consideration of content's type can play an important role for accurate characterization of content's growth pattern. Most videos exhibit their peak viewing day early in their lifetime and there is a sharp decay afterwards. Relative trends of viewing patterns of videos within categories over the first 5 months of their lifetimes were observed. Some categories contain a non-trivial number of videos that are still popular 5 months after the time of uploading, whereas other categories have viewing patterns that dwindle to nothing. The confirmation of Zipf distributions for the total views of popular videos in nearly every category indicates that caching would be effective at reducing server loads, especially if regional viewing data confirmed the details of the viewing patterns persisting across the regions. This analysis is vital in order to model the category-specific introduction of new content over time to drive simulations and/or prototype content distribution networks to evaluate different policies for storing data across a set of distributed caches.

Chapter 5

Towards a Workload Generator

The differences in growth pattern and overall popularity of YouTube categories has been illustrated in the previous chapter. This gives the intuition that earlier methods for modeling growth patterns of YouTube videos might not be effective for category-specific modeling [10, 48]. This chapter first analyzes the futility of two well-known methods in modeling growth pattern of YouTube categories. Then a generalized solution for workload generation, using time-series clustering, is offered that seems to work for four of the selected categories. These four categories, News, Music, Film, and People exhibit different popularity trends. News, Music, and Film are different in their growth patterns although these are the most popular categories in the dataset. On the contrary, People videos are selected to represent the unpopular categories in YouTube. Moreover, it is the category that contains most of the uploaded videos in the recent dataset. Workload generation is not performed for Sports videos because of their very similar nature to News videos. Success in modeling for these categories gives the confident that time-series clustering would be effective for all other categories.

5.1 Predicting Future Popularity

Figure 4.5, 4.6 and Table 4.2 illustrate the fact that different categories in YouTube have different proportions of popular videos, and their popularity evolves differently over time. In order to model an appropriate caching mechanism as well as design effective advertisement policies, it is important to predict the future popularity of a video at its early age.

As the first approach to predict future popularity, Pearson's correlation coefficient [57], as in equation 5.1, is calculated between the added views[1] at different snapshots of the measurement period.

$$r_{xy} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{5.1}$$

A high correlation coefficient between early views and and rest of the period implies that prediction of future views of individual videos is achievable [48].

Although Borghol *et al.* [10] claimed that it is not possible to predict the future popularity of very young YouTube videos, this is not the case if video categories are considered. Acceptable correlation is found

[1] Added views is only the number of views on a particular day

between different pairs of snapshots (Table 5.1) for some of the categories. For instance, the correlation coefficient between the first day's added views and next 7 days' added views for individual News videos is 0.8, which reflects strong positive linear correlation [10]. On the contrary, for the same two snapshots, the correlation coefficient for Music videos is only 0.35. This may be due to the different active life spans of News and Music videos.

Table 5.1: Correlation Coefficient between Different Snapshots

Category	**Snapshot 1**	**Snapshot 2**	**Coefficient**
Howto	1	2-149	0.80
Film	10-20	21-30	0.85
Entertainment	1	2-7	0.90
Tech	1	2-149	0.92
People	1-7	8-14	0.80
Games	1	2-149	0.90
Comedy	1-7	8-14	0.80
Music	1-10	11-149	0.92
Travel	1-2	3-149	0.94
Sports	1	2-149	0.99
News	1	2-8	0.80

The correlation coefficient for Music videos is 0.92 between first 10 days' views and the rest of the measurement period, confirming the stable relative popularity of Music videos. For some categories—Sports, Travel, Tech, Games and Howto—the first one or two days' views are sufficient to predict the popularity for the entire measurement period. On the contrary, for Entertainment, People, Comedy and News videos, only the immediate future popularity can be predicted. This suggests that the list of popular videos in these categories changes over time. For Film, Autos, Animals, Education and Nonprofit videos, even immediate popularity cannot be predicted by observing their early views, although popularity becomes predictable for older videos.

It is a bit surprising to observe that Music and Film exhibit significantly different correlation coefficient patterns in spite of very similar growth rate of these two categories. To analyze this issue, view changes in two different snapshots are observed for Music and Film along with Sports that shows very good correlation coefficient for even very distant measurement periods. Figure 5.1 shows how the set of popular film videos changes over time. Knowing the first day's popularity is not sufficient to estimate the immediate next day's popularity for Film videos. As a result, it is impossible to use the first day's views for the rest of the measurement period. As videos become older, however, immediate future popularity of Film videos can be predicted successfully.

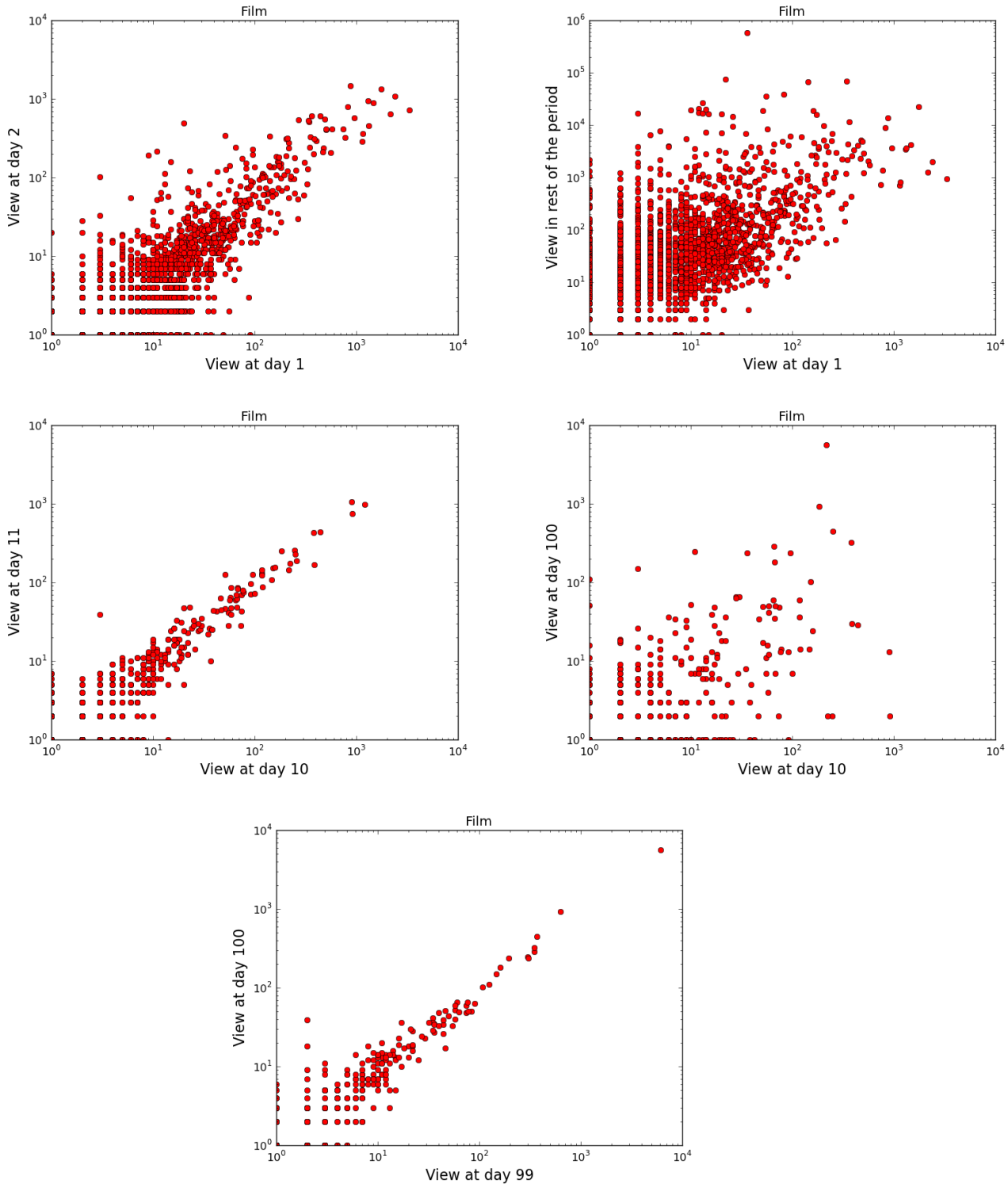


Figure 5.1: View Changes of Film Videos between Different Snapshots

A completely different scenario is observed for Sports videos. Figure 5.2 shows the changes in popularity of Sports videos in two different sets of measurement period. The view changes between day one and two are very negligible and thus produce a very high correlation coefficient. Although some view changes are observed when the second snapshot is taken as the rest of the measurement period, this is too little to change the overall result for Sports videos; Sports videos do not retain popularity for a long period of time after their publication.

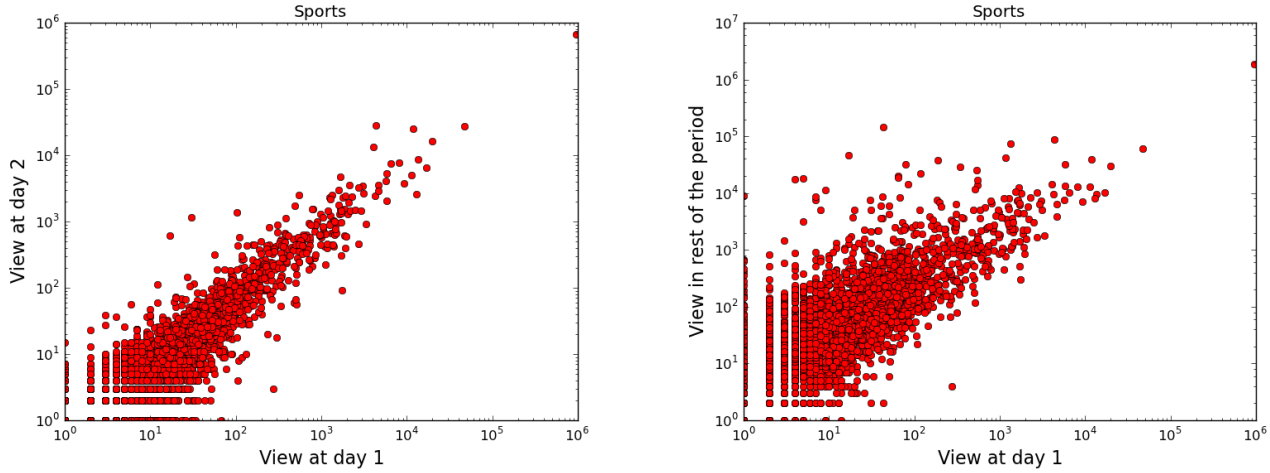


Figure 5.2: View Changes of Sports Videos between Different Snapshots

Music videos exhibit worse performance than Sports, but much better than Film videos in retaining popularity over time (Figure 5.3). Immediate future popularity of Music videos can be predicted even for very young videos, although this information becomes invalid for the prediction of very distant future popularity.

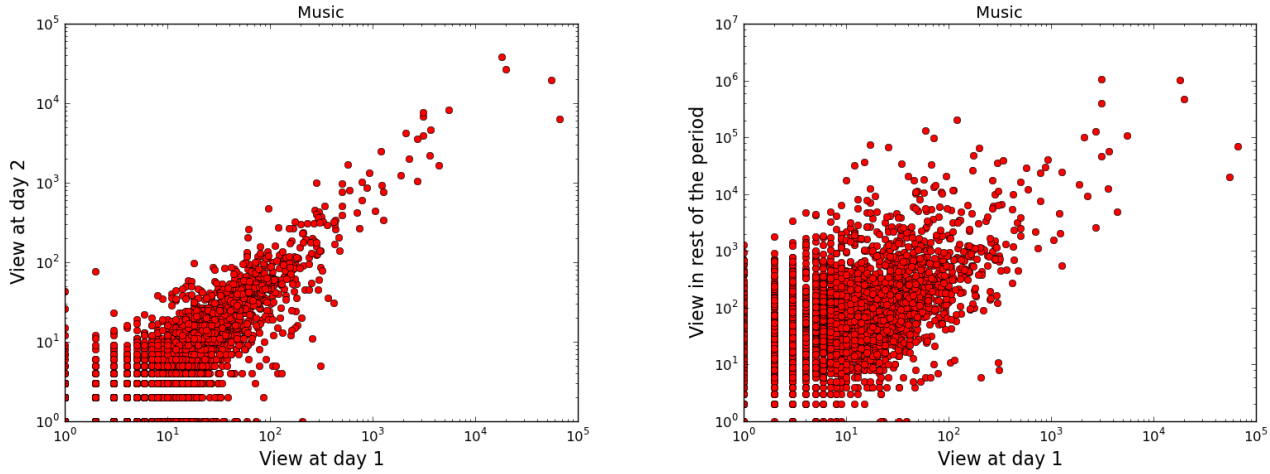
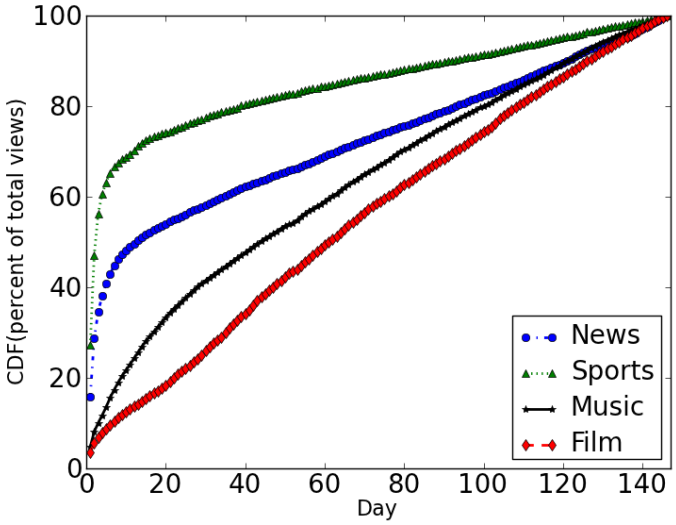
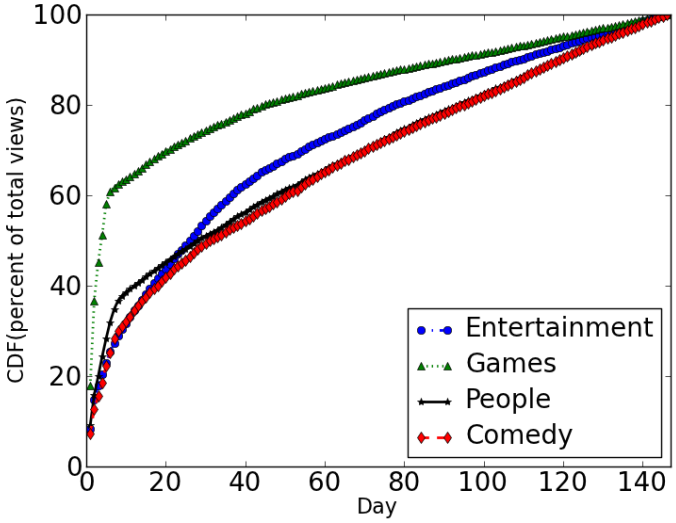


Figure 5.3: View Changes of Music Videos between Different Snapshots

This enormous variation among the categories in predicting their future viewing patterns confirm that the Base-line model offered by Szabo *et al.* [48] can not be applied in general to model the growth patterns of all the YouTube categories. However, in spite of the difficulties for the very young Film videos, information of the older videos can still be helpful for caching and advertisement policies. Many views for Film, and Music videos come even after a long time of their uploading. Figure 5.4 provides more insight into the correlation between snapshots and predictability.



(a) News, Sports, Music and Film



(b) Entertainment, Games, People and Comedy

Figure 5.4: CDF of Percentage of Total Views Over Time

Over 75% of the views of Sports videos take place within the first 4 days (over 50% on the first two

days). However, at day one, less than 20% of the total views of Sports videos is observed. This means up to 80% hit rate is still achievable by using the first days's video rank information. Although very young videos' information are not helpful for categories like Film, only approximately 10% views of Film videos come within first 10 days in the measurement period. As immediate future popularity becomes predictable after that period, up to 90% hit rate can be obtained for Film videos. Conclusion can be drawn for other categories as well from Figure 5.4.

5.2 Three-phase Characterization

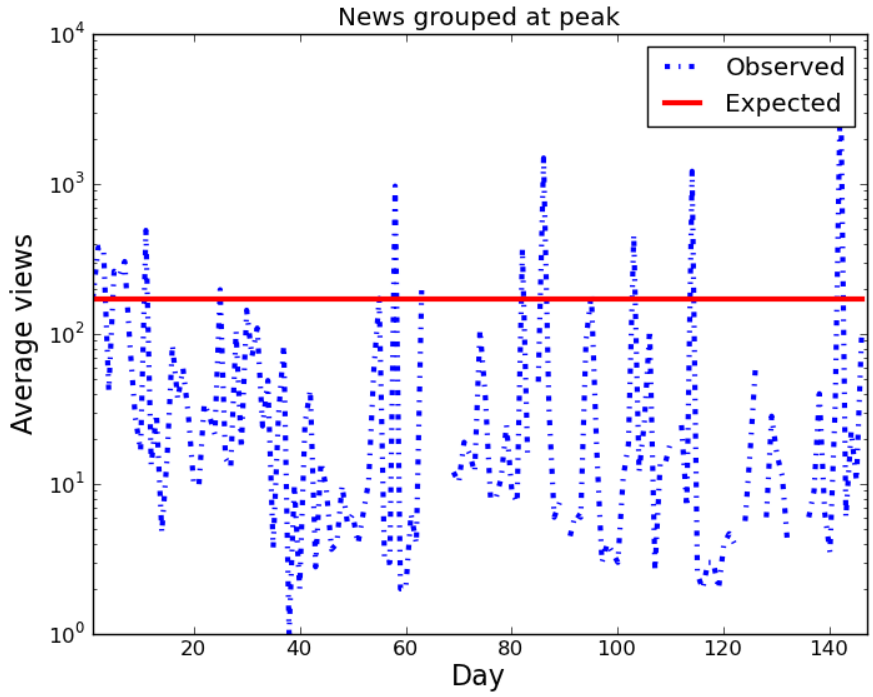
The three-phase characterization of Borghol *et al.* [10] considers that the average viewing rate over time is constant when the videos are grouped into at-peak, before-peak or after-peak on a particular day. This is because of the similar view distributions found in all over the measurement period. This approach is fairly simple as only three fixed distributions along with the fixed peak distribution are needed for the entire modeling. This section examines the appropriateness of the proposed model in a category specific environment.

Figures 5.5 to 5.8 show the average viewing rate for some of the selected categories grouped at their peak phases. Showing result for only one of the three phases is enough, as three-phase characterization method can only be applied when constant rate is found for all the three phases. This clearly suggests that the viewing rates over time are not constant for any of the selected categories. The high and highly variable average views for News videos at the end of the measurement period is because of the very few number of videos that reach peak popularity around that time. Otherwise, a decay in viewing rate is observed for first two months, thus contradicting the time-invariant nature observed previously [10]. For some of the days, no videos were at their peak popularity as depicted in the figure. For Music videos, however, no specific trend is observed, as many music videos reach at peak popularity much later than News videos. Nevertheless, three-phase characterization cannot be applied for any of the categories. The following hypotheses can be made to describe the absence of three-phase behaviour in category based environment.

