Popularity Characterization and Modelling for User-generated Videos

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, Saskatchewan, Canada

By
M. Aminul Islam

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

User-generated content systems such as YouTube have become highly popular. It is difficult to understand and predict content popularity in such systems. Characterizing and modelling content popularity can provide deeper insights into system design trade-offs and enable prediction of system behaviour in advance.

Borghol *et al.* collected two datasets of YouTube video weekly view counts over eight months in 2008/09, namely a "recently-uploaded" dataset and a "keyword-search" dataset, and analyzed the popularity characteristics of the videos in the recently-uploaded dataset including the video popularity evolution over time. Based on the observed characteristics, they developed a model that can generate synthetic video weekly view counts whose characteristics with respect to video popularity evolution match those observed in the recently-uploaded dataset.

For this thesis, new weekly view count data was collected over two months in 2011 for the videos in the recently-uploaded and keyword-search datasets of Borghol *et al.* This data was used to evaluate the accuracy of the Borghol *et al.* model when used to generate synthetic view counts for a much longer time period than the eight month period previously considered. Although the model yielded distributions of total (lifetime) video view counts that match the empirical distributions, significant differences between the model and empirical data were observed. These differences appear to arise because of particular popularity characteristics that change over time rather than being week-invariant as assumed in the model.

This thesis also characterizes how video popularity evolves beyond the eight month period considered by Borghol *et al.*, and studies the characteristics of the keyword-search dataset with respect to content popularity, popularity evolution, and sampling biases. Finally, the thesis studies the popularity characteristics of the videos in the recently-uploaded and keyword-search datasets for which additional view count data could not be collected, owing to the removal of these videos from YouTube.

Acknowledgements

I would like to express my sincere appreciation to Dr. Derek Eager for all his invaluable guidance, support and encouragement as my supervisor. His endless energy and enthusiasm in research had motivated me and provided me the rhythm of doing research. In addition, he was always accessible, motivative and willing to help me in my research even in his tight schedule. As a result, my research life became smooth and rewarding for me. It has been my great pleasure and honour to have worked with him. I would also like to thank Dr. Niklas Carlsson and Dr. Anirban Mahanti for their encouraging advice, guideline and feedback on my research output. I would also like to thank the members on my supervisory committee, Dr. Dwight Makaroff and Dr. Nadeem Jamali, as well as my external examiner Dr. Khan A. Wahid, for their helpful comments and constructive suggestions.

I would also like to express my gratefulness to all the faculty, staff and graduate students in the Department of Computer Science, for their caring and support during my study at the University of Saskatchewan.

Finally, I would like to thank my family and friends for their continuous encouragement and help, especially my parents who sacrificing the companionship of their only son for this research work. I am grateful for their continuing understanding and countless support throughout my entire life.

Contents

List of Tables

List of Figures

List of Abbreviations

VoD	Video on Demand
UGC	User-generated Content
HBO	Home Box Office
DVR	Digital Video Recorder
OECD	Organization of Economic Co-operation and Development
P2P	Peer-to-Peer
IPTV	Internet Protocol Television
CDN	Content Distribution Network
OSN	Online Social Network
CDF	Cumulative Distribution Function
CCDF	Complementary Cumulative Distribution Function
MLE	Maximum Likelihood Estimation

Chapter 1

Introduction

In the past few years, video sharing services have become very popular. Among these services, the most widely used is YouTube. The quality and the popularity of YouTube videos varies vastly. Traditional video on demand (VoD) systems generally offer professionally created content and the quality of those videos is controlled by content providers. Content popularity in such traditional VoD systems is more predictable and is expected to generally be very high, while it is highly unpredictable for user-generated content (UGC) [8][17]. Sections 1.1 and 1.2 present the basic ideas of video on demand systems and user-generated content systems respectively. Popularity characterization and popularity modelling are introduced in Sections 1.3 and 1.4. The motivation for the thesis research is presented in Section 1.5 and a summary of the thesis contributions is given in Section 1.6. Finally, Section 1.7 presents the organization of the remainder of the thesis.

1.1 Traditional Video on Demand (VoD)

Video on demand (VoD) systems allow users to watch multimedia content at times of their own choosing. In traditional VoD systems, videos are supplied to the content providers by a limited number of media producers or licensed broadcast companies. These videos are provided in, or converted to, a common format that enables clients to play them without installing any additional codec. The content providers can deliver the videos in two ways: (1) videos can be sent to specific requesting clients via unicast transmission or (2) videos can be multicast to multiple customers simultaneously, those customers tuning in to a particular IPTV channel for example [39].

Video on demand service was first commercially launched in Hong Kong in 1990, but the service was not commercially successful.[1] In the mid-1990s, a number of video on demand trials were launched in the United States and the United Kingdom. Video on demand systems were first introduced in the United States in an "interactive TV" trial launched by *Time Warner* in Orlando, Florida in December 1994 [36]. At the beginning, the system faced some technical challenges to integrate system software and transmission of video and audio synchronization. Another video on demand trial was launched by *Bell*

[1] `http://www.wisegeek.com/what-is-video-on-demand.htm`; accessed 10-December 2012

Atlantic Corporation in Fairfax County, Virginia in May 1995. Other early trials in the United States were launched by *TeleCommunications Inc.* and *Microsoft* in Redmond, Washington in March 1995, by *US West* in Omaha, Nebraska in August 1995, and by *Pacific Telesis* in San Jose, California in December 1995.

Large-scale successful commercial deployments of video on demand did not occur until several years later. For example, *Home Box Office (HBO)* launched the "HBO on demand" video on demand service in July 2001.[2] HBO on demand service is now available on many cable and satellite providers around the globe. Locally, in 2003, SaskTel introduced a new technology, Max Front Row that provides video on demand service on television and computers via high speed Internet. SaskTel claims that they were the first telecommunications company in North America to offer a video on demand service.[3]

Videos in a VoD system are typically stored in a number of video servers in different locations. These servers allow clients to play videos at any time upon their request. Requested videos are delivered through a high speed communication network. Therefore to achieve the video transmission smoothly, the system should have sufficient network bandwidth. Ma *et al.* [29] observed four major requirements for a traditional VoD system: (1) a VoD system should be able to provide service for long periods of time, because in a traditional VoD system the video length is typically at least 30 to 60 minutes which is substantially longer than that of user-generated content videos [9], (2) a VoD system should have high bandwidth to forward content to the clients, for example 1.5 Mbps for an MPEG-1 stream and 25 to 35 Mbps for Blu-ray videos, (3) a VoD system should support DVR or "trick play" functionality such as skipping forward or backward, pausing, and skipping advertisements, and (4) a VoD system should provide high-quality videos with low transmission latency.

1.2 User-generated Content (UGC)

User-generated content (UGC) refers to content that is created by ordinary people or end-users rather than media producers. UGC is also called consumer-generated media.[4] UGC can include all kinds of content including photos, music, videos, text, design, graphics, messages, information, blogs, research, photography and other digital resources. A study on UGC in 2006 sponsored by the Organization of Economic Co-operation and Development (OECD) stated that a distinguishing characteristic of UGC systems is that content should have a certain amount of creative effort by amateur people rather than just being directly uploaded from other sources.[5]

The advent of UGC systems revolutionized the online content delivery market for both content producers and consumers. UGC systems allow clients content upload opportunity with great variability and flexibility. Also, UGC systems allow consumers to upload different versions of the same content (clones)

[2] `http://en.wikipedia.org/wiki/HBO`; accessed 10-December 2012
[3] `http://www.sasktel.com/about-us/company-information/history/index.html`; accessed 10-December 2012
[4] `http://www.digital-marketing-course.co.nz/resources.php?Glossary-8/`; accessed 10-December 2012
[5] `http://www.oecd.org/internet/interneteconomy/38393115.pdf/`; accessed 10-December 2012

without any constraints. Users contribute content to UGC systems mainly for four reasons: prestige, self-expression, recognition and connection with people.[6] Although content production in UGC systems is usually non-profitable work, still large numbers of people every day contribute content to UGC systems.

UGC systems are expanding in such a way that different groups of people are attached with different UGC systems for their own purposes. For example, socially active people engage with Facebook or Flickr to connect with each other by sharing their status, comments, pictures, videos and also by chatting. In e-commerce sites such as Amazon and eBay, reviews and ratings by the users play a major role in the choice of products by the customers. Blog news has become very competitive to conventional news sources such as CNN or BBC or Yahoo news. For example, at the time of the Pacific tsunami in 2004, people that survived from the tsunami shared their experience through online social media or blogs that provided more detailed information than provided by BBC or CNN.[7]

In a December 2006 report on UGC from the Organization of Economic Co-operation and Development (OECD), it was stated that the involvement of a huge number of people in UGC systems is driven by four major aspects, namely technological, social, economical and legal aspects.[8] The upload and download of some types of UGC content requires a high speed Internet connection which is also associated with multiple hardware and electronics goods such as hub, switch, digital camera, microphone, headphone, etc. Therefore, the technological aspect plays a major role as a driver of UGC systems. Social media that engages large numbers of people of different ages and interests is also significantly responsible for the distribution of UGC content. Because of the huge impact of UGC content with respect to its technological and social aspects, various new business ideas and platforms are created and launched by different business organizations. Those organizations create new fields of advertisement and financial marketplaces. Therefore, all aspects are linked with each other to drive the UGC systems as a whole. Apart from other aspects, the legal aspect is concerned with taking care of the copyright issues by ensuring the ownership of the content.

Online video sharing services are important examples of UGC systems in which users upload video clips that they have created and watch videos uploaded by others. Users are free to upload and download videos of their choosing and therefore there is a high degree of availability with respect to videos of differing topics, qualities, and popularities.

YouTube, launched in 2005, is the most popular UGC system. People can watch videos without installing any YouTube-specific client program and without creating any client account. However, only "subscribed users" can upload their videos. Due to frequent upgrades, a simple interface and its user-friendly nature, YouTube videos received more than two billion views per day as of May, 2010.[9] As of May 2011, YouTube videos received more than three billion views per day, and over four billion views per day as of

[6] http://ict4peace.wordpress.com/2007/10/27/web-20-wikis-and-social-networking-oecd-study-on-user-generated-content/; accessed 10-December 2012

[7] http://news.nationalgeographic.com/news/2005/01/0126_050126_tv_tsunami_blogs.html/; accessed 10-December 2012

[8] http://ict4peace.wordpress.com/2007/10/27/web-20-wikis-and-social-networking-oecd-study-on-user-generated-content/; accessed 10-December 2012

[9] http://en.wikipedia.org/wiki/YouTube/; accessed 10-December 2012

January, 2012.[10] [11] Subscribed users in YouTube can "like", "comment" and "rate" videos. Subscribed users can also create their own video lists for quick access and to automatically play videos one after another.

1.3 Popularity Characterization

If not carefully designed, online video sharing services could be overwhelmed by a large number of requests from clients. The systems could experience service bottlenecks due to insufficient resources and inefficient content management. To address such situations, multiple content servers can be established at different locations and videos distributed to them. A suitable content distribution policy can also increase service quality in terms of transmission latency by serving from nearby content servers. However, the content distribution policy requires information concerning video popularities and how these popularities evolve over time. Analysis of popularity characteristics is therefore important for the avoidance of improper content management and system bottlenecks.

Considerable user interest has moved from conventional web applications to multimedia applications during the last decade [21]. UGC systems opened a new era in the history of online video services by allowing millions of producers and consumers. Web 2.0 features (favourites, comments, ratings etc.) provide more interaction and motivate people to upload more videos. Due to the usual nature of UGC systems, content popularity in such systems is relatively hard to predict. Popularity characterization is therefore an important problem in such systems, so as to determine efficient system design trade-offs for user satisfaction and maximum viability.

Clients mostly find YouTube videos using Google search or the YouTube internal searching mechanism, from a large video storage system into which approximately 72 hours of video were being uploaded every minute by the end of 2011.[12] An understanding of video popularity characteristics may be useful in development of better systems for helping clients find videos of interest to them. Similarly, lower latency and lower network load could be achieved by use of efficient caching mechanisms, development of which also requires an understanding of video popularity characteristics. It has been reported that only 30% of the videos receive 99% of the total views, and relative popularities shift over time.[13] Understanding popularity characteristics, and popularity evolution in particular, could be important for development and evaluation of caching policies as well as policies for distributing videos across different content servers.

[10] `http://www.telegraph.co.uk/technology/google/8536634/YouTube-users-uploading-two-days-of-video-every-minute.html`; accessed 10-December 2012

[11] `http://www.reuters.com/article/2012/01/23/us-google-youtube-idUSTRE80M0TS20120123`; accessed 10-December 2012

[12] `http://www.youtube.com/t/press_statistics`; accessed 10-December 2012

[13] `http://en.wikipedia.org/wiki/YouTube/`; accessed 10-December 2012

1.4 Popularity Modelling

Modelling is a technique that presents the system architecture in such a way so as to express its fundamental building blocks and complex system aspects.[14] The main idea in the design of a system model is the decomposition of system components and visualizing the interaction of components, followed by development of the most fundamental components separately. In popularity modelling for multimedia content, the request rate for each content item is considered the fundamental building block [5].

Popularity modelling can aid in understanding in-depth workload characteristics for online content sharing services. Different content sharing services such as UGC systems, VoD systems and P2P systems have different content popularity characteristics corresponding to their different service nature and resources. These distributed systems have different system design strategies and service policies, and show different popularity skewness and dynamics. A proper popularity distribution model can provide significant insight into the content interests of the system users, as well as into design trade-offs. For example, Yu *et al.* [43] present a measurement study on user behaviour and content access patterns; based on their analysis they claimed that they rectified some common misunderstandings about popularity distributions in VoD systems, enabling more accurate models of system behaviour.

Predicting the future popularity of content is a challenging task. Szabo *et al.* [41] proposed a model that attempts to estimate future popularity as a function of current popularity. The model uses logarithmically transformed view counts; with such a transformation the authors find a strong correlation between popularity early in the content lifetime and the later popularity. Content that tends to become unpopular quickly was found to have more predictable popularity than content that can remain popular for long time periods. A similar popularity prediction model was designed by Lee *et al.* [27] who attempted to predict the maximum popularity videos can achieve instead of predicting the actual popularity. However, it is not surprising that high error rates can occur in these models when predicting future behaviours.

1.5 Thesis Goals

The outstanding success of online content sharing services during the last decade has resulted in large numbers of clients and substantial interaction of these clients through web 2.0 features (rating, comment, view count etc.). The huge usage of these systems requires careful system design so as to provide quality service and continued success. Designers must know the in-depth system behaviour to ensure quality-of-service and to identify potential bottlenecks, and to best exploit strategies such as content caching. Understanding popularity characteristics can provide in-depth knowledge about system behaviour and efficient design criteria. Similarly, popularity modelling can provide deeper insights into system design trade-offs and enable

[14] `http://www.comp.lancs.ac.uk/projects/renaissance/RenaissanceWeb/project/Acrobat/System_Modelling.pdf/`; access 10-December 2012

prediction of system behaviour in advance. The main goal of this thesis is to contribute to improved characterizations and models of video popularity evolution in UGC systems such as YouTube.

Borghol *et al.* [5] collected two YouTube empirical datasets, namely "recently-uploaded" and "keyword-search", and they mainly observed the popularity characteristics of the videos in the recently-uploaded dataset. Based on the observed empirical behaviour of the videos in the recently-uploaded dataset, they developed a model that can generate synthetic datasets in which key properties of the popularity evolution dynamics match those observed in practice, for the first 8 months of video lifetime (time since uploaded). A primary goal of this work is to investigate whether this model can be applied for much longer time periods.

A second goal concerns the work by Borghol *et al.* on characterizing basic properties of popularity evolution (such as degree of churn in relative video popularity) over the first 8 months of video lifetime. Of interest is how those popularity characteristics may change, as videos age further beyond 8 months. A third goal concerns the characteristics of the "keyword-search" dataset obtained by Borghol *et al.*, and similar such datasets obtained through keyword searches. It is known that such datasets are biased towards inclusion of more popular videos, but a detailed study of such biases has not been done previously. Such a study is a topic of interest in this thesis research.

For the purpose of addressing the above goals, new view count data was collected for the videos in the recently-uploaded and keyword-search datasets of Borghol *et al.* For some videos this was not possible since they had been removed from YouTube during the intervening time period. Figueiredo *et al.* [17] claimed that videos removed from YouTube reach their peak popularity at an early age. A long period of time has passed since the Borghol *et al.* datasets were collected, therefore a large number of videos have been removed. A study of the popularity characteristics of these removed videos (for the first measurement period) is also a part of the thesis research.

1.6 Thesis Contributions

This thesis presents the popularity characteristics and popularity modelling of user-generated videos using two different datasets (recently-uploaded and keyword-search) collected from YouTube. View counts for the videos in these datasets were collected using a YouTube API call that returns a video's meta-data. Among several sampling approaches, the *recently-uploaded* dataset contains videos that were uploaded very recently to YouTube at the time of the initial data collection, and appears to contain an unbiased sampling of such videos. The *keyword-search* dataset was collected by searching for videos through keywords. Using these datasets, this thesis makes the following contributions.

- The first contribution is the separation of available and removed videos in the datasets collected by Borghol *et al.* [5] in 2008/09, and the collection of the weekly view counts for the available videos over a nine week period. Using the YouTube API, a crawler performed the available and removed

video separation task, and it was observed that 67.13% of the keyword-search videos and 55.23% of the keyword-search videos were still available. The weekly views of the available videos were collected by another crawler using the YouTube API.

- The second contribution is the analysis of video view count distributions for both datasets and both measurement periods. For the recently-uploaded videos, one surprising result is that the average weekly view count for these videos had not decreased by the time of the second measurement period, in comparison to the middle and later portions of the first measurement period. For the keyword-search videos, the analysis suggests that even though this dataset is biased towards more popular videos, it could be used in studies of popularity evolution if videos are binned according to both their current age (time since upload) in the week when their popularity is considered, and their popularity some number of weeks previous to this week. The observed popularity evolution may be representative of what one would see with randomly selected videos in the same age and popularity bins.
- Analysis of popularity dynamics and churn over long periods of time is also a significant contribution of the thesis. One of the observations is that as videos age, popularity churn decreases. One reason for popularity churn is differences in the rate at which videos attain their peak popularity. The majority of the videos in the recently-uploaded dataset (approximately three quarters) reached their peak popularity (as defined by weekly view count, and considering only those weeks belonging to the first or second measurement period) during the first six weeks since upload. Interestingly, however, there is a considerable number of videos that reached their peak popularity at much older ages, during the second measurement period.
- The most significant contribution of the thesis is the analysis of the accuracy of the Borghol *et al.* model for the longer time span covered by the beginning of the first measurement period up to the end of the second measurement period. Both the "basic" and "extended" variants of the model defined by Borghol *et al.* were implemented and evaluated. Although a good match was observed between the model and empirical data for distributions of the total (lifetime) video view count, significant differences were observed for other metrics. These differences appear to arise because of popularity characteristics that change over time rather than being week-invariant as assumed in the Borghol *et al.* model.
- Lastly, the thesis carries out an analysis of the popularity characteristics of videos that were in the datasets collected by Borghol *et al.*, but that were removed from YouTube by the time of the second measurement period. The analysis shows that the removed recently-uploaded videos had a substantially higher average weekly view count in the first measurement period than the recently-uploaded videos that were still available during the second measurement period. Such a difference is not observed for the keyword-search videos. Also, for both datasets, the removed videos tended to experience lower popularity churn (compared to the videos that were still available) during the first measurement period.

1.7 Thesis Organization

The rest of the thesis is organized as follows. Chapter 2 presents the related work concerning content popularity, first for traditional content such as web pages, and then for UGC systems. Prior work concerning modelling and predicting popularity evolution is also discussed. Chapter 3 presents the data collection methodology, including the tools, techniques and policies used in the data collection. Chapter 3 also summarizes the main characteristics of the recently-uploaded and keyword-search datasets, including the new data collected as part of this thesis research, and describes the challenges and limitations of the data collection. Chapter 4 contains a popularity characteristics analysis of the datasets. Chapter 5 describes the popularity characteristics of synthetic data generated using the popularity model developed by Borghol *et al.* [5], and compares these with the empirical data. Chapter 6 contains a popularity characterization of the "removed" videos that were available when Borghol *et al.* collected their datasets, but not when the data collection for this thesis was carried out. A summary of the thesis and of its contributions, and a discussion of future work, are presented in Chapter 7.

Chapter 2

Related Work

Content distribution is a vast application area in the Internet that includes web content delivery, peer-to-peer file sharing systems, IPTV systems, video on demand, one-click hosting services, and video and picture sharing services. This chapter presents some previous work concerning the popularity characteristics of on-line content (Section 2.1) and attempts to model popularity evolution (Section 2.2). Some of this work is quite old, but still relevant to current work concerning content popularity, and still widely cited owing to the fundamental nature of the observed popularity characteristics.

2.1 Popularity Characteristics of On-line Content

2.1.1 Web Pages

Several researchers have studied the applicability of Zipf's law (i.e., the Zipf distribution) for modeling the relative frequencies of requests to web pages [6][35][16]. If the relative popularity of each page in a collection of N pages follows a Zipf distribution, then the probability that a random request is to the i'th most popular page is $(1/i^{\alpha})/\sum_{j=1}^{N} 1/j^{\alpha}$, where α is a parameter typically between 0.5 and 1.5. Breslau *et al.* [6] analyzed the applicability of Zipf's law to web pages, and implications for web caching performance. Using six different web proxy cache traces, they found that web page relative popularities followed a Zipf distribution with parameter α ranging between 0.64 and 0.83. They then applied a simple model with independent requests and reference probabilities following a Zipf distribution, and found asymptotic cache hit ratio and "temporal locality" properties consistent with those observed in practice. The authors also compared four cache replacement algorithms using trace-driven simulation, and found that the algorithm suggested by their simple model did indeed perform best.

Sunghwan *et al.* [24] analyzed website popularity characteristics over the 100,000 most popular URLs in the United States and China using traces collected during 2006 to 2010. Their popularity distribution analysis revealed that the popular URLs became increasingly popular, and the list of unpopular URLs became larger, resulting in a long tail phenomenon in the popularity distribution. Their analysis showed that the proportion of requests for the most popular URLs increased from 0.08-0.12% in 2006 to 0.28-0.41% in 2010. Similarly, 76.9-83.3% URLs were requested only once in 2006, which increased to 84.6-87.8% in 2010.

Johnson *et al.* [25] analyzed the popularity rank of worldwide websites on a rural network in Zambia during a two week period in 2010, and found that Content Distribution Network (CDN) sites became requested the most frequently (13.61% of the total requests), while Google (7.88%), Facebook (6.65%), and Yahoo (4.65%) were the next most requested websites.

The increases in web traffic can dramatically degrade web performance by increasing the page-load latency. Therefore, page owners apply several techniques such as caching, prefetching and server-based push to improve web performance. However, use of such techniques mostly depends on the content popularity characteristics and the rate at which users make requests. With such motivation, Padmanabhan *et al.* [35] analyzed web page popularity characteristics for the MSNBC news site (one of the popular websites in the Internet) including popularity dynamics and the possible applicability of a Zipf-like popularity distribution. Their analysis showed that web pages are mostly popular when they are new and fresh but some pages are always popular; mostly the always popular pages are index pages (default pages). The authors also analyzed repeated access of pages by dividing them into modified and non-modified pages and observed that over 50% of the total accesses were to the modified pages, but the number of accesses was much higher than the number of modifications. Then they analyzed popularity stability by looking at the amount of overlap in consecutive days among a certain percentage of popular pages. The authors found that approximately 60% of popular pages were overlapping in any two consecutive days. Also, the analysis of popularity stability with a 10 months separation period showed that a significant percentage (20%) of pages were overlapping. The analysis of the potential applicability of a Zipf-like distribution showed that the page popularity followed a Zipf distribution (with α in the range of 1.4 to 1.6) reasonably well except for the most popular pages. They observed similar behaviour for the datasets collected on two distinct dates 10 months apart.

Analysis of popularity characteristics is significantly important to design efficient caching policy. However the presence of web cache and its size can reveal the popularity characteristics differently. Therefore, Doyle *et al.* [16] analyzed the popularity characteristics of web content and the impact of the web cache on the popularity distribution (if the content access from cache is not taken into account) based on the traces collected from IBM's main web server. The authors plotted the number of references to each object by rank in log-log scale and claimed that the original popularity distribution curve (without the presence of the cache) follows Zipf's law except for the most unpopular content. However, the presence of cache makes the popularity distribution curve flatter for the most popular content, e.g., there are 1035 reference to the most popular content whereas, 816 references are observed for the content with rank 1000. The authors also analyzed the content hit rate as a function of popularity rank for different cache size and observed that the web cache hit rate rapidly falls as the content popularity decreases. They also compared the content hit rate with a synthetic trace for $\alpha = 0.6$ and claimed that the effect of cache in the real dataset is more pronounced.

2.1.2 Video on Demand

Almeida *et al.* [2] analyzed the workload characteristics of two educational media servers (eTeach and BIBS) that provided videos of course lectures to students. The authors analyzed the file access frequency for the first 100 files for both of the servers and observed that the distribution curve (log-log plot of view count versus file rank) is approximately linear for the first few files and linear for the rest of the files with a different slope than for the first few files. Therefore, the distribution of file access frequency in their analysis can be modelled as the concatenation of two Zipf-like distributions. The authors analyzed relative file popularity in different days as well as different hours in a day and observed that videos significantly changed their relative popularity. They also claimed that the popularity of the files was stable for very few periods for both of the servers. The authors also analyzed the access frequency of different segments of the same file. For example, considering the most popular, 4th most popular and 6th most popular files, the segment access frequency for ten-second segments was similar for the most popular and 4th most popular files but for the 6th most popular file, the segment access frequency was skewed toward the earlier segments. A significant observation from their analysis is that a large number of files were accessed only once and many files, once accessed, were not accessed again during the next eight or more hours. Like Almeida *et al.*, Huang *et al.* [23] observed different behaviour in the popularity distribution curve for the top most popular videos, in comparison to the rest of the videos. They analyzed the popularity distribution for 59,000 on-demand videos on the MSN sites over nine months. The log-log plot of the view counts as a function of the file rank showed that the curve is almost flat for the most popular ten files but approximately a sloped straight line for the rest of the files. The reason for such behaviour is that, on any given day, some videos are highly popular and their view counts are similar. This behaviour of the popularity distribution is similar for the minimum, medium and maximum traffic day during the trace collection period.

A similar analysis is done by Chesire *et al.* [14] where they characterized streaming-media workloads using a trace collected from the University of Washington. The authors observed that the popularity distribution follows a fairly straight line on a log-log plot, and therefore could be modelled using a Zipf-like distribution, with $\alpha = 0.47$. They also observed that approximately 78% of the media objects were accessed only once and only 1% of the objects were accessed ten or more times. Out of 23,738 video objects, the most popular 12 objects were accessed 100 or more times. The authors also analyzed the number of unique viewers for each object and found that 84% of the objects were accessed by only one unique client and 1.6% of the objects were accessed by five or more individual clients. The authors also observed strong temporal locality during the peak hours.

Acharya *et al.* [1] analyzed video download characteristics using a trace collected from a video on demand system at the Lulea University of Technology, Sweden. The analysis showed that the most popular ten percent of the videos accounted for 50% of the accesses. Video popularity was found to significantly deviate from a Zipf-like distribution. The authors also analyzed the number of unique machines from which

videos were accessed and observed that 59% of the videos were accessed from only 10% of the most active machines and 74% of the video accesses were accounted for by 20% of the most active machines. The authors analyzed the locality of machines and observed that 67.4% of the unique machines were local machines. The authors also observed high temporal locality in their traces, wherein several requests for the same video occur within a short time span. In summary, video requests were concentrated mostly on a relatively small fraction of the videos, were generated mostly from a relatively small fraction of the active machines, and were bursty in time.

Video on demand systems provide videos to users according to their need, but true video on demand systems also provide the features that are present in DVR control (forward, backward etc.). Griwodz *et al.* [19] analyzed movie popularity based on traces collected from VodeoWoche magazine and a video rental store and proposed an architectural model for true video on demand systems. First, they analyzed video popularity with respect to age based on two long-term popular movies, Highlander 3 and The Lion King. Over approximately 28 weeks, analysis showed that both the movies followed a decreasing trend of popularity with the increase of age, and both of the movies achieved their peak popularity at an early age (first 3-4 weeks). The authors also analyzed the possible applicability of the Zipf distribution for modelling the popularities of the 250 most popular movies on each of a number of days and plotted curves for the days with the lowest and highest request rate for the top 10 movies (over a one month period). They showed that a Zipf distribution curve overestimates the popularities of the most popular movies.

2.1.3 One-click Hosting

File hosting services provide a convenient way to disseminate content over the Internet. There is no specialized software required to upload or download content using file hosting services. Further, file hosting services provide advantages in content availability, download performance and copyright policy [3]. A number of file hosting services are responsible for a significant amount of content dissemination and Internet traffic. Through further analysis file hosting services can perhaps be made more efficient.

Mahanti *et al.* [31] presented a detailed study of the five most popular file hosting services (RapidShare, Megaupload, ZSHARE, MediaFire and Hotfile) using traces collected over a one year period (Jan-Dec 2009) from a university's 400 Mbps Internet access link. They analyzed content popularity in terms of content downloaded by both free and premium users. Their analysis revealed that except for zSHARE, over 95% of the downloaded files were downloaded only once, whereas 83% of the zSHARE downloaded files were downloaded only once. zSHARE provides a streaming service for media file download which was considered the main reason for this difference. For all the services, no files were downloaded more than 10 times. They also analyzed the number of links to downloaded files that were found at each of a number of third-party sources, using the HTTP Referrer header for this analysis. The distribution of these links per referrer shows

a heavy tail nature and could be fit by a power law distribution with exponential cutoff.[1]. The authors also analyzed content download by free and premium users and showed that most of the downloads were by premium users for the Megaupload system, while for RapidShare system the numbers of downloads by free users and premium users were similar. All, or almost all, of the downloads were by free users for the other file sharing services. They also analyzed the content download growth throughout the year; for example, RapidShare and MediaFire had 28% and 14%, respectively, fewer downloads in the fall university semester compared to in the following winter semester. Overall, 46% of the observed downloads occurred during the winter semester at the end of the measurement period. They observed that except for MediaFire most of the file downloads, about 60%, occurred during the evening.

In a subsequent paper, Mahanti *et al.* [30] analyzed the service popularity of four file hosting services (RapidShare, MediaFire, Megaupload, Hotfile) and two P2P torrent discovery sites (Pirate Bay and Mininova). They also analyzed content popularity for specific publishers. Their analysis showed that in the case of monthly unique users, the service popularity followed an increasing trend during the first half of their data collection period (May/08 to May/11), and after a temporal peak during May/10, Pirate Bay, MediaFire and Megaupload followed an increasing trend and the rest followed a decreasing trend. The number of monthly total users also followed similar trends. File hosting services do not publicly provide the number of downloads of each file, but the authors were able to obtain datasets from three anonymous RapidShare publishers that included file download counts. They categorized these publishers as a large, a medium and a small publisher. The authors analyzed the possible applicability of several variations of power law distribution (Lavalette, Zipf-Mandelbrot, and Tsallis), as well as the stretched exponential Laherrere distribution, for modelling the file popularity for each of the publisher's datasets. They observed that the file popularity for the large and medium publishers could be well-modelled by a Lavelette distribution and the file popularity for the small publisher by a Laherrere distribution. They claimed that the file popularity for none of the publishers followed a Zipf distribution. The datasets were very small, however, with the largest containing download counts for only 3,525 files.

Antoniades *et al.* [3] explored content popularity characteristics with respect to the number of downloads per file for the popular RapidShare file hosting site. They also compared the popularity of RapidShare with the popularity of BitTorrent in terms of content download throughput. One-click file hosting services can contain the same content in multiple files with different names, which has a strong impact on the observed popularity characteristics. Thus considering individual files, the file popularity analysis showed that 75% of the files were downloaded only once and a very small percentage (0.05%) of the downloaded files were downloaded more than five times during their 5 month data collection period. Very few files were downloaded more than ten times. The authors also analyzed the number of files downloaded by individual clients, as distinguished by IP address. The analysis showed that 57% of the clients that downloaded at least one file

[1]A power law distribution has a probability density or mass function f(x) that is proportional to $x^{-\alpha}$ for large x, for some constant α A power law distribution with exponential cut off has a probability density or mass function f(x) that is proportional to $x^{-\alpha}e^{-\beta x}$ for large x, for constants α and β.

on a particular day, downloaded more than one file that day, whereas only 23% of these clients downloaded more than 10 files that day. They also showed that the distribution of the number of downloads in a day by a client that has downloaded at least one file that day can be approximated by a Pareto distribution with a shape parameter of 0.66.