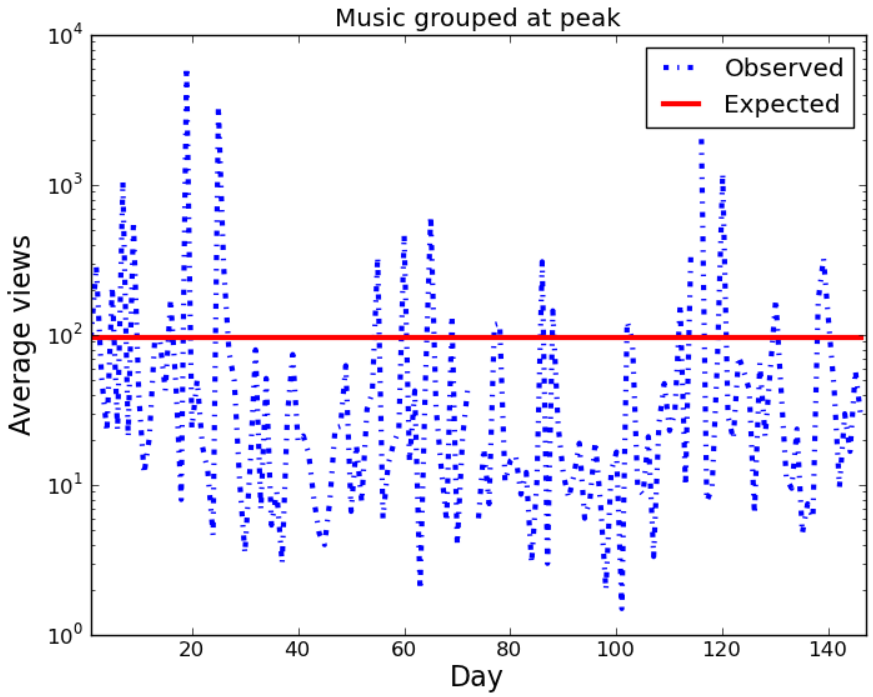
- YouTube categories, as observed earlier, are conspicuously different in their popularity patterns. As a result, a model that works for a specific or group of categories might not exhibit similar performance for others.

- Time invariant nature of YouTube videos' request pattern can only be observed in per-week granularity.

- The proposed three-phase model is not scalable to small/large number of videos or length of the measurement period.

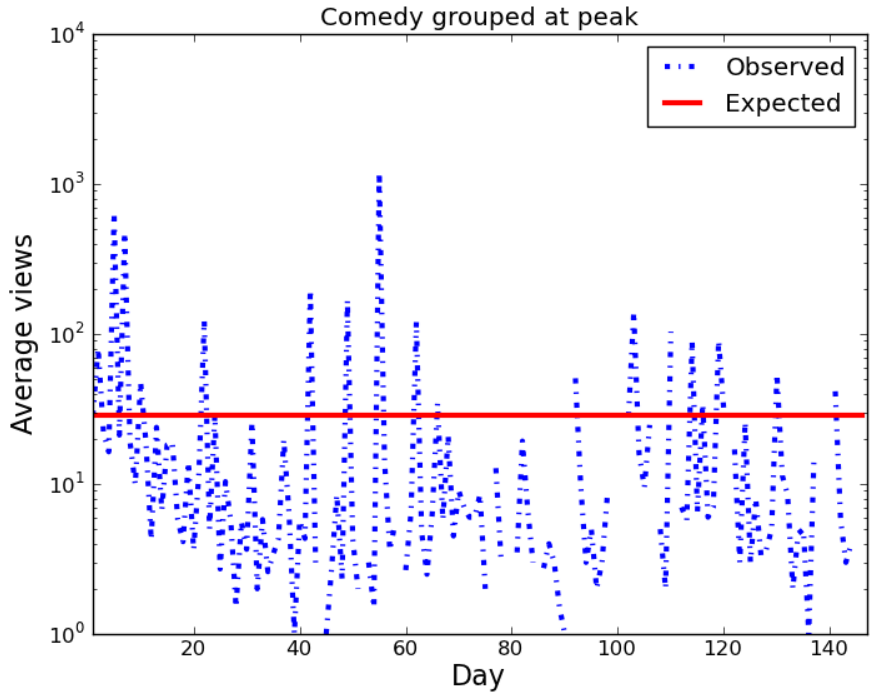


(a) News Videos with Peak at Day X

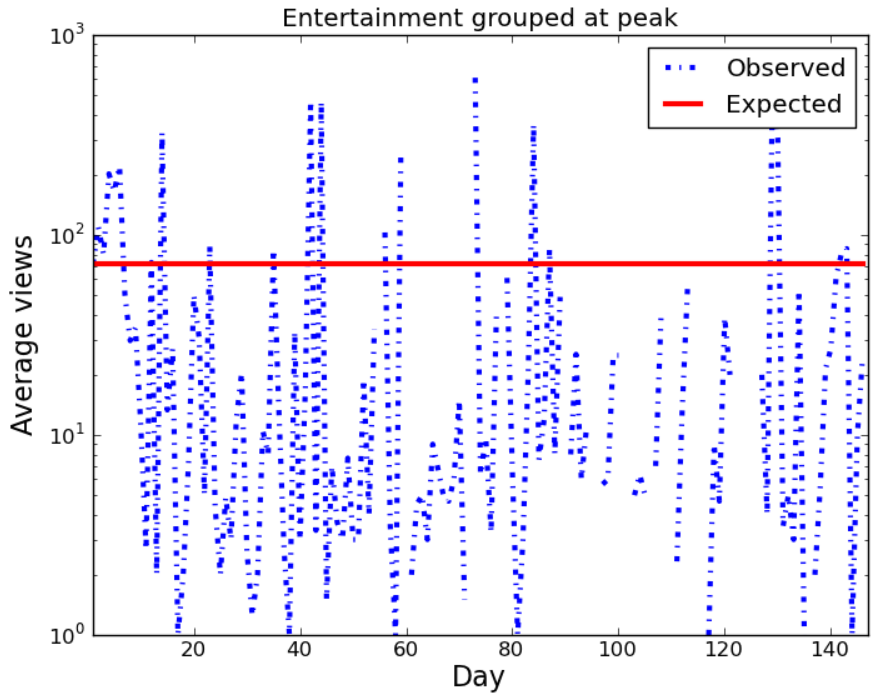


(b) Music Videos with Peak at Day X

Figure 5.5: Average Views Over Time for News and Music Videos at Peak

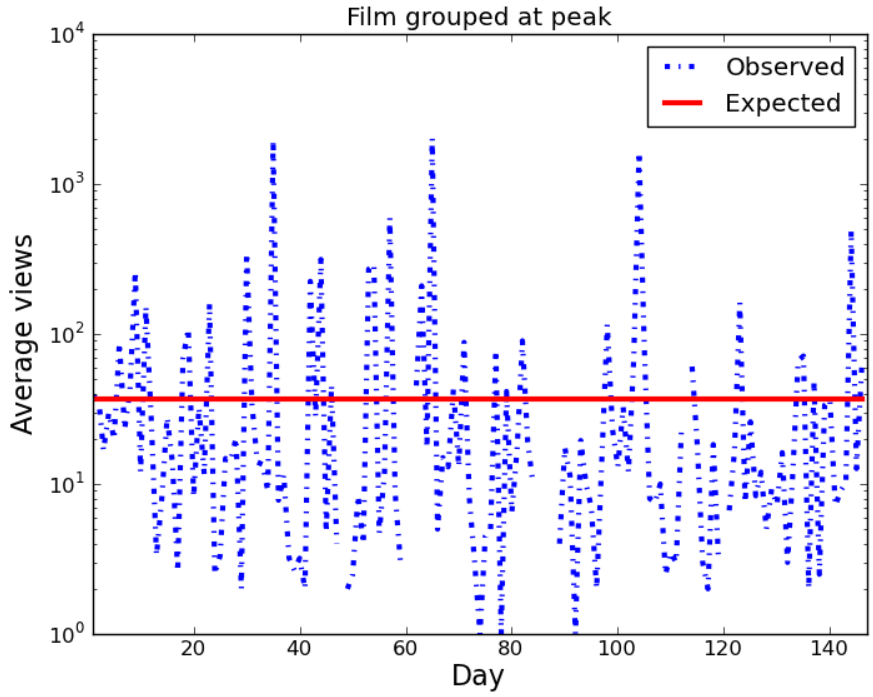


(a) Comedy Videos with Peak at Day X

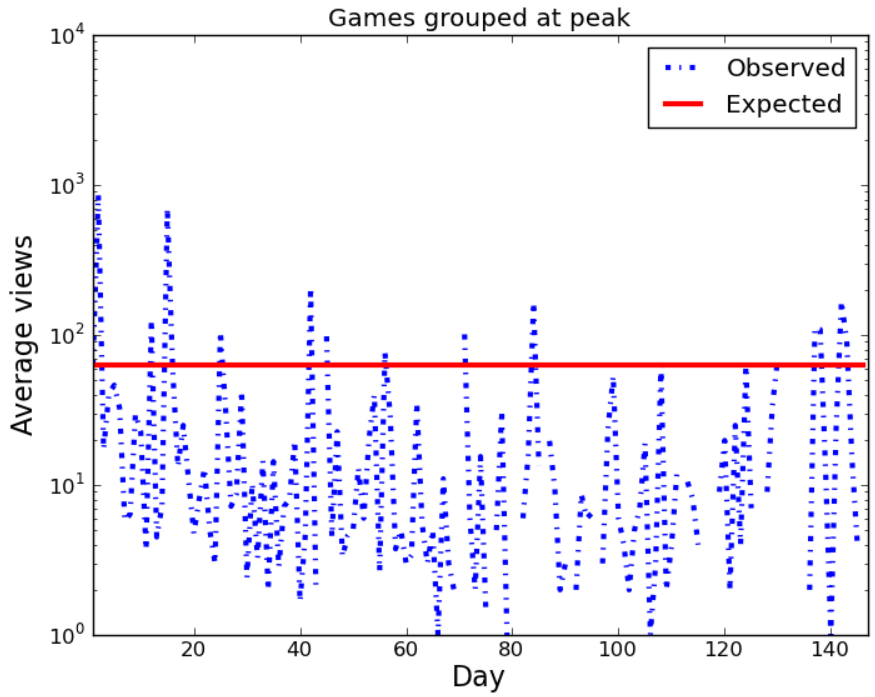


(b) Entertainment Videos with Peak at Day X

Figure 5.6: Average Views Over Time for Comedy and Entertainment Videos at Peak

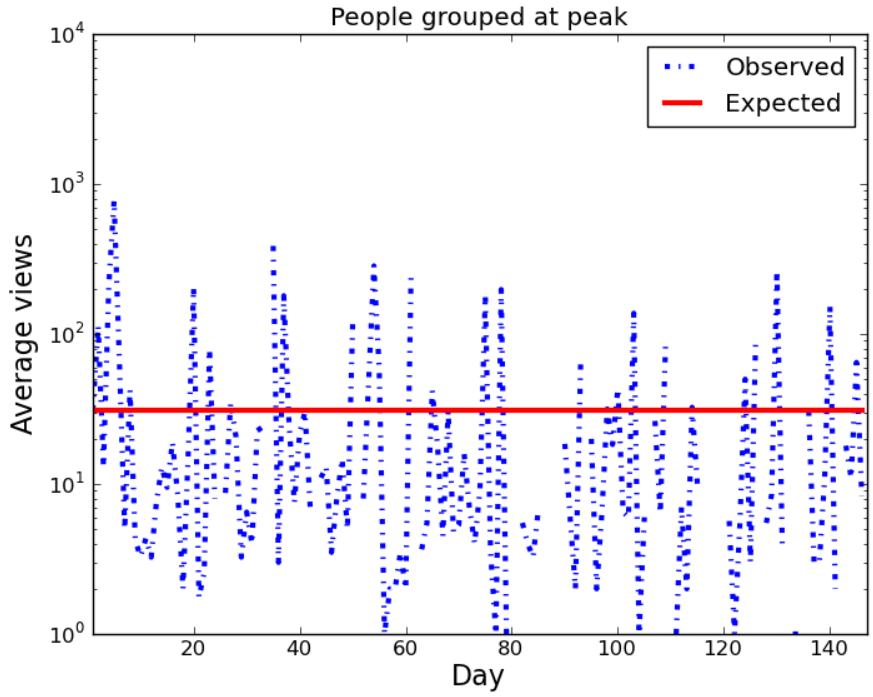


(a) Film Videos with Peak at Day X

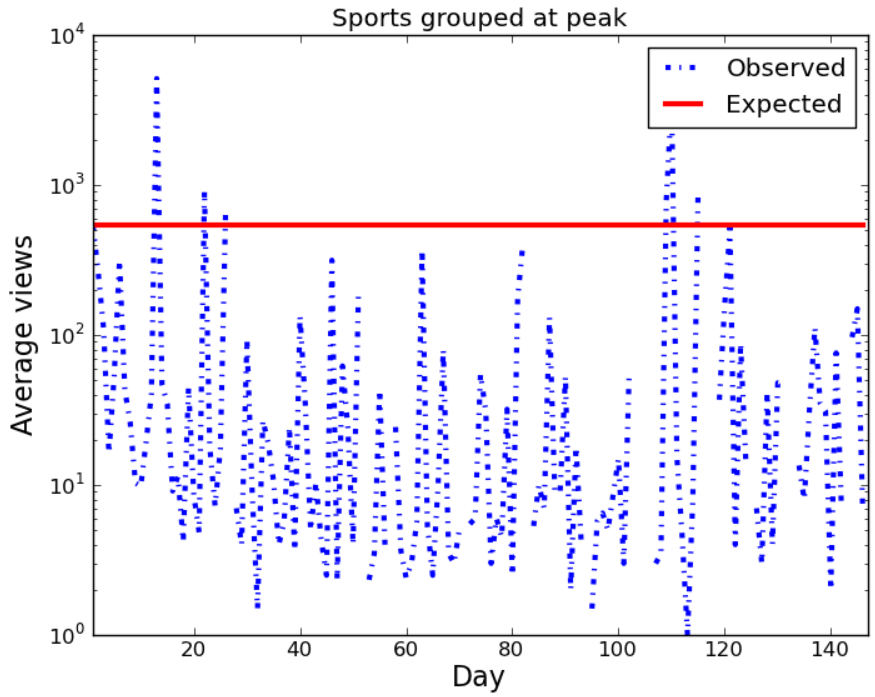


(b) Games Videos with Peak at Day X

Figure 5.7: Average Views Over Time for Film and Games Videos at Peak



(a) People Videos with Peak at Day X



(b) Sports Videos with Peak at Day X

Figure 5.8: Average Views Over Time for People and Sports Videos at Peak

5.3 Time-Series Clustering of Growth Patterns

The futility in modeling growth pattern for all the categories by the aforementioned techniques means that another technique is required for category-specific modeling. One possible technique is to analyze the popularity evolution of videos in a specific category to determine if they follow similar shapes, and thus, can be modeled using fixed number of clusters. This approach can be considered as a typical time-series clustering problem. This becomes a reasonably challenging problem as different videos reach peak popularity at different times. Inspired by a study on viral videos [12], all the time-series are translated so that x-axis is centered on the peak day. This is because of the observations that most of the significant events happen around the peak periods of YouTube videos. This ensures clustering videos based on the most important period of their lifetime.

Another challenging issue is to select the appropriate time-series clustering algorithm. The main objective is to identify the similar shapes of the views per day, regardless of the time to peak, i.e., a shift-invariant algorithm. Moreover, the algorithm should not be affected much by the outliers as some videos in each category with very atypical growth patterns were observed.

K-Spectral (K-SC) clustering, offered by Yang *et al.* [52], is selected as the potential solution for this problem. It has been found to be accurate in identifying the growth patterns of other Web content. Unlike the more traditional K-Means clustering, the centroids of the K-SC clusters are not distorted by the outliers. Instead of considering Euclidean distance between the curves of videos growth, K-SC applies a scale and shift invariant distance metric, initially proposed by Chu *et al.* [18]. The clustering was performed only for the top 2000 videos in each category in order to present more accurate results. Videos with a small number of requests do not show a significant pattern over time.

Figures 5.9 to 5.13 show the centroids/clusters for some of the selected categories as found by K-SC. In fact, for all other categories, very similar clusters are found; the Howto category is not considered for this part of analysis as the number of available videos is less than 2000. This is one of the most important findings that a video in YouTube, regardless of its category, follows one of the six depicted growth profiles. This can be vital for any kind of resource optimization problem ranging from caching to advertisement mechanisms. Forcing K-SC to select fewer than six clusters drops the accuracy significantly, as some of the interesting patterns are lost. More than six clusters, however, does not significantly improve the accuracy as repetition of similar clusters take place with very little differences. The figures suggest that there is not a significant number of YouTube videos that reach peak popularity slowly but become unpopular very rapidly. Based on that observation, the growth profiles can be divided into three broad categories.

- Videos that slowly reach peak popularity and stay popular afterwards (cluster (b)).
- Videos that become popular suddenly, but retain popularity for a long period (cluster (c) and to a lesser extent cluster (a)).