This previous work on file hosting services has shown that power law distributions are commonly observed for some characteristics of these systems, as is also the case for the web and video on demand systems considered in the previous sub-sections. A particularly interesting aspect of file hosting services, however, is the impact of clones (files with the same content) on the file popularity characteristics.

2.1.4 P2P File Sharing

Gummadi *et al.* [20] analyzed the fundamental properties of multimedia file sharing systems that use peer-to-peer networks. They collected traces of over 20 terabytes of Kazaa P2P file sharing system traffic at the University of Washington. They observed significantly different characteristics between P2P traffic in Kazaa and web traffic. Their analysis revealed that a much higher percentage of clients download any specific Kazaa object only once, in comparison to accesses of web objects. Specifically, 94% of the time clients downloaded a Kazaa object only once and 99% of the time clients downloaded a Kazaa object at most twice. However, 57% of times clients accessed Web pages (e.g. CNN or Yahoo pages) only once. The reason for such differences between Kazaa files and web pages is that Kazaa files are immutable and most popular Kazaa multimedia files are periodically replaced by totally new files; however web pages are periodically updated with new content (e.g., CNN or Yahoo news). For both small and large Kazaa objects, the authors compared the most popular 10 and the most popular 100 objects for the first month of their trace collection with the most popular 10 and 100 objects respectively in the last month of their data collection. (The total trace collection period is 6 months.) They observed that for the most popular 10 small Kazaa objects, there is no overlap experienced and only one overlap for the most popular 10 large Kazaa objects. When considering the 100 most popular objects, only 5 small objects and 44 large objects overlapped. Also, 72% of requests for the large objects were to "old" objects (at least one month since the first request to that object in the trace) while only 52% of requests for small objects were to "old" objects. The authors also analyzed the possible applicability of the Zipf distribution for Kazaa objects and observed that the popularity distribution of the Kazaa objects is highly deviated from a Zipf distribution. The reason for such behaviour is the absence of the basic characteristics of a Zipf distribution; because of the "fetch-at-most-once" behaviour. For example, out of 10000 Kazaa objects only 47 objects received more than 100 requests and the most popular object received only 272 requests.

P2P-XL is a special purpose server designed by Leibowitz *et al.* [28] to observe the workload characteristics of newly uploaded content in peer-to-peer file sharing systems. They installed their server in a major ISP with 10,000 users and an average of 2,000 simultaneous content downloads. They analyzed workload characteristics in terms of content popularity based on both overall downloads and file category specific

downloads. In the one month period of data collection, they observed that 80% of the overall downloads were of only 20% of the files. The authors separated files into six subcategories (song, movie, application, picture, document, others) and found that the majority of the downloads were for song files (but due to the file size, the majority of the traffic was due to downloads of movie files). The authors also determined the types of the most popular 100 files and observed that close to 50% of them were movie files and the rest were mostly music and application files in roughly equal proportions.

2.1.5 IPTV

IPTV networks are an important type of content distribution system. Therefore, many researchers have focused their interest on workload characteristics in IPTV networks so as to support the design of more efficient architectures and functionalities. Qiu *et al.* [37] identified IPTV channel popularity characteristics as well as channel popularity dynamics, and also designed a popularity model to capture channel popularity for different groups of users. This work used a dataset collected in June 2008 from a large scale IPTV service provider in the USA that had over one million users and over five hundred TV channels. The authors observed that channel popularity was highly skewed, with about 10% of the channels receiving 90% of the accesses. They found that the relative channel access frequencies could be accurately modelled by a Zipf distribution that did not depend on either the time period over which access frequencies were measured, or on the time of day that was considered (although which channels occupied which spots in the popularity ranking did depend on these factors). They also observed a strong correlation between the channel access frequency and the channel dwell time (amount of time the channel is tuned into) and found a Spearman rank correlation value of 0.98 and a Pearson correlation coefficient value of 0.97. The authors also observed the popularity dynamics and found that these dynamics could be modelled by a stationary stochastic process, with the exception of non-stationary behaviour due to diurnal patterns that were found to be significant for some channels. The authors also analyzed the behaviour of different categories of users and observed some interesting findings such as that 28% of users (as identified by a particular set-top box) watched on average more than 12 hours of TV daily whereas 36% of users watched on average less than 1 hour daily; 31% of users were "daytime watchers" (watch TV for at least twice as long during the day - from 6am to 6pm - than during the night, on average) and 39% of users were "nighttime watchers"; and 24% of users switched channels over 200 times a day on average while 12% of users switched less than 10 times a day on average. The authors used a Zipf distribution to model long term channel popularity, a *mean reversion model* for modelling popularity dynamics, and a multi-class model for capturing non-stationary behaviour. The models were evaluated through comparison with the real data and found to capture channel popularity as observed in the real dataset reasonably well.

Another analysis presented by Cha *et al.* [11] on IPTV channel popularity characteristics and its dynamics considered both channel popularity for individual geographic regions, and overall popularity. Their analysis revealed that the top 10% of the channels experienced approximately 80% of the viewers. This

property was consistent for different times in a day. The authors also determined the channel popularity distribution at a short time scale, by finding the average number of viewers of each channel over a 15 minute period. The resulting popularity distribution for the top 100 channels showed that the popularities of these channels could be modelled by a Zipf-like distribution. However, the popularities of channels ranked over 100 (i.e., were not ranked within the most popular 100 channels) failed to follow a Zipf-like distribution. Channels ranked over 100 included channels that users had to pay for individually, foreign news channels, local channels for small regions, and channels carrying audio or reality shows. The authors also analyzed the popularity dynamics of channels by observing the rank changes during their measurement period when channel popularities were measured over 5 minute intervals, and observed significant non-stationarity in channel popularity for different times of the day. However, the popularity dynamics pattern was similar for the same time on different days. The popularities of channels for different genres were found to vary according to the geographical location, with viewing preferences for different genres varying by up to 20%. However, free (free-to-air broadcast channels), mixed (collections of comedy, soaps and reality shows) and kid's channels were found to be consistently popular in all regions (12 regions in total for the analysis).

Hei *et al.* [22] presented a measurement study on peer-to-peer based IPTV system named PPLive, where they analyzed channel popularity for both popular and less-popular channels. Their analysis showed that the channel popularity pattern in P2P based IPTV varied similarly to the regular TV in terms of channel watch time, and channel switch behaviour (people watch more channels on the weekends and in the evening, but switch less in the evening according to the Cha *et al.* [11] analysis). Also, channel popularity mostly depended on the popularity of the program rather than the popularity of the channels itself. They analyzed this characteristic based on the annual Spring Festival Gala on Chinese New Year, where they found a sharp jump in the number of viewers right before and after the program. They also analyzed the channel popularity and its dynamics by observing channel watch and change characteristics. Their analysis revealed that 90% of the peers stay connected to the IPTV system less than 1.5 hours and they usually show less switching channel churn in the middle of any program. They also observed that a small number of peers share a large amount of content in the system. Their analysis revealed that P2P based IPTV can increase channel popularity by providing short start-up delay, higher streaming rate and by reducing playback lags. Hei *et al.* also observed channel popularity as following a Zipf-like distribution for the most popular channels but with popularity decaying quickly for the least popular channels. Similar applicability of a Zipf distribution is observed by Cha *et al.* [12] where they analyzed the channel popularity of the world's largest Telco-managed IPTV network. Using their measurement data, the authors studied the possible integration of IPTV and peer-to-peer video delivery using TV set-top boxes or home gateways. Cha *et al.* found that the popularity distribution was Zipf-like for the popular channels but that the popularity decayed faster for channels ranked 30th and below. This behaviour was similar for different times of the day.

2.1.6 Video Sharing

The fundamental difference between the traditional VoD systems and UGC systems is that in traditional VoD systems the content is professionally produced, and has quality and popularity that is more controlled and predictable than in UGC systems. Therefore, the popularity distribution for VoD and UGC content might be significantly different. YouTube is the most popular site for sharing of UGC videos. One of the first studies of UGC video popularity was carried out by Cha *et al.* [8], who performed an in-depth study of the content popularity characteristics of UGC systems using traces collected from the YouTube and Daum video sharing services. Their analysis showed that the videos aged below one month experienced slightly more average view counts than the older videos. Some older videos also account significantly larger number of view counts. Interestingly, 80% of the requested videos in any given day are older than one month, but 50% of the top twenty videos are in any day are new. The authors also analyzed popularity shift by observing the change of popularity rank and shows that, videos frequently change their popularity rank at their early age than their later age. Approximately 1% of young videos experience high rank shift indicating their availability in the most popular list. Interestingly, some very old videos also significantly increase their popularity rank. In video popularity characteristics, the authors analyzed the Pareto Principle[2] for both UGC and non-UGC video, and mentioned that 10% of the most popular videos received about 80% view counts; this statistic is differently expressed in Yu *et al.* [43] where 10% of the most popular videos account for 60% of the requests with the trace collected from a China Telecom dataset. The authors also analyzed the distribution of video requests among popular and non-popular videos. For this purpose, the authors categorized videos into four sub-categories (YouTube Ent., YouTube Sci., Daum Travel, Daum Food category). The log-log plot of the popularity distribution showed that all four sub-categories followed power law behaviour but, YouTube Sci and Daum Food categories showed a sharp decay for the most popular videos. The authors also analyzed long-tail behaviour in the YouTube Sci sub-category and observed that the truncated tail in the distribution curve is best fit by a Zipf distribution with an exponential cut off and the second best fit is with a lognormal distribution.[3] Additionally, the authors presented a comparison between the UGC and non-UGCs content in terms of popularity characteristics using datasets from Netflix, Lovefilm and Yahoo! Movies. Their analysis showed that the UGC and non-UGC videos showed a strong correlation coefficient between popularity and rating (0.8 for YouTube and 0.87 for Yahoo videos).

Zink *et al.* [45] analyzed the popularity distribution of YouTube content by collecting network traces from the interconnection between a university campus access network and the Internet. They also analyzed the global popularity of this local trace and compared with local popularity behaviour. The global view count analysis showed that a significantly large number of videos (approximately 77%) are viewed only once

[2]Also known as 80-20 rule where 80% of the effects come from 20% of the cases. For example, 80% of the view counts experienced by 20% of the videos.

[3]The lognormal distribution has a probability density function of $f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(lnx-\mu)^2}{2\sigma^2}}, x > 0$

in the local set, whereas only 23% of videos are requested for two or more times. They compared popularity between global and local popularity by calculating the correlation coefficient and observed very low correlation coefficient in all three sets of traces, which are 0.04 in a 12-hour trace, 0.06 in a 3-day trace and also 0.06 in 4-day trace.

Benevenuto *et al.* [4] characterized users interaction via YouTube video objects. Their analysis showed that 20% of most responsive users contributed with 65% of all video responses, whereas 84% of all responsive users posted less than five video responses each. They also analyzed view counts for the fitting of power law distribution and observed that both the number of views responsible by each user and by each video fitted with power law for $\alpha = 0.782$ and $\alpha = 0.741$ respectively. They also assessed the the accuracy with a linear regression model and found that, in both cases, the value of R^2 is above 0.91 ($R^2 = 1$ implies perfectly accurate). The authors also analyzed video response in terms of geographical location and found that 40% of the videos received over 60% of the response from the same country as the uploader. The authors also analyzed the contribution of the uploader him/herself in video response who uploaded the video and observed that, 25% of all responses are self-response and roughly 35% of the responded videos received at least one self-response, and around 12% of them received only self-response.

Figueiredo *et al.* [17] analyzed YouTube video popularity on three different YouTube datasets: 1) YouTube top list videos ("Top" videos) 2) videos removed from YouTube called YouTomb videos (collected by YouTomb project at MIT) and 3) set of videos collected using the YouTube search API ("Random" videos). Their analysis focused on the popularity evolution of the videos in these three datasets and showed that the popularity growth patterns in all three datasets are different. Copyright protected videos received their peak popularity at their early age of lifetime. They also claimed that videos receive their peak popularity when appeared at the first page of YouTube. The authors identified fourteen referrers to reach YouTube videos and claimed that Google search and YouTube internal searching mechanism play an important role to make video popularity high. Their analysis revealed that YouTomb (removed dataset from YouTube) videos received most of their popularity at their early age and Random videos received most of their popularity at the later age of their lifetime. Similarly, 50% of the Top, YouTomb and Random videos received 90% of their views in 65%, 21% and 87% of their lifetime respectively. Again, 31% of the Top videos received 10% of their views in 20% of their lifetime which is for 18% videos in Random dataset. Similarly, 50% of the Top videos and 40% of the YouTomb videos received 50% of their views in their 1st peak week. Random videos are insignificant for this criterion because of their different age limit. The authors also distinguished the evolution process of video popularity in four different types: 1) Memoryless (video popularity is stable), 2) Viral (popularity propagation through Online Social Network (OSN)), 3) Quality (popularity due to external events) and 4) Junk (popular due to spamming). Based on the fraction of views (t) on the most popular day, the authors claimed that for t = 20%, 77% of Random videos and 44% of YouTomb videos were viral. A significant number of videos were also in quality and junk category. For $t \leq 30\%$, most videos were in quality category and for $t > 30\%$, most videos were viral. The authors also analyzed referrers' contribution to video

popularity by identifying fourteen types of referrers that appeared in the dataset and claimed that Google search and YouTube's internal mechanisms are mainly responsible for users finding YouTube videos, which is also claimed by Oliveira *et al.* [34]. Figueiredo *et al.* also claimed that 75% of the Top and YouTomb videos received their first referrer access during the first quarter of their lifetime. Again, 9% of the Top and 25% of the YouTomb videos had their first referrer after 40% and 70% of their lifetimes respectively. The authors also claimed that search, internal, external and social referrers were mainly responsible for views to videos in the random dataset.

Popularity distribution with respect to the total view counts might not be an appropriate metric for significant caching decision, because cache hit ratio depends on the current popularity of videos instead of popularity that videos may have had a long time ago. Therefore, Mitra *et al.* [33] analyzed video popularity characteristics on four video sharing services, Dailymotion, Yahoo, Veoh and Metacafe based on both total views popularity and viewing rate popularity (since uploaded and most recent two weeks data collection). They also analyzed Zipf and power law fitting with all four video sharing services using different values of the shape parameter (α). Their popularity analysis for both total view counts and average viewing rate showed high skewness and the 80-20 rule. Over 80% of the total view counts were to 20% of the videos which was similar for average viewing rate. However, video popularity in terms of total view counts for Metacafe and video popularity in terms of average viewing rate for Dailymotion experienced significantly more skewness compared to the other services. The analysis of the Zipf and power law matching showed that the total view distribution for all four video sharing services followed a Zipf-like distribution with exponential cut off for least popular videos. The complementary cumulative distribution of the total view counts showed heavy-tailed nature and followed power law behaviour throughout a significant range of views. The authors also analyzed total view count distribution with power law, power law with exponential cut off and lognormal distribution to find the best fit and found that the middle region of the distribution curve showed the best match with power law and the left region and the fraction of the middle region showed best match with lognormal distribution or power law with cut off. However the popularity distribution in terms of average viewing rate more clearly showed Zipf-like behaviour than was shown in terms of total view count distribution. For both Dailymotion and Metacafe, the CCDF of the average viewing rate since video upload was best fit by a power law distribution, with $\alpha = 1.93$ and $\alpha = 1.46$ respectively. However, for Yahoo and Veoh the best fits were with lognormal and power law with exponential cut off distributions. However, the average viewing rate over the two weeks period for Dailymotion videos fit best with power law ($\alpha = 2$).

Broxton *et al.* [7] classified view counts of YouTube videos into social and non-social categories. Videos viewed by users through external links like blogs, emails, instant message etc. were categorized as social videos and videos viewed by YouTube internal mechanisms, YouTube searching mechanism and external search engine were categorized as non-social videos. Videos that became highly popular through such social or non-social links sharing were classified as viral videos. The authors analyzed the characteristics of viral videos on YouTube in terms of social or non-social views. By comparing data of 1.5 million random YouTube

videos that were uploaded between April 2009 to March 2010, excluding videos that did not receive at least 100 views in the first 30 days of publishing. The analysis showed that the view counts through social link rose to, and fell from, their peak popularity more quickly than view counts through non-social links. The authors claimed that in the first 30 days of viewing, 42.2% views were accounted as social views, but overall only 25% of views were captured as social views. However, 20% of the videos had social views greater than 65%. Almost all of these videos reached their peak popularity at their 5th day of views. They also observed social views of videos on different video categories and observed that music category had the highest percentage (18.2%) of overall social views. Facebook and Twitter referral videos had a higher fraction of view counts but Twitter referral videos had more social views than Facebook referral videos. The authors also analyzed the short-term popularity and long-term popularity characteristics and observed that most of the views of short-term popular videos came from more social links and most of the views of long-term popular videos came from less social links. Highly social videos behave differently than less social videos. Highly social videos achieved high popularity in short time period but could not keep up over the long term. Viral videos are a subset of the highly social videos.

YouTube separates their videos into different categories based on the video types and also maintains different lists based on users' interactions. Cheng *et al.* [13] analyzed YouTube video characteristics using statistics for 2.6 million videos of 27 different YouTube datasets, obtained by crawling the YouTube site. Their crawls were started using sets of videos in lists such as "Recently Featured", "Most Viewed", "Top Rated", "Most Discussed", as maintained for the different time periods of "Today", "This Week", "This Month" and "All Time". Among the videos found in their crawls, Music category videos were the most numerous, accounting for 22.9% of the total videos. Entertainment and Comedy category videos were the second and third most numerous and accounted for 17.8% and 12.1% of the total videos respectively. They analyzed video popularity characteristics with respect to the possible applicability of a Zipf-like distribution using a trace collected on one day, containing total view counts for more than 100,000 videos. All 27 datasets were not used since these were collected over a three month period, and the authors wanted to observe the distribution of total view counts at one point in time. Their analysis showed that the total view counts as a function of the video popularity rank (by views) was a straight line for popular videos but decayed very fast for the unpopular videos. Their analysis also showed that a large number of videos were very unpopular. Note that Cheng *et al.* considered total view counts for their analysis but did not consider video age which has a major impact on total view counts, nor did they consider the current rates at which the videos were acquiring views.

Zhou *et al.* [44] recently proposed a *random prefix sampling* method to estimate the total number of YouTube videos, and to obtain a random sampling of such videos. In this method they repeatedly used the YouTube API with search strings that specified random prefixes within the character string space from which YouTube video ids were drawn at the time of the study. For each search, the API call returned a list of matching ids for existing YouTube videos (if any). Using their method, the authors were able to conclude

that as of May 2011, there were approximately 500 million YouTube videos. They were also able to derive, as a function of the number of random prefixes that are searched for, a confidence interval for the estimate of the total number of videos. The authors then used a random sampling of videos obtained using their method, to demonstrate that collections of YouTube videos obtained by previously-used breadth-first search techniques result in substantially biased samples. According to statistics derived from their random sample of YouTube videos, only 14% of videos had a total view count of more than 1000. For two collections of videos obtained from breadth-first search (following "related video" links), however, the corresponding percentages were 89% and 52%, respectively. The average view count for the videos in the random sample was 3898, whereas for the two collections of videos obtained from breadth-first search the average view counts were 32046 and 9928, respectively. This result is highly relevant to one of the motivations of the work in this thesis. Often, it is difficult to obtain a random sampling of user-generated content items. Without a random sampling, results for popularity distributions and other characteristics may be biased. One of the goals of this thesis is to study the biases, with respect to popularity characteristics, of videos obtained using keyword search.

2.1.7 Picture Sharing

Clients of an Online Social Network share their content from their home, but on the server side everything is centrally controlled in the OSN server. Therefore, content is manipulated through presentation in different formats and qualities; as well, clients lose full control over their content. Marcon *et al.* [32] analyzed the feasibility of providing full control of the content to the clients. They also studied content popularity based on the weekly viewing rate. For their analysis they collected trace for 1,324,080 publicly accessible photos from Flickr for 19 days and 1,251,492 publicly accessible videos for 166 days. The authors also compared the popularity rate between the Flickr photos and YouTube videos and observed that YouTube videos in general received more requests than the Flickr photos. This popularity difference between Flickr and YouTube content was expected because, at the time of their study, 22% of global internet users visited YouTube while only 2.5% of global Internet users visited Flickr [32]. Their analysis also showed that a significant amount of content (97% Flickr photos and 44% YouTube videos) for both the sites had not requested during a one day period. Similarly, almost all Flickr photos received less than 1,000 weekly requests, whereas each of the most popular 1,000 YouTube videos received over 10,000 weekly views. The log-log plot of the view counts as a function of the file rank shows that the curve is almost straight for all the Flickr photos but decays fast for unpopular YouTube videos.

Cha *et al.* [10] analyzed the popularity growth of 11 million Flickr photos and observed that millions of pictures exhibit fewer than 10 view counts and only very few pictures experienced very high popularity. The view count distribution curve shows heavy tailed nature with approximately flattened for the unpopular videos. The authors also compared picture popularity among local versus global picture by picking most popular 100 local pictures in 1-hop, 2-hop, 3-hop and 4-hop neighbourhood. The analysis showed that the set of popular videos was highly different in different local regions. In 1-hop neighbourhood, out of 250 region

no overlap observed in 133 regions. However, in 2-hop, 3-hop and 4-hop neighbourhood, on average 8, 39 and 70 popular pictures were overlapped respectively. The authors also analyzed the picture popularity growth pattern over age by analyzing the popularity growth of four most popular pictures and observed that the popularity growth curve was different for different pictures. Pictures usually followed three different growth patterns: 1) view counts were approximately uniformly distributed throughout the age, 2) view counts stayed very steady over a long period and suddenly spiked as result of winning any award and again became steady, and 3) view counts were uniformly distributed with some spike due to being featured on the front page or in the explore page. The authors also observed the picture popularity over the long period and found that the view count for most of the pictures was high during the first few days since the picture was uploaded and approximately linear during the rest of the picture lifetime. This work is relevant also to work on video popularity evolution, since different types of user-generated content (pictures and videos, for example), may share similar popularity evolution patterns.

2.2 Modelling and Predicting Popularity Evolution

Borghol *et al.* [5] analyzed the popularity characteristics of UGC videos using two YouTube datasets called the recently-uploaded (newly uploaded videos in YouTube) and keyword-search (videos retrieved by searching on keywords) datasets, both containing weekly view counts recorded over an eight month measurement period. Their analysis mainly focused on the popularity dynamics and churn characteristics, and the three-phase viewing rate characteristics (before-peak: popularity characteristics before reaching their peak popularity, at-peak: popularity characteristics at their peak popularity, and after-peak: popularity characteristics after passing their peak popularity), and based on the three-phase characteristics, the authors developed a model that can capture the popularity evolution of recently-uploaded videos. To observe the popularity dynamics and churn, the authors plotted the weekly view counts in consecutive weeks and also the change in popularity rank of the recently-uploaded videos and found that the videos had highly non-stationary popularity early in their lifetime which became more stable as the videos aged. The authors also analyzed the time until videos reached their peak popularity ("time-to-peak"), when considering only the view counts in the weeks covered by the measurement period, and found that approximately three-quarters of the videos reached their peak popularity during the first six weeks of their lifetime since they were uploaded. The three-phase characteristics analysis showed that the view count distributions for each of the phases were heavy-tailed and approximately week-invariant. For example, the peak weekly view count of a video was approximately independent of the video age at which that peak was achieved. Based on the three-phase characteristics, the authors developed a basic model assuming the week-invariant view count distribution for each of the phases and determining the number of videos that reach their peak popularity during each week using an analytic distribution that fit the empirical time-to-peak distribution. The comparison between the recently-uploaded dataset and the model generated synthetic views showed a good match for both the

weekly view count distribution and the total view count distribution. The model did not, however, yield a similar level of video popularity churn as observed in the empirical dataset; therefore the authors extended the model using a scheme for exchanging video view counts in such way that the three-phase characteristics and the time-to-peak distribution are not affected. The extended model introduces additional churn and was found able to exhibit the churn observed in the empirical dataset.

Ratkiewicz *et al.* [38] analyzed the popularity dynamics of online content based on two model systems namely Wikipedia and Chilean web, and proposed a model that can capture the dynamics observed in the real datasets. To analyze popularity characteristics, they considered the number of clicks associated with a particular page and number of web-links pointing to the page. They also considered temporal information of each page that provides popularity burst of each content. The popularity burst analysis of each page shows that almost all pages show a sudden popularity burst at the beginning of their lifetime. The distribution of such popularity burst in logarithmic scale shows heavy-tailed nature. The authors claimed that the popularity dynamics and the heavy-tailed nature observed in the real data could be achieved by a model. The rich-get-richer behaviour could be achieve by preferential attachment[4] model and ranking model [40][18]. The authors applied both of the models to achieve the long tail behaviour, but none of the model were able to exhibit that phenomenon. Therefore, the authors implemented a re-ranking probability model, where at each iteration each item forward in a front position in the list by equal probability to move for each item in the list at each iteration. The authors analyzed the accuracy of their implemented model by comparing with the popularity distribution observed in the real datasets and claimed that their implemented model can excellently exhibit the empirical long tail behaviour.

Szabo *et al.* [41] proposed a model to predict long term popularity characteristics of online content based on the empirical popularity evolution. To observe empirical popularity evolution, they collected two empirical datasets from Digg and YouTube and observe strong linear correlation in popularity between the early and later age after some initial periods. The authors described this linear correlation by a linear model using two arbitrary time points and a noise term. They applied a logarithmic transformation to show popularity evolution at early and later content ages. Also the fluctuation of popularity throughout its age is obtained using the noise term. Based on this correlation and popularity fluctuation, they developed two models and combined them with another existing model that can predict popularity evolution of videos in future age. The authors also compared their proposed popularity prediction model with two online portals YouTube and Digg and claimed that the popularity observation of Digg videos during first two hours allows them to predict popularity during next thirty days. The same prediction for YouTube videos requires ten days observation. The authors claimed significant model accuracy when compared with empirical data.

The popularity characteristics of web content differ depending on the length of the time period considered. Most of the previous researchers analyzed popularity characteristics based on short-term traces.

[4]Also known as "rich-get-richer", "cumulative advantage" and "yule Process" where some form of credits are distributed among some objects in proportion to how much they have already received. For example, popular videos may achieve additional views in proportion to the number of views they have already received.

Tang *et al.* [42] collected two traces from two HP servers: HP Corporate Media Solutions Server(HPC) and HPLabs Media Server(HPL), and analyzed popularity characteristics and proposed a popularity prediction model over the long-term period. Like other research, content popularity is highly skewed in this research too: 14% of files in HPC server and 30% of files in HPL server account 90% of the access. The analysis of Zipf-like distribution including popular and unpopular files reveals that the distribution of the file popularity follow Zipf distribution except for the most popular and the most unpopular files. Therefore to fit the file popularity for both popular and unpopular files, authors proposed a generalized Zipf-like distribution using k-transformation with Zipf law. If x is the file rank and y is its access frequency then according to the k-transformation x and y will be calculated as x = (x + k_x - 1)/k_x and y = (y + k_y - 1)/k_y, where k_x and k_y are the scaling parameter (scaling parameter's value is 12 for HPC and 7 for HPL).

The popularity characteristics of online content can be measured by popularity metrics such as the number of views, comments, or links. Lee *et al.* [27] proposed a popularity prediction model by inferring those popularity metrics whose observation early in a content item's lifetime can be used to predict whether or not the popularity of that content item will eventually exceed some threshold. To develop the model, the authors considered the correlations among the popularity metrics and, based on these correlations, a few selected metrics are taken into account for predicting future popularity. The authors tested their approach using two datasets collected from online discussion forums on the sites forums.dpreview.com and forum.myspace.com. They were able to accurately predict whether thread lifetime would exceed some specified threshold, as well as whether the number of comments would exceed some threshold.

2.3 Summary

This chapter has surveyed prior work concerning the popularity characteristics of on-line content, and attempts to model the popularity evolution of such content. Content distribution in the Internet includes web content delivery, peer-to-peer file sharing systems, IPTV systems, video on demand, one-click hosting services, and video and picture sharing services. Work in each of these areas is relevant to the work in this thesis, since on-line content of different types has frequently been shown to share common popularity characteristics.

The popularity analysis on the web pages showed that popular web pages become more popular and the number of unpopular web pages increases over time. The news web pages are mostly popular when they are fresh and new. The popularity distribution of web pages follows a Zipf distribution except some most popular pages. The popularity characteristics analysis of VoD systems showed that a large number of videos are viewed only once, whereas a small percentage of videos account for a large fraction of the total view counts. Significant differences are observed in the applicability of Zipf's law for VoD content. Most of the analyses showed the applicability of Zipf's law with some deviations. However, Almeida *et al.* analyzed the most popular 100 files on two education media servers and found that the popularity distribution curve showed a

good match with a concatenation of two Zipf-like distributions. Acharya *et al.* [1] analyzed a trace collected from a VoD system at the Lulea University of Technology, Sweden, and observed significant deviation from a Zipf-like distribution. In one-click hosting systems, file popularity is strongly impacted by the presence of clones (files containing the same content), resulting in a large proportion of the file references observed in measurements being one-time references. The popularity distribution for the Kazaa P2P file sharing system showed that a small number of files experience extremely high popularity, therefore significantly deviated from a Zipf distribution. Channel popularity in the IPTV systems mostly depends on the popularity of the programs and is highly skewed towards the most popular channels. Channel popularity follows a Zipf-like distribution for the most popular files but decays fast for the unpopular files. Similar applicability of a Zipf-like distribution is observed for the picture popularity, but the view counts for the most popular pictures are far lower than the view counts for the most popular UGC videos. Pictures usually experience most of their view counts in the first few days and become steady over the long period with some sudden popularity bursts.

Content popularity in video sharing systems is highly skewed towards the most popular content and follows the Pareto Principle. A significant fraction of the content has very low view counts. Videos typically experience higher view counts at their early ages than their later ages. Figueiredo *et al.* [17] observed that YouTube videos received their peak popularity when they appeared at the first page of YouTube and Removed (from YouTube) videos received most of their popularity at the early age and Random (from YouTube) videos received most of their popularity at their later age. The popularity distribution in UGC systems follows a Zipf distribution for most popular files but decays fast for the least popular files. It is often difficult to obtain a random sampling of user-generated content items. Without a random sampling, results for popularity distributions and other characteristics may be biased.

Based on the empirical time-to-peak distribution and the three-phase characteristics, Borghol *et al.* [5] developed a model that can generate synthetic view counts. The model assumes that video view counts in each of the three video lifetime phases (before-peak, at-peak, and after-peak) can be modelled using week-invariant distributions. The authors also extended their model to introduce additional churn in the synthetic view counts. Szabo *et al.* [41] and Tang *et al.* [42] proposed models to predict long term popularity characteristics using logarithmic transformation and k-transformation with a Zipf's law respectively. Ratkiewicz *et al.* [38] proposed a model that can capture the popularity dynamics of web pages. The model uses temporal information of web pages to achieve popularity bursts, and preferential attachment and ranking model to achieve rich-get-richer behaviour.

Chapter 3

Data Collection

Popularity dynamics analysis of user-generated videos requires a large amount of empirical data covering a long period of time. Since it is the most popular user-generated video system, YouTube[1] is the most suitable site to use for popularity dynamics analysis. However, collecting data from the YouTube site is challenging due to its large scale and service-specific limitations. Section 3.1 describes the methodologies used for data collection including development of a crawler using the YouTube API[2] and the hardware/software platform used for data collection. Section 3.2 presents a summary of the datasets. Section 3.3 describes the challenges and limitations of the data collections.

3.1 Methodologies

Two empirical datasets were collected from YouTube using YouTube API on two different sampling approaches. These datasets are termed the recently-uploaded and keyword-search datasets. These datasets were collected in two phases.

The first phase of data collection was carried out by Borghol *et al.* from 27 July 2008 to 29 March 2009. They retrieved meta-data (including video ID, total number of views and number of minutes since uploaded) for two large sets of YouTube videos by developing a crawler using the YouTube API, written in Python. Meta-data for each video was collected once each week throughout those eight months. A one-week sampling period was chosen so that each video's data collection was always on the same day of the week at approximately the same time, so as to avoid potential day-of-the-week effects [5]. This procedure resulted in 35 "snapshots" for each video's meta-data, including the "seed" snapshot obtained during the first week of data collection. Each snapshot for a video contains the total number of views received by that video. Thus from the total view count at each snapshot i ($1 < i <= 35$) one can determine the number of views each video received during the one-week time period between snapshot i and i-1.

The YouTube API enables different sampling approaches. For the recently-uploaded dataset, Borghol *et al.* used a YouTube API call that returns video IDs for 100 "recently uploaded" videos. The videos returned by this call were at most one week old, and appear to be randomly selected among videos within

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

[2] `Website:http://code.google.com/apis/youtube/overview.html/`

this age range. In total, 29,791 distinct videos were found through repeated use of this API call, during the first week of data collection, and tracked for the eight month data collection period.