- Videos that suddenly reach peak popularity and become unpopular very quickly (all other clusters)

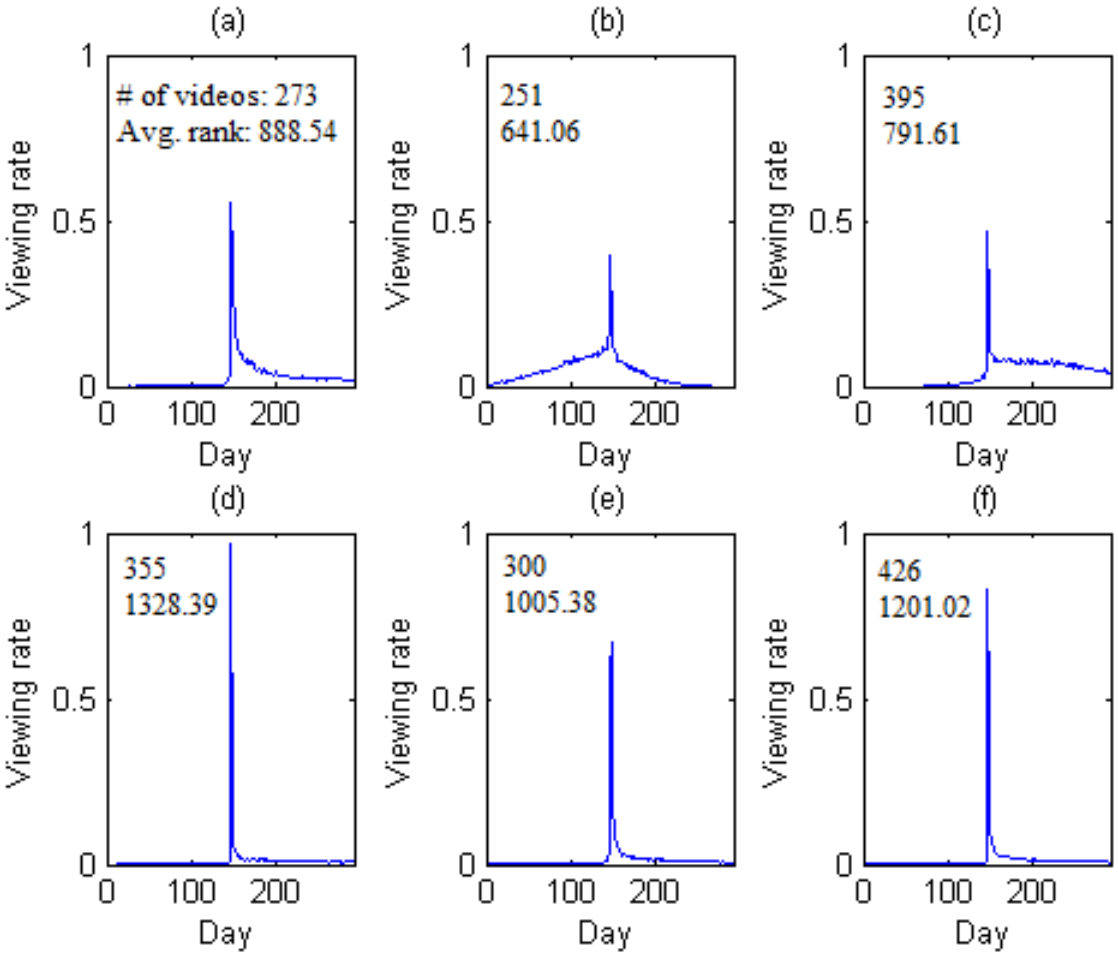


Figure 5.9: Growth Curves of Music-clusters

Another important question that must be answered, besides identifying the number of videos in each cluster, is that whether a particular cluster is more biased to the popular videos than others. This can be answered by taking the average of the rank values of all the videos in a cluster. According to the central limit theorem, the average rank of the videos in each cluster should be around 1000—as the first 2000 videos are considered—if it is not biased to the popular or unpopular videos. Each cluster from Figures 5.9 to 5.13 contains this information (first value is the number of videos in the cluster and the second value is the average rank value of the cluster). The number of videos and average rank for each cluster of all the categories are summarized in Table 5.2 for comparison.

These observations do not contradict the earlier findings, in spite of sharing common growth patterns by different categories. The number and rank of videos in a specific cluster differ significantly among the categories. The shapes of the centroids for News videos, as depicted in Figure 5.10, are very similar to Music (except very little difference between cluster (a) in Figure 5.9 and 5.10 respectively). However, the numbers

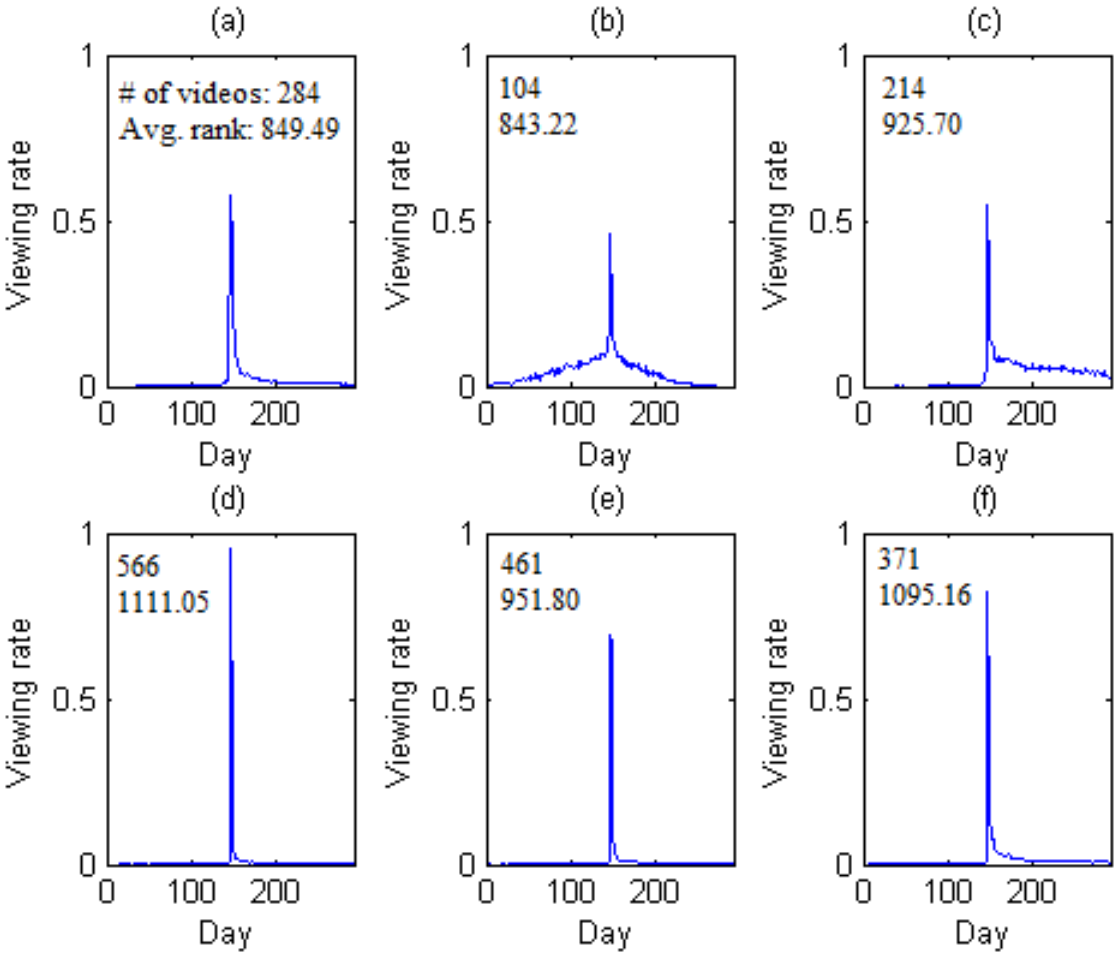


Figure 5.10: Growth Curves of News-clusters

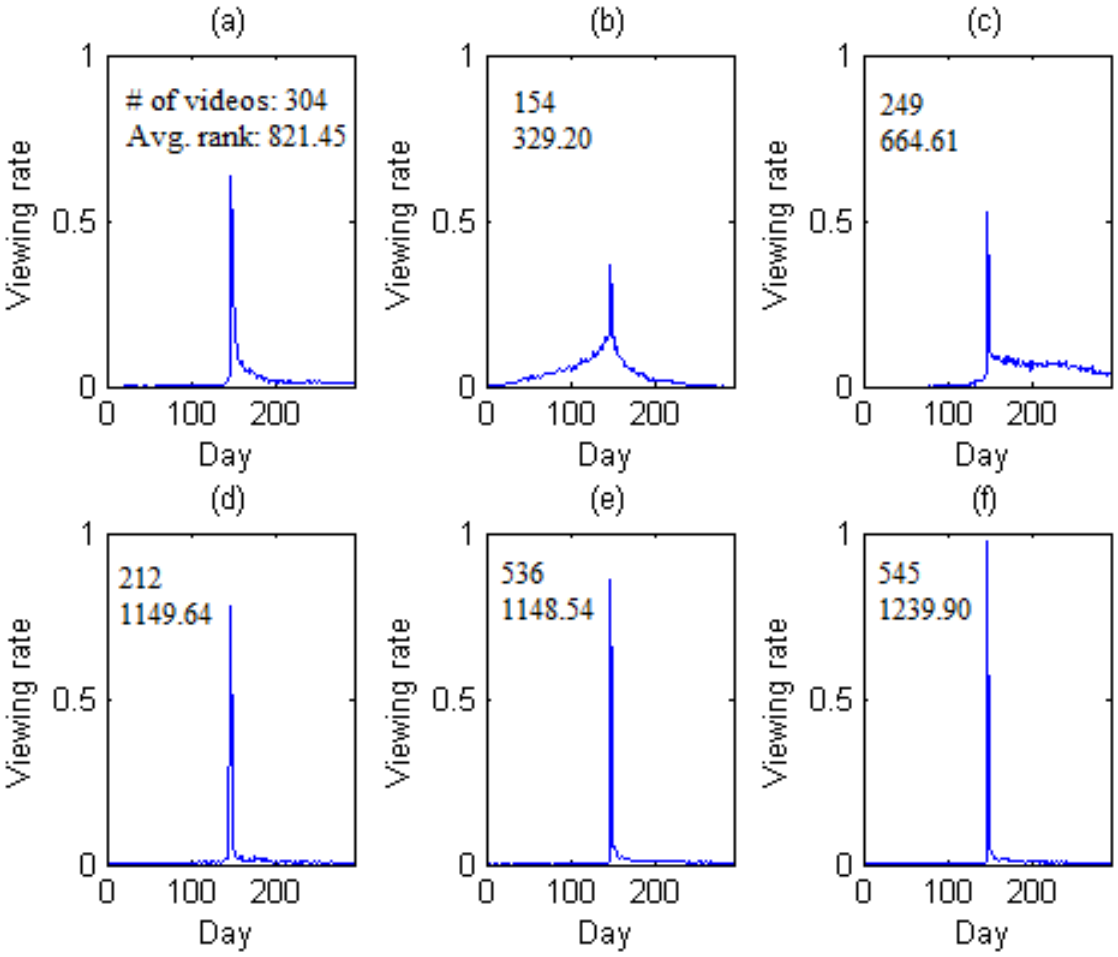


Figure 5.11: Growth Curves of Film-clusters

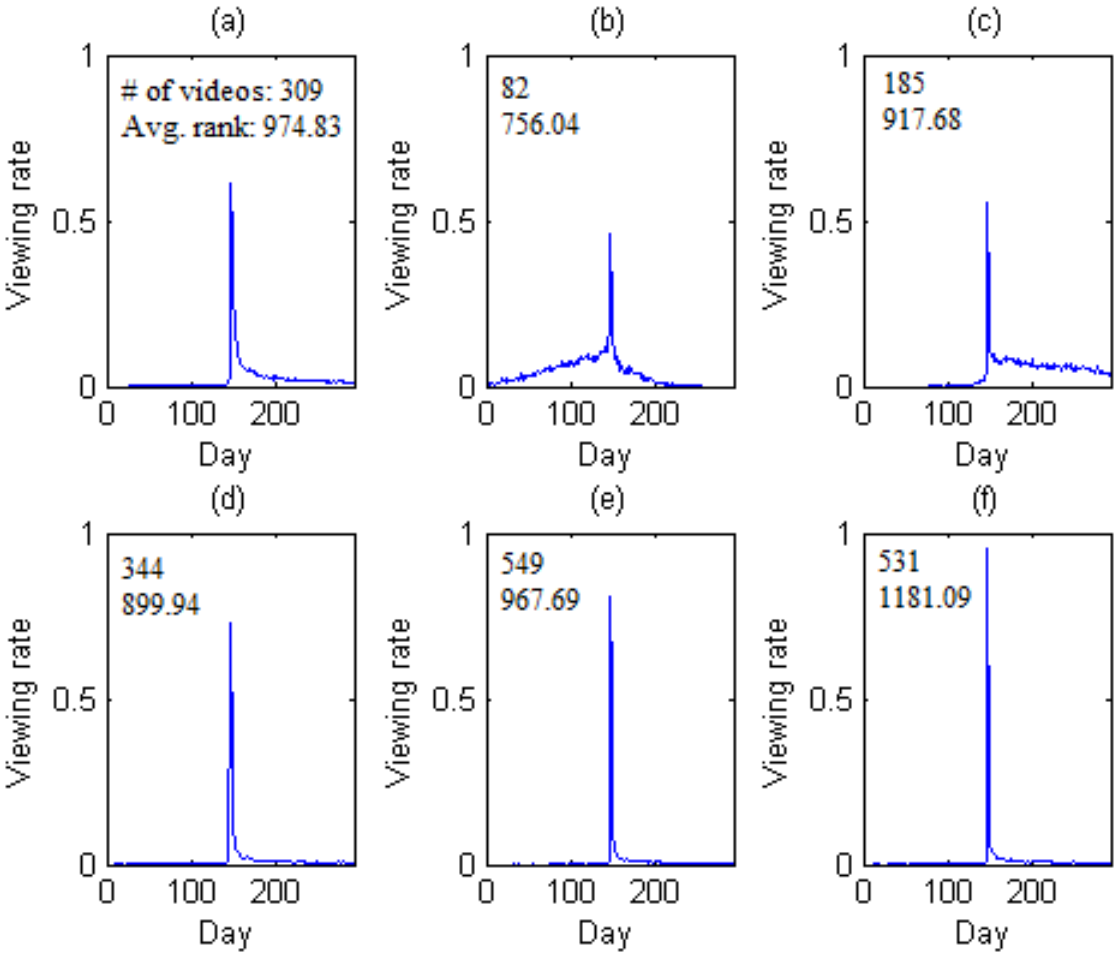


Figure 5.12: Growth Curves of Sports-clusters

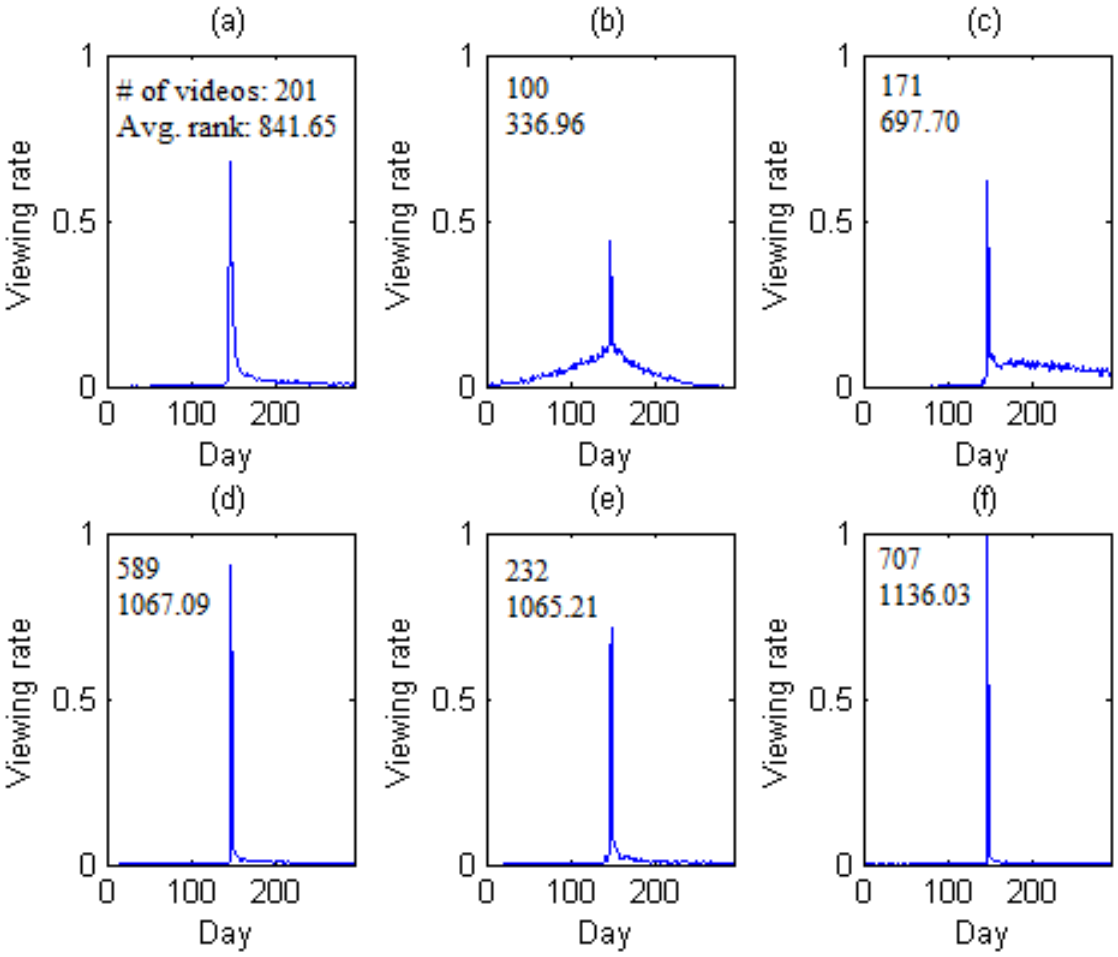


Figure 5.13: Growth Curves of People-clusters

of videos in each cluster differ conspicuously between these two categories. The number of Music videos contained in cluster (b) is approximately 2.5 times more than News videos. These findings complement the earlier findings shown in Chapter 4.

Similarly, the rank values for News videos are found to be very similar for each cluster and does not indicate any significant bias; even for cluster (b), the average rank value is 843 which is not significantly less than 1000. This value is 641 for Music and 329 for Film videos. This is not surprising, as seen earlier that views for Film videos come very steadily. Entertainment videos, as in Table 5.2, follow very similar growth patterns to Film videos. The effect of containing more popular members in cluster (b) can be best described by Figure 5.14, which shows the average viewing rates, centered on peak, of three different categories. Entertainment and Film videos do not exhibit a sharp rise/decay around the peak day, compared to Sports videos; most of the very popular videos in these two categories are contained in cluster (b), as suggested by average rank value in Table 5.2.

Table 5.2: Cluster Information

Category	Cluster (a)		Cluster (b)		Cluster (c)		Cluster (others)	
	No of videos	Avg rank	No of videos	Avg rank	No of videos	Avg rank	No of videos	Avg rank
News	284	849.49	104	843.22	214	925.70	1398	1054.31
Sports	309	974.83	82	756.04	185	917.68	1424	1030.89
Music	273	888.54	251	641.06	395	791.61	1081	1188.55
Film	304	821.45	154	329.20	249	664.61	1293	1187.22
Entertainment	303	851.16	107	382.13	228	638.17	1362	1142.94
Animals	262	839.04	126	554.92	257	750.04	1355	1120.65
Autos	281	841.17	182	577.14	243	641.67	1294	1162.02
Games	272	885.79	996	433.19	206	530.96	1423	1120.85
Nonprofit	323	911.41	183	864.13	270	922.61	1224	1061.97
People	201	841.65	100	336.96	171	697.70	1528	1098.70
Travel	285	792.97	161	677.32	264	739.91	1290	1140.00
Tech	281	939.97	190	490.92	274	667.43	1255	1163.91
Education	238	887.41	308	728.05	330	868.54	1124	1137.83
Comedy	222	903.94	190	503.73	288	946.71	1300	1101.50

The peak distributions of the clusters in a category, however, are not same. As expected, slower cluster shows slower peak distribution. Figure 5.15 shows the distribution for two of the categories. Cluster (b) exhibits the slowest growth rate, complementing earlier observations. This observation is found common for all of the selected categories.

5.4 Workload Generation and Performance of K-SC

In order to evaluate the performance of K-SC, a synthetic workload generator is developed for News, Music, Film and People videos. The People category is selected to see if K-SC can also be applied for modeling unpopular categories. News, Music and Film are not only the popular categories, but also have been found to

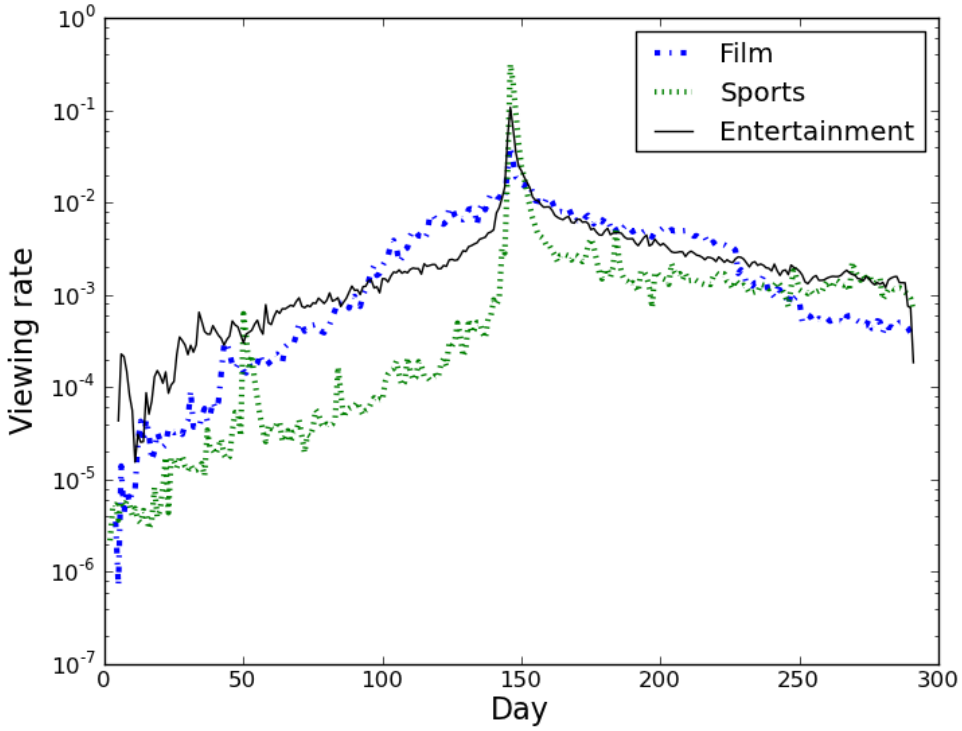
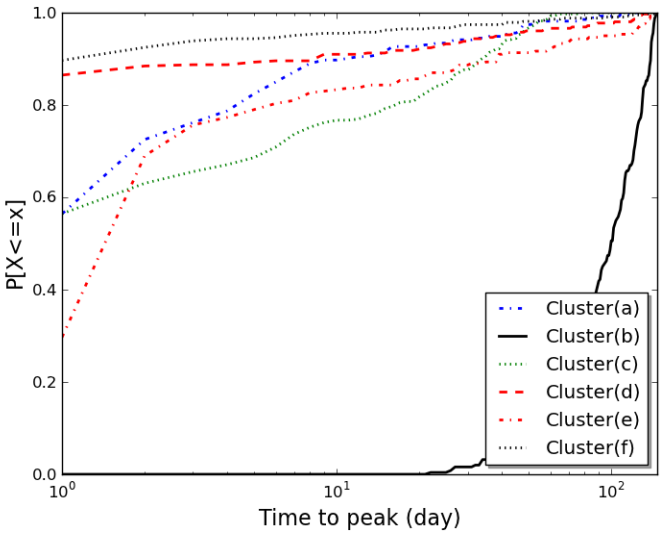


Figure 5.14: Viewing Rate of Top 2000 Videos (Centered on Peak)

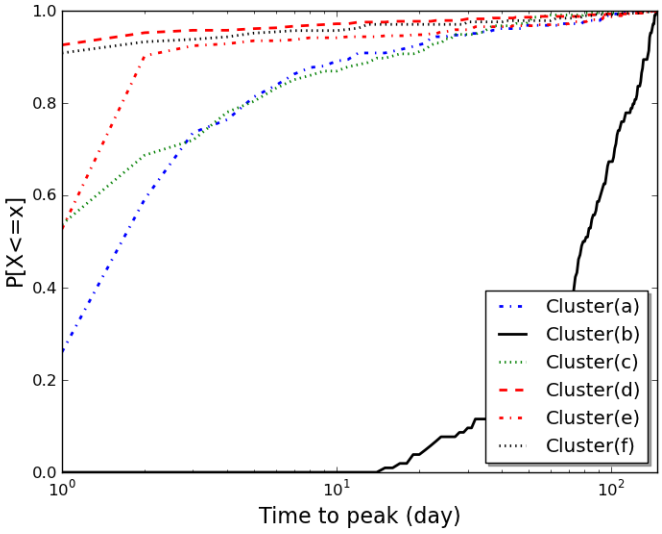
exhibit three different growth rates. The synthetic data should show similar characteristics to the empirical YouTube data if the clustering of K-SC is accurate. The mechanism of the workload generator can be described as follow.

A rank value is assigned to each of the 2000 videos as suggested by Zipf distribution, except for News. Weibull distribution is followed for News videos. Then the video is assigned to one of the six centroids/clusters based on the distribution observed earlier. Bias is imposed for the popular videos before selecting the appropriate cluster in order to match the observed average rank value, which involves a time consuming calibration process. As the peak distributions are conspicuously different among the clusters in a category, each of them are considered separately in the request generator, so that the accuracy of K-SC can be verified. The simulation is run only once for all the selected categories. This is because of the probabilistic assignment of peak, cluster and rank to the videos that produce different results for individual videos in each run and thus taking averages will not generate appropriate results. However, the observed deviation in each run is very little.

Similarity between the synthetic and empirical data is examined from four different perspectives: 1) The total view distribution, 2) time-to-peak distribution, 3) Average daily views over time, and 4) 95^{th} percentile of views over time. Very similar characteristics between the two datasets, especially for metrics 3 and 4, is unexpected because of the following reasons:



(a) Peak Distributions of Music-clusters



(b) Peak Distributions of News-clusters

Figure 5.15: Peak Distributions of Clusters

- Instead of following fixed distributions for views, Zipf/Weibull distributions are applied as observed earlier.
- The cluster for a video is selected randomly although bias was involved in the selection process to have similar average rank values.
- Peak day was selected randomly for a video in a cluster.
- Outliers that exist, more or less, in videos growth patterns, were largely ignored by K-SC.

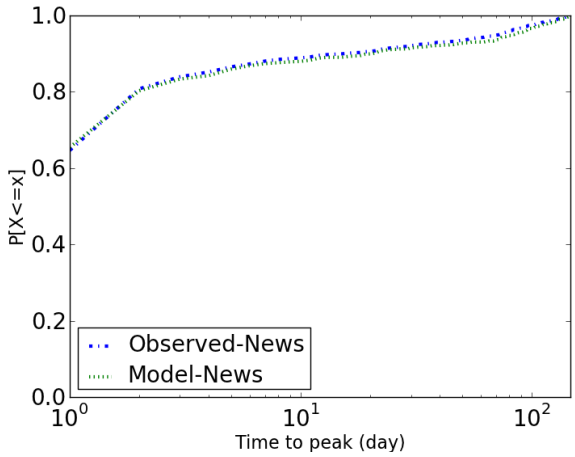
Figures 5.16 to 5.18 indicate very good matches between synthetic and empirical data for metrics 1 (peak distribution) and 2 (total view distribution), which is not surprising and does not in itself indicate high accuracy of K-SC; the distributions, fixed peak and Zipf/weibull for total views, were imposed in the simulation. For example, popularity distribution of News videos was found to be best described by Weibull distribution. Thus, following the same Weibull distribution for generating synthetic views will definitely produce very similar results to the empirical dataset.

However, the matches for metrics 3 and 4 for all of the examined categories confirm that K-SC algorithm exhibits very good performance in identifying growth patterns of groups of videos in YouTube. Similar daily average views on a particular day indicates that view distribution among videos on that particular day are similar both in the empirical and synthetic data. These results also supports the existence of few outliers in both of the categories. If there were no outliers in the original dataset, better matches could have been expected. On the contrary, with large number of outliers, the centroid in a cluster should be distorted so much so that unacceptable similarity would be observe.

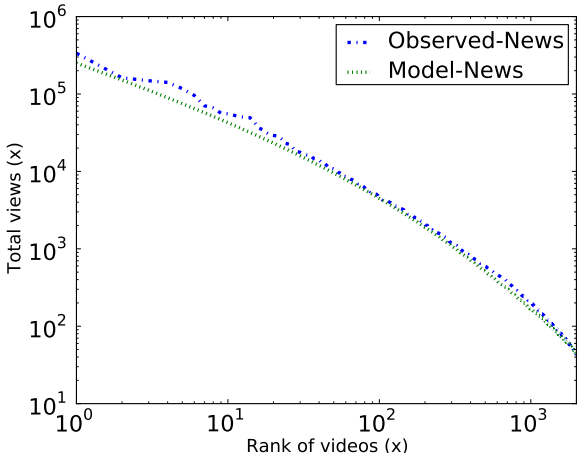
Describing growth curves with only six clusters implies that YouTube videos growth pattern, in most of the cases, are not totally random. These findings can be helpful to the network administrators and on-line marketers [5, 36, 48].

5.5 Summary

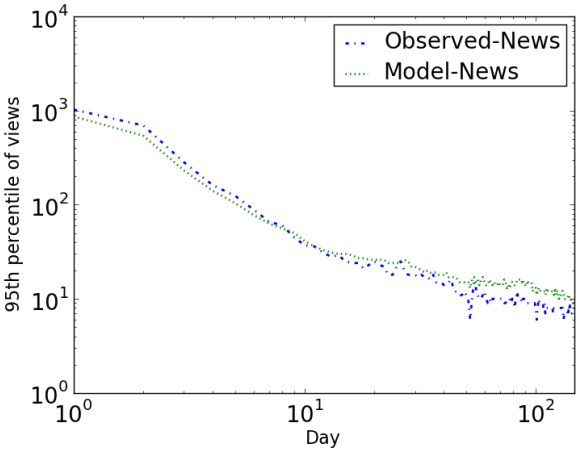
This chapter mainly focused on understanding and modeling popularity evolution of YouTube videos. Results confirm that characterization is more accurate when content type is considered. Early view patterns can be successfully used to estimate future popularity for some of the categories. This has direct applications in on-line marketing and other effective services like recommendation and searching [14]. For example, identification of popular content can be used in ranking policies and thus improving the quality of searching and recommendation [42]. However, this phenomenon is not observed for all the YouTube categories. In categories like Music and Film, many videos become popular much later than the uploading time. As a result, observing the first couple of days' views has very little predictive power regarding future popularity. Three-phase behaviour also fails to model YouTube categories. Finally, it has been found that time-series clustering algorithm, like K-SC, can be used to model popularity evolution of YouTube categories.



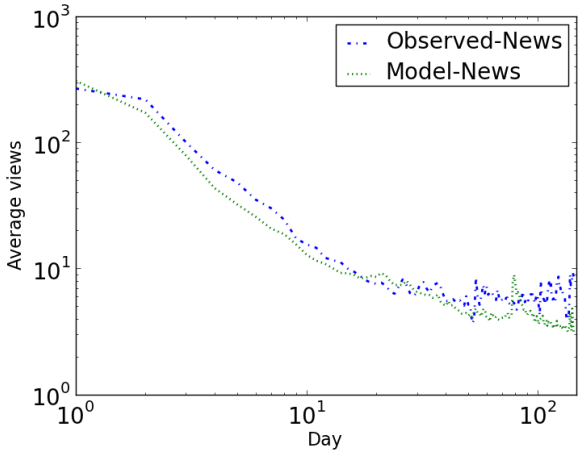
(a) Peak Distributions (Synthetic vs. Observed)



(b) Total Views (Synthetic vs. Observed)

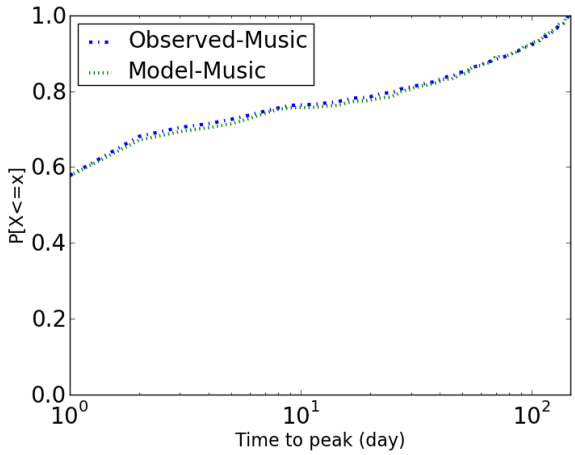


(c) 95th Percentile of Views (Synthetic vs. Observed)

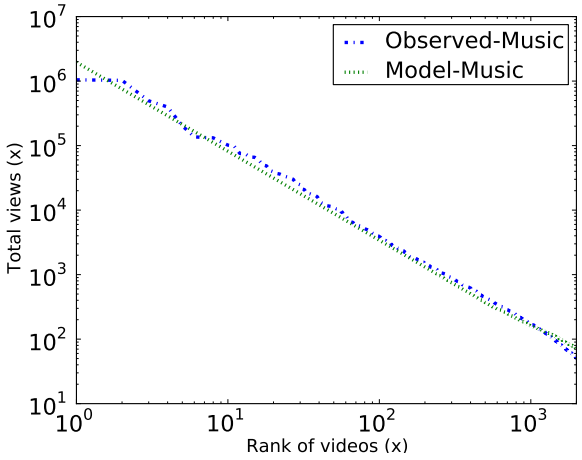


(d) Added Views (Synthetic vs. Observed)

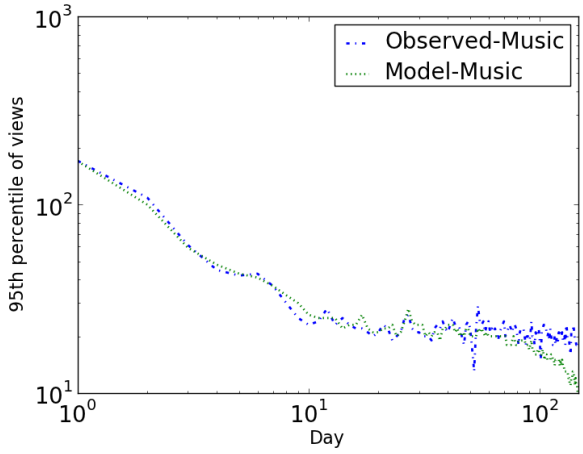
Figure 5.16: Synthetic vs. Empirical (News)



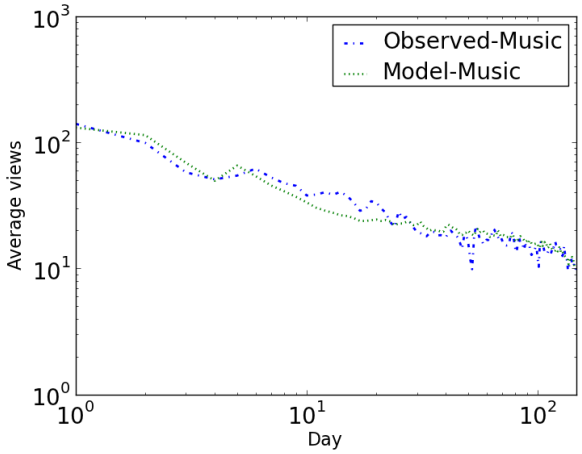
(a) Peak Distributions (Synthetic vs. Observed)



(b) Total Views (Synthetic vs. Observed)

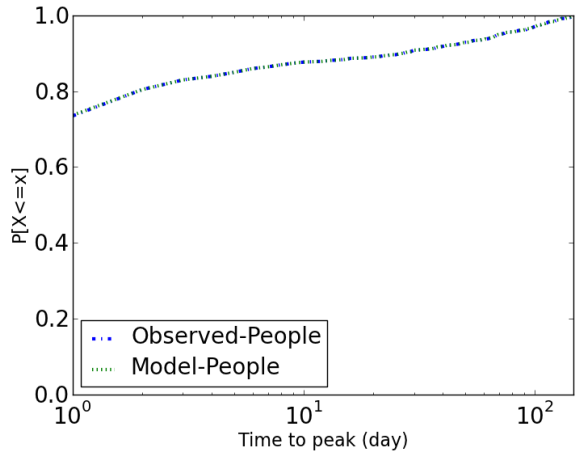


(c) 95th Percentile of Views (Synthetic vs. Observed)

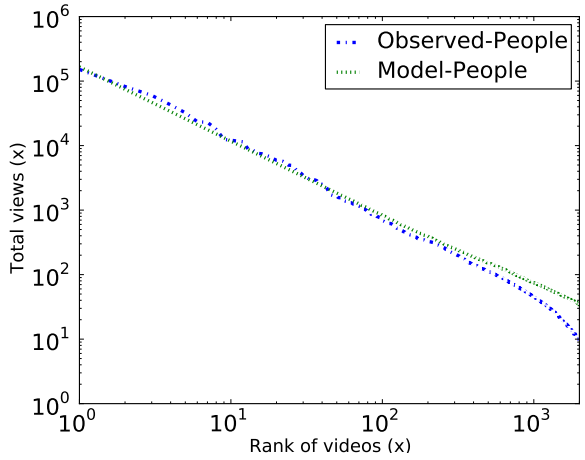


(d) Added Views (Synthetic vs. Observed)

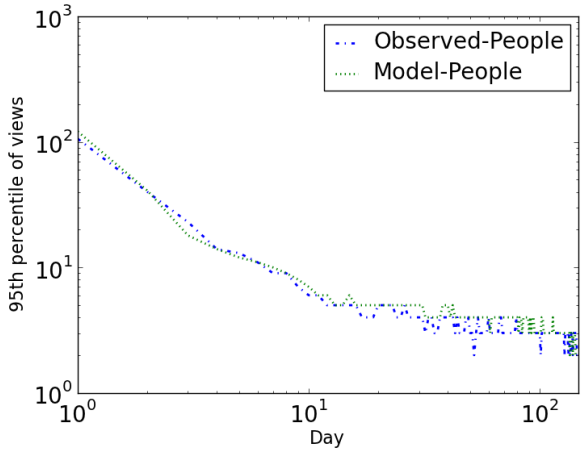
Figure 5.17: Synthetic vs. Empirical (Music)



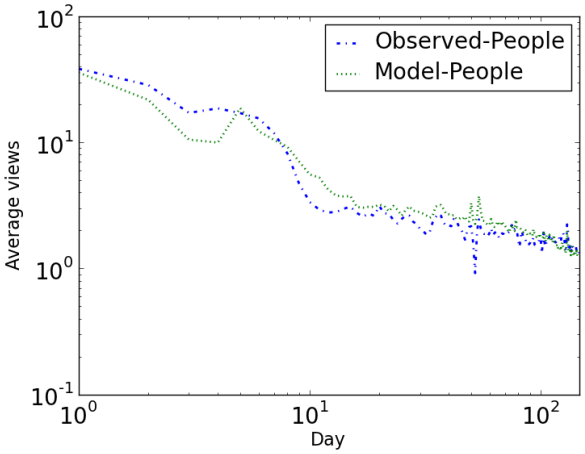
(a) Peak Distributions (Synthetic vs. Observed)



(b) Total Views (Synthetic vs. Observed)