The YouTube API also allows retrieval of videos using keyword search where the search returns a set of relevant videos that is sorted according to "relevance". Borghol *et al.* chose keywords randomly from a dictionary. Some searches returned over 500 videos, and in such cases only the first 500 videos were added to the set of videos used for the keyword-search dataset. In total, 1,135,253 distinct videos were found through repeated use of keyword search, and tracked for the eight month period of data collection.

The second phase of data collection was carried out as part of the thesis research from 25 October to 27 December in 2011. In this phase, an attempt was made to collect meta-data for the same videos for which data was collected by Borghol *et al.* from 27 July 2008 to 29 March 2009. For this second phase of data collection, 11 different computers were used from a cluster computer system in the Discus laboratory at the University of Saskatchewan. Two crawlers were developed using the YouTube API, written in Python. The first crawler ran from 16 October to 23 October 2011 for the purpose of identifying the videos from the first phase that were still available. Video availability was checked by attempting to retrieve video meta-data for each video, using the video ID as an argument to a YouTube API call. The IDs for the available videos were then separated into 11 sets (for 11 computers) and each set again separated into 7 files (each file for each day in a week). Then the second crawler ran simultaneously in all 11 computers from 25 October to 27 December 2011 using the YouTube API to obtain the total view count for each video, each week. The crawler was designed in such a way that data collection started shortly after midnight on October 24, and after finishing each day's data collection (for the videos in the file corresponding to the day), the crawler waited until the beginning of the next day for the next day's data collection. Note that the total view count was collected for each video once per week, on the same day of the week and at a similar time. This procedure resulted in 9 "snapshots" for each video's meta-data.

YouTube videos became unavailable mainly for three reasons: 1) violation of YouTube terms and conditions, 2) making videos private by uploader, and 3) deletion by uploader of the uploader's account or of selected videos. The videos that were available in the first phase of data collection but not throughout the second phase of data collection were included in a "removed videos" dataset. Removed videos are distinguished by whether they come from the keyword-search dataset or the recently-uploaded dataset. For the recently-uploaded dataset, out of 29,791 videos only 20,000 were available throughout the second phase of the data collection, the remaining 9,791 videos were placed into the removed videos dataset. Similarly for the keyword-search dataset, out of 1,135,253 videos, 627,002 were available throughout the data collection and the remaining 508,251 videos were placed into the removed videos dataset.

Table 3.1: Summary of data from 1st phase of data collection

27 July 2008 to 29 March 2009		35 Snapshots	
Dataset	Recently-uploaded		Keyword-search	
Status	still available	removed	still available	removed
Videos	20,000	9,791	627,002	508,251
Views (start)	765,564	438,191	21,816,635,175	18,277,879,332
Views (end)	20,223,336	18,805,848	34,192,809,485	29,827,097,541

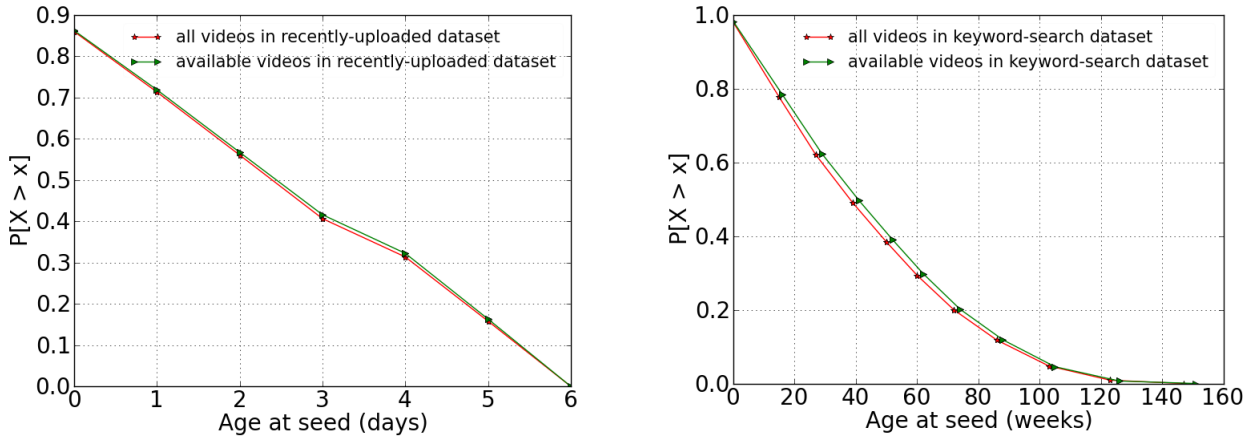
Table 3.2: Summary of data from 2nd phase of data collection

25 October to 26 December 2011	9 snapshots	
Dataset	Recently-uploaded	Keyword-search
Videos	20,000 (67.13%)	627,002 (55.23%)
Views (start)	99,421,245	75,000,355,063
Views (end)	104,269,966	76,802,553,854

3.2 Summary of Datasets

Summaries of the collected datasets are presented in Tables 3.1 and 3.2. From Table 3.1, the 29,791 recently-uploaded videos in the recently-uploaded dataset received approximately 38 million additional views during the first phase of data collection. Similarly, the 1,135,253 keyword-search videos acquired approximately 24 billion additional views. Table 3.1 also shows that the 20,000 recently-uploaded videos that were available throughout the second phase of data collection received approximately 19 million additional views during the first phase of data collection, while the 9,791 removed videos from the recently-uploaded dataset received approximately 18 million additional views. Similarly, the 627,002 keyword-search videos that were available throughout the second phase of data collection received approximately 13 billion additional views during the first phase of data collection, while the 508,251 removed videos from the keyword-search dataset received approximately 11 billion additional views.

From Table 3.2, as noted previously, 20,000 recently-uploaded and 627,002 keyword-search videos were available throughout the second phase of data collection, which is 67.13% and 55.23% of the total numbers of videos in those datasets respectively. Perhaps the higher average age of the keyword-search videos may explain, at least in part, the lower percentage of available videos. During the second phase of data collection, the 20,000 recently-uploaded videos received almost 5 million additional views, and the 627,002 available keyword-search videos receive about 1.8 billion additional views.



(a) Recently-uploaded

(b) Keyword-search

Figure 3.1: Age bias in the datasets

3.3 Challenges and Limitations of Data Collection

1. Data Collection Challenges: Meta-data collection for a representative set of random videos from YouTube via a crawler is challenging due to the huge and continually-increasing number of available videos. There is no YouTube API call that returns video IDs randomly selected over the entire video collection, so some other sampling approach must be used.

The choice of a one-week sampling period instead of a day or a month is also a critical consideration. Too short a sampling period may result in difficulties owing to API call rate limitation policies adopted by Google. On the other hand, a one month sampling period may be too long to observe some critical characteristics like time-to-peak popularity.

There are some other critical issues that make continuous data collection challenging for longer periods of time such as the possibility of computer, network, or power failures interrupting the data collection. Fortunately no such failures were experienced during the data collection for this thesis. Analysis of the datasets is also complicated by the long time interval between the first and second phase of data collection.

2. Potential Biases: Popularity and age biases are observed in the keyword-search dataset. As suggested by the fact that the keyword-search videos receive new views at a higher average rate than the recently-uploaded dataset videos (as shown in Tables 3.1 and 3.2), the keyword-search dataset is biased towards more popular videos. The keyword-search dataset is also biased towards younger videos, and popularity bias increases as age (time since uploaded) increases. The recently-uploaded dataset, however, appears to contain randomly-selected videos at most one week old. The age distribution of videos in the two datasets

is shown in Figure 3.1.

3. Clone Issues: Video popularity could be affected by the existence of related videos, or "clones" with essentially the same content. These effects are not considered in the analysis.

Chapter 4

Popularity Characterization

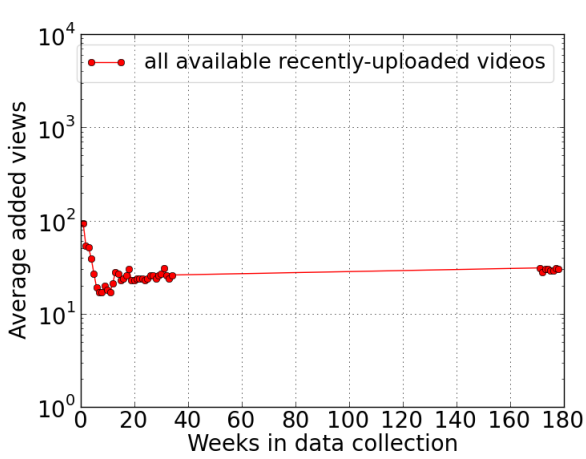
This chapter presents the popularity characterization on the recently-uploaded and keyword-search datasets. Obviously, only videos available in the second measurement period (as well as the first) are considered. For the recently-uploaded dataset, the main properties examined are the popularity evolution over age, popularity dynamics and churn and the time-to-peak distribution. The results from study of the time-to-peak distribution are also shown in Chapter 5, since they are used there for the development of the Borghol *et al.* model. For the keyword-search dataset, of particular interest is the popularity evolution as videos age for both different age bins and popularity bins, and implications with respect to possible biases in datasets obtained using keyword search.

4.1 Recently-uploaded Dataset

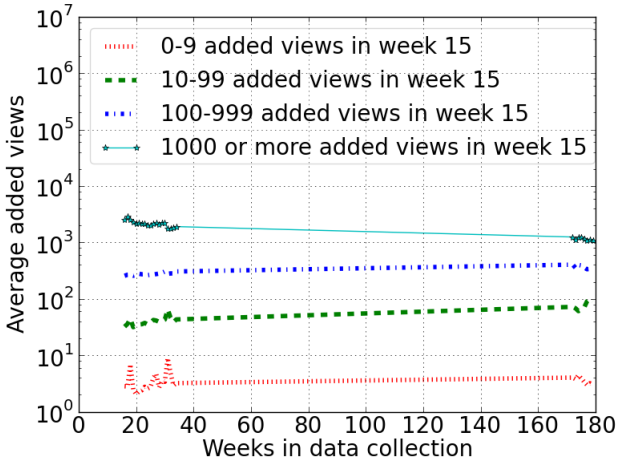
Some basic popularity characteristics for the recently-uploaded dataset are already observed by Borghol *et al.* [5]. However, those analyses are done on the data collected in the first measurement period only. The main goal of studying those characteristics again is to observe whether or not they remain similar after a long period has passed since the videos were uploaded.

4.1.1 Average View Count for Each Week

This section presents the average added views of the recently-uploaded videos in different snapshots. The main goal for this analysis is improved understanding of the dataset and the observation of high-level popularity characteristics and their evolution from the first measurement period to the second measurement period. Figure 4.1(a) shows the average added views at different snapshots. Note that points are plotted only for the weeks for which snapshots were taken, i.e., the weeks of the first and second measurement periods. The points are connected by a line, spanning the first and second measurement periods as well as the gap in between, so as to more clearly show rates of change. This figure shows that the average viewing rate for the videos in the recently-uploaded dataset has not decreased by the time of the second measurement period. In fact, it slightly increased. This is somewhat surprising; one might expect that the average viewing rate of these videos would decrease as the videos age, particularly in the light of the vast amount of new content



(a) Average added views in each week



(b) Average added views for videos binned by added views in week 15

Figure 4.1: Average view count for the recently-uploaded videos

uploaded over this time period. A possible contributing factor is growth in the YouTube user population.

To gain a better understanding of the evolution of the average viewing rate, videos are separated into bins according to their view count during week 15. The main reason that week 15 is chosen is to avoid the higher popularity churn at earlier video ages. The results shown in Figure 4.1(b) are again somewhat surprising. It appears that the average viewing rate of videos with intermediate popularity increases, evidentially resulting in the increased overall average viewing rate. The average viewing rate of the highly popular videos, however, decreases.

4.1.2 View Count Distribution

Examining the view count distribution (rather than just the average) enables a more detailed look at video popularity evolution throughout the time since the videos were uploaded. The main goal is to observe how the popularity distribution is impacted by video age and the changes in the popularity distribution between the first and second measurement period.

Figure 4.2 shows the complementary cumulative distribution function (CCDF) of the added views for the recently-uploaded videos at weeks 2, 5, 10, 25, 171 and 179, using a logarithmic scale for each axis. The figure shows that during the first measurement period, videos tend to receive more views at an early age than when they are older. The heavier right tail of the curve shows an order of magnitude more new views in week 2 than in weeks 10 and 25. However, when the video age is very high (week 171, 179) the heavier right tails of the curves again show substantially more new views for highly popular videos than occur in weeks 10 and 25.

For further analysis of the popularity distribution, videos are separated into various bins according

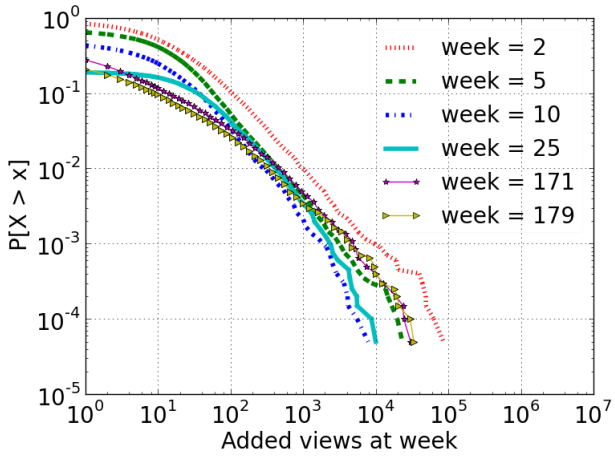


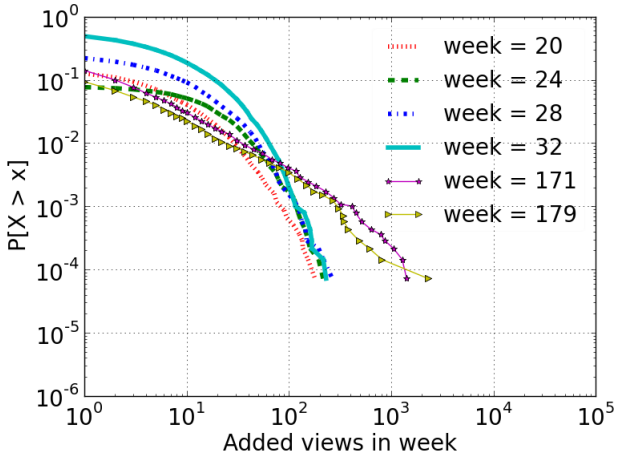
Figure 4.2: Distribution of added views for the recently-uploaded videos at different ages

to their views in week 15. As before, week 15 is chosen to avoid the higher popularity churn at earlier video ages. Separating the videos into bins allows study of video popularity evolution for videos of differing popularities. Figure 4.3 presents the CCDF of the added views in weeks 20, 24, 28, 32, 171 and 179 using a logarithmic scale for each axis, for the videos in the different bins. Most notably, this figure shows that for all but the [1000-9999] popularity bin, the highest video viewing rates during the considered weeks of the second measurement period (weeks 171 and 179) are an order of magnitude higher than those of the first measurement period weeks.

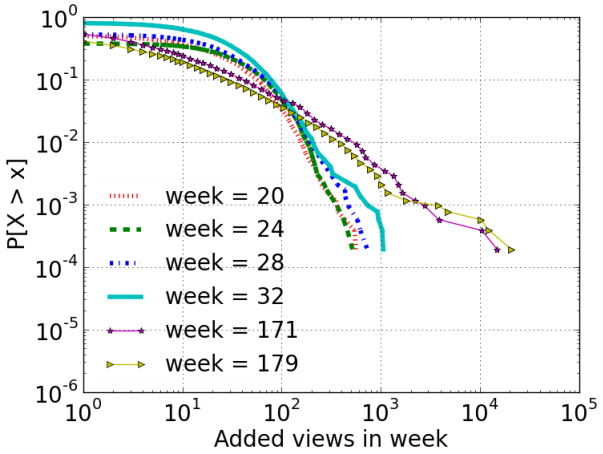
4.1.3 Popularity Dynamics and Churn

Popularity dynamics and churn analysis is an important study of this thesis. Of interest is the observation of how popularity dynamics patterns and churn change between the first and second measurement period. The accuracy of future popularity prediction depends on the stationarity of the current popularity observation [41], therefore observing churn is important. Also, if the popularity dynamics patterns are not significantly changed between the first and second measurement period, then the popularity evolution model of Borghol *et al.* [5] may be applicable across a much longer time period than considered in that work.

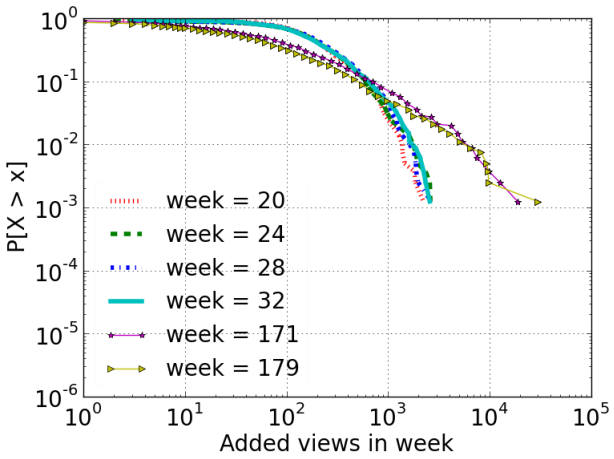
Figures 4.4 and 4.5 illustrate the popularity churn in the recently-uploaded dataset by looking at the added views in adjacent weeks. Figure 4.4 shows scatter plots of added views at different pairs of consecutive weeks. Note that the points are spread out more for early video ages than for the later ages. For example, highly unpopular videos during week 2 could be highly popular in week 3, and vice versa. Videos with less than 100 added views in week 2 could receive more than 1,000 added views in week 3. As the videos age, churn decreases, and becomes lowest at the maximum age considered (Figure 4.4(f)). This implies that view counts are highly variable from one week to the next early in the video's lifetime, but become more stable as videos age.



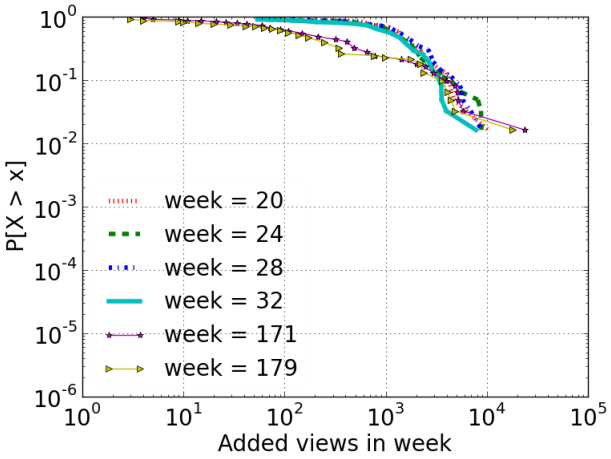
(a) 0-9 added views in week 15



(b) 10-99 added views in week 15

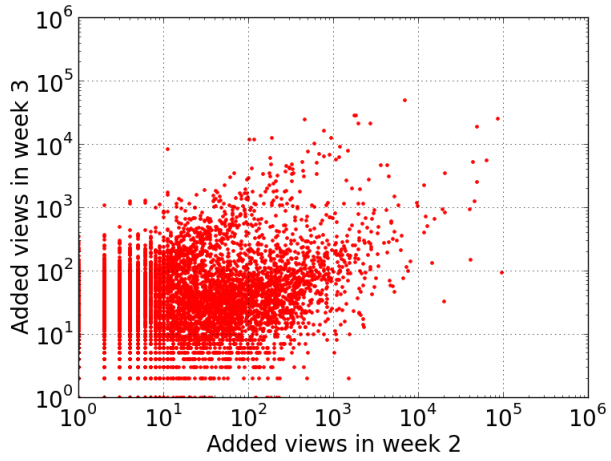


(c) 100-999 added views in week 15

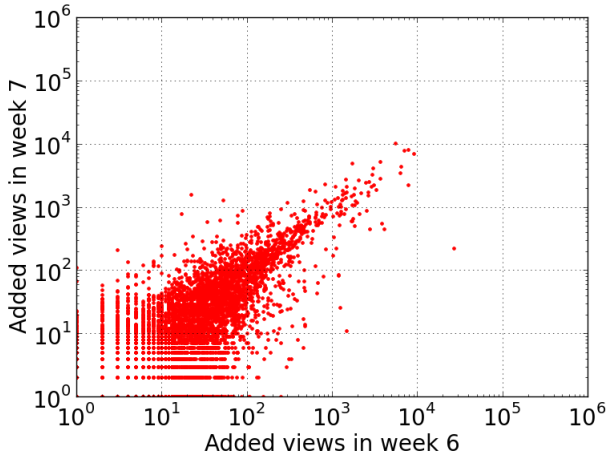


(d) 1000-9999 added views in week 15

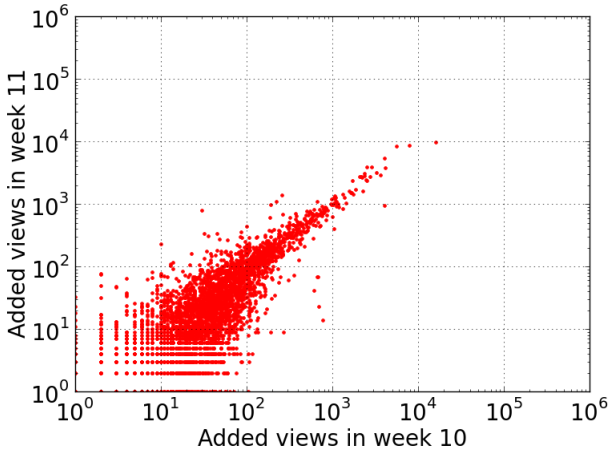
Figure 4.3: Distribution of added views for the recently-uploaded videos binned by added views in week 15



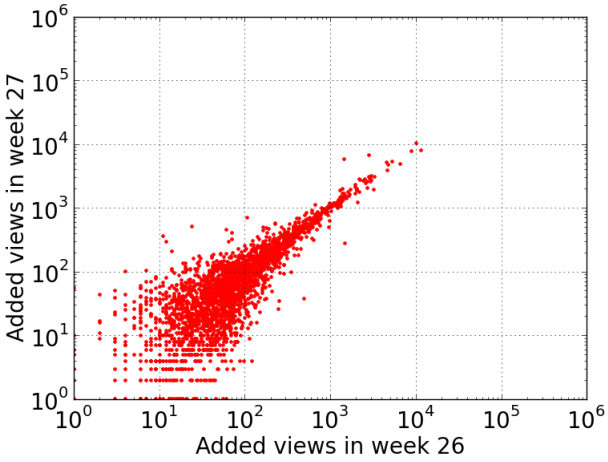
(a) Week 2 vs 3



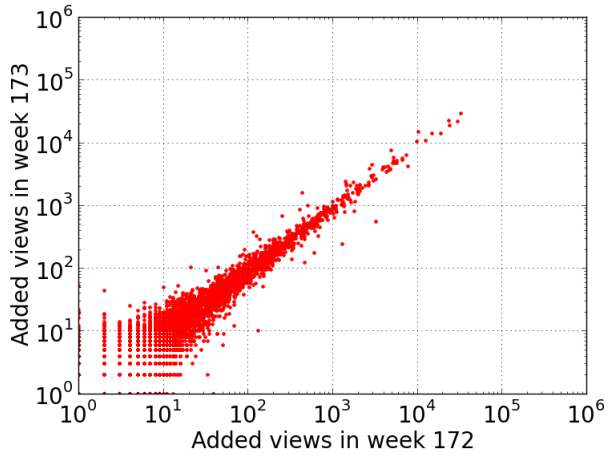
(b) Week 6 vs 7



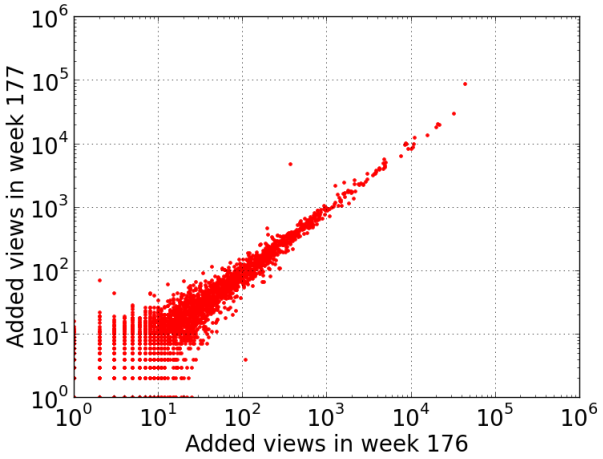
(c) Week 10 vs 11



(d) Week 26 vs 27

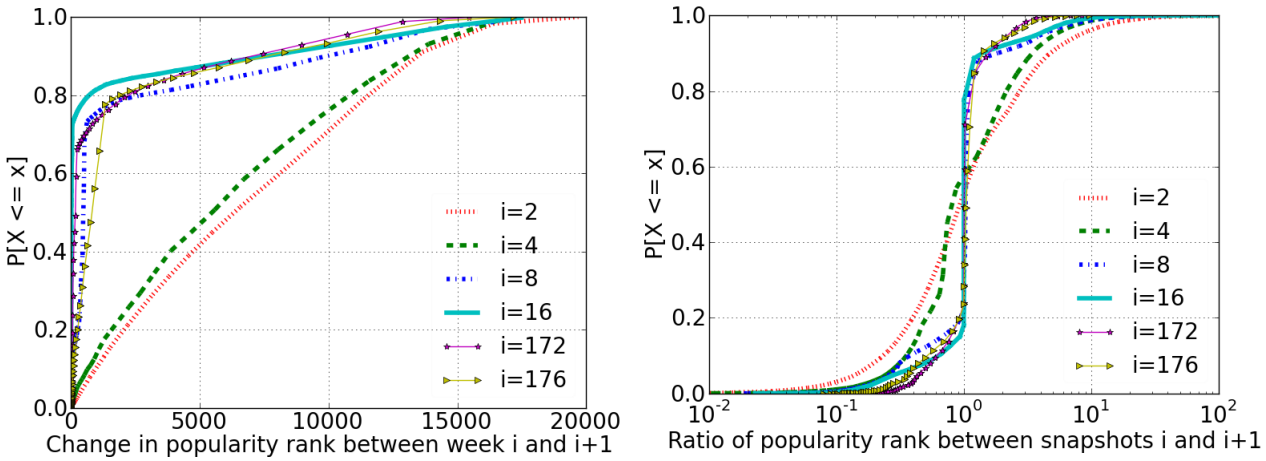


(e) Week 172 vs 173



(f) Week 176 vs 177

Figure 4.4: Scatter plot of added views for the recently-uploaded videos in week i vs week i+1



(a) Absolute change in popularity rank

(b) Ratio of new to old popularity rank

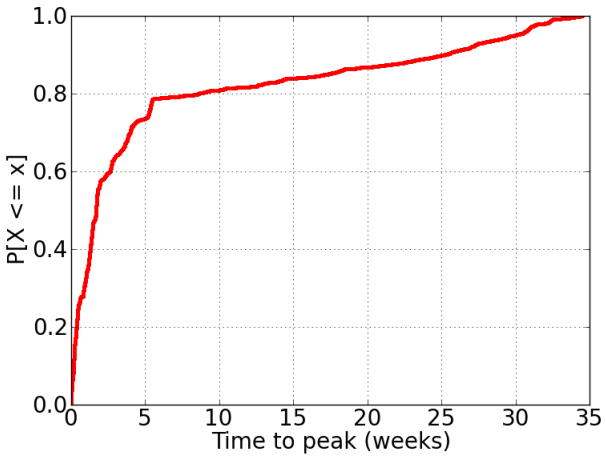
Figure 4.5: Distribution of change in popularity ranks of the recently-uploaded videos

In Figure 4.5(b), popularity churn is observed by looking at the distribution of the absolute value of the video popularity rank change across consecutive weeks. For each week in the measurement period videos are ranked according to the number of views added in that week, with the video with the most added views at rank 1, the video with the next highest number of views at rank 2, and so on. Videos with the same number of added views are ranked according to their ordering in the list of the videos as created by Borghol *et al.* [5]. Figure 4.5(a) shows that popularity churn is greater when videos are newly-uploaded. As shown in this figure, close to 30% of the videos change rank by more than 10,000 positions between week 2 and 3 of the measurement period. This amount of rank change is observed for less than 10% and 8% of the videos between weeks 8 and 9, and weeks 176 and 177, respectively.

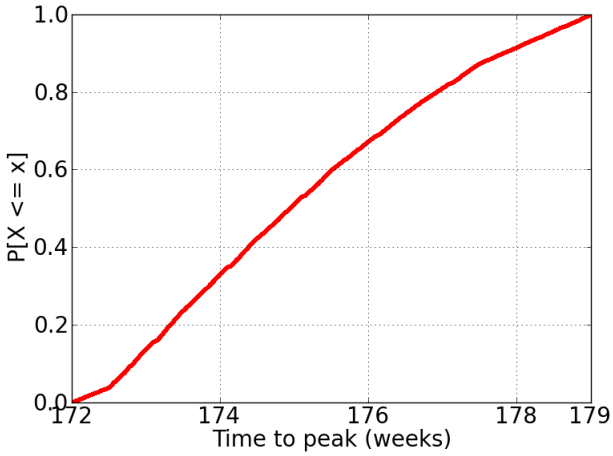
Figure 4.5(b) shows the distribution of the ratio of popularity ranks, across consecutive weeks. Looking at ratios rather than differences is motivated by the fact that a change from rank 1 to rank 10, for example, is much more significant than a change from rank 10,000 to rank 10,010. As in Figure 4.5(a), in Figure 4.5(b) greater churn is seen for younger video ages. For example, over 50% of the videos increase, or decrease, their popularity rank by at least a factor of two between week 2 and 3. The corresponding percentage for weeks 176 and 177 is less than 20%.

4.1.4 Time-to-peak Distribution

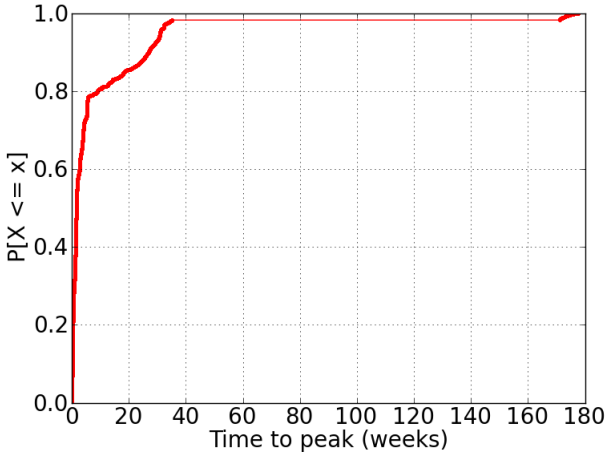
Looking at the popularity dynamics and churn, videos experience higher churn at their early ages than when older, and the degree of popularity churn decreases as video age increases. Widely differing rates at which videos achieve their peak popularity could be a major reason for such popularity churn. Figure 4.6 shows the CDF of the time-to-peak (i.e., the time until a video achieves its highest observed weekly number of added views) for each measurement period individually as well as across both measurement periods. The



(a) Time-to-peak for first measurement period only



(b) Time-to-peak for 2nd measurement period only



(c) Time-to-peak for both measurement periods

Figure 4.6: Time-to-peak distribution of the recently-uploaded videos

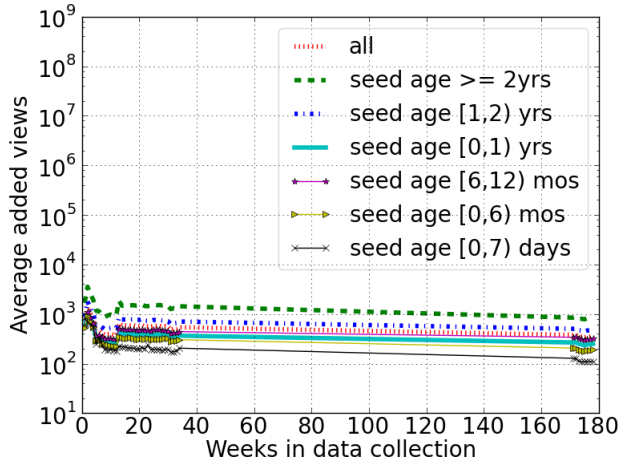
time-to-peak is calculated for each video by comparing weekly view counts and keeping track of the week with the maximum weekly view count. Ties are broken by randomly picking one of these weeks. After determining a week in which the video attains its maximum weekly view count, the time-to-peak value is calculated using the video age, in the same way as described by Borghol *et al.* [5]. The initial partial week, from the video upload until the seed time, is also handled as described by Borghol *et al.* [5]. Figure 4.6(c) shows that most videos (almost 80%) reach their peak popularity (over the weeks in the first and second measurement period) within their first six weeks since upload. The time to achieve peak popularity for the remaining 20% of the videos is approximately uniformly distributed throughout the rest of the observed weeks. Note that some videos received their peak weekly number of views long after upload, during the second measurement period (from week 171 to 179). Considering just this second measurement period, the time-to-peak is approximately uniformly distributed as shown in Figure 4.6(b). It should be emphasized that only the weekly view counts in the measurement periods are considered in Figure 4.6; any of the videos could have its actual peak weekly view count during some week outside of these measurement periods.

4.2 Keyword-search Dataset

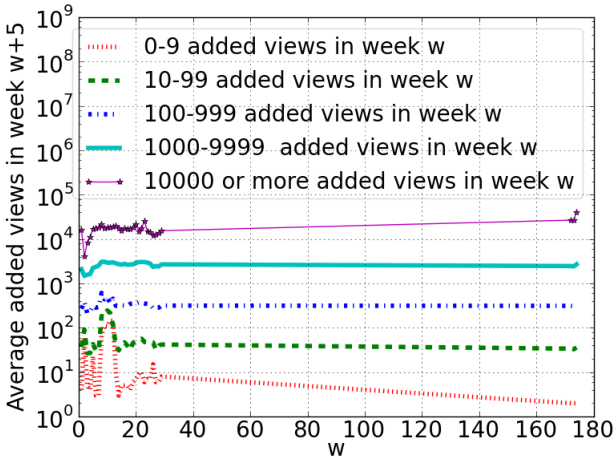
The keyword-search dataset is biased towards more popular videos. Here the popularity characteristics of the keyword-search dataset are analyzed in terms of the distribution of the weekly added views, and popularity dynamics and churn. Of particular interest is a detailed study of the popularity biases in the dataset.

4.2.1 Average View Count for Each Week

Before further analysis of the popularity characteristics of the keyword-search dataset, the average added views at each snapshot is observed. The average added views at each snapshot for both measurement periods is presented in Figure 4.7(a), with videos binned based on their age at seed time. This figure shows that the older videos in the dataset experience significantly higher viewing rate throughout the measurement periods than the young videos. Evidently, the older videos in the keyword-search dataset are not "typical" old videos, but instead ones with substantial popularity, with higher weekly view counts (on average) than both the young videos in the keyword-search dataset, and those in the recently-uploaded dataset. It is interesting to note that at the beginning of the first measurement period, average viewing rates are substantially higher than for subsequent weeks, even for the older videos that would be expected to have more stable popularity. This suggests that there is not only bias in the keyword-search dataset towards videos with higher long-term popularity, but also bias towards videos that happened to be particularly popular at the seed time. The highly non-stationary average viewing rates at the beginning of the first measurement period appear to have stabilized by week 15; as seen in Figure 4.7(a), the average weekly view counts are initially relatively high,



(a) Average added views in each week for videos binned by age at seed time



(b) Average added views in week w+5 for videos 25 weeks old at seed time, binned by added views in week w

Figure 4.7: Average added views for the keyword-search videos

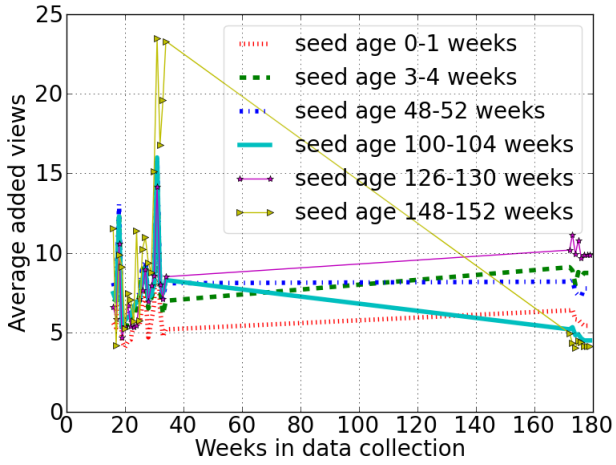
then drop sharper lower, before rising again and finally stabilizing (by week 15) with a gradually decreasing trend.

Nonstationarity in the average viewing rates early in the first measurement period is also shown in Figure 4.7(b), which presents the average added views in week w+5 for videos 25 weeks old at seed time, binned according to their added views during week w, as a function of w. Nonstationarity is particularly evident for the unpopular videos (0-9 and 10-99 view bins). Further analysis on the average added views at each snapshot is conducted by binning on both age at seed time and popularity in week 15, with results shown in Figure 4.8. Since popularity of the videos is highly non-stationary early in the measurement period, week 15 is considered for separating videos into popularity bins. The figure shows that the average added views for the bins is highly variable during the first measurement period, and much less variable during the second measurement period. For some view bins (0-9, 1000-9999 and 10000 or more), the seed age 148-152 bin has highly variable added views during both the first and second measurement period. This is due to the limited number of videos in that category.

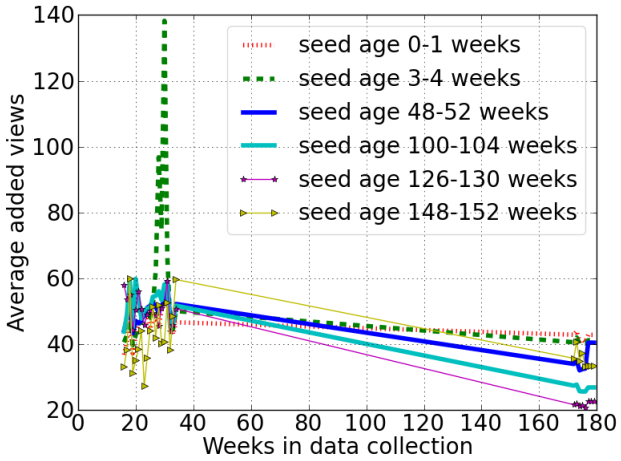
4.2.2 View Count Distribution

View count distribution analysis considers the video popularity distribution and how it changes over the measurement periods. Such an analysis can shed further light on the biases in the keyword-search dataset.

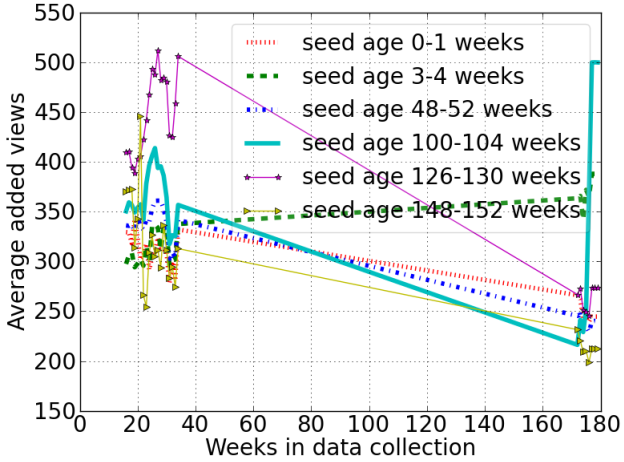
Figure 4.9 presents the CCDF of added views for different weeks of the measurement period, using a logarithmic scale for each axis. As seen in the figure, the most popular videos early in the first measurement period (weeks 2 and 5) have considerably more added views than the most popular videos later in the measurement periods. This is consistent with the results in Figure 4.7(a). Note for the later weeks (weeks



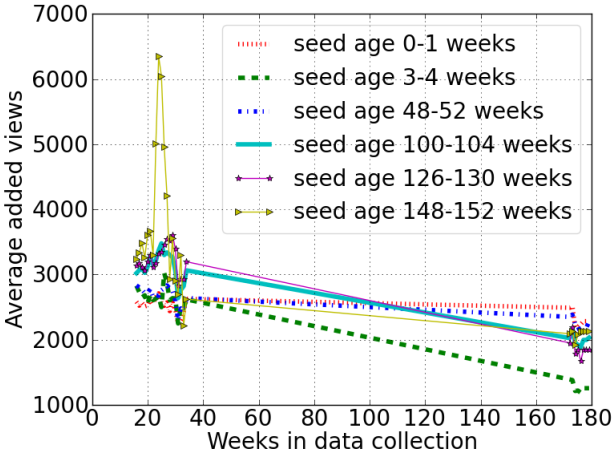
(a) 0 to 9 added views in week 15



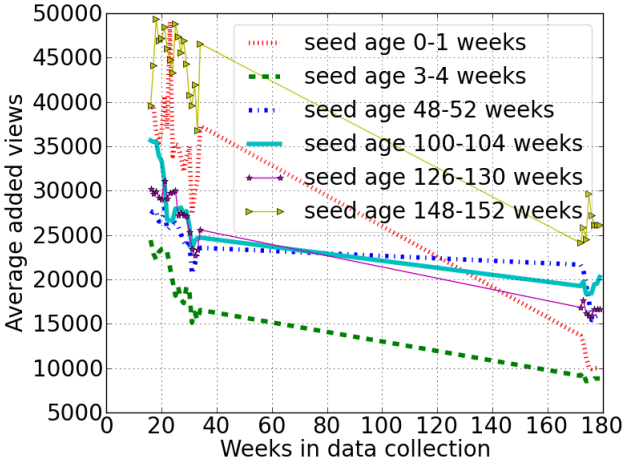
(b) 10 to 99 added views in week 15



(c) 100 to 999 added views in week 15



(d) 1000 to 9999 added views in week 15



(e) 10000 or more added views in week 15

Figure 4.8: Average added views for the keyword-search videos binned by added views in week 15, and by age at seed time

10, 25, 171 and 179), the numbers of added views for the most popular videos in these weeks are very similar.

For further analysis of the distribution of added views and possible changes between the first and second measurement periods, Figure 4.10 shows the CCDF of added views for weeks 15 and 171, with the videos binned by age at seed time. Consistent with Figure 4.7(a), Figure 4.10(a) shows that the older videos in the keyword-search dataset tend to have higher viewing rates, evidently owing to biased selection of these videos. Comparing Figure 4.10(a) and 4.10(b), similar trends are observed, but a significantly larger fraction of the keyword-search videos in each age bin receive at most one view in week 171, compared to in week 15. Note that the curves for the seed age 139-143 and seed age 152-156 bins do not extend as far as the others, owing to the relatively small numbers of videos in these bins.

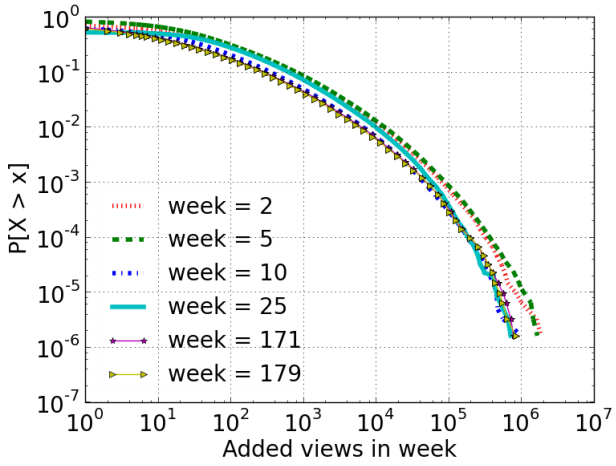
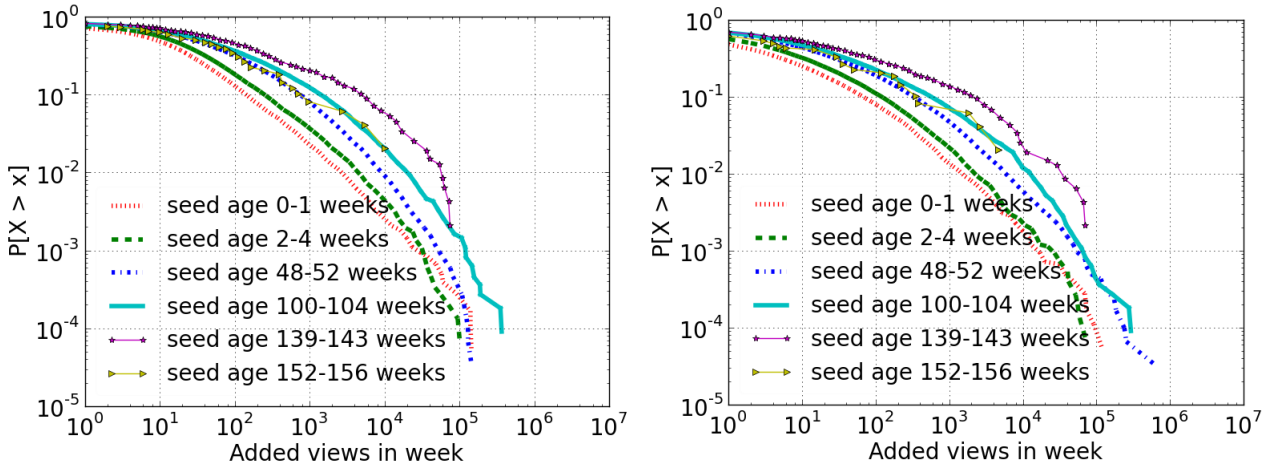


Figure 4.9: Distribution of added views for the keyword-search videos



(a) Added views in week 15

(b) Added views in week 171

Figure 4.10: Distribution of added views for the keyword-search videos binned by age at seed time

4.2.3 Popularity Dynamics and Churn

Popularity dynamics and churn analysis is a significant analysis on the keyword-search dataset, since it can give further insight into the biases in this dataset. Figure 4.11 shows the popularity stationarity in the keyword-search videos by a scatter plot of added views in adjacent weeks. The figure shows the change of added views between consecutive weeks. The scatter plot shows that there is greater variability in the added views from week to week shortly after the seed time (Figure 4.11(a)), compared to much later after the seed time (Figure 4.11(e)). Shortly after seed time, a large number of videos experience significant popularity variation from one week to the next week. Highly unpopular videos can become highly popular at the next week, and vice-versa. Therefore, the relative popularity is highly non-stationary. Another significant observation is the presence of two different clusters when added views are plotted for weeks 2 and 3 (Figure 4.11(a)). This may be due in part to young videos in this dataset, some of which have increasing popularity (upper cluster) and some of which have already reached their peak and have decreasing popularity (lower cluster). It may also reflect the presence of videos in this dataset that have elevated short term popularity, which quickly decreases (lower cluster).

The non-stationarity is also observed by looking at the rank change in adjacent weeks. Videos are ranked for each week according to the number of added views, with videos that have the same added view count ranked based on their ordering in the video list. Since video age may be an important factor in the rank change behaviour, Figure 4.12 presents the CDF of the absolute change in popularity rank for different age at seed time bins. Figure 4.12 shows that young videos experience much greater rank changes than old videos (note the differing scales on the x-axis). Also, for the same video age at seed time, videos experience greater rank changes shortly after seed time (i=2 or 4). This is seen even for the old videos (Figure 4.12(c) and 4.12(d)), reflecting their elevated short term popularity at seed time and the bias in this dataset. Interestingly, there are significant rank changes observed for old videos even long after the seed time. This may be due to large numbers of videos having very few added views, implying that small changes in the added view counts of these videos can cause large rank changes.

The rank change ratios are also observed by plotting the CDF of the ratio of new to old popularity rank for different age bins in Figure 4.13. As noted previously for Figure 4.5(b), this analysis is motivated by the fact that a change from rank 1 to rank 10, for example, is much more significant than a change from rank 10,000 to rank 10,010. As in Figure 4.12, larger rank changes are observed for young videos, and for shortly after the seed time. As expected, the rank changes are approximately symmetrically distributed between increasing and decreasing changes. For example, for 0-1 weeks seed age and measurement weeks 2 and 3 (Figure 4.13(a)), approximately 40% of the videos gain or lose a factor of 2 or more in popularity rank, compared to approximately 20% for weeks 172 and 173 for the same seed age bin. However for old videos (e.g. 100 weeks or older), and measurement weeks long after the seed time (e.g. weeks 172 and 173), video popularity is almost stationary. As noted for Figure 4.12, even in these cases there is some observed churn

in ranks, probably owing to the large number of videos with very few added views.

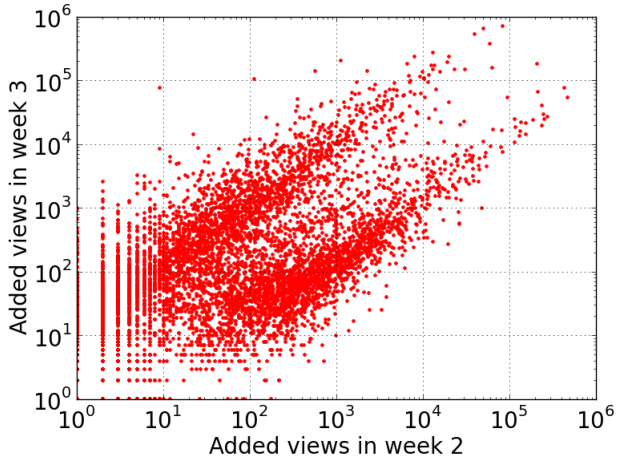
4.2.4 Impact of Age Biases

The keyword-search dataset likely contains more young videos than would be obtained with a random sampling (Figure 3.1). Understanding the impact of age biases requires an understanding of the impact of age on video popularity evolution.

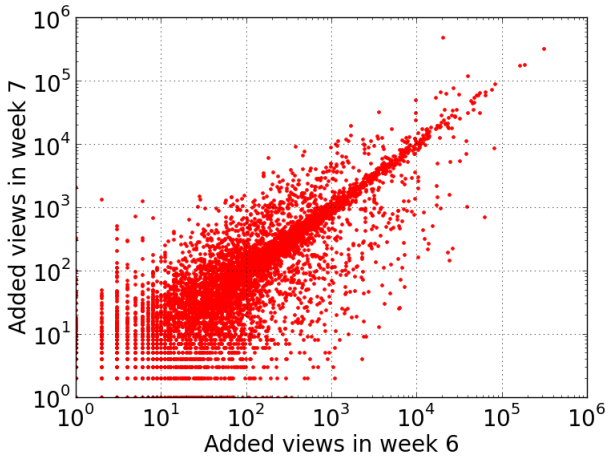
Figure 4.14 presents the CCDF of added views in weeks 2, 5, 10, 25, 171 and 179 with videos binned according to the age at seed time. For all age bins, generally in the later measurement weeks, the video added view distribution is weighted towards lower view counts, compared to in the earlier measurement weeks, although there are a few exceptions (such as week 10 for the videos 48 to 52 weeks old at seed time). This is particularly evident for the videos 0 to 1 weeks old at seed time (Figure 4.14(a)), where during the initial measurement weeks the most popular videos get an order of magnitude more new views than during the later measurement weeks. Such a big difference is not observed for the older age bins.

Of interest next is how the popularity evolution for videos in different age ranges varies for videos with different levels of popularity. As a first step in this direction, Figures 4.15 and 4.16 present the CCDF of added views in weeks 5, 10, 15, 25, 171 and 179 for videos in different popularity bins according to the added views at week 2 and at week 15 respectively, without yet binning on age. Video popularity is less stable early in the measurement period, owing to the presence in the dataset of videos with elevated short term popularity, as observed in the popularity dynamics and churn section. For this reason, videos are binned according to their added views in week 15, as well as week 2. Perhaps most noteworthy from Figure 4.15, where videos are binned according to their added views in week 2, is that for all bins the highest added view counts (in the weeks considered in figure) occur during the second measurement period, and/or in the earliest week among those considered in the first measurement period (week 5). In Figure 4.16, where videos are binned according to their added views in week 15, the highest added view counts mostly occur during the second measurement period. For two of the bins (0 to 9 added views and 1000 to 9999 added views), the differences between the highest added view counts (in the weeks considered in the figure) during the first versus the second measurement period are quite substantial. This characteristic of Figures 4.15 and 4.16 may reflect the impact of videos that take a long time to reach their peak popularity, and/or the growth in popularity of the YouTube site between the first and second measurement periods.

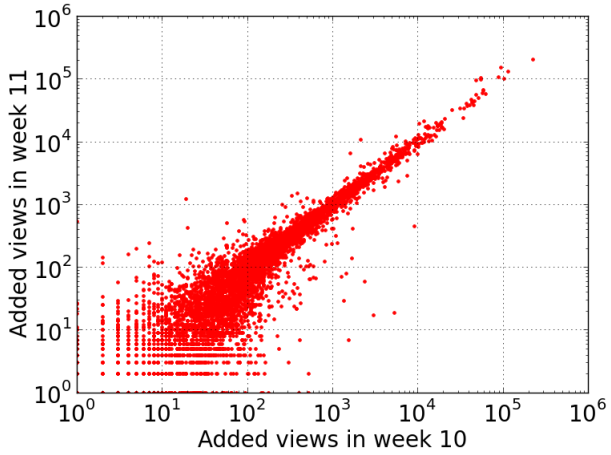
Figures 4.17, 4.18 and 4.19 present the CCDF of added views in weeks 20, 24, 28, 32, 171 and 179 with videos binned according to both age at seed time, and their added views during week 15. Results are shown only for the age bins 0-1 weeks (Figure 4.17), 3-4 weeks (Figure 4.18), and 48-52 weeks (Figure 4.19). Similarly as in Figure 4.16, for the older videos (Figure 4.19) the highest added view counts (in the weeks considered in the figures) mostly occur during the second measurement period. This is also the case for the younger videos in the first three or four popularity bins (Figures 4.17 parts (a), (b) and (c), and Figure 4.18



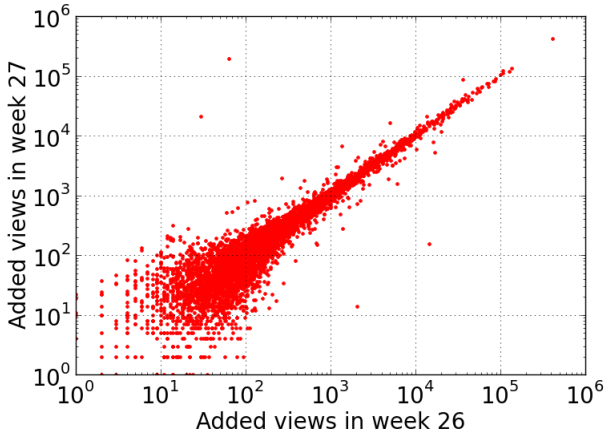
(a) Week 2 vs 3



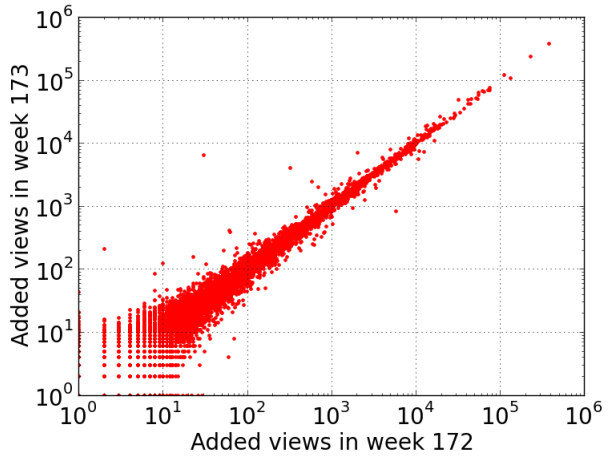
(b) Week 6 vs 7



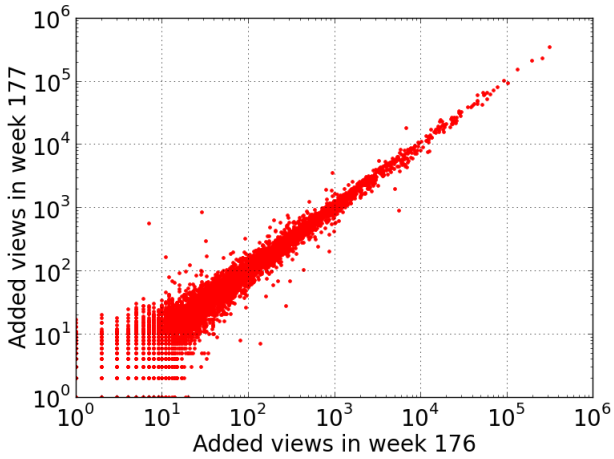
(c) Week 10 vs 11



(d) Week 26 vs 27

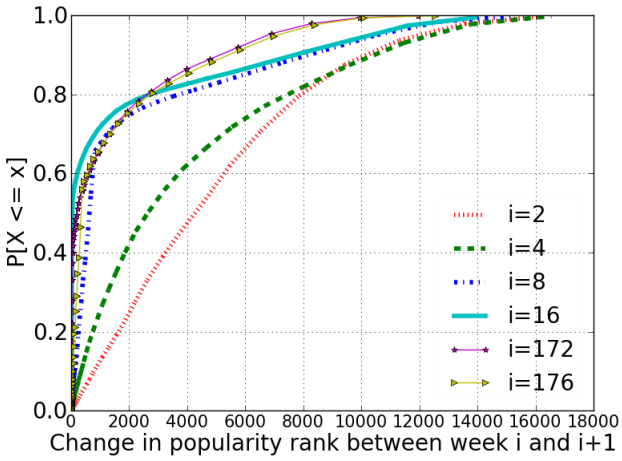


(e) Week 172 vs 173

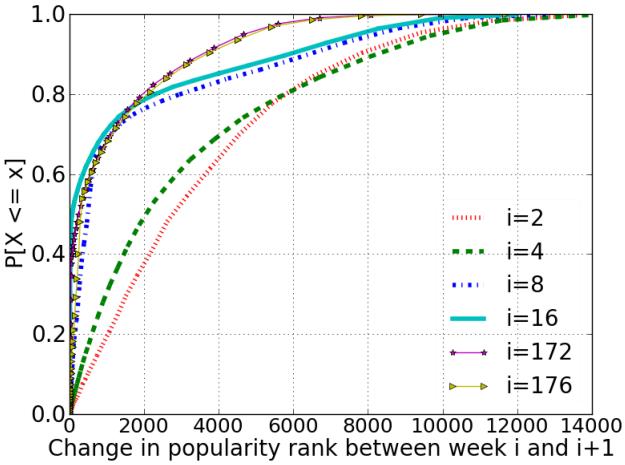


(f) Week 176 vs 177

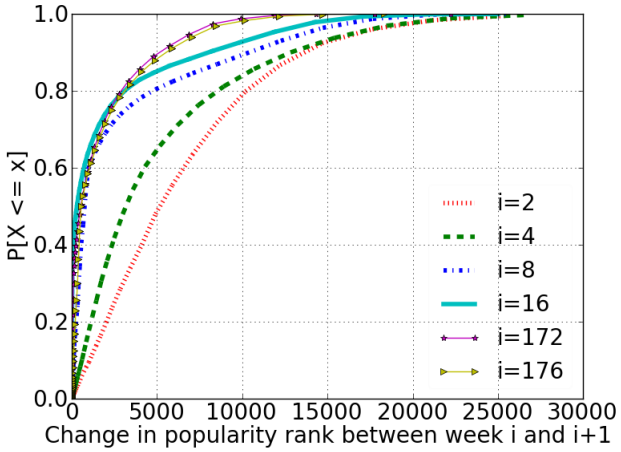
Figure 4.11: Scatter plot of added views for the keyword-search videos in week i vs week i+1



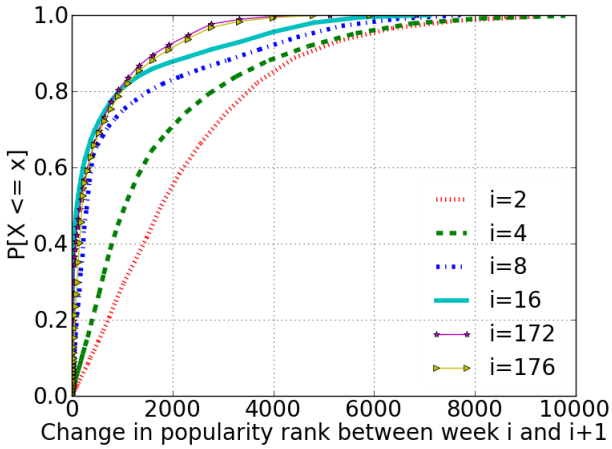
(a) 0 to 1 weeks old at seed time



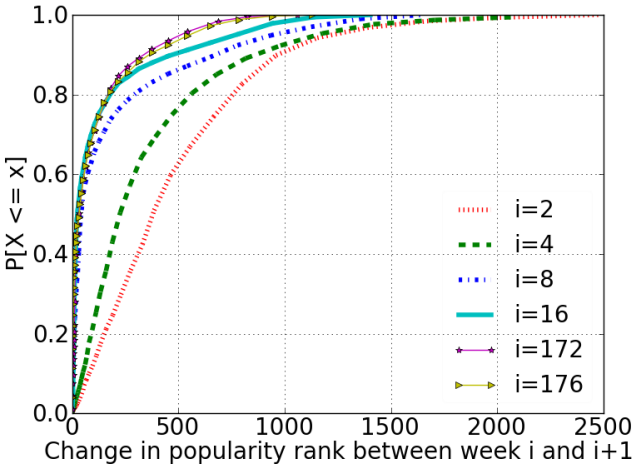
(b) 3 to 4 weeks old at seed time



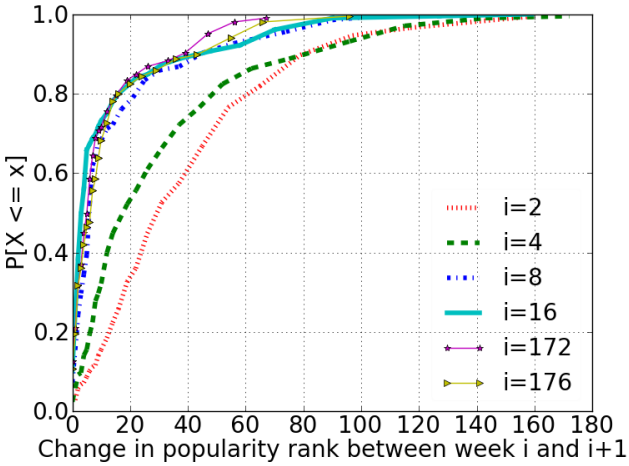
(c) 48 to 52 weeks old at seed time



(d) 100 to 104 weeks old at seed time

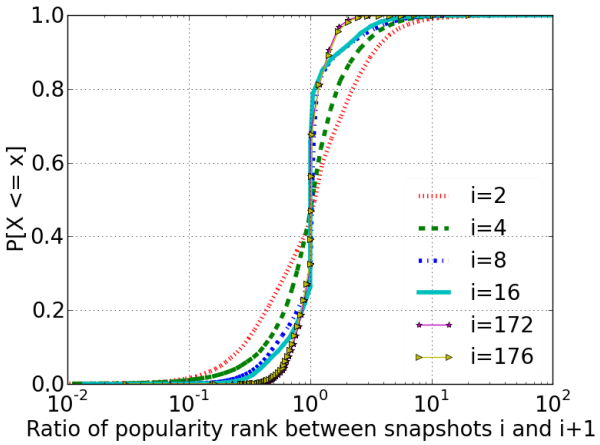


(e) 126 to 130 weeks old at seed time

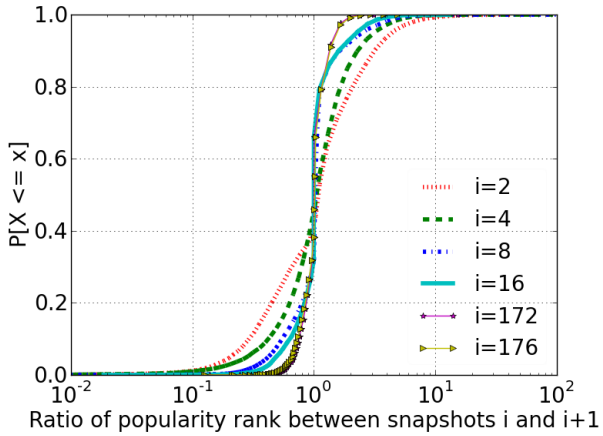


(f) 146 to 156 weeks old at seed time

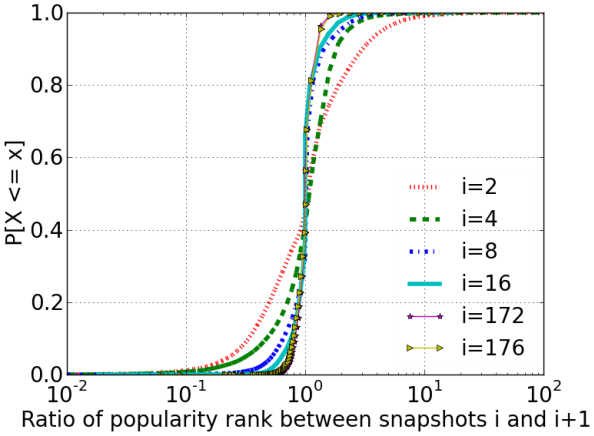
Figure 4.12: Distribution of absolute change in popularity rank for the keyword-search videos binned by age at seed time