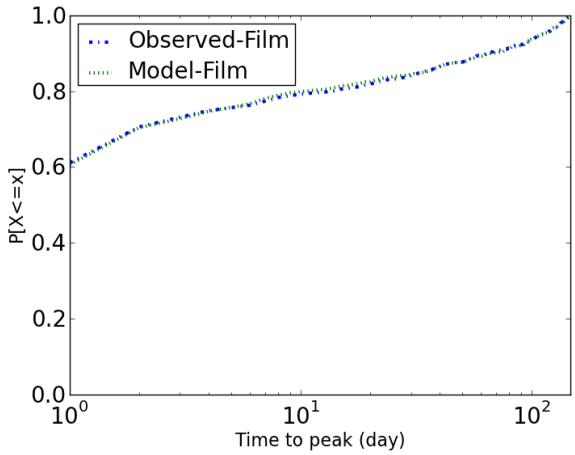


(c) 95th Percentile of Views (Synthetic vs. Observed)

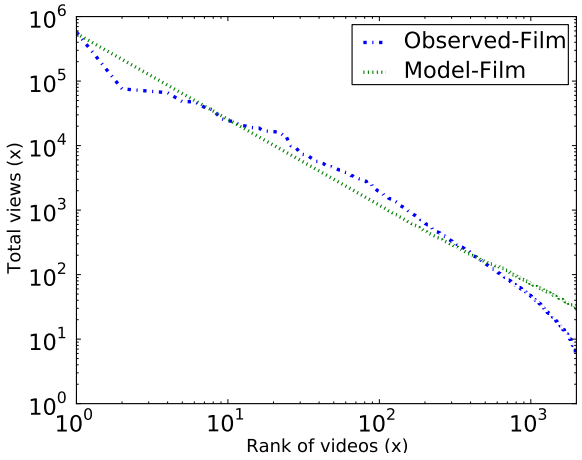


(d) Added Views (Synthetic vs. Observed)

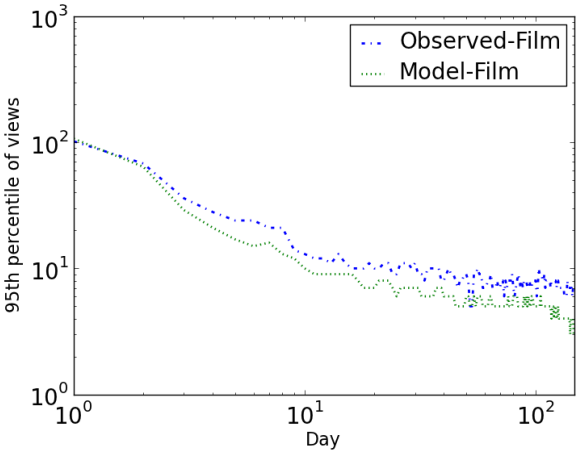
Figure 5.18: Synthetic vs. Empirical (People)



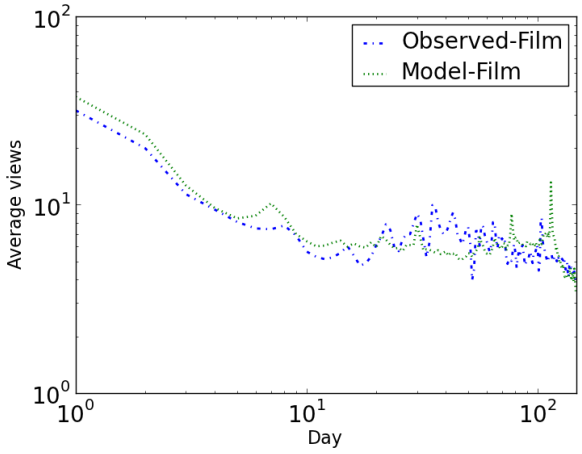
(a) Peak Distributions (Synthetic vs. Observed)



(b) Total Views (Synthetic vs. Observed)



(c) 95th Percentile of Views (Synthetic vs. Observed)



(d) Added Views (Synthetic vs. Observed)

Figure 5.19: Synthetic vs. Empirical (Film)

Chapter 6

User Interaction with YouTube Videos: Implications for Local Caching and Video Recommendation

Identification of user interaction patterns with YouTube videos can be vital in two specific areas: developing and deploying effective caching mechanisms and video recommendation systems. YouTube does not share the list of users that watched a particular video, which might be one of the reasons behind the lack of research in this particular topic. This chapter explores these two important areas using a dataset that was collected at the University of Massachusetts (UMass) [57].

The problem with LRU cache replacement policy for P2P video distribution, as discussed in Chapter 2, can be minimized by employing an algorithm that emphasizes each user's local utility (predicted) of a video than the global utility in the network of the same video. Such a mechanism ensures the availability of the very frequently accessed videos by each user, while maintaining the advantages of the P2P technique. It is difficult to predict, in advance, which videos are going to be requested multiple times by the same users. The difficulty of this problem can be reduced substantially by observing if there is a relation between video type and repeated views.

Another important area of interest in video distribution systems is to design an effective video recommendation algorithm. YouTube uses a user's recent activity (e.g., watched, liked, favorited videos) to construct the set of recommended videos for the user [21]. From the perspective of individual YouTube videos, video recommendation using related video lists has been found as the most effective technique to increase video popularity [55]. The major problem in formulating the related video lists is that most of the YouTube videos often have no or very poor metadata [21]. However, similarities among the YouTube users in watching videos is not examined by the previous studies, which could be a potential approach for YouTube video recommendation system.

The basic contribution of this chapter is twofold.

1) Relationship between repeated views and video category is analyzed. Outcomes of this analysis can be used directly in designing caching approaches for P2P video distribution.

2) A new promising video recommendation technique for YouTube is presented, which encompasses the idea of forming communities among YouTube users based on their similar video watched list. Furthermore, small-world network's properties is found in the network established through YouTube users' interests.

6.1 Description of the Dataset

Zink et al. [57] collected six different datasets in six different periods from the network of University of Massachusetts, referred as T1 to T6. When a campus user made a request to the YouTube server, the request was captured with the specific video_id and user's IP address. Among all the datasets, T5 is used in this thesis, as the measurement period of this trace collection was the longest (14 days), from 01/29/08 to 02/12/08. Each record in the dataset contains video_id, user's IP address, and the date and time the video was requested. A total of 16,336 unique users is reported in the trace collection, assuming each IP address corresponds to a unique user, with the accompanying limitation that an IP address could correspond to a lab computer. In this thesis, each machine is considered as a unique user, and it is expected that students usually do not use lab computers for watching YouTube videos. This specific assumption, however, does not have any impact on caching strategies, as cache hit ratio depends on the patterns of repeated requests of videos from the same sources (machines/IP addresses). The number of unique videos accessed from the network within that trace period is 263,970.

Unfortunately, the category of a video was not captured which is very necessary for this research. Another crawler was implemented that collects the category name for a given video_id. After the data collection, 133,722 videos with category names were obtained for further analysis; much video metadata is unavailable because of the deletion from the YouTube repository. Table 6.1 shows the number of unique videos in each category, and Figure 6.1 shows the relative quantity for the videos that have not been deleted.

Music is at the top position, followed by Entertainment and Comedy, in case of number of accessed unique videos, whereas News is at the seventh position. Interestingly, a very similar observation was found for University of Calgary campus [27]. The category with the least number of videos in the sample dataset is Nonprofit; this is also one of the most unpopular categories in the global dataset (Table 4.2). While the number of accessed unique videos does not necessarily indicate how popular a category is, Table 6.2 shows the viewing rates of different categories with different threshold values.

Among the non-deleted videos, 36% of the videos were accessed more than once within the measurement period of 14 days (40% for deleted videos). Most importantly, the number of requests made by the campus users (considering both the deleted and non-deleted videos) is 611,968, which means up to 56% hit ratio (as in equation 6.1) can be obtained by employing an intelligent caching algorithm; the number of unique videos for equation 6.1 is 263,970. Although the number of accessed unique Tech videos is very small, it is the most popular category under the campus network, considering the percentages of videos with more than 50 and 100 views. On the contrary, even under the university network, no Education video was watched more than 50 times. The percentage of popular Comedy videos, within the measurement period, is significantly larger than Music and Entertainment videos, which is completely different than the findings of the global analysis (Table 4.2). This is further discussed in Section 6.3. While only 0.029% videos, from the non-deleted videos, were accessed more than 100 times within the measurement period, this ratio is much higher for News videos.

Table 6.1: Videos/Category

Category	**No. of videos**
Music	42223
Entertainment	31484
Comedy	14942
People	9749
Sports	9747
Film	8100
News	5295
Animals	3066
Howto	2720
Autos	2228
Travel	1737
Education	816
Tech	784
Games	474
Nonprofit	357
Total	**133722**

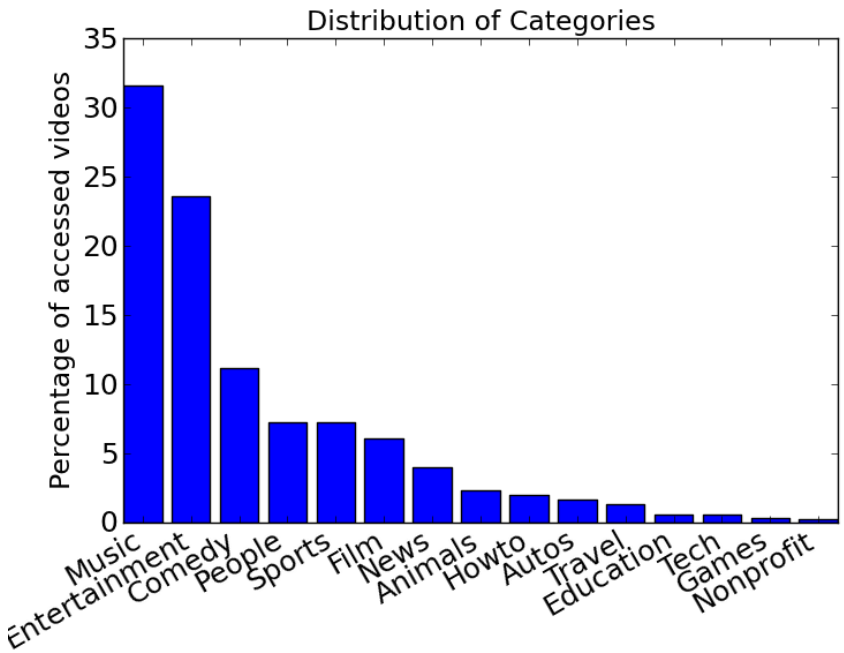


Figure 6.1: Category Distribution

Surprisingly, two People videos have the property of being watched more than 100 times, in spite of the low global popularity of this category. This suggests that some videos can be very popular regardless of their categories.

$$h = \frac{(total\_requests - number\_of\_unique\_videos) * 100}{total\_requests} \tag{6.1}$$

Table 6.2: Percent of Popular Videos in UMass 2008 (Snapshot T5)

Category	$>$ 1 views		$\geq$ 10 views		$\geq$ 50 views		$\geq$ 100 views	
	videos	Pct	videos	Pct	videos	Pct	videos	Pct
Music	16131	38.2	1204	2.85	44	0.1	8	0.02
Entertainment	11127	35.34	601	1.91	38	0.12	6	0.02
Comedy	5985	40.05	526	3.52	51	0.34	13	0.09
People	3407	34.95	178	1.83	12	0.12	2	0.02
Sports	2944	30.2	113	1.16	7	0.07	0	0.0
Film	2932	36.2	143	1.77	4	0.05	1	0.01
News	2073	39.15	190	3.59	14	0.26	7	0.13
Animals	1037	33.82	77	2.51	5	0.16	0	0.0
Howto	941	34.6	32	1.18	1	0.04	0	0.0
Autos	594	26.66	14	0.63	0	0.0	0	0.0
Travel	539	31.03	23	1.32	0	0.0	0	0.0
Tech	307	39.16	37	4.72	3	0.38	1	0.13
Education	300	36.76	25	3.06	0	0.0	0	0.0
Nonprofit	149	41.74	16	4.48	0	0.0	0	0.0
Games	121	25.53	8	1.69	1	0.21	0	0.0
Total	48587	36.33	3187	2.39	180	0.13	38	0.029
Deleted	52066	40.05	3930	3.02	205	0.16	67	0.05

6.2 Similarity to Zipf Distribution

Unlike the tail section in the access frequency distribution observed with global access patterns, YouTube video popularity in a local network has been shown to follow a Zipf distribution [27]. The same general observation is found if the category of a video is considered; a comparatively weaker fit for News, Animals, Autos, and Games is observed, as in Figures 6.2 to 6.5. The deviations of these categories mainly occur in the head sections of the distributions. It is possible that deletion of videos may affect the observations of these four categories; the deviation for the entire dataset mainly occurs in the head section (Figure 6.6),

again possibly supporting earlier findings that popular videos suffer more from deletion rate than unpopular videos. Moreover, the videos in this dataset are not of similar ages, which can also be responsible for this phenomenon; a video that is too young, or was first accessed in the local network very recently did not have enough time to be a member of the popular video list [29]. Similarity to Zipf distribution suggests that caching videos according to their rank can be a successful approach even in a confined network.

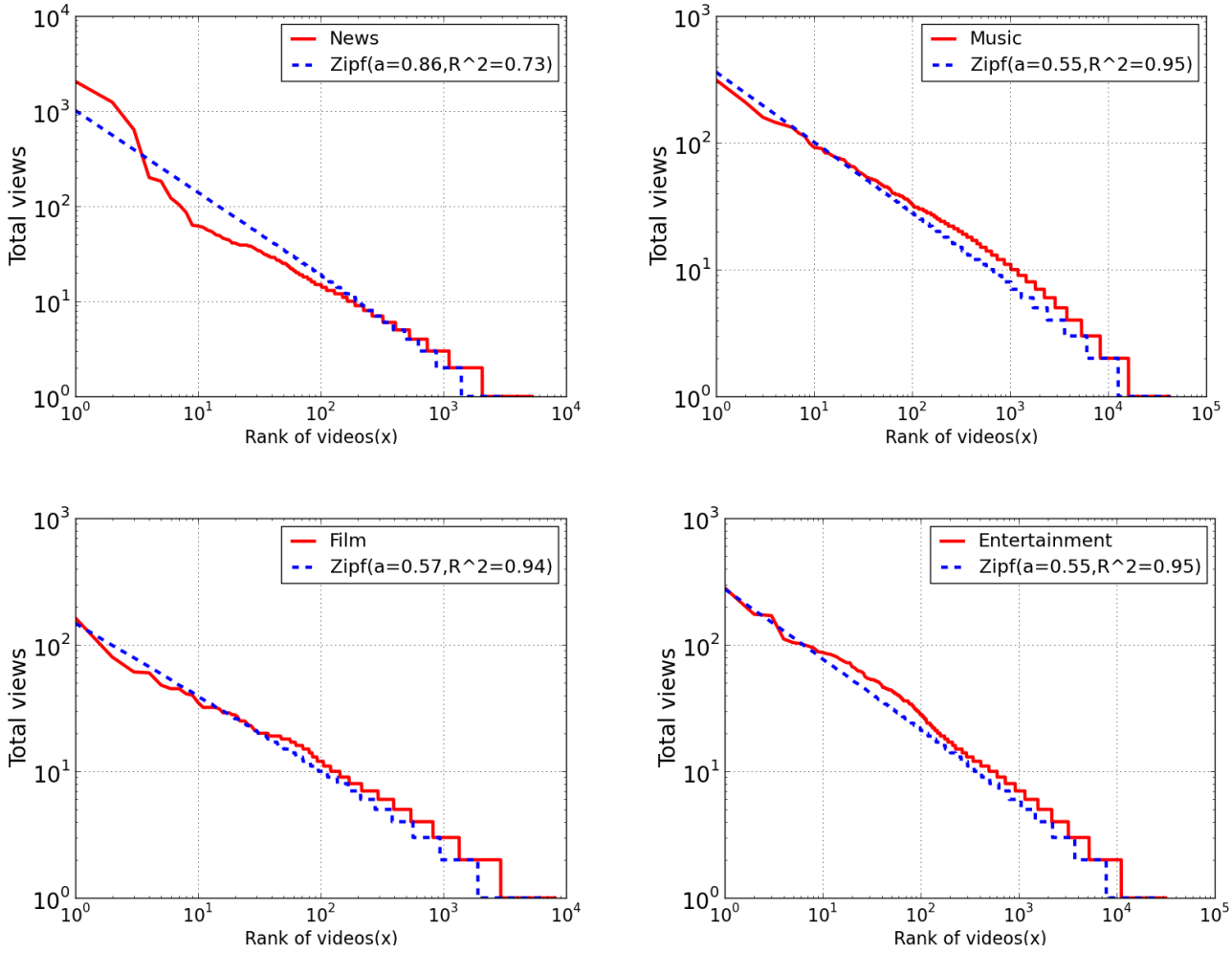


Figure 6.2: Number of Views Against Rank (Selected Categories)

6.3 Repeated Views in Categories

This section analyzes the influence of repeated views on popularity in different categories. This explains if videos in a category become extremely popular due to a large number of users, or due to attracting the same set of users multiple times, assuming each machine as a unique user. A video that exhibits both/either of the phenomena, can become a candidate to be a member of the set of YouTube's most popular videos. In

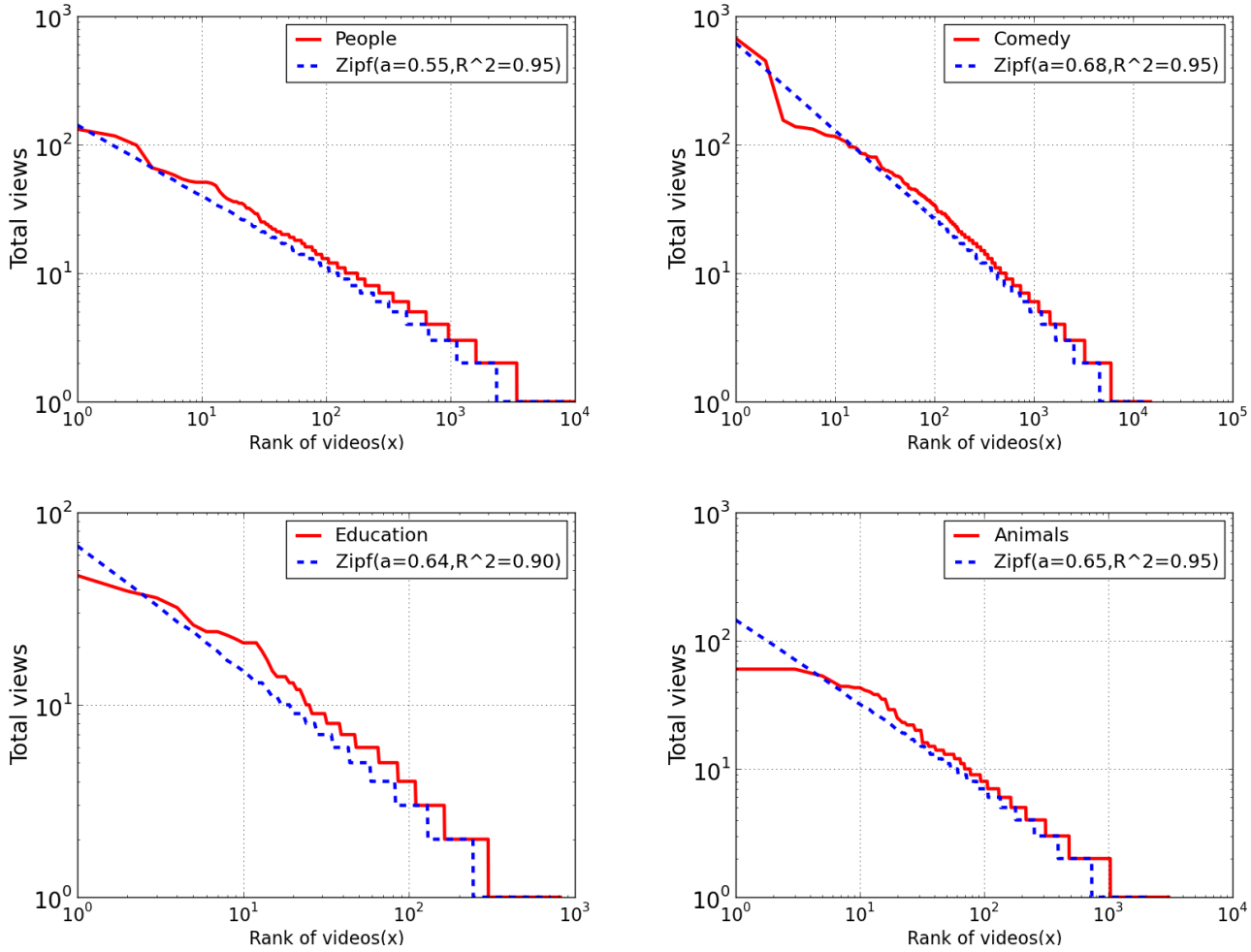


Figure 6.3: Number of Views Against Rank (Selected Categories)