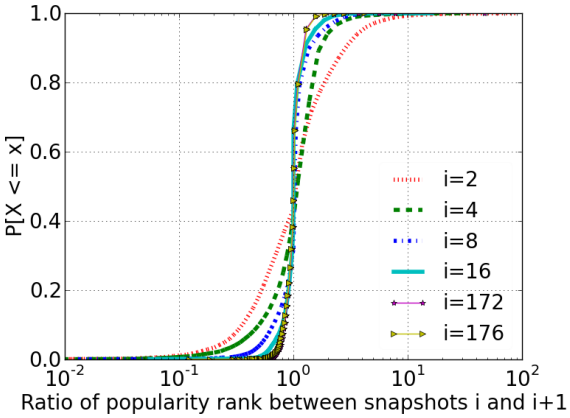
(a) 0 to 1 weeks old at seed time



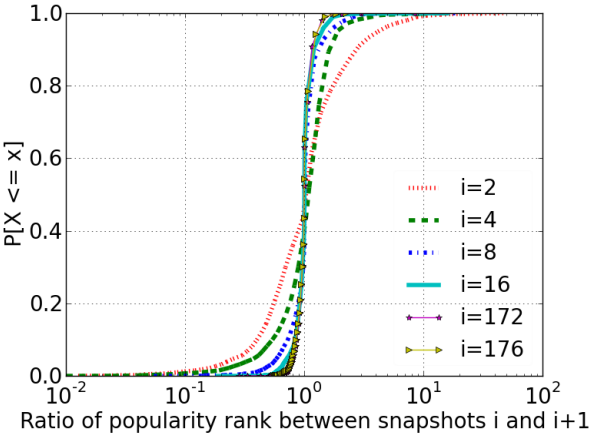
(b) 3 to 4 weeks old at seed time



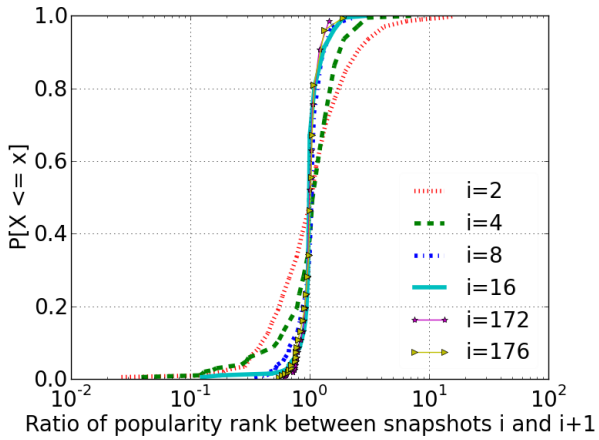
(c) 48 to 52 weeks old at seed time



(d) 100 to 104 weeks old at seed time

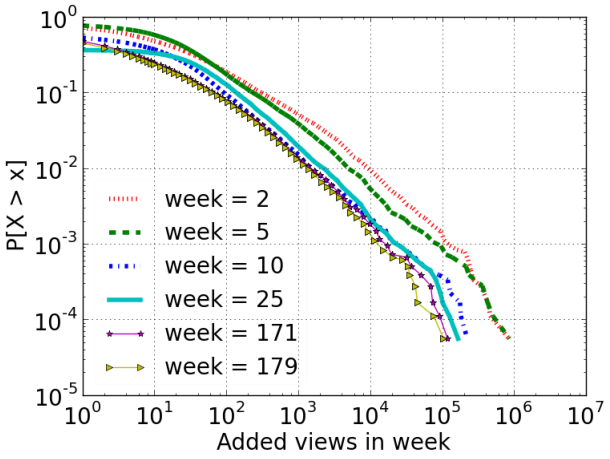


(e) 126 to 130 weeks old at seed time

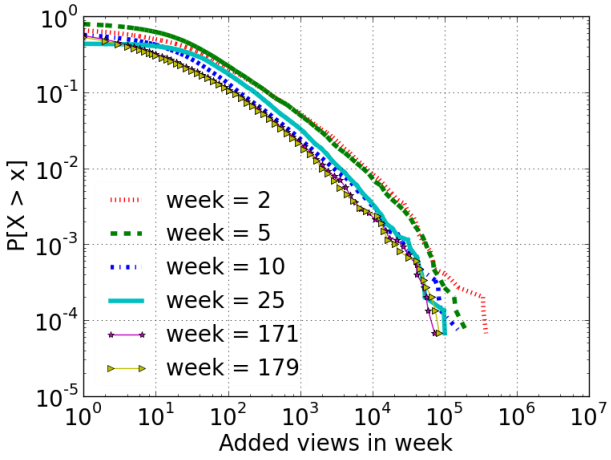


(f) 146 to 156 weeks old at seed time

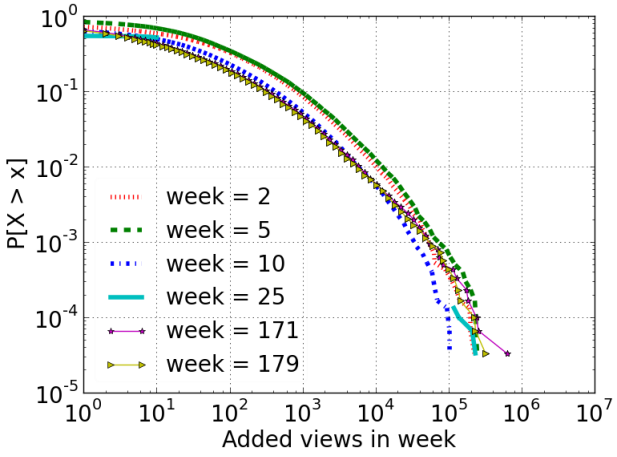
Figure 4.13: Distribution of ratio of new to old popularity rank for the keyword-search videos binned by age at seed time



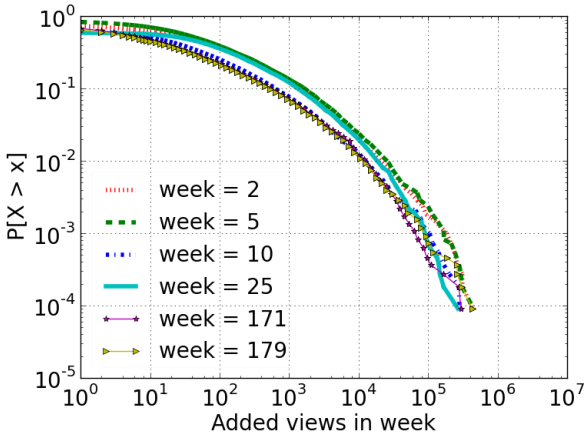
(a) 0 to 1 weeks old at seed time



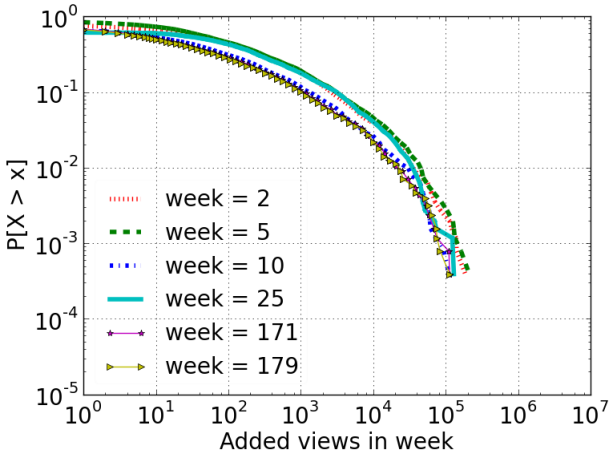
(b) 3 to 4 weeks old at seed time



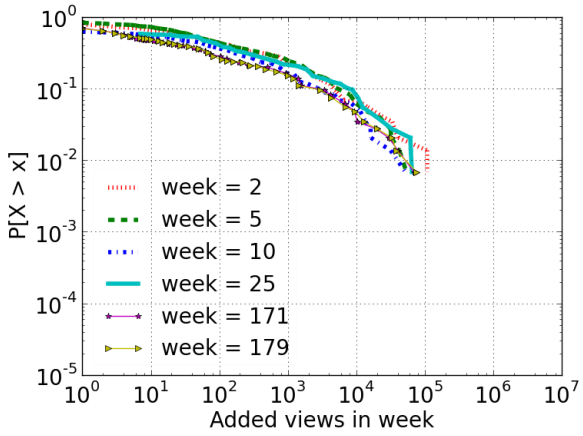
(c) 48 to 52 weeks old at seed time



(d) 100 to 104 weeks old at seed time

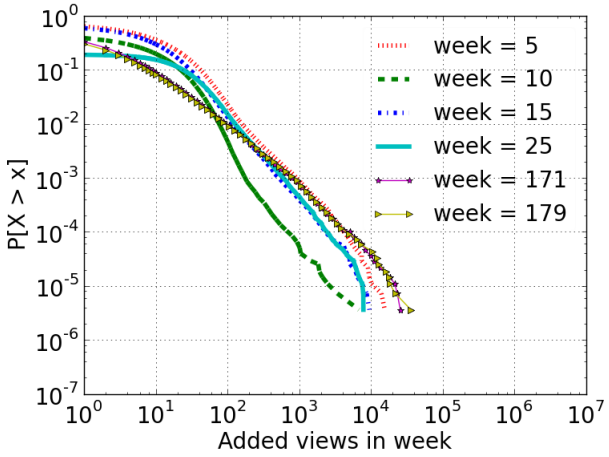


(e) 126 to 130 weeks old at seed time

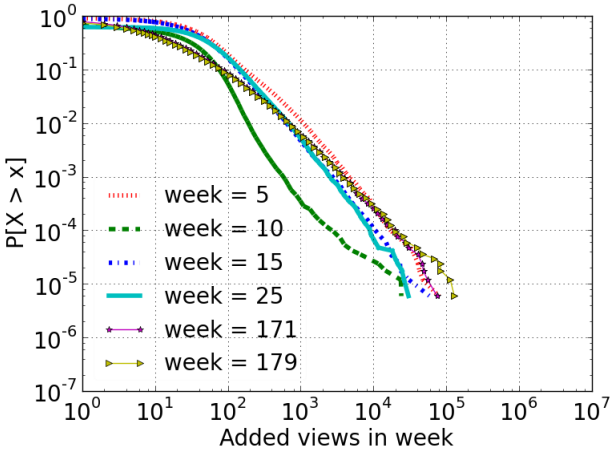


(f) 146 to 150 weeks old at seed time

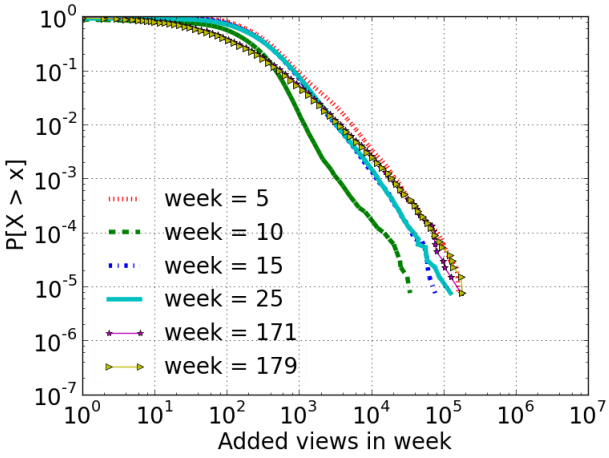
Figure 4.14: Distribution of added views for the keyword-search videos binned by age at seed time



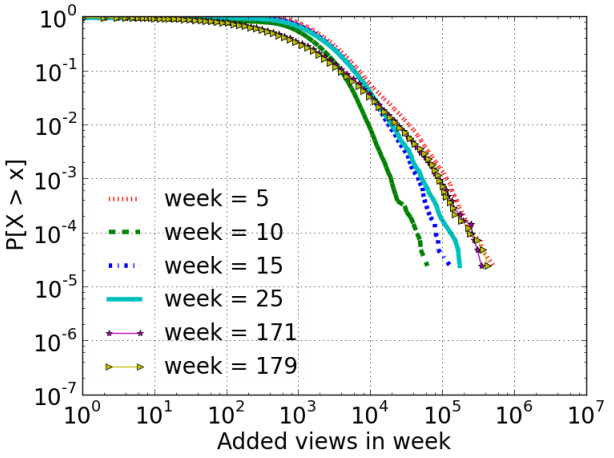
(a) 0 to 9 added views in week 2



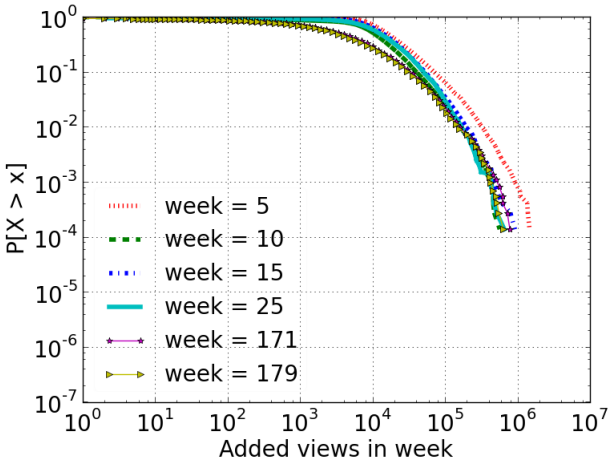
(b) 10 to 99 added views in week 2



(c) 100 to 999 added views in week 2

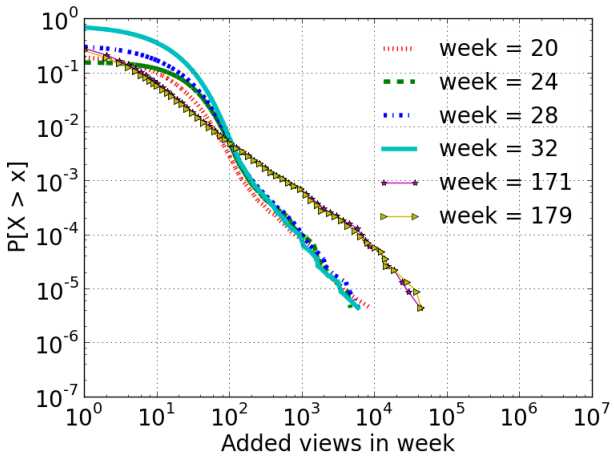


(d) 1000 to 9999 added views in week 2

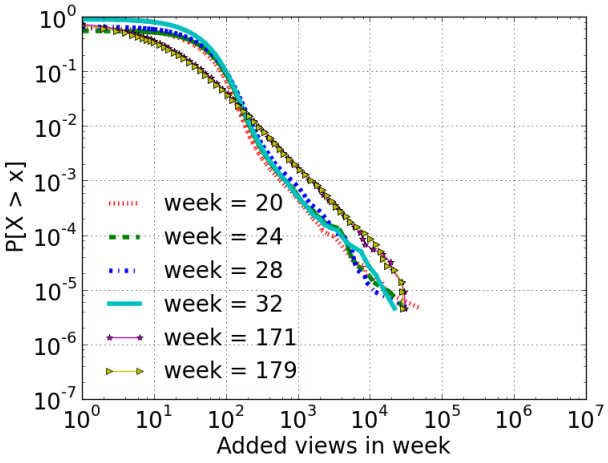


(e) 10000 or more added views in week 2

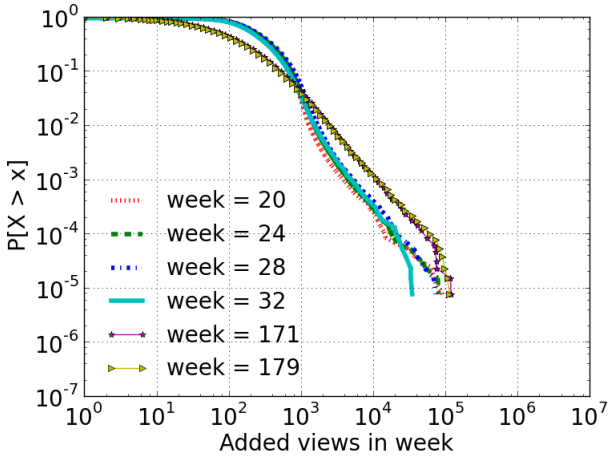
Figure 4.15: Distribution of added views for the keyword-search videos binned by added views in week 2



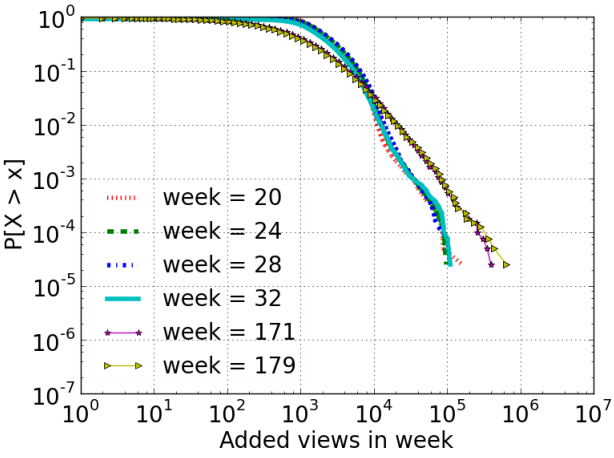
(a) 0 to 9 added views in week 15



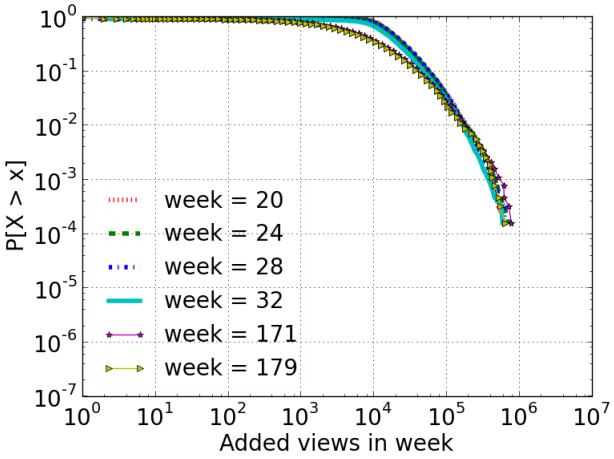
(b) 10 to 99 added views in week 15



(c) 100 to 999 added views in week 15



(d) 1000 to 9999 added views in week 15



(e) 10000 or more views

Figure 4.16: Distribution of added views for the keyword-search videos binned by added views in week 15

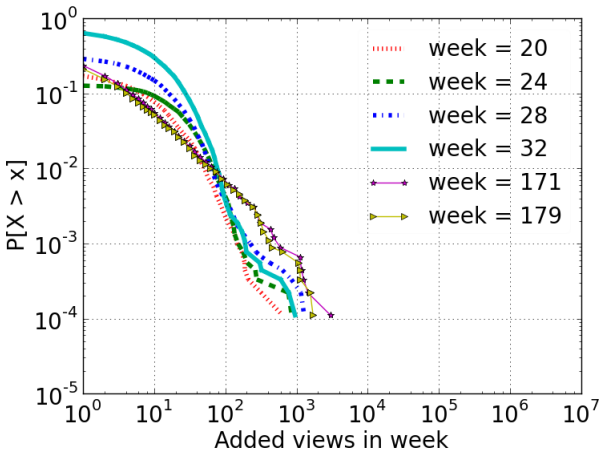
parts (a), (b), (c) and (d)).

For the most popular young videos, however (Figures 4.17 parts (d) and (e) and Figure 4.18 part (e)), the highest added view counts occur during the first measurement period. These latter videos may have experienced their peak popularity early in their lifetime, whereas the former videos may experience higher added view counts during the second measurement period owing to having taken a long time-to-peak and/or owing to the growth in popularity of the YouTube site.

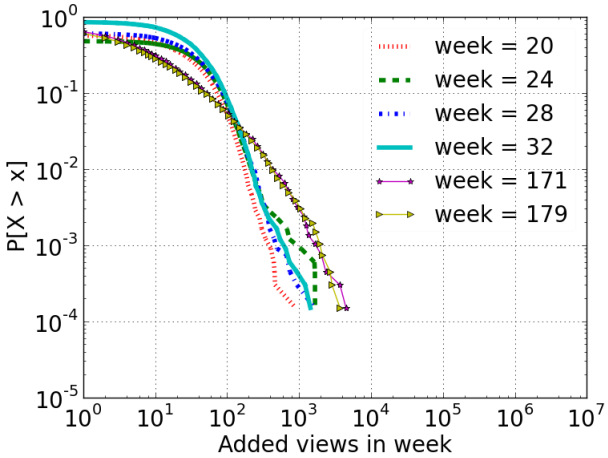
4.2.5 Independence of Popularity Biases and Video Age at Seed Time

This section considers the question of whether the longer-term popularity biases observed in the keyword-search dataset (i.e., after the initial 14 or 15 weeks of the first measurement period) are dependent on the video age at seed time. For this purpose, Figures 4.20, 4.21, 4.22 and 4.23 show the CCDF of added views for videos of approximately the same age in the measurement week for which their added view count is considered. For example, Figure 4.20 includes curves for videos 2 weeks old at seed time during the measurement week 22, 4 weeks old at seed time during the measurement week 20, and so on. In all cases, the added view counts are for videos 24 weeks old. Interestingly, although there is some variability between the curves (most notably, there are evidently some particularly popular videos in the dataset that were 6 weeks old at seed time), there appears to be no consistent trend with respect to the impact of the video age at seed time.

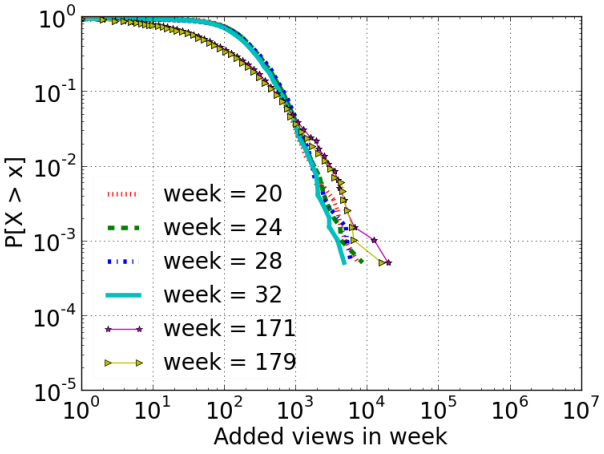
The impact of the video age at seed time is explored further in Figures 4.24 and 4.25. Similarly as in Figures 4.20 - 4.23, these figures show the CCDF of added views for videos of approximately the same age in the measurement week for which their added view count is considered. Now, however, videos are binned according to their added view count five weeks previous to the measurement week. It is observed that regardless of the seed age, for videos in the same popularity bin and with approximately the same age, the further popularity evolution is approximately the same (with some minor deviations, e.g., for the case of very unpopular videos in Figure 4.25(a)). These results suggest that the keyword-search dataset, even though it is biased towards more popular videos, could be used in studies of popularity evolution if videos are binned according to both age in the measurement week, and popularity some number of weeks previous. The fact that popularity evolution appears to be independent of the seed age suggests that the observed popularity evolution is representative of what one would see with randomly selected videos in the same age and popularity bins.



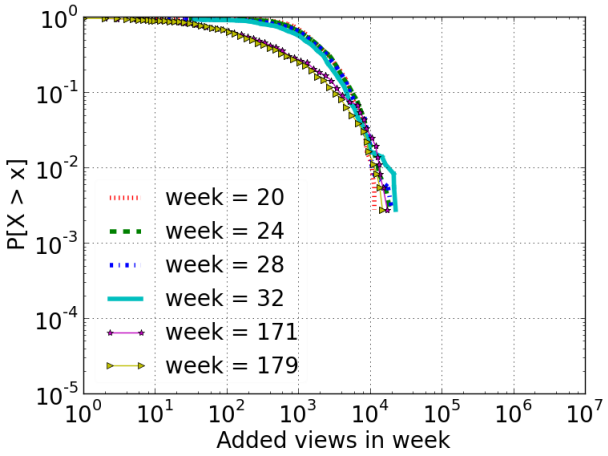
(a) 0 to 9 added views in week 15



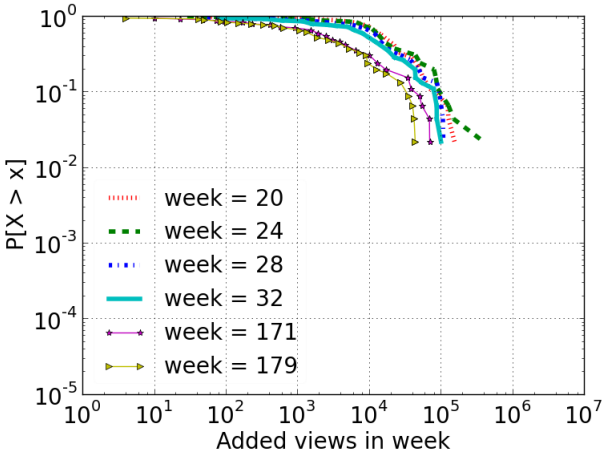
(b) 10 to 99 added views in week 15



(c) 100 to 999 added views in week 15

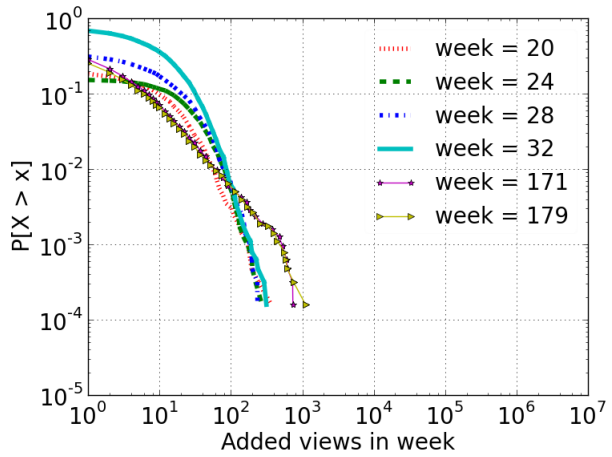


(d) 1000 to 9999 added views in week 15

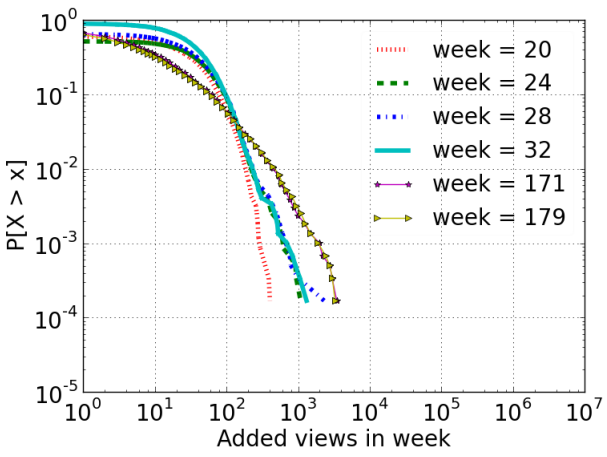


(e) 10000 or more added views in week 15

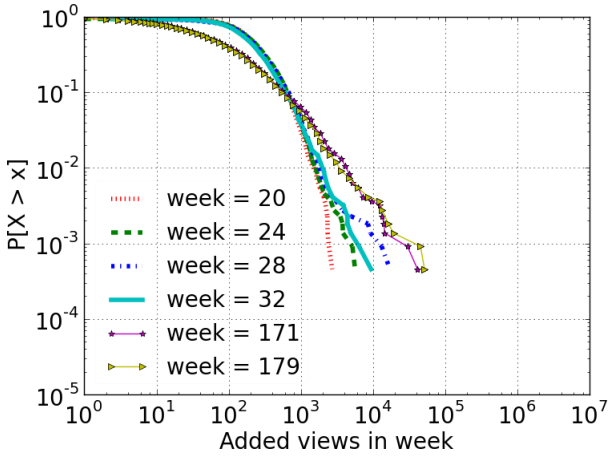
Figure 4.17: Distribution of added views for the keyword-search videos 0-1 weeks old at seed time, binned by added views in week 15



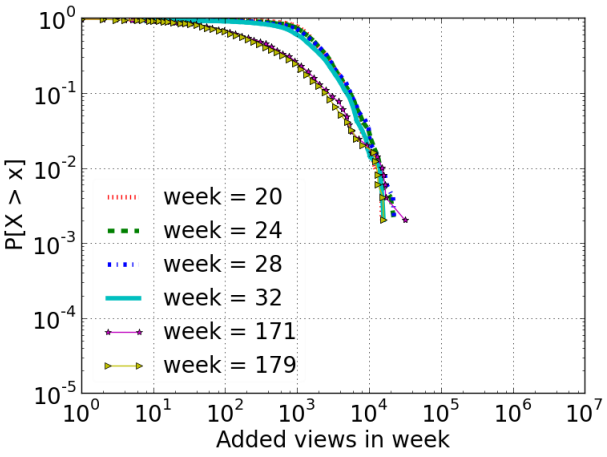
(a) 0 to 9 added views in week 15



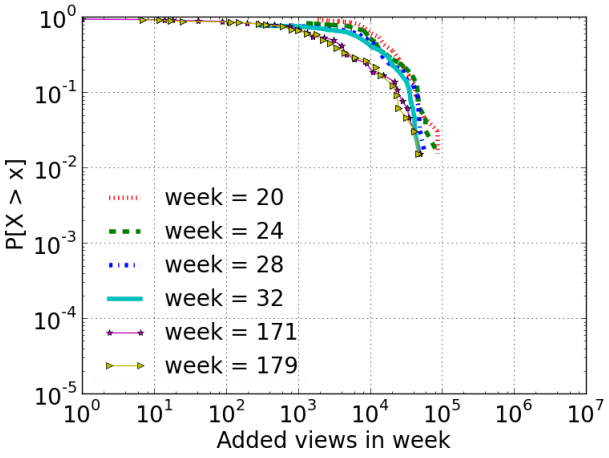
(b) 10 to 99 added views in week 15



(c) 100 to 999 added views in week 15

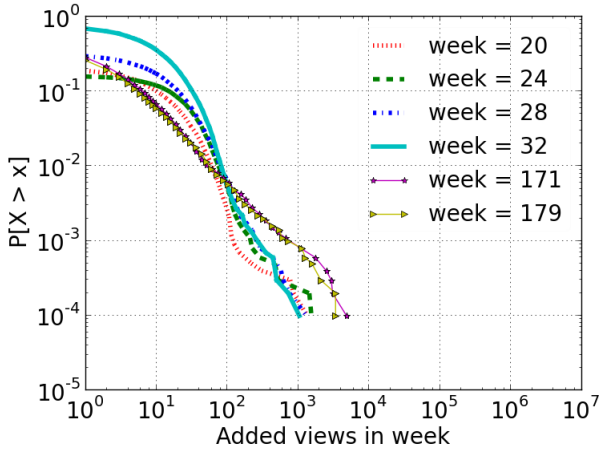


(d) 1000 to 9999 added views in week 15

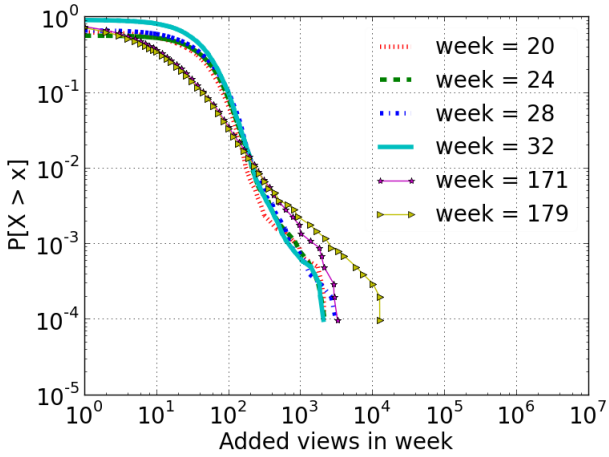


(e) 10000 or more added views in week 15

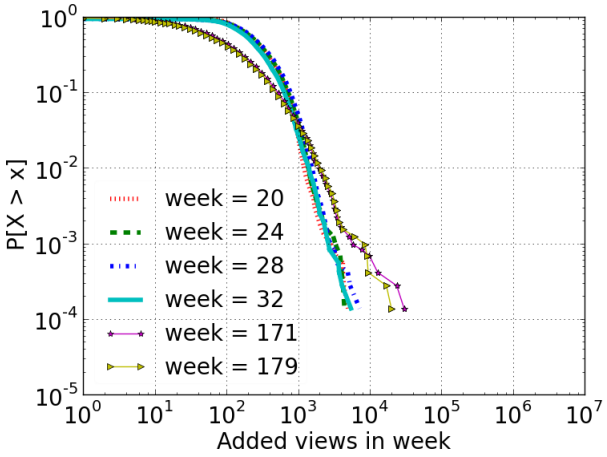
Figure 4.18: Distribution of added views for the keyword-search videos 3-4 weeks old at seed time, binned by added views in week 15



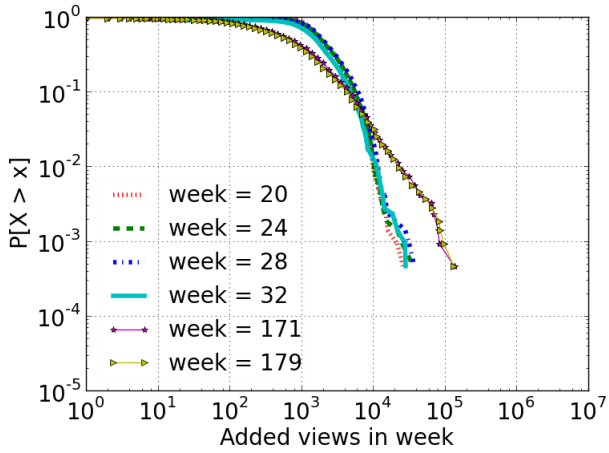
(a) 0 to 9 added views in week 15



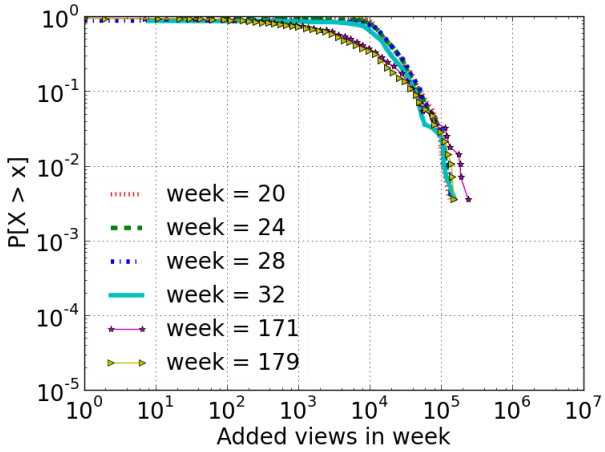
(b) 10 to 99 added views in week 15



(c) 100 to 999 added views in week 15

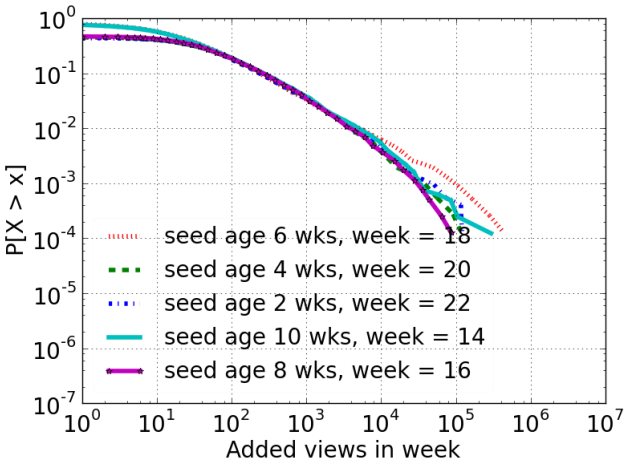


(d) 1000 to 9999 added views in week 15



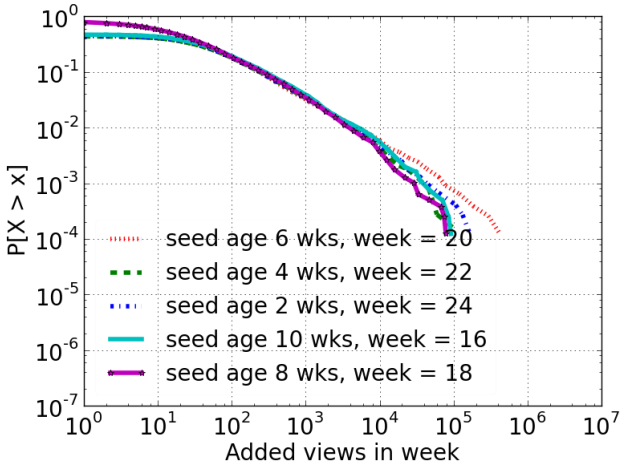
(e) 10000 or more added views in week 15

Figure 4.19: Distribution of added views for the keyword-search videos 48-52 weeks old age at seed time, binned by added views in week 15



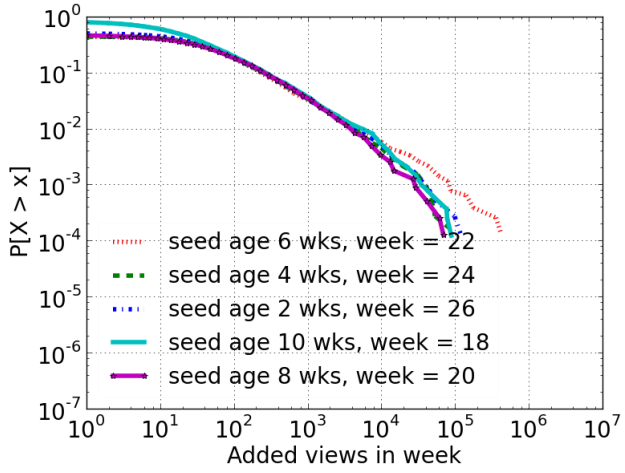
(a)

Figure 4.20: Distribution of added views for the keyword-search videos 24 weeks old



(a)

Figure 4.21: Distribution of added views for the keyword-search videos 26 weeks old



(a)

Figure 4.22: Distribution of added views for the keyword-search videos 28 weeks old

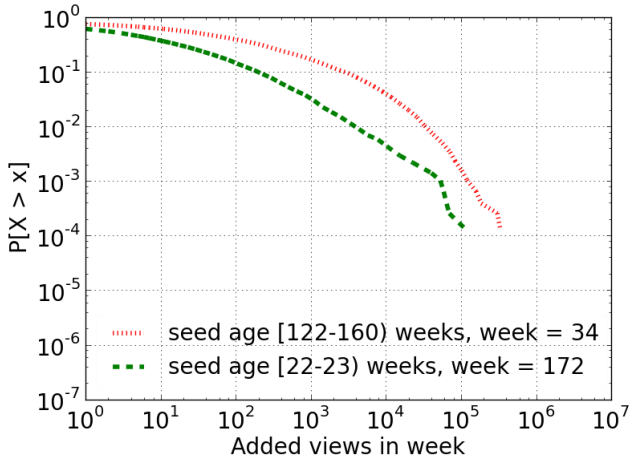
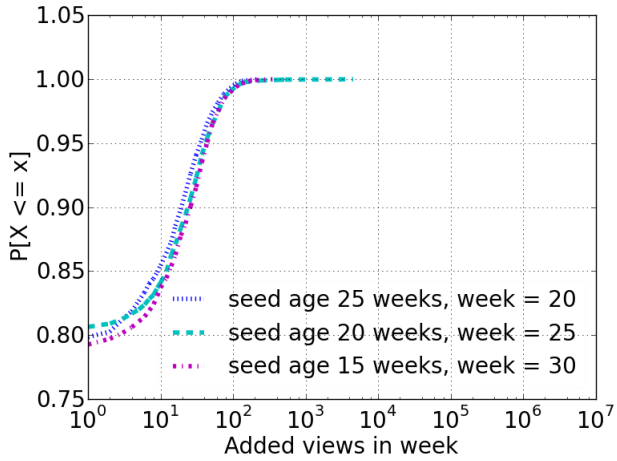
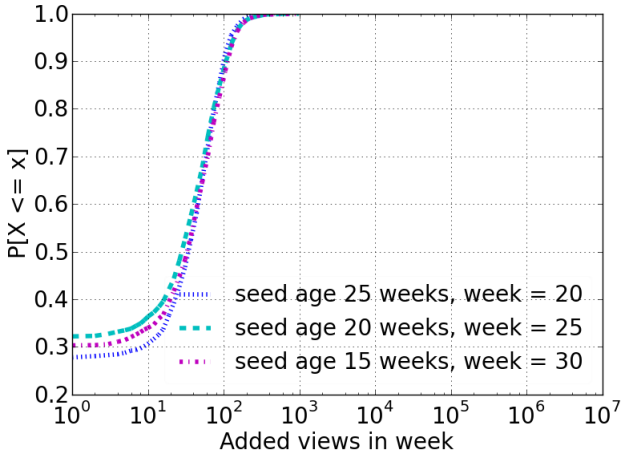


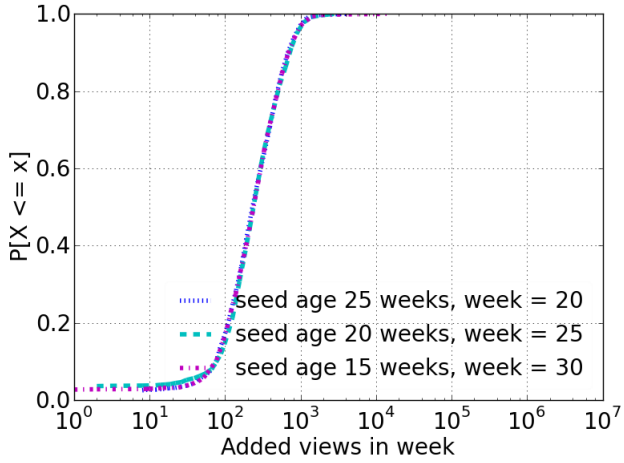
Figure 4.23: Distribution of added views for the keyword-search videos between 156 to 194 weeks old



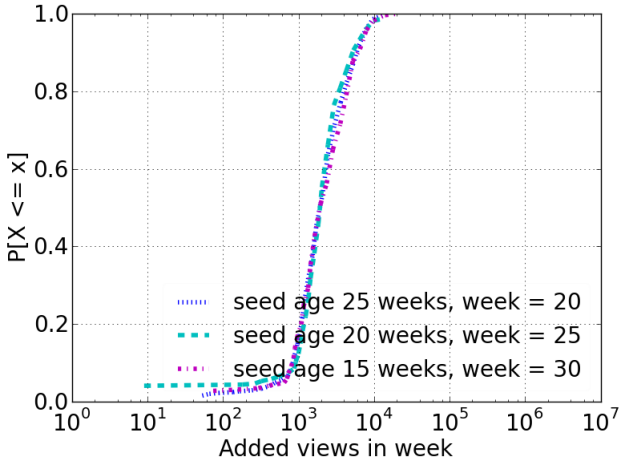
(a) 0 to 9 added views when 40 weeks old



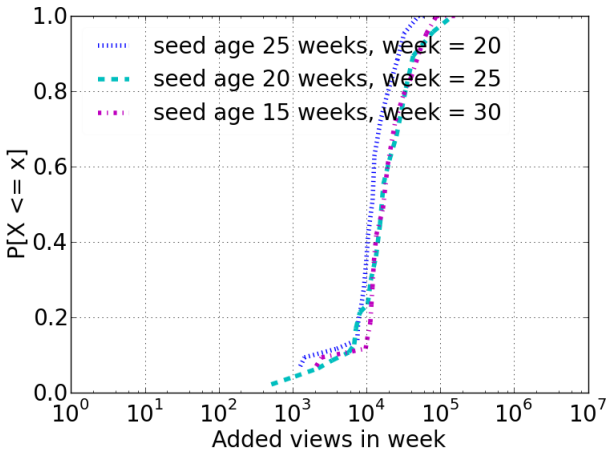
(b) 10 to 99 added views when 40 weeks old



(c) 100 to 999 added views when 40 weeks old

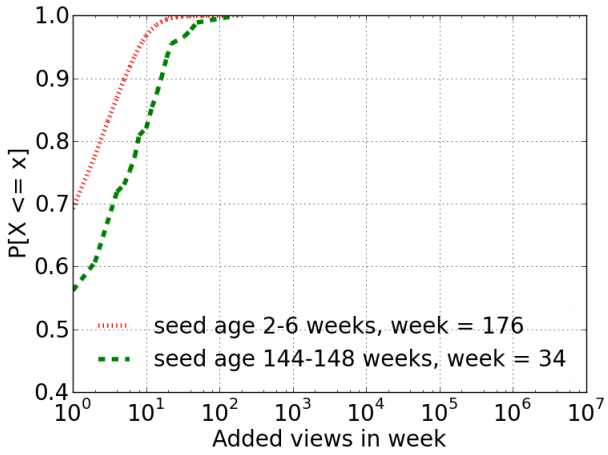


(d) 1000 to 9999 added views when 40 weeks old

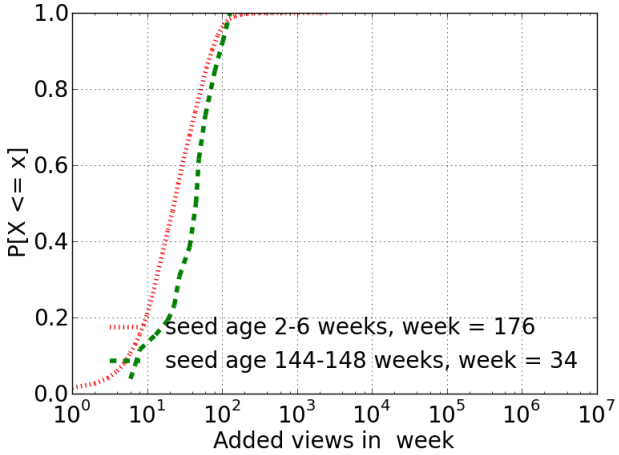


(e) 10000 or more added views when 40 weeks old

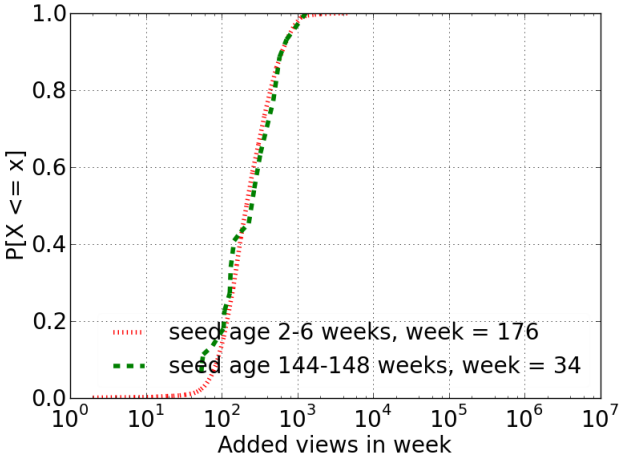
Figure 4.24: Distribution of added views for the keyword-search videos 45 weeks old, binned by added views when 40 weeks old



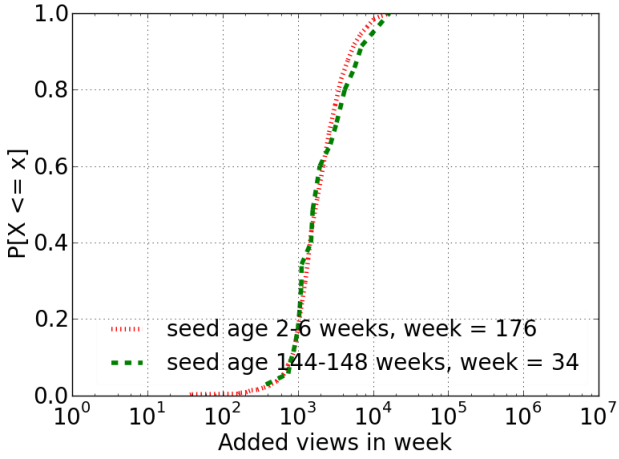
(a) 0 to 9 added views when 173 to 177 weeks old



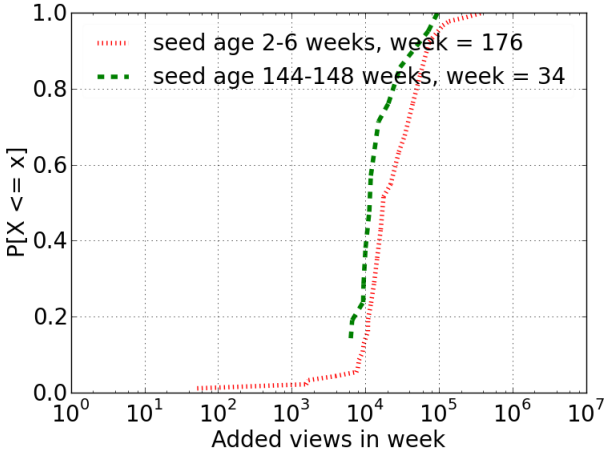
(b) 10 to 99 added views when 173 to 177 weeks old



(c) 100 to 999 added views when 173 to 177 weeks old



(d) 1000 to 9999 added views when 173 to 177 weeks old



(e) 10000 or more added views when 173 to 177 weeks old

Figure 4.25: Distribution of added views for the keyword-search videos between 178 to 182 weeks old, binned by added views when 173 to 177 weeks old

Chapter 5

Modelling Popularity Evolution

This chapter presents the basic and extended popularity evolution models developed by Borghol *et al.* [5] and evaluates the accuracy of these models for a much longer time span than considered by Borghol *et al.* The accuracy of the basic model is evaluated by comparing model and empirical total view count distributions, weekly view distributions, and popularity dynamics and churn, using the data from the first and second measurement periods for the recently-uploaded videos. The extended model is evaluated with respect to its ability to more accurately model popularity churn than the basic model. Section 5.1 presents the basic model developed by Borghol *et al.* Applying the model over the time span up to the end of the second measurement period requires determining a "time-to-peak" distribution over this time span, a problem that is considered in Section 5.2. The model also requires distributions for weekly view counts of videos that have not yet "peaked" (i.e., have not yet achieved their maximum weekly view count), for videos that are peaking in the current week, and for videos that have peaked in some prior week. These distributions (for the longer time span considered here) are examined in Section 5.3. Section 5.4 describes the basic characteristics of the synthetic view counts generated using the Borghol *et al.* basic model for the longer time span considered in this week. The basic model is evaluated through comparison with empirical data in Section 5.5. Section 5.6 presents the extended model and evaluates its accuracy with respect to popularity churn characteristics.

5.1 Basic Model of Borghol *et al.*

Based on observations from their recently-uploaded dataset, Borghol *et al.* [5] developed a model that generates synthetic weekly view counts with characteristics similar to those observed for newly-uploaded videos as they age. The model generates the weekly view counts for each video within a collection of synthetic newly-uploaded videos using a three phase characterization of popularity evolution, in which each video is either "before-peak" (i.e., has not yet attained its highest weekly view count), "at-peak", or "after-peak". The number of synthetic videos whose popularity peaks in any particular week after video upload is determined using a time-to-peak distribution parameterized from the empirical data.

The model uses three view count distributions, one for each of the "before-peak", "at-peak", and "after-peak" phases. For each modelled week after upload, view counts sampled from the before-peak and at-peak distributions will be assigned to videos that were in their before-peak phase during the previous week

(according to which of these videos are modelled as now being "at-peak" and which are still "before-peak"), and views sampled from the after-peak distribution will be assigned to videos that were in their at-peak or after-peak phase during the previous week. In more detail, the main components of the view count generation procedure are as follows:

1. Generating the number of videos in each phase

The model requires a time-to-peak distribution to determine the number of synthetic videos that peak in each modelled week. Borghol *et al.* [5] obtained this time-to-peak distribution by fitting an analytic distribution to an empirical time-to-peak distribution. The empirical distribution was obtained using data from the first measurement period of the recently-uploaded dataset (Figure **????**). Note that it is assumed here that the number of modelled weeks will be the same as the number of weeks covered by the empirical time-to-peak distribution. Note also that the "time-to-peak" refers to the time required to achieve the highest weekly view count that is attained during the measured or modelled weeks only, and thus the longest possible "time-to-peak" is the length of the measurement or modelling period. Borghol *et al.* [5] used an analytic distribution in which approximately 3/4 of the videos peak during the first 6 weeks according to an exponential distribution with parameter λ, and the remaining 1/4 of the videos peak at a time uniformly distributed from week 6 to the last considered week d. $\lambda = 0.598$ was determined using the *Maximum Likelihood Estimation (MLE)* method.

Note that for any week i, the total number of synthetic videos N = $n_i^{before} + n_i^{at} + n_i^{after}$, where n_i^{before}, n_i^{at} and n_i^{after} denote the number of videos "before-peak", "at-peak", and "after-peak", respectively. Therefore, once the number of videos that peak in each week is determined from the analytic time-to-peak distribution, the number of videos in each other phase can be determined iteratively using $n_i^{before} = n_{i-1}^{before}$ - n_i^{at} and $n_i^{after} = n_{i-1}^{after} + n_{i-1}^{at}$, for i > 1 and i $\leqslant$ d, and n_1^{before} = N - n_1^{at}, $n_1^{after} = 0$.

2. Generating synthetic view counts

Borghol *et al.* [5] made the approximation that the probability distribution for the weekly view counts of videos in their before-peak phase, as well as that for videos at-peak, and that for videos in their after-peak phase, are all week-invariant, i.e., are the same regardless of which week is considered. Analytic distributions were obtained by separately fitting the body and tail of the empirical before-peak, at-peak, and after-peak weekly view count distributions (as obtained using data from the first measurement period of the recently-uploaded dataset). The body and tail of each empirical distribution were separated by a threshold view count value, chosen for each distribution so that the probability of a view count greater than the threshold is approximately 10%. Borghol *et al.* considered both power law and lognormal distribution fits for the tail and, using the log-likelihood ratio test [15], determined that lognormal distributions provided better fits. Specifically, they used the "tail-method", and fit the tail of each distribution by the right tail of a lognormal distribution. The distribution bodies were fit by four-parameter beta distributions.

The parameters used for the lognormal distributions are presented in Table 5.1. The parameters μ and σ were estimated by direct maximization of tail-conditional log-likelihood [15]. Table 5.2 presents the

Table 5.1: Parameters used by Borghol *et al.* [5] for lognormal distributions

Phase	x_{thresh}	μ	σ
Before-peak	119	2.000	2.135
At-peak	297	-3.826	3.477
After-peak	30	-0.356	2.533

Table 5.2: Parameters used by Borghol *et al.* [5] for beta distributions

Phase	x_{min}	x_{thresh}	α	β
Before-peak	0	119	0.191	1.330
At-peak	4	297	0.543	2.259
After-peak	0	30	0.077	0.968

parameters used for the beta distributions. There is no closed-form for the maximum likelihood estimates for α and β, and instead α and β were estimated using the method of moments [5]:

$$\alpha = \tilde{x} \times (\frac{\tilde{x} \times (1 - \tilde{x})}{v} - 1),$$

$$\beta = (1 - \tilde{x}) \times (\frac{\tilde{x} \times (1 - \tilde{x})}{v}) - 1,$$

with

$$\tilde{x} = \frac{mean - x_{min}}{x_{thresh} - x_{min}}$$

$$v = \frac{var}{(x_{thresh} - x_{min})^2}$$

Here, the variables *mean* and *var* are the empirical mean and variance, respectively.

3. Assigning view counts to videos

In the basic model of Borghol *et al.*, view counts sampled from the before-peak, at-peak, and after-peak distributions are assigned to the synthetic videos in the respective phases so as to preserve the relative popularities of videos in the same category [5]. Specifically, the view counts sampled from the at-peak and before-peak distributions for week i are assigned to the videos that were in the before-peak phase during week i - *1* (since those are the videos that will now be either at-peak or still before-peak) in such a way that their relative popularity rank is the same as that during week i-*1*. Those synthetic videos assigned the views sampled from the at-peak distribution are then considered to have peaked, and will be moved to their after-peak phase for the subsequent weeks. Similarly, the view counts sampled from the after-peak distribution for week i are assigned to the videos that were at-peak or after-peak during week i - *1*, in such

a way that their relative popularity rank is the same as that during week *i - 1*.

Borghol *et al.* [5] use their model to generate view counts for 29,791 videos (the same number of videos as meta-data was collected for in the first measurement period for the recently-uploaded dataset) and d = 34 weeks. The model evaluation shows that the synthetic view count distributional characteristics match well with the empirical data. An extension of the model (described in Section 5.6) also yields popularity dynamics such as hot set churn similar to those observed empirically. The model of Borghol *et al.* is implemented in this thesis with N = 20,000 videos and d = 179 weeks.

5.2 Time-to-peak Distribution

Popularity dynamics and churn are impacted by how quickly videos reach their peak popularity since they are uploaded. Time-to-peak is defined as the age at which a video receives its highest weekly viewing count, over some measured of modelled period. Figure 4.6 shows the CDF of the time-to-peak distribution for the first measurement period, the second measurement period and over both measurement periods. Considering both measurement periods, a large fraction of the videos (approximately three-quarters) reach their peak popularity within the first six weeks since they are uploaded. The remaining videos reach their peak popularity at a time approximately uniformly distributed throughout the remainder of the measurement periods. Figure 5.1 shows the highly variable number of empirical videos that reach their peak popularity in different weeks in both the first and second measurement periods. Although the number that reach their peak popularity during the second measurement period is somewhat less than a uniform distribution would predict, it is nonetheless surprising that so many videos peak in the second measurement period given that a long period of time has elapsed since those videos were uploaded.

Based on these results, an analytic time-to-peak distribution is developed which is used as an input for the basic model to generate the number of videos that peak in any arbitrary week. In the analytic time-to-peak distribution, approximately three-quarters of the videos peak within the first six weeks and the remaining videos peak at a time drawn from a uniform distribution U(6,d), where d is the duration of the measurement period. Borghol *et al.* [5] estimated the rate parameter $\lambda = 0.598$ of the exponential part of the distribution using the Maximum Likelihood Estimation (MLE) method.

5.3 Three-phase Characteristics

The basic model developed by Borghol *et al.* generates synthetic weekly view wants in a week-invariant manner. In the generation of synthetic view counts, the most complex part is the achievement of churn in the relative video popularities similar to that seen in the empirical data (Chapter 4). Empirically, different videos achieve their peak popularity at different times with highly variable rate. Therefore, it is apparent that the time-to-peak distribution can play an important role in the appearance of popularity churn. Generating

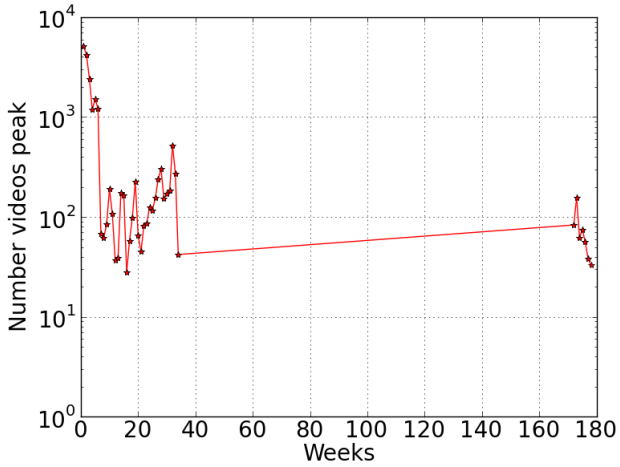


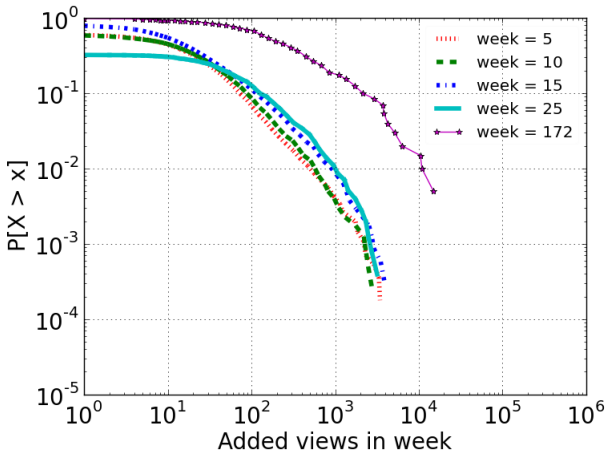
Figure 5.1: Number of videos achieving their peak popularity in different weeks for the empirical recently-uploaded videos

realistic synthetic view counts for videos in different phases (before-peak, at-peak and after-peak) requires analysis of the empirical view count characteristics of videos in those different phases.

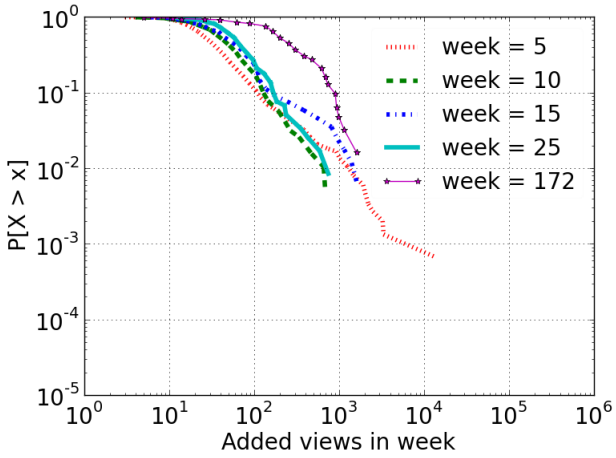
Figure 5.2 presents the CCDF of weekly views in each phase, for the videos in the recently-uploaded dataset, using logarithmic scale on each axis. Note that the distribution of weekly views in each phase is heavy-tailed even in week 172, which is long after the videos were uploaded. It is also observed that the view count distribution for before-peak videos in week 172 (Figure 5.2(a)) as well as for at-peak videos (Figure 5.2(b)) is quite distinct from that for the other weeks. In contrast, except for unimportant differences for videos with lower view counts, the after-peak distribution (Figure 5.2(c)) appears week-invariant. The Borghol *et al.* model assumes that all three distributions are week-invariant, and so a topic of future work would be to modify the model to make the before-peak and at-peak view count distributions week dependent.

Assuming week invariance as in the Borghol *et al.* model, Figure 5.3 presents the CCDF of weekly views in each phase when data is aggregated across all weeks. As in the previous figure, the view count distribution for each phase is heavy tailed, and it is apparent that videos tend to have higher view counts in the at-peak phase, lower view counts in the before-peak phase, and the lowest weekly view counts when they move to the after-peak phase. The same lognormal and beta distribution fits for the tail and body of each distribution, respectively, as used by Borghol *et al.* are used in this week. Although the at-peak and before-peak distributions show significant differences in the second measurement period, relatively few videos are in these phases in the second measurement period, and so the impact on the distributions in Figure 5.3 is quite small.

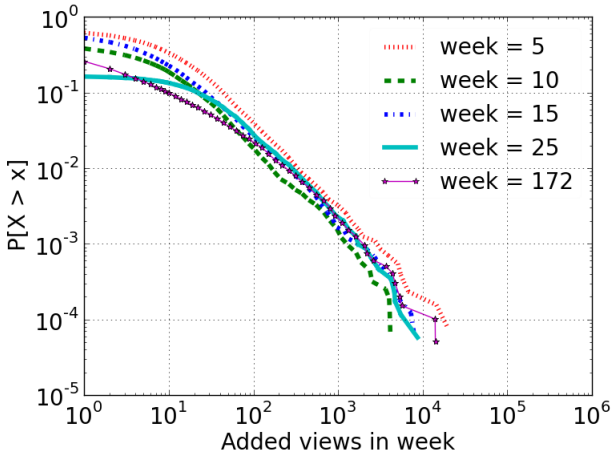
Figure 5.4(a) presents the average weekly view count of videos in each phase throughout the measurement period. One goal in presenting this figure is the observation of how much variability there is in the average weekly view count for all three phases. As seen in the figure, the average weekly view count of videos in their at-peak phase shows a higher degree of variability than does the average view count of videos in their



(a) Before-peak phase



(b) At-peak phase



(c) After-peak phase

Figure 5.2: Distribution of weekly views for the empirical recently-uploaded videos in the before-peak, at-peak and after-peak phase

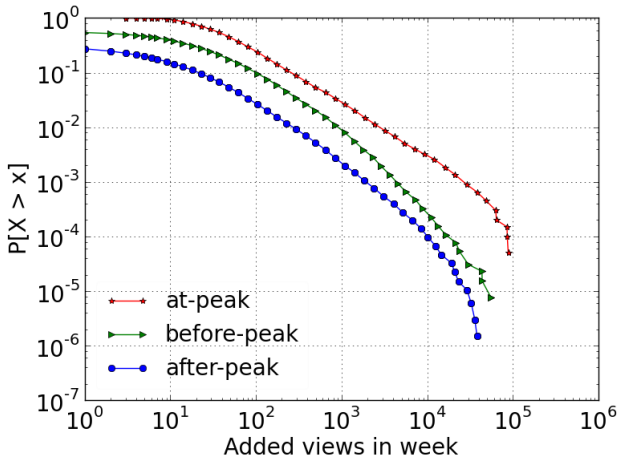


Figure 5.3: Distribution of weekly views for the empirical recently-uploaded videos in their before-peak, at-peak and after-peak phase, when data is aggregated across all weeks

before-peak or after-peak phase. This can be explained by the relatively small number of videos that are in their at-peak phase in any given week.