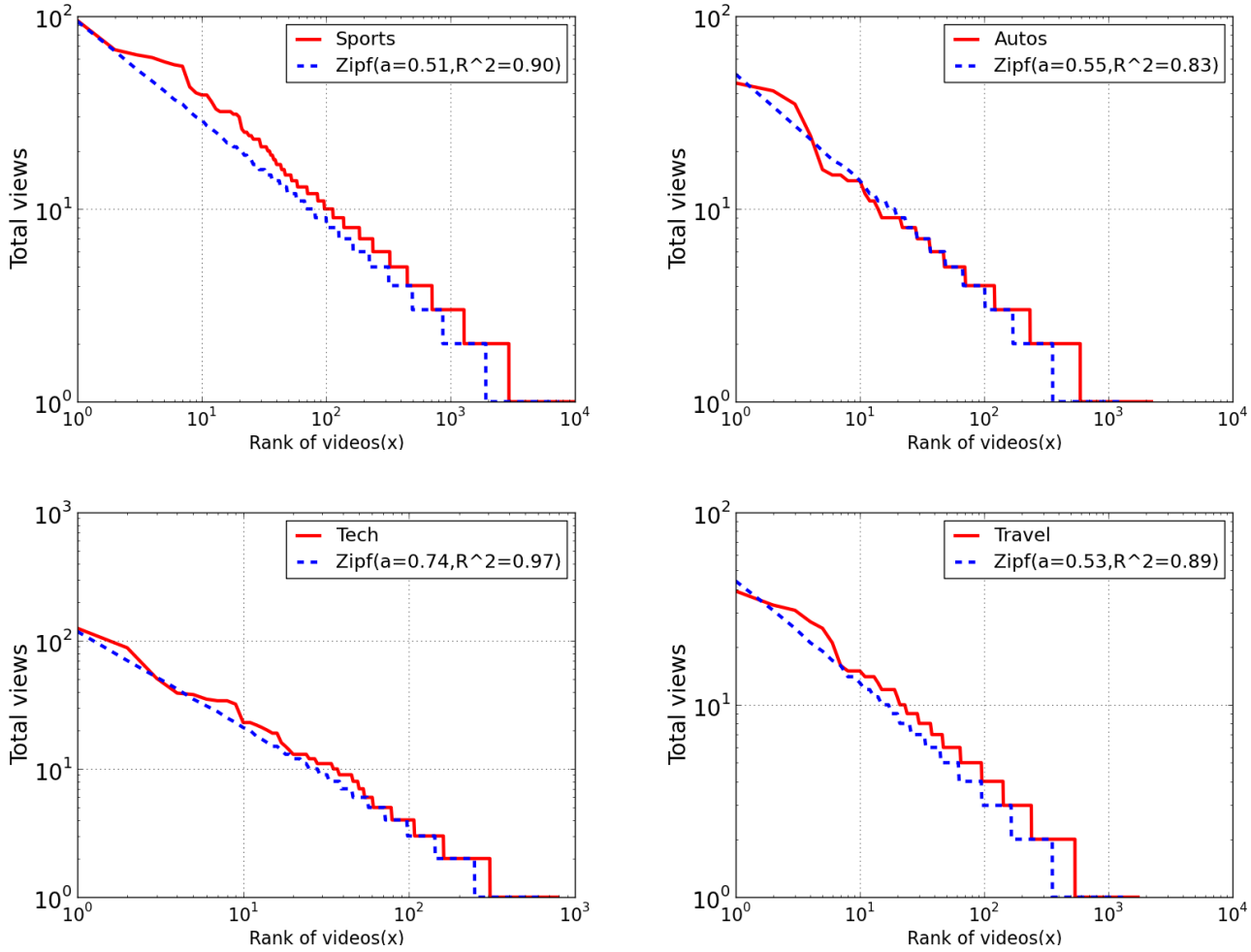


Figure 6.4: Number of Views Against Rank (Selected Categories)

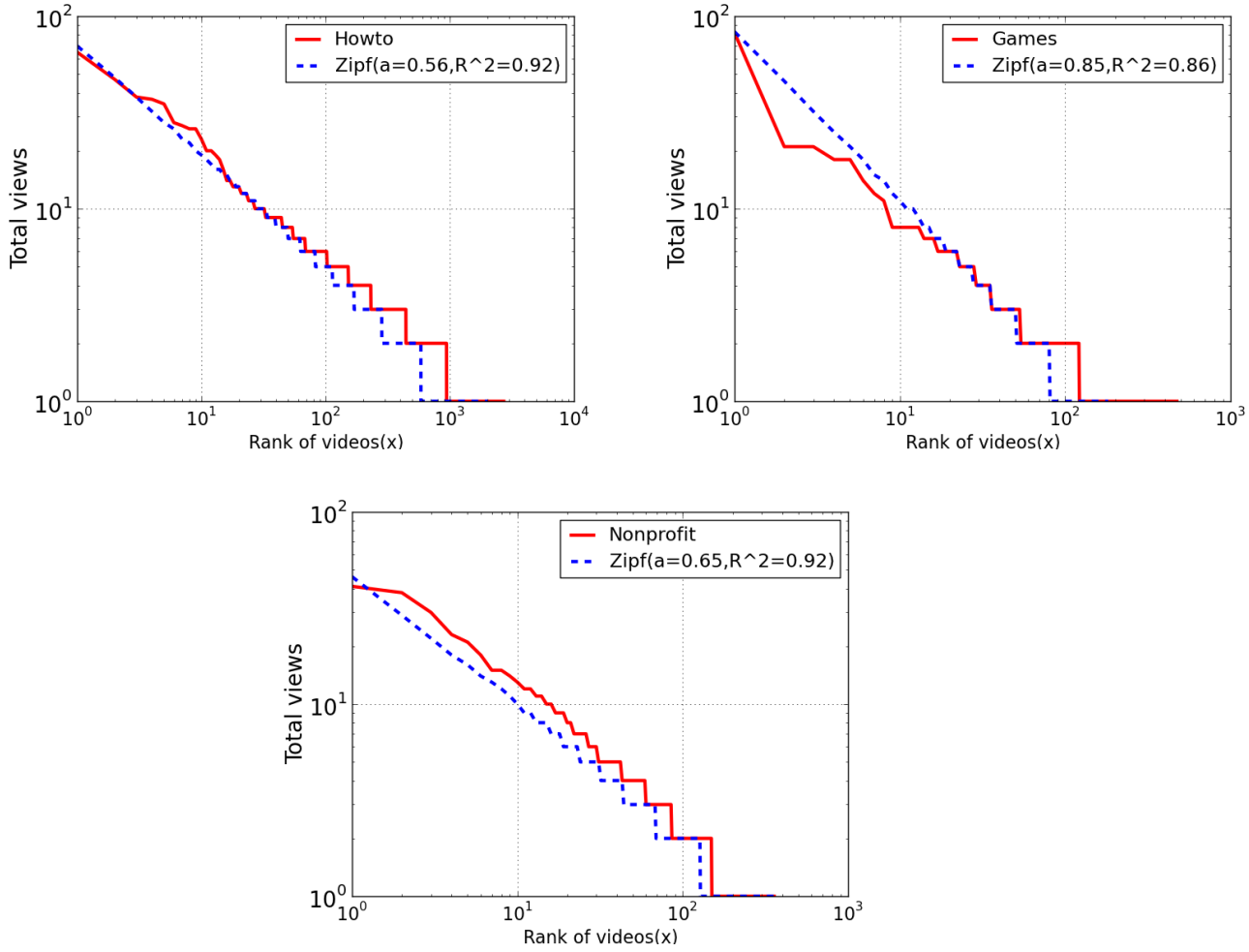


Figure 6.5: Number of Views Against Rank (Remaining Categories)

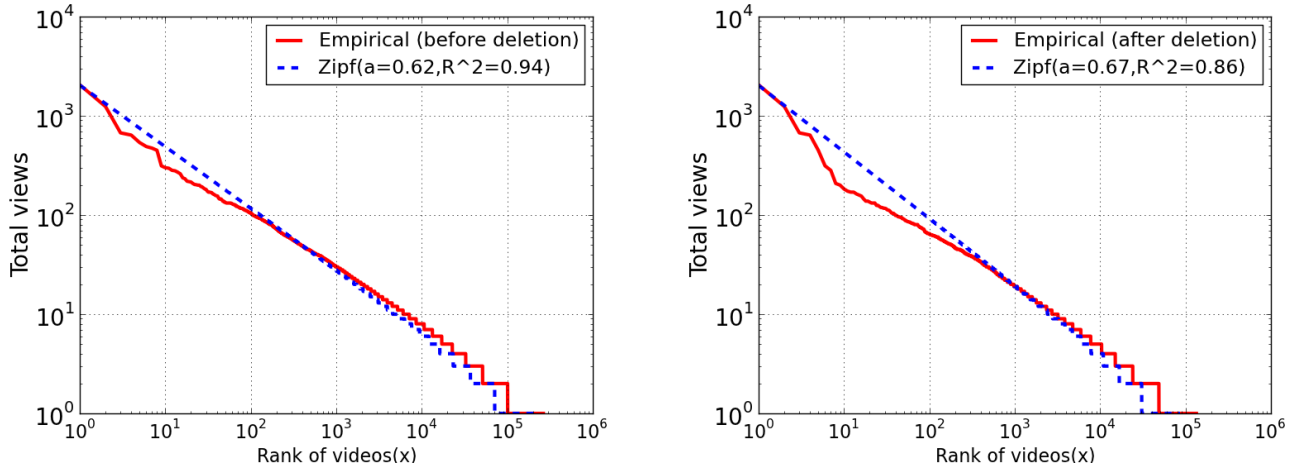


Figure 6.6: Number of Views Against Rank Before and After Deletion

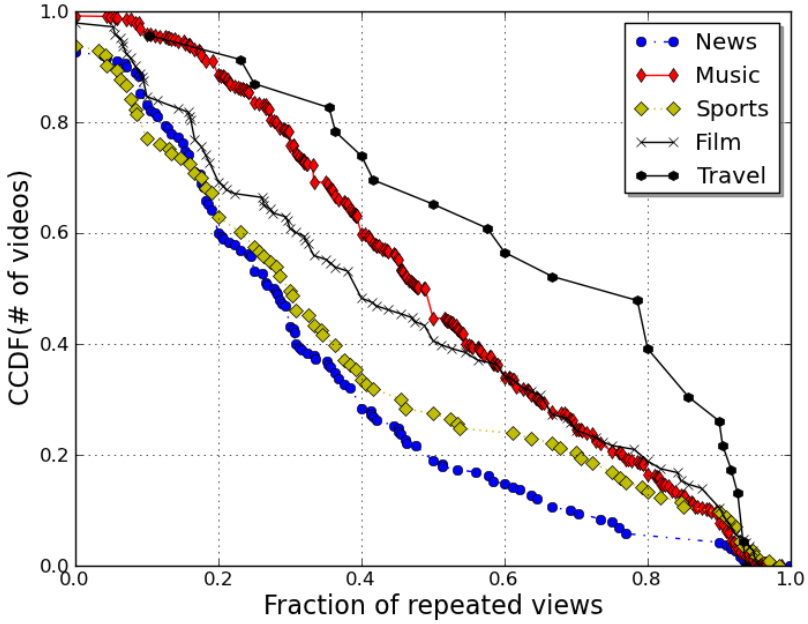
order to have an accurate observation, videos with at least 10 views are considered for this part of analysis.

Table 6.3 depicts the average views of users for different categories. The number of average views/user for each video is calculated for each of the categories. Then three different statistics are shown to observe the outliers effects: mean of average views, median of average views and maximum of average views. Although the average metric shows much higher viewing rate by the users for News than Music, the median for Music videos is higher than News. This is because of the small number of outliers as can be seen from the maximum of average views. Interestingly, the median for Travel videos is four, indicating that the few users that are interested in this category watch the same video multiple times. In subsequent analyses, it is shown that Music videos, in general, enjoy much more repeated views than News videos.

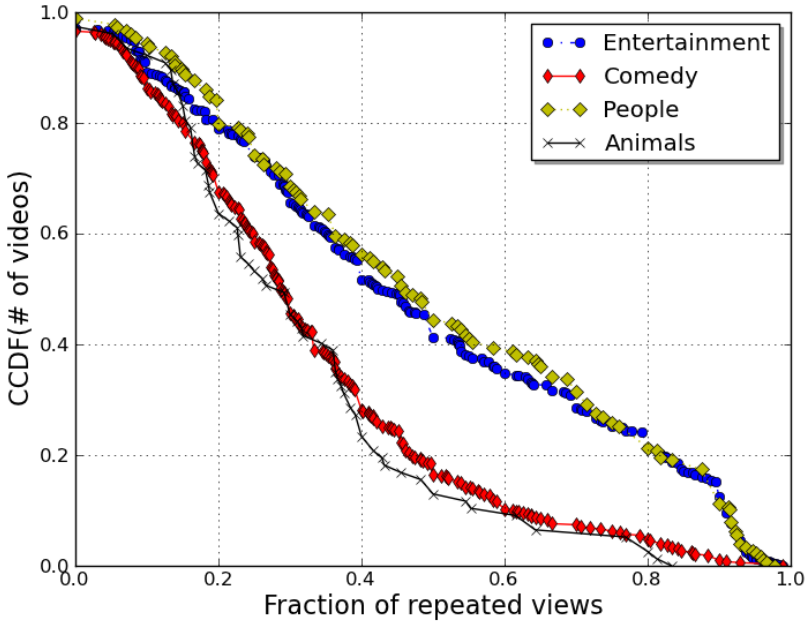
Table 6.3: Mean, Median and Maximum of Average Views for Users (With $\geq$ 10 views)

Category	Mean avg. views/user	MED. avg. views/user	MAX. avg. views/user
News	4.62	1	513
Sports	3.43	1	63
Music	3.2	2	80
Film	3.52	1	23
Entertainment	3.96	1	85
Games	4.5	2	21
People	4.08	1	43
Comedy	1.76	1	95
Tech	2.43	1	19
Travel	5.87	4	16
Autos	2.29	1	15
Education	4.08	2	17
Animals	1.35	1	6
Nonprofit	2.13	1	7
Howto	4.66	1	28

Figure 6.7 depicts the fractions of videos in selected categories that experience different scales of repeated views. The fraction of repeated views is plotted in X-axis (as in equation 6.2), where the Y-axis shows the CCDF of number of videos. In equation 6.2, views (v) is the total number of views of a video v and users (v) is the total number of unique users that watched video v. The Travel category enjoys a significant number of repeated views for most of its videos. This explains the higher median of average views shown earlier in Table 6.3. The significantly more repeated views of Music than News clarifies the observation that there are more extremely popular Music videos than News videos which was found in the global analysis. More than 90% of the Music videos observed 20% repeated videos, which is true only for 60% of the News videos.



(a) Repeated Views (Selected Categories)



(b) Repeated views (Selected Categories)

Figure 6.7: Repeated Views of Categories

This difference could be more significant if the measurement period is longer; the global analysis shows longer active lifespans of Music videos than News. Film videos, on the other hand, show similar patterns to Music for the videos that experience most of their views from the same users. This could be one of the reasons behind the longer lifespans of Film videos.

$$fr_v = \frac{views(v) - users(v)}{views(v)} \tag{6.2}$$

Figure 6.7 (b) shows that Entertainment videos experience significantly more repeated views than Comedy videos. This can play a vital role for a significantly longer measurement period, as observed from the difference in popularity between the global and local analysis of these two categories. People videos, in spite of very similar patterns to Entertainment, was found to be one of the most unpopular categories in the global analysis. This is because of the smaller number of unique users that are interested in this category.

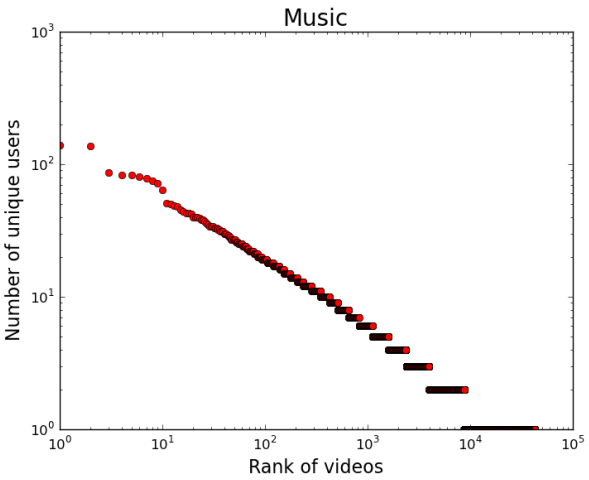
The importance of the number of users attracted by a category is depicted in Figure 6.8. Although Travel videos experience most of the views from the same set of users, the number of unique users in this category (compared to Music) is extremely low. In case of Music, even the video in position 10^{th} attracted approximately 60 unique users, whereas the same ranked video had only six unique users in the Travel category. A similar conclusion can be drawn for Entertainment and People videos, explaining the higher popularity of Entertainment videos than People videos, in spite of their similar repeated viewing patterns.