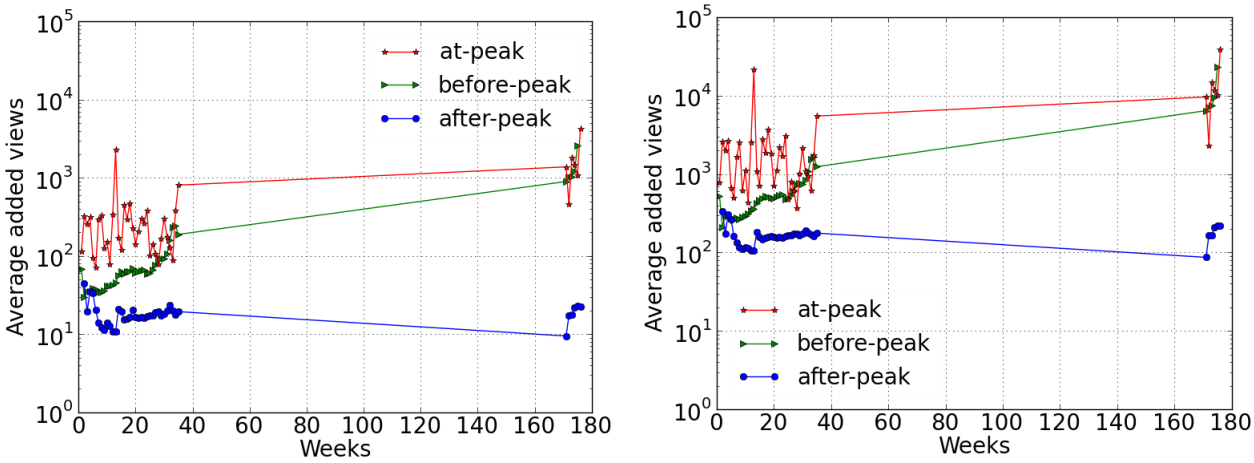
Figure 5.4(a) also shows that the average weekly view count in the at-peak and before-peak phases exhibits an increasing trend in the second measurement period compared to the first measurement period, consistent with Figure 5.2. Furthermore, the difference in viewing rate in the after-peak phase compared to that in the before-peak and at-peak phases is smaller in the first measurement period compared to in the second measurement period.

Figure 5.4(b) shows the average weekly view count for videos in the tail of the at-peak, before-peak and after-peak distributions. The separation point between body and tail is defined by a threshold view count above which videos are considered to be in the tail. The threshold values are chosen so that approximately 10% of the videos fall into the tail of each distribution as done by Borghol *et al.* [5]. The threshold values for the before-peak, at-peak and after-peak view count distributions are 116, 296 and 31 weekly views respectively.

Characteristics such as high variability in the at-peak phase and stability in the before-peak and after-peak phases are the same as observed in Figure 5.4(a). Also as in Figure 5.4(a), the average added views for the at-peak and before-peak distributions (specially the tails in this figure) increased in the second measurement period.

5.4 Distribution of Synthetic Views

According to the basic model developed by Borghol *et al.* [5], a set of synthetic weekly views are generated for N = 20,000 (the number of videos in the recently-uploaded dataset) and d = 179 (the time span from the beginning of data collection for the recently-uploaded dataset until the end of the second



(a) Average added views in each week

(b) Average added views in each week for the tail only

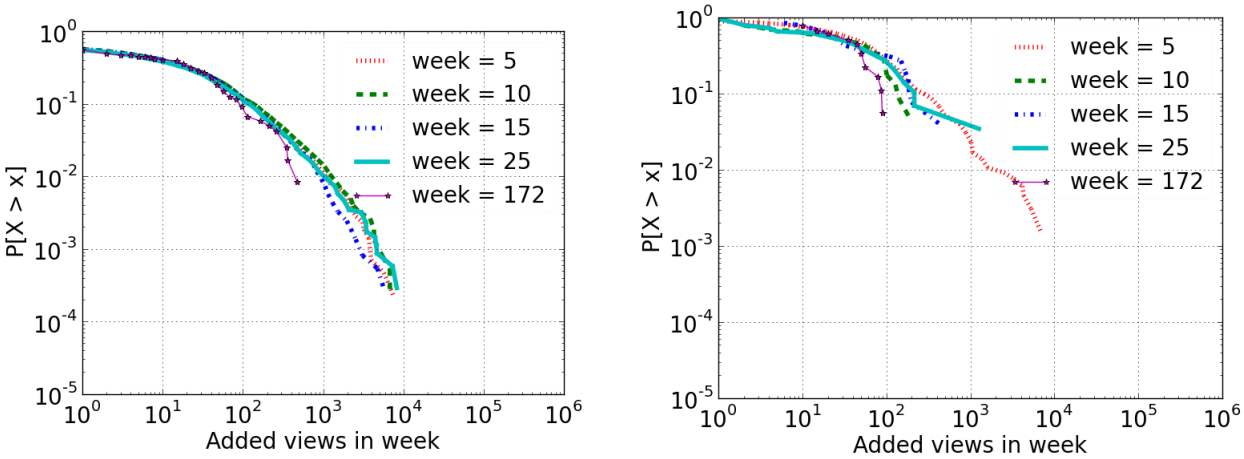
Figure 5.4: Average view count for the empirical recently-uploaded videos

measurement period). Some basic characteristics of these synthetic weekly views are presented in this section. The main goal of this section is the observation of the view count distribution for synthetic videos.

Figure 5.5 presents the CCDF of weekly views for synthetic videos in each phase for a number of example weeks. The heavy-tail nature of the view distribution is observed in all three phases both in the first and in the second measurement period. There are no significant differences observed in view distribution for different weeks. Some of the plotted lines end well before others simply owing to differences in the number of videos in that phase, during the respective weeks. Therefore, as expected, the view distribution in all three phases is week-invariant. Similar to Figure 5.3, Figure 5.6 presents the view count distributions of all three phases using data from all of the weeks. As in Figure 5.3, view counts tend to be highest for videos in their at-peak phase, lower for videos in their before-peak phase, and the lowest for videos in their after-peak phase. The heavy-tail nature is also observed in this figure.

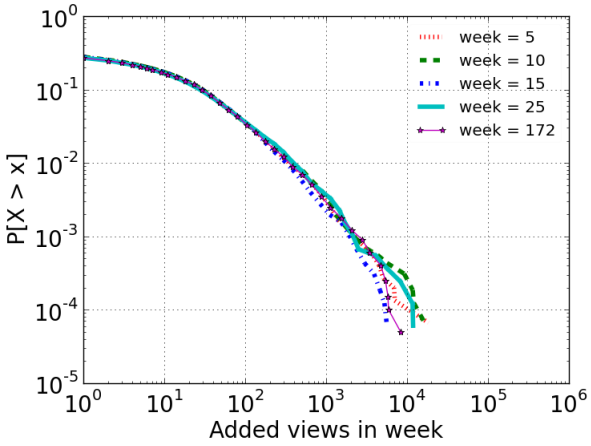
The variability in view counts for the three phases is investigated by plotting average weekly views considering all videos in the respective phase, and also for the tail videos only in Figure 5.7. Since the number of videos that peak in each week is relatively small, excepting for the first few weeks, the average weekly view count for the at-peak phase is highly variable throughout the measurement period. Unlike the at-peak phase, the average weekly viewing rate for videos in the before-peak and after-peak phases is quite stable. Figure 5.7(b) presents the average weekly views for the most popular ten percent of the videos in each phase. The average weekly view count for videos in the tail follows the same characteristics as observed for all videos but with a higher average rate in all three phases.

An important property of video popularity evolution that is observed empirically is churn in the relative video popularities. Figure 5.8 presents a scatter plot of synthetic weekly view counts in consecutive weeks. The figure shows a strong relationship between week i and i+1 view counts in all subfigures. The



(a) Before-peak phase

(b) At-peak phase



(c) After-peak phase

Figure 5.5: Distribution of weekly views for the synthetic recently-uploaded videos in the before-peak, at-peak and after-peak phase

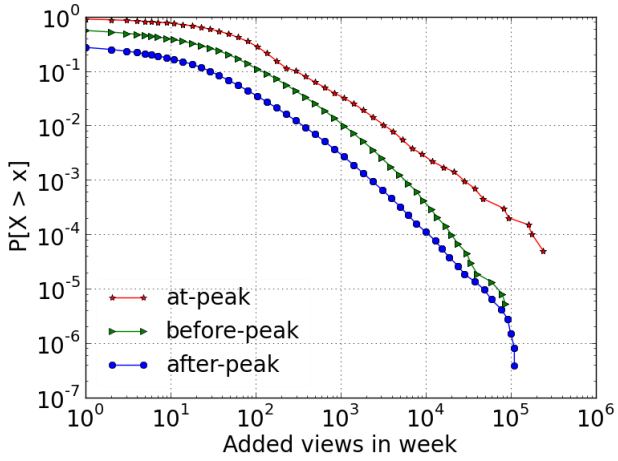
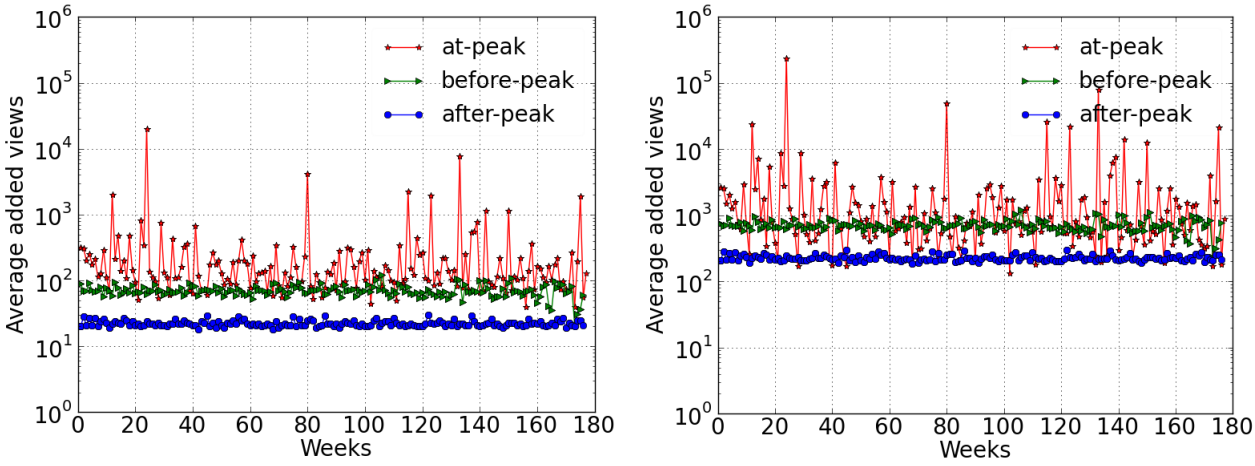


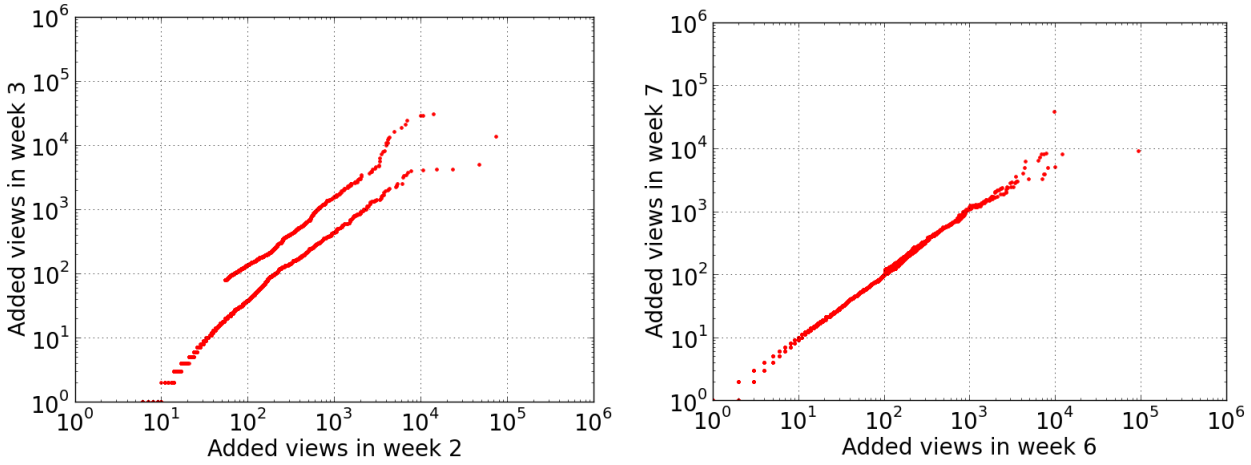
Figure 5.6: Views distribution in a week for the synthetic recently-uploaded videos in their before-peak, at-peak and after-peak phase



(a) Average added views in each week

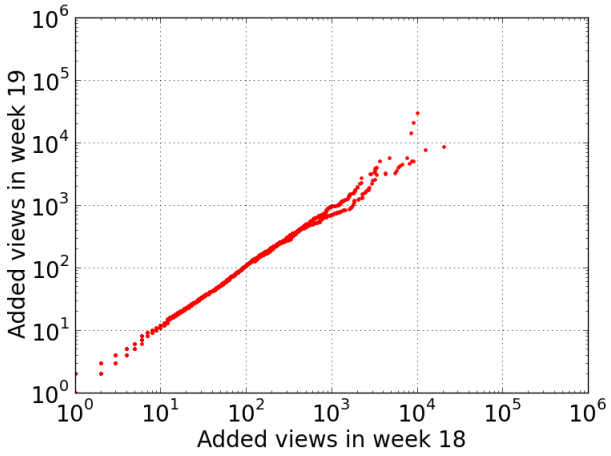
(b) Average added views in each week for the tail only

Figure 5.7: Average view count for the synthetic recently-uploaded videos

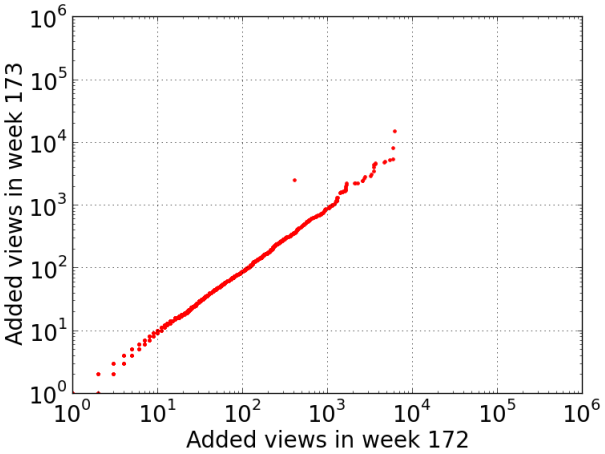


(a) Week 2 vs 3

(b) Week 6 vs 7



(c) Week 18 vs 19



(d) Week 172 vs 173

Figure 5.8: Scatter plot of added views for the synthetic recently-uploaded videos in week i vs week i+1

reason behind this relationship is that sampled view counts are assigned to videos so as to maintain the same relative popularities of the videos within each category, as observed in Section 4.2.3. However, it is noticeable that each of subfigures (a) through (c) shows two different lines, with these lines being most clearly separated in subfigure (a). Each lower line is due to the videos that are in their at-peak or after-peak phase in week i and their after-peak phase in week i+1, and each upper line is due to the videos that are in their before-peak phase at week i and their before-peak or at-peak phase in week i+1. Since only a small number of videos reach their at-peak phase during the later weeks, the lines merge and are indistinguishable by the time of the second measurement period (Figure 5.8(d)). Churn is examined further in Section 5.5, as well as in Section 5.6 where the extended model of Borghol *et al.* is considered.

5.5 Model Evaluation

This section evaluates the accuracy of the model by comparing synthetic data with data for the empirical recently-uploaded videos. The test metrics considered to evaluate the model are: (a) distribution of total views or accumulated views at different video ages (weeks since uploaded) (b) distribution of weekly views at different weeks both overall and when videos are binned, and (c) popularity dynamics and churn in terms of hot set overlap.

Figure 5.9 presents the CCDF of total views acquired by weeks 2, 32 and 172 both for the videos in the recently-uploaded dataset and for the synthetic videos, using a logarithmic scale on each axis. The figure shows excellent matches between the distributions for the synthetic and empirical videos. Note that the model is not parameterized using empirical total view count statistics, but instead the synthetic total view counts are consequences of the view generation algorithm and modelling parameters derived from the model's three-phase characterization of video popularity evolution.

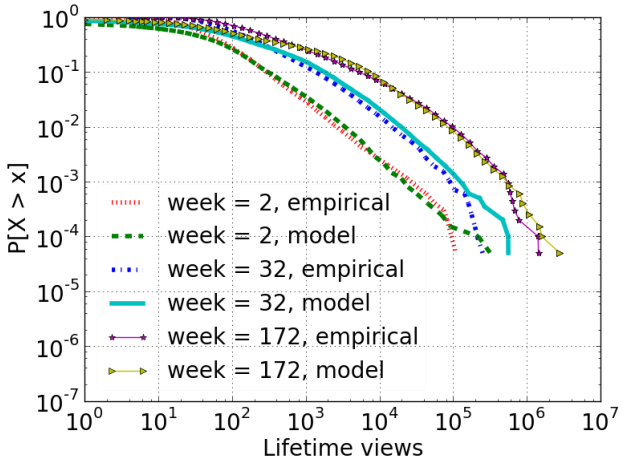
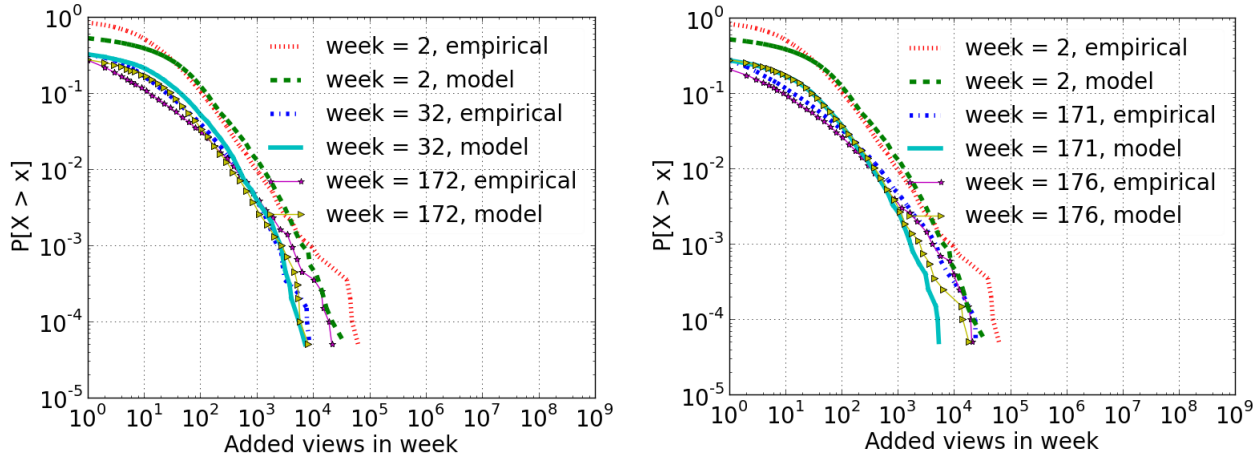


Figure 5.9: Distribution of total views for both the empirical recently-uploaded videos and the synthetic recently-uploaded videos

Figure 5.10 shows the CCDF of weekly views during weeks 2, 32, 171, 172 and 179 for both the videos in the recently-uploaded dataset and the synthetic videos using a logarithmic scale for each axis. Although there is a good match between the general forms of the corresponding distributions for the synthetic and empirical videos, there are some significant differences that are apparent in the figure. The model assumes that the distributions used in its three-phase characterization are week-invariant, and inaccuracies in this assumption may be a cause of these differences. The substantial growth in the YouTube user population between the first and second measurement periods (not accounted for in the model) may also be a factor.

Figure 5.11 shows the comparison of weekly view distribution during week 20, 30 and 172 with videos separated into different bins according to their added views in week 15. As in previous analyses, week 15 is chosen to avoid the higher popularity churn at earlier video ages. Again, significant differences are observed,



(a) Distribution of added views

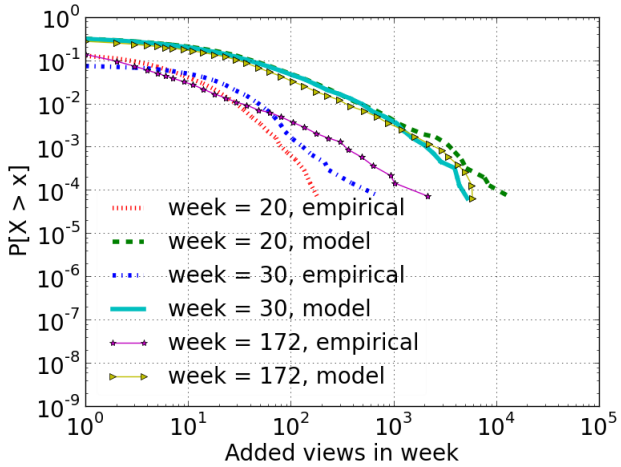
(b) Distribution of added views

Figure 5.10: Distribution of added views for both the empirical recently-uploaded videos and the synthetic recently-uploaded videos

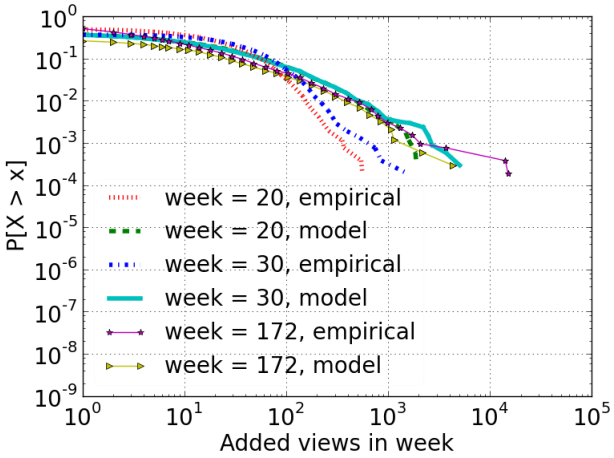
possibly owing to the assumptions of week invariance that are made in the model, with the growth in the YouTube user population being another possible factor.

Popularity dynamics and churn can have a substantial impact on the performance of caching. As was done by Borghol *et al.* [5], hot set overlap is used as another metric for evaluating the accuracy of the model. For this purpose, the most popular 10% and 1% of the videos in a week (two differently-sized "hot sets" for that week, where the notion of "most popular" is based on the new views acquired in that week only) are compared to the correspondingly-sized hot sets for some other week, and the overlap in videos is determined.

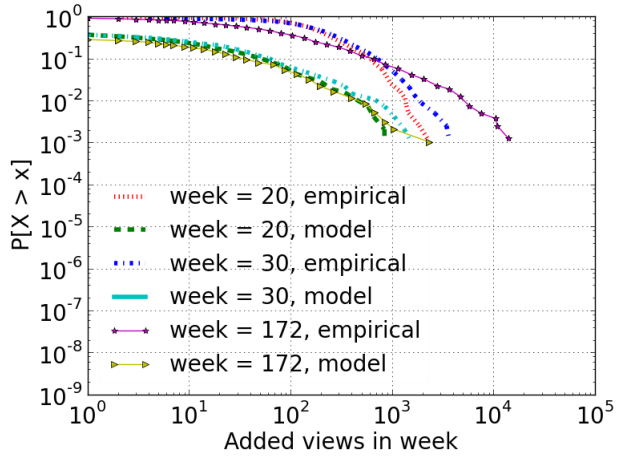
Figure 5.12 shows the fraction of overlap for both sizes of hot set. Note that for the empirical videos, data points exist only for the weeks within the first and second measurement periods, and so these plots show straight lines across the weeks between these two periods. Figure 5.12(a) presents the fraction of overlap for hot sets in consecutive weeks. The figure shows that the empirical hot set overlap ranges between 0.40 and 0.90 for the larger hot set size, and between 0.60 to 0.95 for the smaller hot set size. The churn generally decreases as the videos age. Figure 5.12(b) presents the fraction of overlap in hot sets for each week i, compared to week 2. Although the model could not exhibit the empirical churn, however both of the figures shows that the model can exhibit the trend (higher churn at early ages but lower churn at the later ages) at which videos experienced popularity churn. This lack of accuracy for hot set churn is consistent with the results of Borghol *et al.* [5], which led these authors to develop the extended model described next.



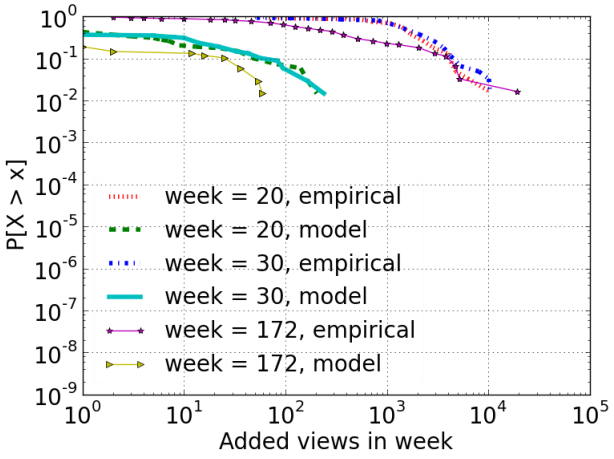
(a) 0 to 9 added views in week 15



(b) 10 to 99 added views in week 15

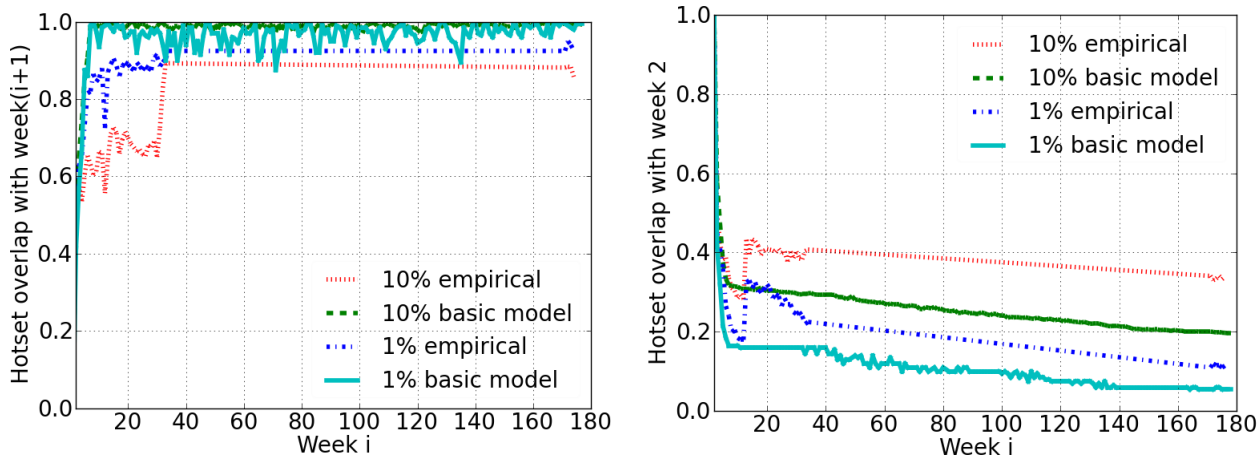


(c) 100 to 999 added views in week 15



(d) 10000 to 9999 added views in week 15

Figure 5.11: Distribution of added views for both the empirical recently-uploaded videos and the synthetic recently-uploaded videos binned by added views in week 15



(a) Week i vs week i+1 hot sets

(b) Week 2 vs week i hot sets

Figure 5.12: Churn in video popularity for both the empirical recently-uploaded videos and the synthetic recently-uploaded videos

5.6 Extended Model

The extended model introduces additional churn in video popularity. The basic model achieves popularity churn only by moving videos among the before-peak, at-peak and after-peak phases. However, the extended model achieves popularity churn also by repeatedly exchanging the added view counts for a randomly chosen week and pair of synthetic videos subject to certain constraints. Views are exchanged in such a way that the three-phase characteristics and the time-to-peak distribution remain unchanged.

Specifically, in the extended model the added view counts of synthetic videos are exchanged by repeatedly choosing a random week i and two videos u and v such that both videos are either in their before-peak phase, or both videos are in their after-peak phase. Then windows W_i^u and W_i^v are calculated for these two videos. If the currently assigned added view counts of both videos for week i fall within each other's window then their added view counts for that week can be exchanged. The window for a video v and a week i is defined as follows:

$$W_i^v = [\frac{x_i^v}{g}, min(x_i^v \times g, x_{max}^v)], g \in [1, \infty]$$

Here, x_i^v is the added views assigned to video v for week i by the basic model and x_{max}^v is the maximum weekly views assigned to the video v during its lifetime by the basic model. g is a model parameter that determines the amount of possible churn.

For the results in this thesis, a random week and a potentially eligible pair of videos are picked 5 million times for each of several choices of the model parameter g ($1 <= g <= \infty$); note that only a fraction of these selections will result in an exchange of added view counts owing to the constraint imposed by the

windows. Note also that the weekly views generated by the extended model for g =1 are the same as the weekly views generated by the basic model.

Figure 5.13 presents the hot set overlap between adjacent weeks, for hot sets comprised of the most popular (according to their count of added views in that week) 10% of the videos, for different values of the model parameter (g). The figure shows that the popularity churn increases with the increase of the value of g, and that the maximum churn when g = ∞ results in approximately 40% hot set overlap between adjacent weeks throughout the considered period.

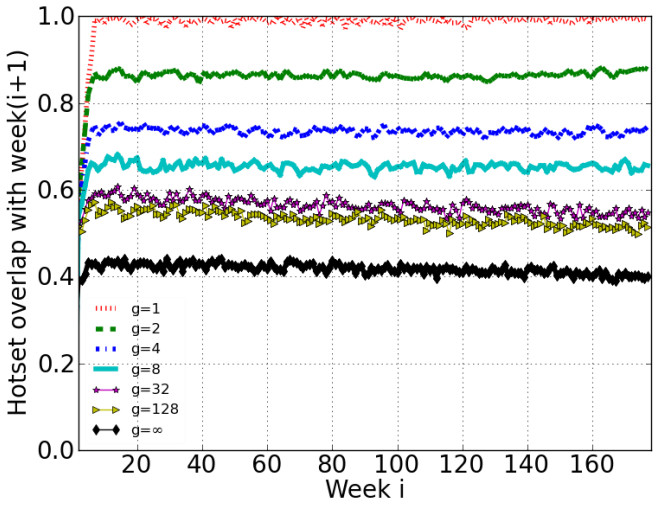
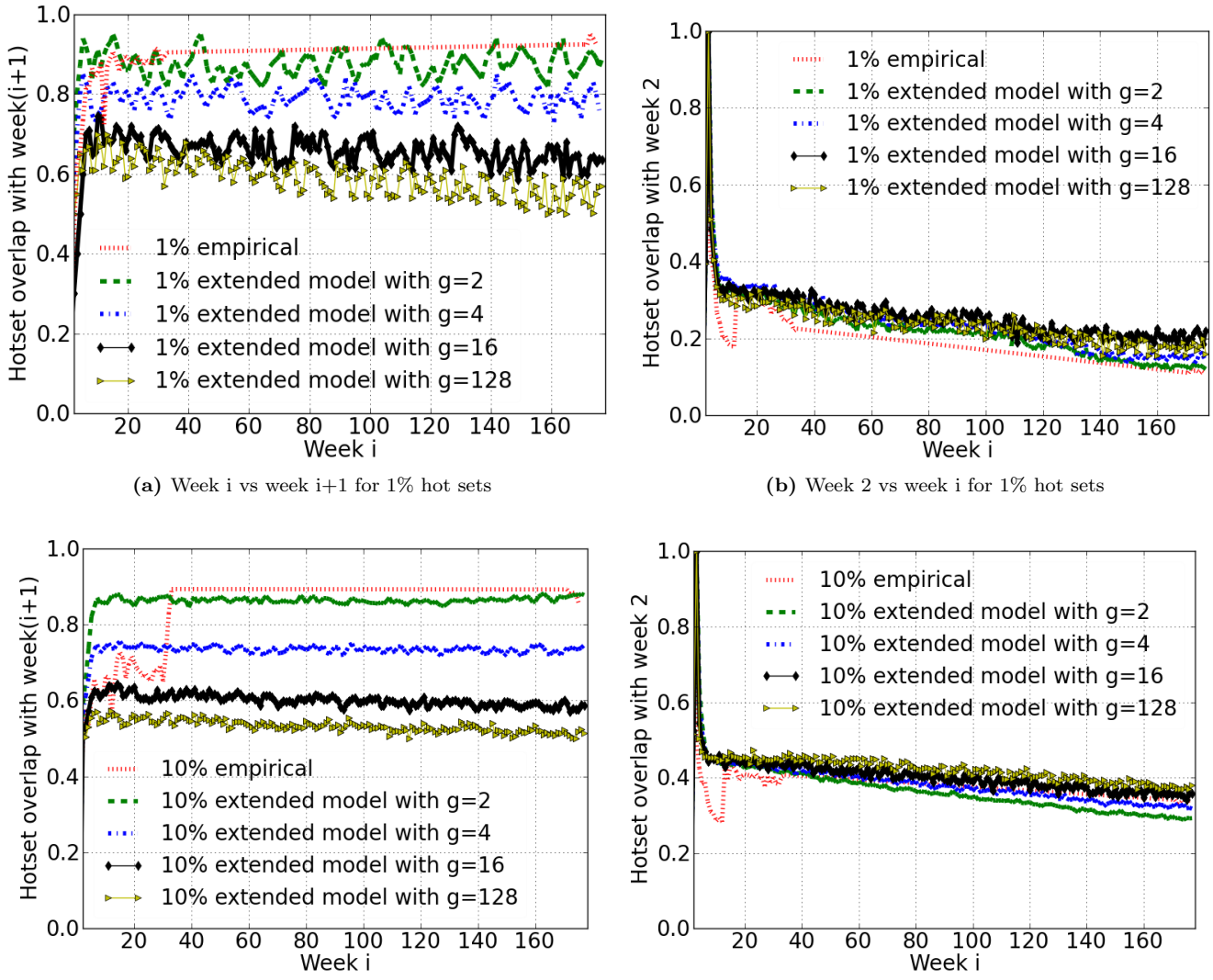


Figure 5.13: Impact of churn modelling parameter in the extended model

Figure 5.14 compares the hot set overlap in the recently-uploaded dataset and the hot set overlap in the extended model for different g values. For the hot set overlap between adjacent weeks (Figure 5.14(a) and Figure 5.14(c)), a better match is observed with the extended model for a relatively high value of g, for both 10% and 1% hot sets, for young video ages, while as the video age increases (i.e, for later weeks) the best value of g decreases. For the time period and values of g considered, the best value of g changes from g=16 for the initial weeks to g=2 for the weeks corresponding to the second measurement period. The hot set overlap with week 2 (Figure 5.14(b) and Figure 5.14(d)) shows approximately similar churn for different g values, and similar behaviour as the empirical results.