6.4 Singleton Views

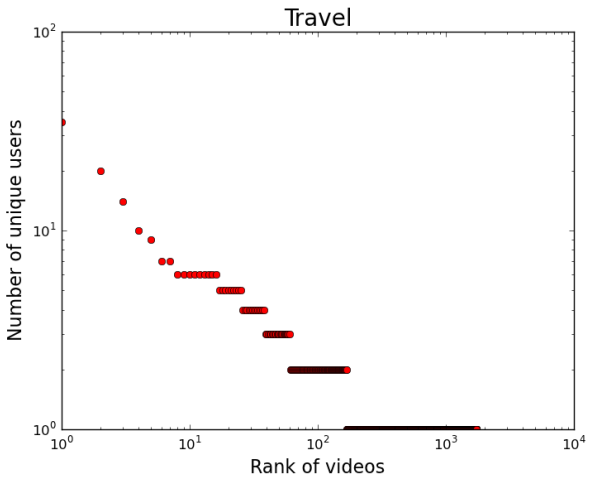
The previous measurement of repeated views, however, could be misleading in some cases, as the evaluation can be affected to a large degree by very few outliers. For example, the fraction of repeated views is more than 85% for a video which was watched by four users only once, but 30 times by a single user. Although most of the users were interested in this video a single time, the statistics draw a different conclusion. The outlier effect can be observed from Figure 6.9. Videos are ranked according to the number of views by a single user. For a better comparison, only first 1737 videos are considered for the categories in Figure 6.9 (a), as the number of remaining Travel videos is 1737. The same procedure is followed for 6.9 (b).

Except for a single News video, Music videos attract the same users to watch the same video much more than any other categories. No user, in spite of the significant repeated viewing patterns, watched a single Travel video more than 20 times. In the second group of videos, Entertainment videos are in top position in case of maximum views by a single user.

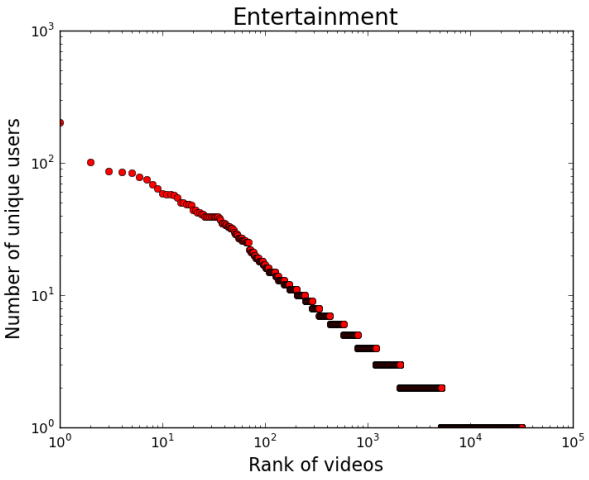
The News video mentioned previously is the most popular video in the entire data set. Upon manual inspection, this video is found to be a Music video about a News story. Moreover, the video was requested almost exclusively by one IP address (2045/2052 requests), 100 of which are separated in time by less than 10 seconds, and span 325 hours of the trace. This is obviously some kind of programmatic access or problem in the logging data, as the video is 270 seconds long and this IP address is the second most active user (4141



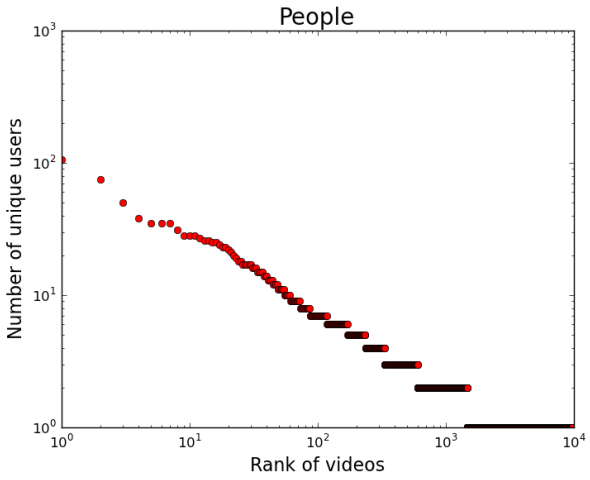
(a) Number of Unique Users vs. Rank (Music)



(b) Number of Unique Users vs. Rank (Travel)

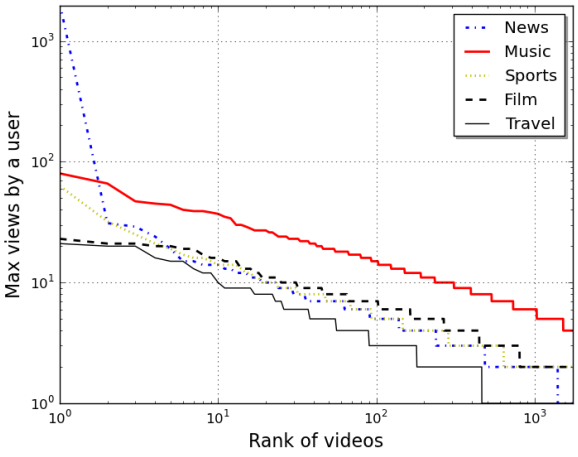


(c) Number of Unique Users vs. Rank (Entertainment)

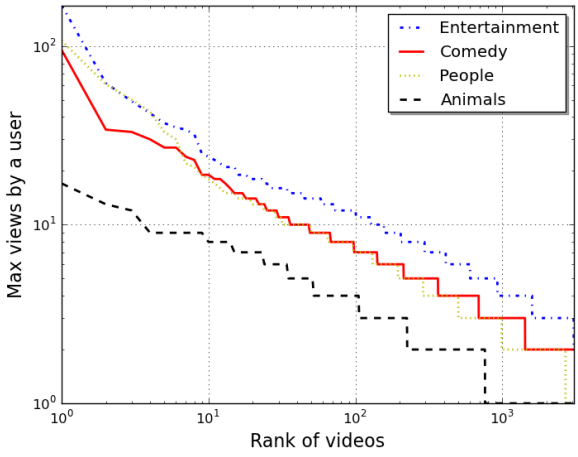


(d) Number of Unique Users vs. Rank (People)

Figure 6.8: Number of Unique Users vs. Rank (Selected Categories)



(a) Max Views by a User vs. Rank



(b) Max Views by a User Vs. Rank

Figure 6.9: Outliers in Categories

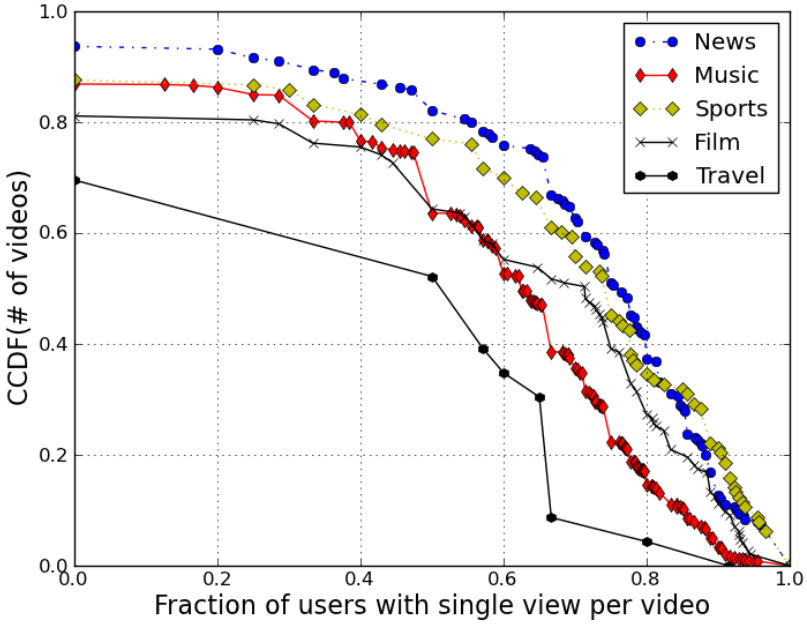
requests) with 65 inter-request times less than 2 seconds. How to deal with such anomalies in general is left as future work. These outliers, however, do not play significant role in the analysis, as the subsequent analyses deal mostly with fractions of videos/users.

This is important to verify if the outliers effect supersedes the observations that some of the categories, indeed, observe many more repeated views than others. This can be accomplished by observing the fractions of singleton views of videos in different categories. Fraction of singleton views is calculated according to equation 6.3, where users (v) is the total number of unique users that watched video v and repeat_users (v) are the users that watched video v more than once.

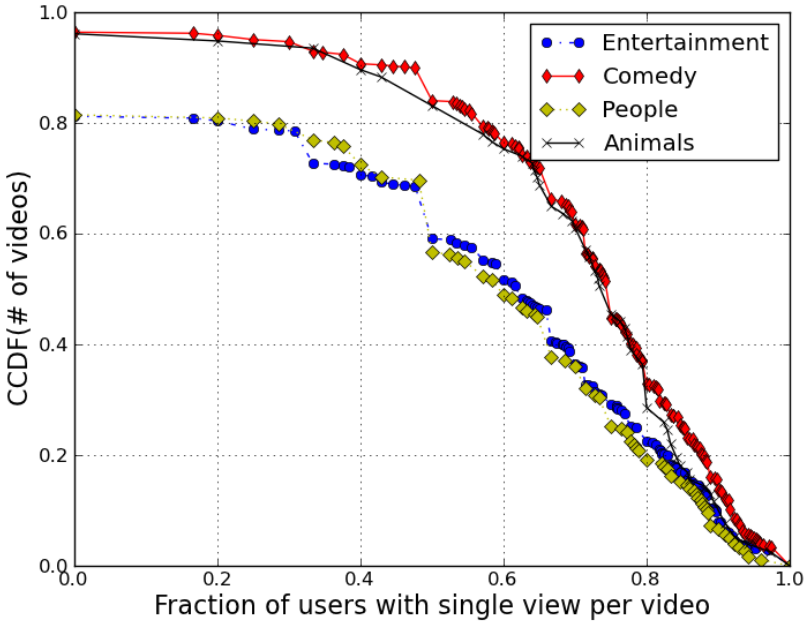
$$fs_v = \frac{users(v) - repeat\_users(v)}{users(v)} \tag{6.3}$$

Figure 6.10 shows the CCDF of number of videos against the fraction of singleton views for selected categories. This confirms the earlier findings by showing that categories that exhibit lower repeated views experience more singleton views. News videos (Sports videos to a lesser extent), for example, experience the most *fetch-at-most-once behaviour* [30] among all categories.

For approximately more than 78% of the News videos, 60% of the users watched them only once. This is much different than Music and Film videos. Travel videos, on the other hand, have very few users that watch a video only once, as compared to other categories. These statistics clearly identifies the categories that are strong candidates to be cached in a user's local cache to significantly increase the hit ratio.



(a) Singleton Views (Selected Categories)



(b) Singleton Views (Selected Categories)

Figure 6.10: Fraction of Singleton Views in Different Categories

6.5 Similar Interests among YouTube Users

This section explores the similarities among YouTube users by observing the request patterns in the local dataset (trace T5, UMass dataset). This dataset contains only 16,000 users' watched video lists within a measurement period of 14 days. The number of users captured in the dataset is not even comparable to the size of the YouTube's actual population. A dataset with a significantly large number of users' information could be used in future to offer more reliable results. Similarities among YouTube users are evaluated by observing if the users form different communities according to their interests on common videos and if the users' network exhibits properties of a small-world network.

6.5.1 Communities in YouTube Users

The K-Means algorithm has been widely used for finding clusters/communities in different kinds of contact graphs/networks. The K-Means algorithm is only applicable when the appropriate number of communities is known in advance, which is infeasible when the characteristics of the graph are unknown. Moreover, forcing K-Means to select a predefined number of communities is impractical for this analysis. K-Means divides the YouTube users into the specified number of assigned communities, even if the users do not form any community of interests. The K-SC algorithm, used in the global analysis, can not be used in this analysis for the same reason. The K-Means algorithm, in general, includes users as a part of a community even if they have no/extremely low similarity with others in watching YouTube videos. Such users should be treated as isolated communities to show their significantly different interests from others.

The Louvain algorithm [8] is selected to find clusters of users characterized by similar interests—users with similar set of watched videos. A predefined number of communities is not necessary for this algorithm, as the best number of communities is determined by improving the modularity of the network; modularity is the total weight of the links inside a community compared to the weights among the communities. This algorithm works in two steps. During the first step, each user is considered as an isolated community. Then in each iteration, two of the communities are merged together only if the modularity of the whole graph is improved. The second step considers each community as a node, and then the first step is followed again, which continues until an optimal measurement of modularity is reached. This number of node reductions in each step makes Louvain a fast and simple algorithm. The Louvain algorithm confirms that no communities will be formed with more than one members if there is no significant common interests among the members.

The second problem in identifying communities is the very short measurement period of the dataset. It is infeasible to identify a user's interest with a very few number of watched videos. The difficulty to understand users' interests with a very short measurement period in YouTube is shown in Table 6.4. Similarity of interests between two users is calculated using equation 6.4, where set (U_i) and set (U_j) are the set of videos watched by user U_i and U_j respectively. These common interests among all users are calculated in two different times, week one and week two, and then the correlation is calculated between these two snapshots. A threshold

value of 10 includes only those users that watched at least 10 unique videos in week one as well as in week two. This is to minimize the effect of different rates of activity by a single user in the two considered periods. The correlation coefficient indicates similar interests among users, especially when more active users are considered. This intensifies the confidence of having a stronger relationship between users for a substantially longer measurement period.

Table 6.4: Correlation Coefficient of Users Interest in Two Different Snapshots

≥1 views	0.004
≥10 views	0.09
≥20 views	0.14
≥30 views	0.20
≥40 views	0.24
≥50 views	0.27
≥100 views	0.48
≥200 views	0.56

Based on the observations, only the top 1000 users, ranked by the number of unique watched videos, are considered to find communities among the users. Figure 6.11 shows that the top 1000 users form only 12 communities, indicating significantly similar interests among the users. In fact, three of the communities contain only one member, as these members have no or extremely few common videos that were watched by other users in the network. The first two communities have approximately 300 members in each, showing very similar video interests among the users.

$$link(U_i, U_j) = \frac{set(U_i) \cap set(U_j)}{set(U_i) \cup set(U_j)} \tag{6.4}$$

6.5.2 Small-world Network among YouTube Users

Small-world network phenomenon is considered as one the most interesting characteristics in social networks. This phenomenon was observed in different kinds of networks: URL links in the Web, related video lists in YouTube, etc. [16]. The concept of a small-world network suggests that people are connected to each other by short path of associations. More formally, if a network is neither completely random nor completely regular, but shows an intersection of both of the characteristics, is known as small-world network. The clustering coefficient of a small-world network must be significantly larger than a similar random graph (same number of nodes and average node degree), whereas the average path length of the two graphs are comparable [37]. The clustering coefficient of a node is given by the proportion of edges between the neighbors of the given node divided by the number of all possible edges between them. The average of all nodes' coefficients is called the clustering coefficient of the whole network. Similarly, the characteristic path length of a node is

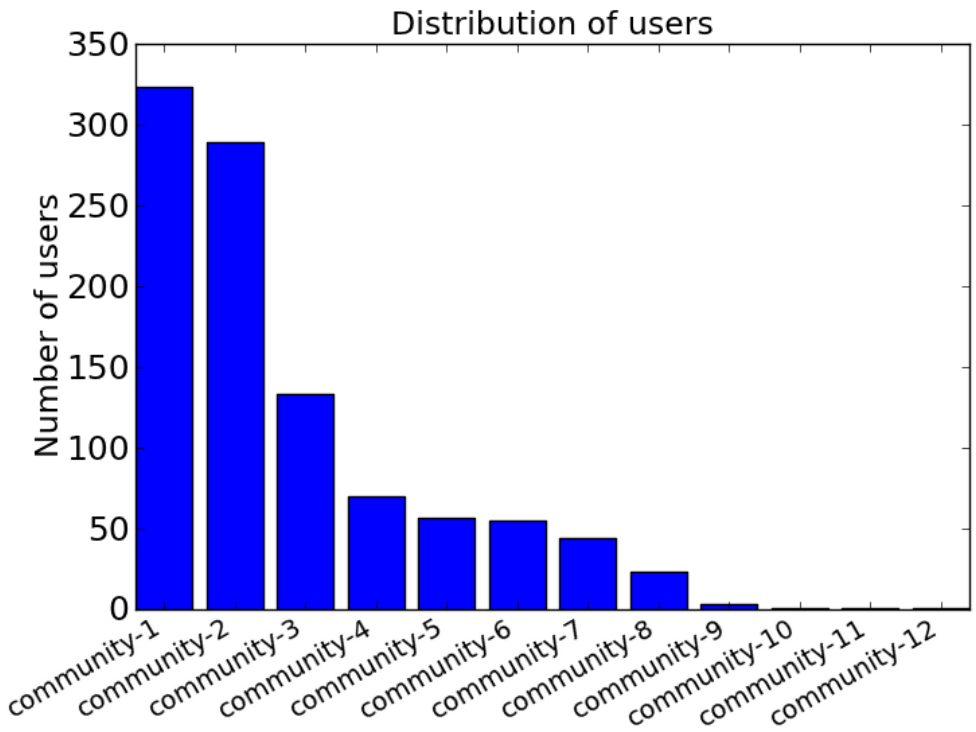


Figure 6.11: Number of Users in Different Communities

the average of all the lengths of the shortest paths to all other nodes. The average of the characteristic path lengths of all nodes in the graph is the characteristic path length of the graph.

In order to compare these two parameters with random graph, the network of YouTube users is constructed first. Two users are considered connected to each other only if they have at least 1% common videos in their watch lists, calculated by equation 6.4. The graph is not strongly connected when a more restricted threshold value, e.g., 10%, is used. The characteristic path length can not be calculated for a graph which is not strongly connected. The threshold value considered here, however, is practical considering the actual size of the YouTube population; it would not be surprising to have a strongly connected graph even with a threshold value of 10%, when all the users information is available. A random graph is generated with the same number of users/nodes and average node degree as the actual users' graph in the dataset.

The clustering coefficient of the users' graph is 0.422, which is significantly larger than the random graph's clustering coefficient (0.15). On the other hand, the characteristic path lengths are 1.88 and 1.84 for actual users' network and random graph respectively. The path lengths of these two graphs are very similar although the clustering coefficients are not, indicating that YouTube users interests' graph exhibits the properties of small-world network. This could be verified in future work, with a significantly larger dataset of YouTube users by capturing their watch lists for a substantially longer measurement period.

6.6 Summary

This chapter concentrated on two of the important issues on YouTube users: patterns of users' interactions with YouTube videos and similarities in video selections among the users. In spite of the limitation of the dataset (very short measurement period), it is clear that better cache replacement policies in P2P can be designed by considering video type, rather than using the simple LRU policy. Among the very popular categories, Music videos, on average, experience much more repeated views than News. The popularity of a News video is primarily governed by the size of its users, whereas the popularity of a Music video is increased by a group of users who watch the same video more than once. The *fetch-at-most-once* behaviour is dominant in News videos compared to Music. This explains the observation of less number of extremely popular News videos than Music in global analysis. Similar conclusions can be drawn for Entertainment and Comedy videos; Entertainment videos were found to be more popular than Comedy in case of global analysis as well as in previous work [27], which might be the result of more repeated views in Entertainment than Comedy (Figure 6.7). Travel videos, on the other hand, show very interesting characteristics. Although very few users watch this type of videos in YouTube, most of these users watch the same videos multiple times. These observations encourage a new hybrid caching algorithm for P2P video distribution. Each peer, instead of having same cache replacement policies for all the videos, will have two different cache spaces. One of the cache segments will be dedicated to the categories that observe significant repeated views. This ensures the availability of a video when the video is requested multiple times by the same user. The other segment is to store the videos that exhibit the *fetch-at-most-once* behaviour. This is to reduce the bandwidth consumption and latency for other users interested on the same videos.

The observation of distinct communities among YouTube users can be helpful for video recommendation policies. A video which is watched/liked by a member of a community, can be recommended to other members in the same community, as they have similar patterns of interests. Besides all other recommendation policies (e.g., related video list), community information can be used to substantially increase the popularity of a YouTube video.

Chapter 7

Conclusion and Future work

The enormous freedom in uploading and watching videos for free has made YouTube one of the most accessed and popular sites in todays' Internet. In YouTube, contributions come mostly from amateurs, making it difficult to approximate the future popularity of a particular video. Accurate prediction of future popularity is urgent for YouTube and similar sites, as view distributions among videos are very unevenly distributed. The performance of YouTube video delivery system has been reported inefficient, especially for resource constrained environments. Moreover, the bandwidth consumption/cost of YouTube is significantly higher than any other Internet sites. These two issues depict the challenges in understanding the popularity characteristics of YouTube videos. In spite of the significant amount of research that have been conducted to unfold this characteristics of content popularity patterns, no study has deeply investigated if there is a relationship between the category of a content and its popularity curve. In this thesis, it is shown that content category can be used towards a better understanding of popularity characteristics and designing a more accurate workload generator for YouTube. Moreover, patterns of users interaction can be identified more precisely when the category of an object is considered.

7.1 Thesis Contributions

This thesis can be summarized as follows.

Many of the notable earlier work on understanding the characteristics of on-line content growth patterns are discussed, emphasizing mainly on YouTube. Different techniques have been employed towards this understanding of YouTube video popularity characteristics. The importance of the peak time of video popularity has been reported in most of the earlier studies. Identifying the trends of peak time is essential for allocating appropriate resources for a specific content. The search and related video features are the two most contributing referrers for the overall popularity of YouTube, which is why social videos usually have very short active life span; videos with most of their views from different social sites are defined as social videos. Although future popularity can be predicted very accurately for significantly older videos, this is not common for most of the young videos. Content duplication is reported as one of the important factor in distorting the popularity of a video. A video's popularity, however, largely depends on its uploader's previous history, such as, number of subscribers, view counts by the previously uploaded videos, etc. The popularity of most

of the YouTube videos is mainly governed by its local region. YouTube videos become popular in their local regions first, before becoming globally popular.

The unique method of data collection for each category is another major contribution of this thesis. The *Most Recent* standard feed, used for the video_id collection, has been found to provide an unbiased set of YouTube videos. This unbiased dataset, along with the substantially long measurement period, intensifies the confidence of having an accurate characterization of YouTube video popularity. Moreover, significant number of videos information have been collected for a period of 5 months to observe the current uploading trend of YouTube.

A category-based study on the characteristics of YouTube videos, based on the collected sample, is provided. News and Sports are the two categories that reach peak popularity faster than any other YouTube categories. Music and Film videos, on the contrary to News and Sports videos, reach peak popularity much slowly. This peak distribution affects popularity dynamics of YouTube videos. Although Music is the most popular category in YouTube, News and Sports videos experience much more popularity than Music for the first couple of days since uploading. Music videos, however, retain popularity for a long period of time, unlike to News and Sports. Peak distribution has been found vital for some of the categories—the categories that reach peak popularity faster. Time to peak is proportional to the active life span of a video; in general, a video that reach peak popularity slowly, does not become unpopular suddenly. In each category, the view distribution is found unevenly distributed. Most of the views in a category come from a very small fraction of videos, as each category contains lots of unpopular videos. The number of very unpopular videos, however, varies according to the category. People videos are the most frequently uploaded videos in YouTube, which was completely different in 2007. People videos, however, still fail to attract the YouTube users as compared to other popular categories. Like the observations made in the earlier work, view distributions do not fit with Zipf's law for any of the categories. Instead, a Weibull distribution is needed to describe the tail sections of the categories.

A method for creating a category-specific workload generator for YouTube and similar sites is presented and evaluated. Performance evaluation is also done for two of the known methods for designing workload generator of UGC systems. These methods fail to model category-specific workload generator, in spite of their success in generating synthetic data when the category is not taken into consideration. The K-SC algorithm exhibits acceptable performance for all of the evaluated categories, which implies that this technique can be used for all of the YouTube categories so that a complete workload generator can be designed. The strong correlation coefficient of video popularity for some of the categories between different snapshots, however, indicates the success in improving cache hit ratio. This is crucial to ameliorate the users' experience in watching YouTube videos as well as to reduce bandwidth consumption.

Patterns of users' interaction with YouTube categories are analyzed. Among the popular YouTube categories, News videos exhibit the *fetch-at-most-once* behaviour more than others. This indicates the potential success of employing P2P video distribution policy rather than local caching only. For category like Travel,

local caching can be very effective in reducing bandwidth consumption, as most of the Travel videos attract the same users multiple times in spite of the less number of users. Music and Entertainment videos, on the other hand, not only experience significant number of repeated views, but also the numbers of users attracted to these categories are significantly larger. The small-world phenomenon along with distinct number of communities observed from the users' interest graph suggest a new video recommendation technique for YouTube and other user-generated video distribution sites.

This thesis suggests that category-based characterization provides better understanding of the actual popularity dynamics of user-generated videos, which is important both for global and local caching of YouTube and similar UGC systems. In addition, a category-specific workload generator facilitates more accurate simulation platform for the actual evaluation of different caching policies.

7.2 Future Work

The characterization of video access patterns and accuracy of the proposed model for workload generator can be improved in future work. Some of the possible improvements are discussed as follows.

The number of videos used in this thesis is small compared to the actual YouTube population. The size of the dataset can be enlarged by applying the same procedure (described in chapter 3) for a significantly longer period. For example, collecting data for a complete week will show if there is any significant difference between video popularity of videos that were uploaded on weekdays and weekend.

Video duration was not considered while designing the workload generator. The earlier work, without considering video category, observed no correlation between video popularity and video length. It would be a worthwhile observation to make if different results can be found after taking video type into consideration. For the evaluation of different caching policies, it would be interesting to find if a very popular video with significantly longer video length should be cached rather than storing more videos with shorter durations and moderate popularity. This is important especially for resource constrained environments, when a choice has to be made for cache replacement.

Many of the YouTube videos are deleted for copyright violation. The proposed workload generator in this thesis did not take this issue into consideration. One of the possible improvements in this area would be to understand the deletion patterns of YouTube videos, which could be attached as a feature in a complete workload generator.

In the proposed workload generator, a user's id was not attached with a video request; YouTube does not share this information publicly. This improvement can be made by incorporating the findings of both the local and global analyses. This requires, however, datasets from different local networks to observe if the access patterns are same across all the networks.

One of the most challenging but very useful research would be to design a hybrid caching policy for user-generated video sites. This is because of the different repeated viewing patterns of different categories.

References

[1] A. Abhari and M. Soraya. Workload Generation for YouTube. *Multimedia Tools and Applications*, 46(1):91–118, January 2010.

[2] Lada A. Adamic. Zipf, Power-laws, and Pareto-a Ranking Tutorial. *Xerox Palo Alto Research Center, Palo Alto, CA (http://ginger.hpl.hp.com/shl/papers/ranking/ranking.html)*, 2000.

[3] V.K. Adhikari, S. Jain, Yingying Chen, and Zhi-Li Zhang. Vivisecting YouTube: An Active Measurement Study. In *31st Annual IEEE International Conference on Computer Communications (INFOCOM 2012)*, pages 2521–2525, Orlando, Florida, March 2012.

[4] C.C. Aggarwal, J.L. Wolf, and P.S. Yu. The Maximum Factor Queue Length Batching Scheme for Video-on-Demand Systems. *IEEE Transactions on Computers*, 50(2):97–110, 2001.

[5] D. Applegate, A. Archer, V. Gopalakrishnan, S. Lee, and K. K. Ramakrishnan. Optimal Content Placement for a Large-scale VoD System. In *6th ACM International Conference on emerging Networking EXperiments and Technologies (Co-NEXT 2010)*, pages 4:1–4:12, Philadelphia, PA, November 2010.

[6] L. Atzori, F.G.B. De Natale, M. Di Gregorio, and D.D. Giusto. Multimedia Information Broadcasting using Digital TV Channels. *IEEE Transactions on Broadcasting*, 43(4):383–392, 1997.

[7] J. Biel and D. Gatica-Perez. Wearing a YouTube Hat: Directors, Comedians, Gurus, and User Aggregated Behavior. In *ACM Multimedia (MM 2009)*, pages 833–836, Beijing, China, October 2009.

[8] V. D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, October 2008.

[9] Y. Borghol, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti. The Untold Story of the Clones: Content-agnostic Factors that Impact YouTube Video Popularity. In *18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD 2012)*, pages 1186–1194, Beijing, China, August 2012.

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

[11] A. Brodersen, S. Scellato, and M. Wattenhofer. YouTube Around the World: Geographic Popularity of Videos. In *21st World Wide Web conference (WWW 2012)*, pages 241–250, Lyon, France, April 2012.

[12] T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer. Catching a Viral Video. *Journal of Intelligent Information Systems*, 40(2):241–259, 2013.

[13] N. Carlsson, G. Dán, A. Mahanti, and M. Arlitt. A Longitudinal Characterization of Local and Global Bittorrent Workload Dynamics. In *13th Passive and Active Measurement Conference (PAM 2012)*, pages 252–262, Vienna, Austria, March 2012.

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

[15] G. Chatzopoulou, C. Sheng, and M. Faloutsos. A First Step Towards Understanding Popularity in YouTube. In *29th IEEE Conference on Computer Communications (INFOCOM 2010)*, pages 1–6, San Diego, CA, March 2010.

[16] X. Cheng, C. Dale, and J. Liu. Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study. Technical report, Cornell University, arXiv e-prints, July 2007.

[17] X. Cheng and L. Jiangchuan. Exploring Interest Correlation for Peer-to-Peer Socialized Video Sharing. *ACM Transactions on Multimedia Computing, Communications and Applications*, 8(1):5:1–5:20, February 2012.

[18] K. K. W. Chu and M. H. Wong. Fast Time-series Searching with Scaling and Shifting. In *18th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 1999)*, pages 237–248, Philadelphia, PA, May 1999.

[19] R. Crane and D. Sornette. Robust Dynamic Classes Revealed by Measuring the Response Function of a Social System. *National Academy of Sciences*, 105(41), 2008.

[20] A. Dan, D. Sitaram, and P. Shahabuddin. Scheduling Policies for an On-Demand Video Server with Batching. In *ACM Multimedia (MM 1994)*, pages 15–23, San Francisco, CA, October 1994.

[21] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, and D. Sampath. The YouTube Video Recommendation System. In *4th ACM Conference on Recommender Systems (RecSys 2010)*, pages 293–296, Barcelona, Spain, September 2010.

[22] Y. Ding, Y. Du, Y. Hu, Z. Liu, L. Wang, K. Ross, and A. Ghose. Broadcast Yourself: Understanding YouTube Uploaders. In *ACM Internet Measurement Conference (IMC 2011)*, pages 361–370, Berlin, Germany, November 2011.

[23] F. Duarte, F. Benevenuto, V. Almeida, and J. Almeida. Geographical Characterization of YouTube: a Latin American View. In *Latin American Web Conference 2007*, pages 13–21, Washington, DC, October 2007.

[24] S. M. Farhad and M.M. Akbar. Multicast Video-on-Demand Service in an Enterprise Network with Client-Assisted Patching. In *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PacRim 2007)*, pages 456–459, Victoria, B.C., Canada, August 2007.

[25] F. Figueiredo, F. Benevenuto, and J. Almeida. The Tube over Time: Characterizing Popularity Growth of Youtube Videos. In *4th ACM International Conference on Web Search and Web Data Mining (WSDM 2011)*, pages 745–754, Hong Kong, China, February 2011.

[26] A. Gember, A. Anand, and A. Akella. A Comparative Study of Handheld and Non-handheld Traffic in Campus Wi-Fi Networks. In *12th Passive and Active Measurement conference (PAM 2011)*, pages 173–183, Atlanta, GA, March 2011.

[27] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube Traffic Characterization: A View From the Edge. In *ACM Internet Measurement Conference (IMC 2007)*, pages 15–28, San Diego, CA, October 2007.

[28] M. Gonzlez-Aparicio, R. Garca, J. L. Brugos, X. G. Paeda, D. Melendi, and S. Cabrero. Video Popularity Characterization Centered on News-on-Demand. *International Journal of Multimedia and Its Applications*, 4(5):19–38, October 2012.

[29] J. Guebert, D.J. Makaroff, and K.M. Patel. Request Generation for a Peer-based PVR. In *20th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV 2010)*, pages 99–104, Amsterdam, The Netherlands, June 2010.

[30] K.P. Gummadi, R. J. Dunn, S. Saroiu, S.D. Gribble, H. M. Levy, and J. Zahorjan. Measurement, Modeling, and Analysis of a Peer-to-Peer File-sharing Workload. In *19th ACM Symposium on Operating Systems Principles (SOSP 2003)*, pages 314–329, Bolton Landing, NY, October 2003.

[31] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang. Measurements, Analysis, and Modeling of BitTorrent-like Systems. In *ACM Internet Measurement Conference (IMC 2005)*, pages 4–4, Berkeley, CA, October 2005.

[32] K.A. Hua, Y. Cai, and S. Sheu. Patching: a Multicast Technique for True Video-on-Demand Services. In *ACM Multimedia (MM 1998)*, pages 191–200, Bristol, UK, September 1998.

[33] F.T. Johnsen, T. Hafsoe, C. Griwodz, and P. Halvorsen. Workload Characterization for News-on-Demand Streaming Services. In *26th International Performance Computing and Communications Conference (IPCCC 2007)*, pages 314–323, New Orleans, LA, April 2007.

[34] S. Khemmarat, R. Zhou, L. Gao, and Zink. M. Watching User Generated Videos with Prefetching. In *ACM Multimedia Systems conference (MMSYS 2011)*, pages 187–198, San Jose, CA, February 2011.

[35] C. Labovitz, S. Iekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian. Internet Inter-Domain Traffic. In *ACM SIGCOMM 2010*, pages 75–86, New Delhi, India, August 2010.

[36] K. Lerman and T. Hogg. Using Stochastic Models to Describe and Predict Social Dynamics of Web Users. *ACM Transactions on Intelligent Systems Technologies*, 3(4):62:1–62:33, September 2012.

[37] G. Liu, M. Hu, B. Fang, and H. Zhang. Measurement and Modeling of Large-Scale Peer-to-Peer Storage System. In *Lecture Notes in Computer Science*, volume 3252, pages 270–277. 2004.

[38] G. Maier, F. Schneider, and A. Feldmann. A First Look at Mobile Hand-held Device Traffic. In *11th Passive and Active Measurement conference (PAM 2010)*, pages 161–170, Zurich, Switzerland, April 2010.

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

[40] J. S. Pedro, S. Siersdorfer, and M. Sanderson. Content Redundancy in YouTube and Its Application to Video Tagging. *ACM Transactions on Information and System Security*, 29(3):13:1–13:31, July 2011.

[41] Y. Peng, C.and Tan, N. Xiong, L. T. Yang, J. H. Park, and S. Kim. Adaptive Video-on-Demand Broadcasting in Ubiquitous Computing Environment. *Personal Ubiquitous Computing*, 13(7):479–488, October 2009.

[42] H. Pinto, J. M. Almeida, and M.A. Gonçalves. Using Early View Patterns to Predict the Popularity of YouTube Videos. In *6th ACM International Conference on Web Search and Data Mining (WSDM 2013)*, pages 365–374, Rome, Italy, February 2013.

[43] B. Qudah and N. J. Sarhan. Towards Scalable Delivery of Video Streams to Heterogeneous Receivers. In *ACM Multimedia (MM 2006)*, pages 347–356, Santa Barbara, CA, October 2006.

[44] S. Saroiu, K.P. Gummadi, R.J. Dunn, S. D. Gribble, and H. M. Levy. An Analysis of Internet Content Delivery Systems. *SIGOPS Operating Systems Review*, 36(SI):315–327, December 2002.

[45] H. Shachnai and P.S. Yu. Exploring Wait Tolerance in Effective Batching for Video-on-Demand Scheduling. *Multimedia Systems*, 6(6):382–394, November 1998.

[46] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. San Pedro. How Useful are Your Comments?: Analyzing and Predicting YouTube Comments and Comment Ratings. In *19th International Conference on World Wide Web (WWW 2010)*, pages 891–900, Raleigh, NC, April 2010.

[47] P. M. Smithson, J. T. Slader, D. F. Smith, and M. Tomlinson. The Development of an Operational Satellite Internet Service Provision. In *IEEE Global Telecommunications Conference (GLOBECOM 1997)*, pages 1147–1151, Phoenix, AZ, November 1997.

[48] G. Szabo and B. Huberman. Predicting the Popularity of Online Content. *Communications of the ACM*, 53(8):80–88, August 2010.

[49] F. Thouin and M. Coates. Video-on-Demand Networks: Design Approaches and Future Challenges. *IEEE Network*, 21(2):42–48, 2007.

[50] A. Vinay, Abhinav Prakash, D. S. Kiran Kumar, K. Nagabhushan, and T. N. Anitha. A Novel and Optimal Video Replication Technique for Video-on-Demand Systems. In *International Conference & Workshop on Emerging Trends in Technology (ICWET 2011)*, pages 344–350, Mumbai, Maharashtra, India, February 2011.

[51] B. Wang, S. Sen, M. Adler, and D. Towsley. Optimal Proxy Cache Allocation for Efficient Streaming Media Distribution. In *21st Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2002)*, pages 1726–1735, New York, NY, June 2002.

[52] J. Yang and J. Leskovec. Patterns of Temporal Variation in Online Media. In *4th ACM International Conference on Web Search and Data Mining (WSDM 2011)*, pages 177–186, Hong Kong, China, February 2011.

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

[54] X. Zhang and H. Hassanein. Video-on-Demand Streaming on the Internet -a Survey. In *25th Biennial Symposium on Communications (QBSC 2010)*, pages 88–91, Kingston, Canada, May 2010.

[55] R. Zhou, S. Khemmarat, and L. Gao. The Impact of YouTube Recommendation System on Video Views. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC 2010)*, pages 404–410, Melbourne, Australia, 2010.

[56] M. Zink, K. Suh, Y. Gu, and J. Kurose. Watch Global, Cache Local: YouTube Network Traffic at a Campus Network - Measurements and Implications. In *15th Annual Multimedia Computing and Networking Conference (MMCN 2008)*, San Jose, CA, January 2008.

[57] M. Zink, K. Suh, Y. Gu, and J. Kurose. Characteristics of YouTube Network Traffic at a Campus Network - Measurements, Models, and Implications. *Computer Networks*, 53(4):501–514, March 2009.