(a) Week i vs week i+1 for 1% hot sets

(b) Week 2 vs week i for 1% hot sets

(c) Week i vs week i+1 for 10% hot sets

(d) Week 2 vs week i for 10% hot sets

Figure 5.14: Churn in video popularity using the extended model compared to the churn for the empirical recently-uploaded videos

Chapter 6

Removed Videos

This chapter presents the popularity characteristics of videos that were available in the first measurement period but unavailable in the second measurement period. There are mainly three reasons for which a video becomes unavailable in Youtube: (a) violation of YouTube terms and conditions, (b) the video is made private by uploader and (c) uploader deletes the video or deletes the entire account. Researchers at the Massachusetts Institute of Technology (MIT) maintain a database to keep track of videos removed from YouTube.[1] Figueiredo *et al.* [17] analyzed the popularity characteristics of the videos in that dataset and claimed that the copyright protected videos attain their peak popularity at an early age. Instead of only copyright protected videos, this chapter presents the popularity characteristics of videos that are removed for any reason. This chapter also compares the available and removed videos in terms of view count distribution and churn. The analysis reveals that the removed recently-uploaded videos have a substantially higher average viewing rate than the available recently-uploaded videos. However, probably due to the popularity bias in the keyword-search dataset, the difference in average viewing rate between the removed and the available keyword-search videos is smaller.

6.1 Removed Keyword-search Videos

The viewing rates of the available and removed keyword-search videos are compared in Figure 6.1 by plotting average weekly views in each week. The figure shows that, on average, the removed videos have somewhat higher weekly view counts than the available videos. However, the difference in the average viewing rate between the available and the removed videos becomes very small by the end of the measurement period. This implies that the removed videos tend to have elevated popularity (compared to the available videos) only during a certain period after the seed time, instead of elevated popularity throughout their lifetime. Interestingly, the increasing/decreasing pattern for the average viewing rate over the first measurement period is similar for both the available and removed videos.

Figure 6.2 presents the CCDF of the added views for the removed keyword-search videos at snapshots i = 2, 5, 10 and 25 using a logarithmic scale for each axis. The figure also shows the distribution of added

[1] `http://en.wikipedia.org/wiki/YouTomb`; accessed 25 - January, 2013

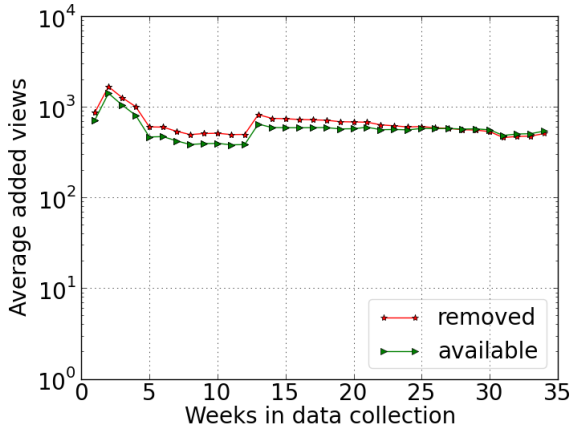


Figure 6.1: Average added views for the available keyword-search and the removed keyword-search videos in each week

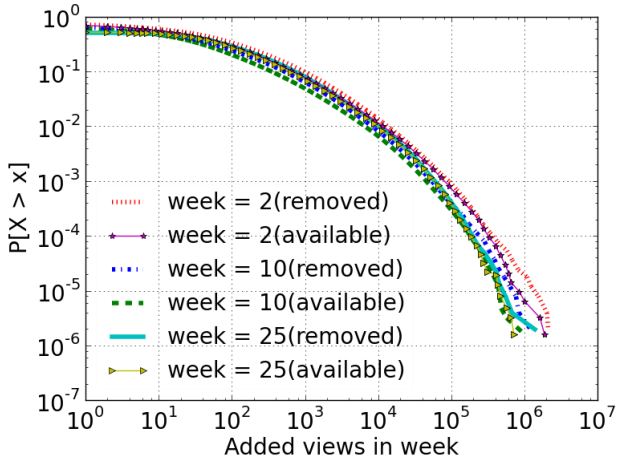


Figure 6.2: Distribution of added views for the available keyword-search and removed keyword-search videos at different data collection weeks

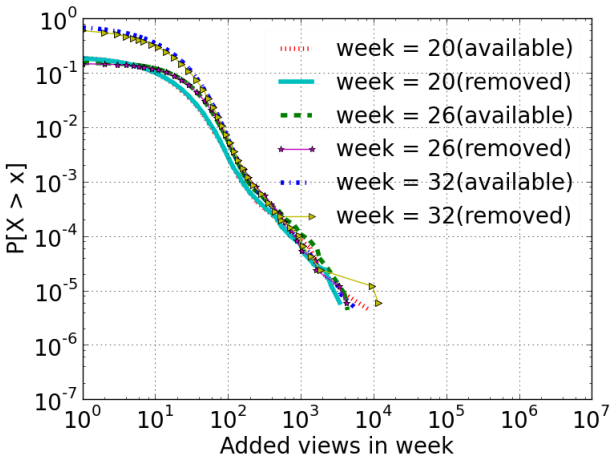
views for the available keyword-search videos for comparison purpose. Like the available keyword-search videos, the removed keyword-search videos exhibit smaller maximum added view counts for the later weeks. Overall, the view count distributions for these two types of videos are very similar.

A close look at the weekly added views is achieved by separating videos into different bins according to their added views in week 15. Week 15 is selected since it is past the time of high popularity churn. Figure 6.3 shows the CCDF of the added views at snapshots 20, 26 and 32 for various bins of the available and removed videos using a logarithmic scale for each axis. No major differences are observed in the corresponding added view distributions for the available versus the removed keyword-search videos, although the highest added views for a given bin and week for the removed videos, tend to exceed those for the available videos. Although previous researchers have found that removed videos tend to be more popular [17], the close distributional similarities observed in the figure are not surprising, since the keyword-search dataset is already biased towards popular videos.

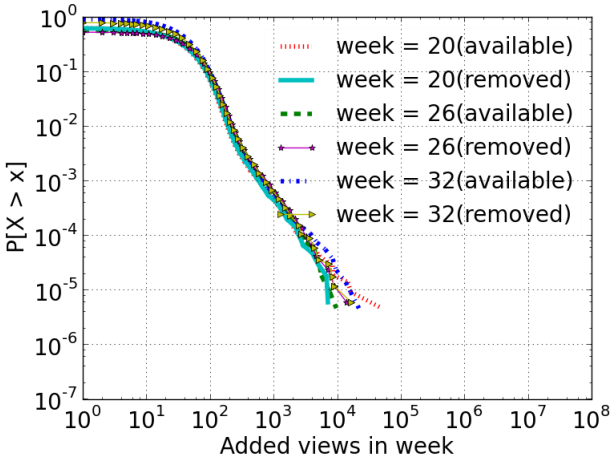
For further analysis of the popularity characteristics of the removed keyword-search videos, Figure 6.4 presents the popularity churn using scatter plots of added views at adjacent snapshots. The main goal for such analysis is the observation of view change patterns for these videos. The figure shows the significant point spread soon after the seed time, with much less point spread long after the seed time, in approximately the same pattern as observed for the available videos (Figure 4.11). As for the available keyword-search videos, soon after the seed time mildly popular removed videos can become highly popular in the next week and vice versa. The two-cluster pattern (Figure 6.4(a)) may be due to the young videos in the dataset, as well as the videos with elevated short term popularity. The upper cluster corresponds to videos with increasing popularity and the lower cluster corresponds to the videos with decreasing popularity.

The nonstationarity in the relative popularities of the removed keyword-search videos is observed by looking at the popularity rank change across adjacent weeks. Figure 6.5 presents the CDF of the absolute change in popularity rank for some example snapshots for the removed keyword-search videos as well as for the available keyword-search videos. For further understanding, videos are binned based on their age at seed time. Significant differences are observed in the rank change probabilities between the removed and available keyword-search videos. In general, for any age at seed time and measurement week, removed videos tend to experience either similar or lower rank change than available videos. The figure shows that the removed and available keyword-search videos have similar rank change probabilities for young videos and early measurement weeks (Figure 6.5(a) and Figure 6.5(b)). However for older videos, removed keyword-search videos tend to experience significantly lower rank change than the available keyword-search videos. This implies that the removed keyword-search videos have similar non-stationarity in their relative popularities as the available keyword-search videos at early ages, but after a certain period their relative popularities become more stable than those of the available keyword-search videos.

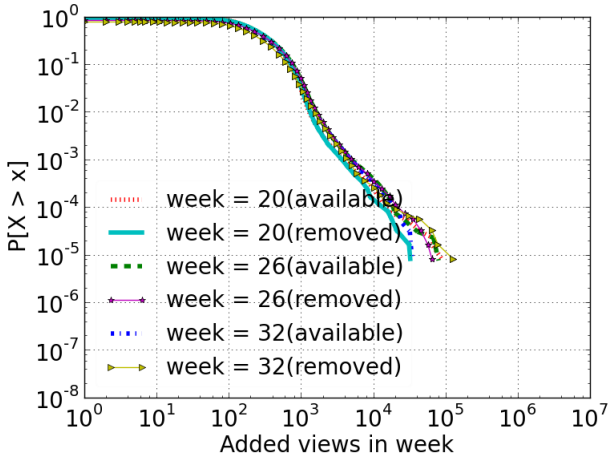
Figure 6.6 complements Figure 6.5 by plotting the CDF of the ratio of new to old popularity rank for some example snapshots for the removed keyword-search videos as well as for the available keyword-search



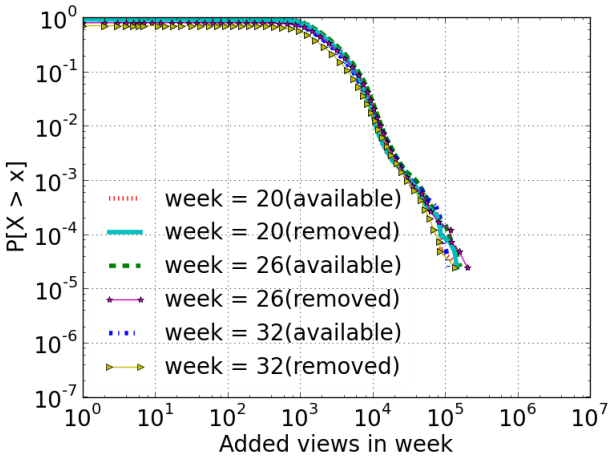
(a) 0 to 9 added views in week 15



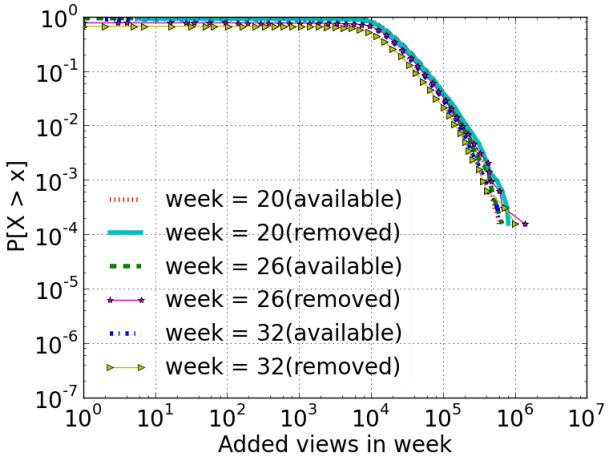
(b) 10 to 99 added views in week 15



(c) 100 to 999 added views in week 15

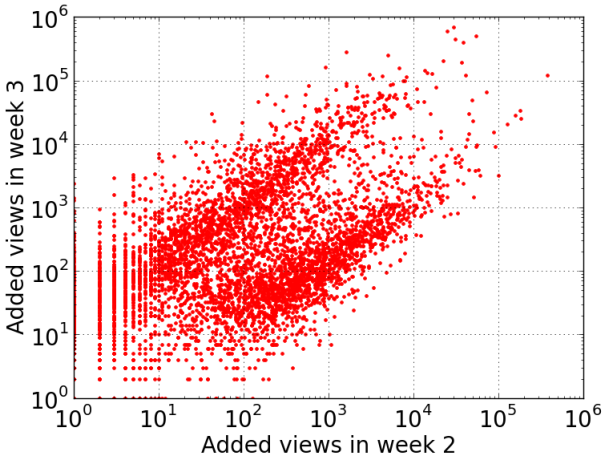


(d) 1000 to 9999 added views in week 15

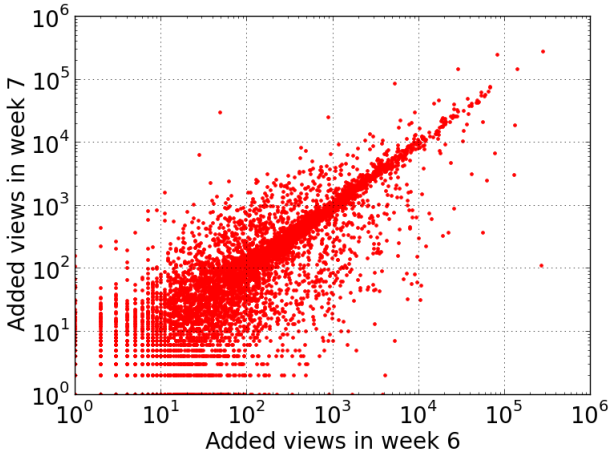


(e) 10000 or more added views in week 15

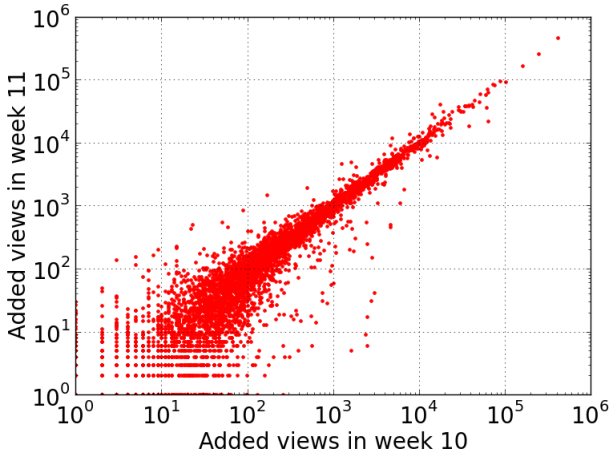
Figure 6.3: Distribution of added views for the available keyword-search and the removed keyword-search videos binned by added views in week 15



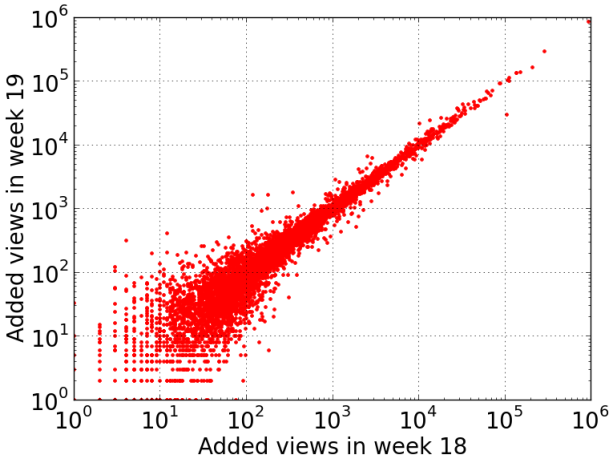
(a) Week 2 vs 3



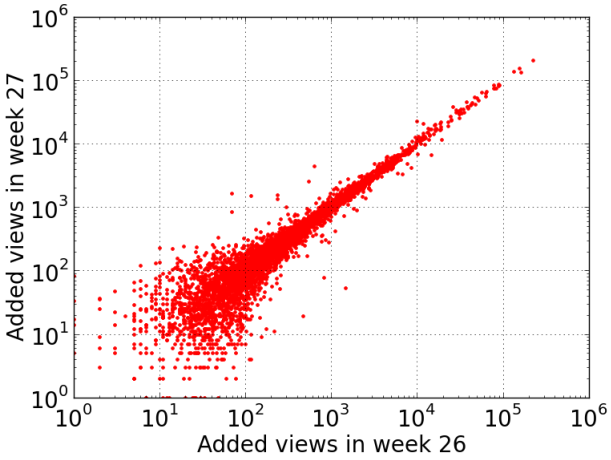
(b) Week 6 vs 7



(c) Week 10 vs 11

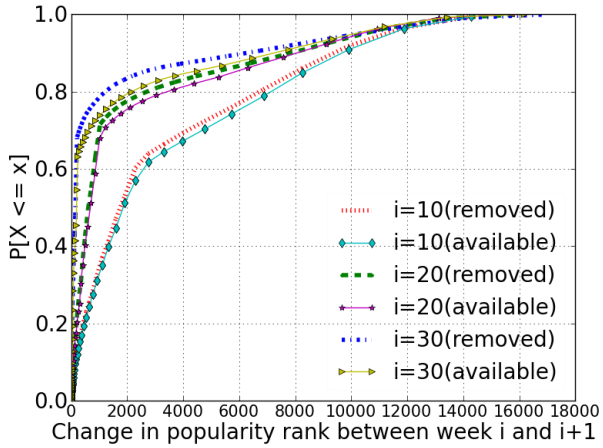


(d) Week 18 vs 19

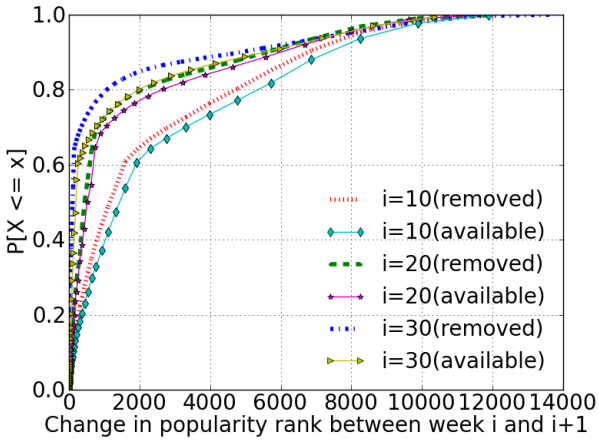


(e) Week 26 vs 27

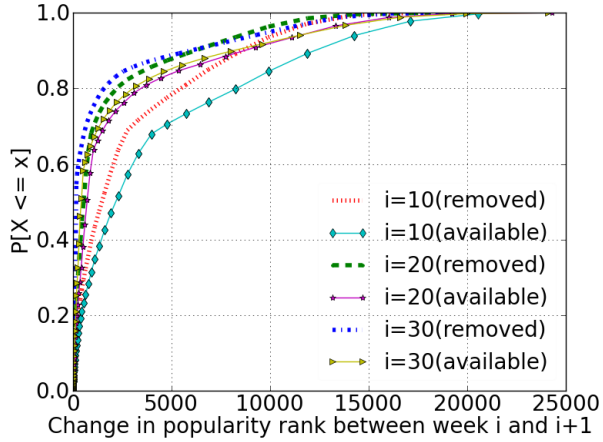
Figure 6.4: Scatter plot of added views for the removed keyword-search videos in week i vs week i+1



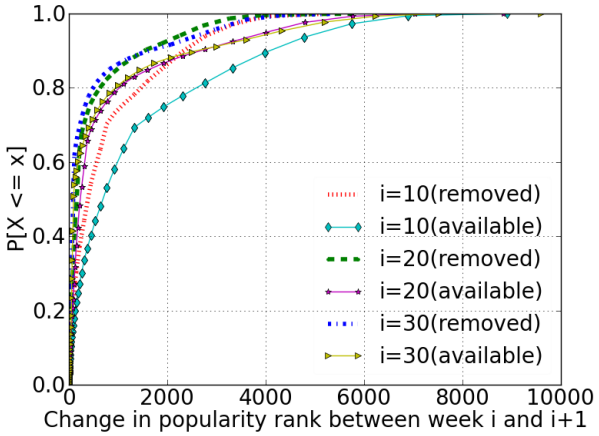
(a) 0 to 1 weeks old at seed time



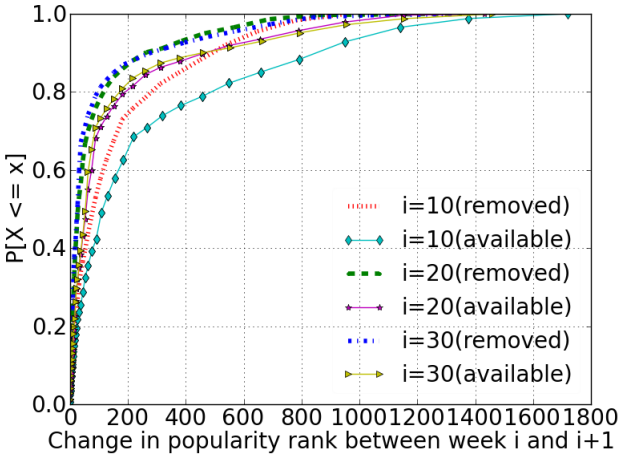
(b) 3 to 4 weeks old at seed time



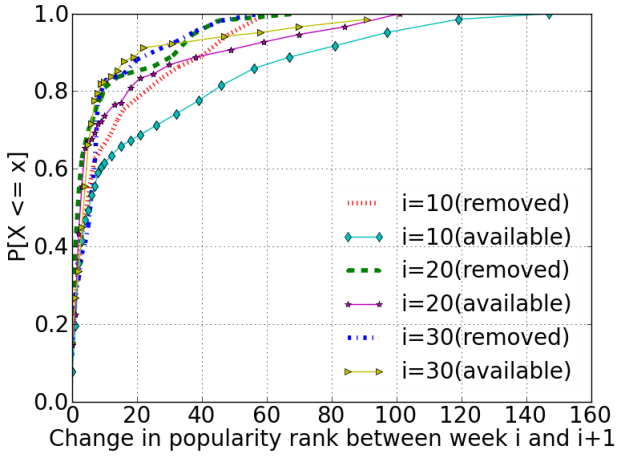
(c) 48 to 52 weeks old at seed time



(d) 100 to 104 weeks old at seed time

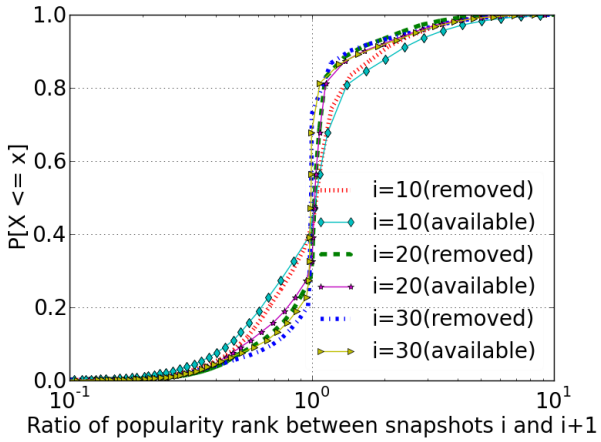


(e) 126 to 130 weeks old at seed time

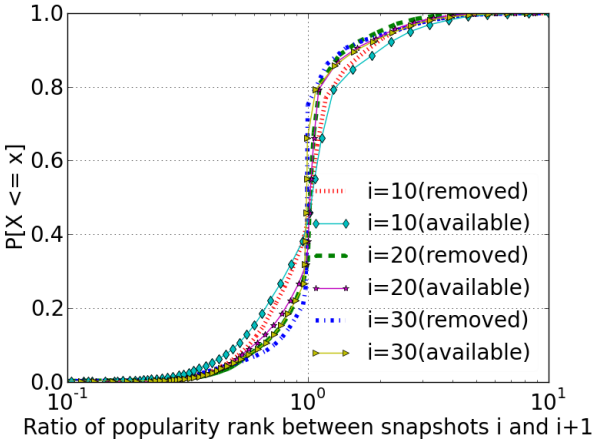


(f) 146 to 156 weeks old at seed time

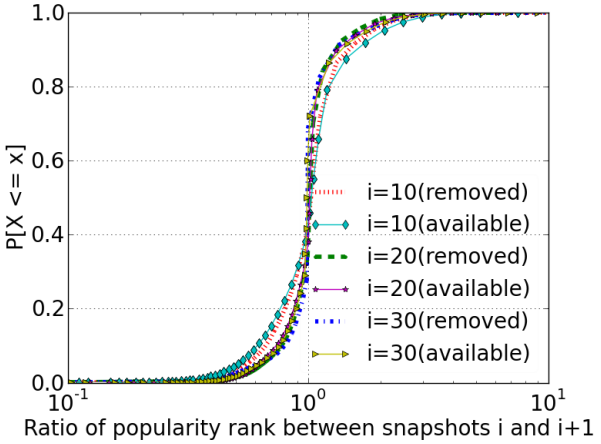
Figure 6.5: Distribution of absolute change in popularity rank for the available keyword-search and removed keyword-search videos binned by age at seed time



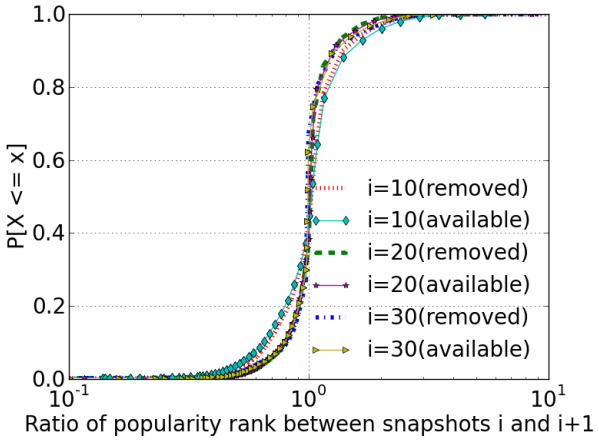
(a) 0 to 1 weeks old at seed time



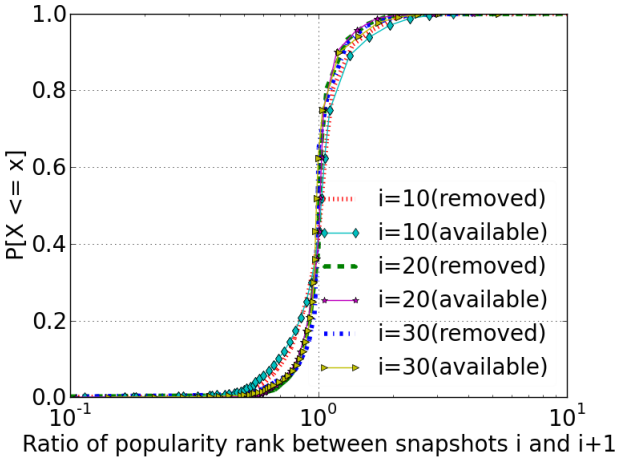
(b) 3 to 4 weeks old at seed time



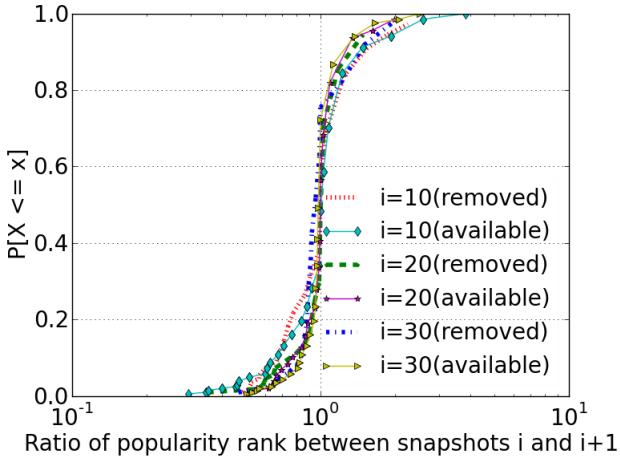
(c) 48 to 52 weeks old at seed time



(d) 100 to 104 weeks old at seed time



(e) 126 to 130 weeks old at seed time



(f) 146 to 156 weeks old at seed time

Figure 6.6: Distribution of ratio of new to old popularity rank for the available keyword-search and removed keyword-search videos binned by age at seed time

videos. For deeper understanding, videos are binned according to their age at seed age. Figure 6.6 shows smaller differences between the removed keyword-search and the available keyword-search videos than those shown in Figure 6.5. However, the keyword-search videos that experience higher rank changes (Figure 6.5) tend to be relatively unpopular videos with large rank values (and so even large absolute changes give small change ratios).

6.2 Removed Recently-uploaded Videos

Figure 6.7 shows the average added views in each week for both available and removed recently-uploaded videos. The average weekly viewing rate for the removed videos is significantly higher throughout the first measurement period, compared to that for the available videos. Interestingly, both sets of videos show similar increasing/decreasing trends of the average added views for the first 15 weeks of the measurement period, but from week 15 to the end, the gap between these two sets narrows since the average added views for the removed videos decreases while that for the available videos remains roughly constant. Recall that analysis of the time-to-peak distribution shows that approximately three-quarters of the recently-uploaded videos experience their peak popularity during the first six weeks, and so it is not surprising that the average viewing rate is higher during that period for both removed and available videos.

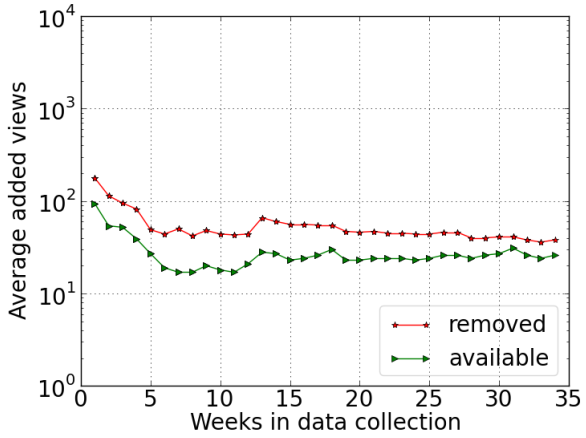


Figure 6.7: Average added views for the available recently-uploaded and the removed recently-uploaded videos in each week

Figure 6.8 shows the CCDF of the added views in weeks 2, 5, 10 and 25 using a logarithmic scale on each axis, for both the available and removed recently-uploaded videos. As can be seen by comparing the plots for the available and removed videos for the same week, the distribution for the removed videos is shifted towards somewhat higher view counts over a substantial portion of its range, although the highest observed view counts for the removed videos are similar to those for the available videos.

A closer look at the weekly added views is achieved by separating videos into different bins according to

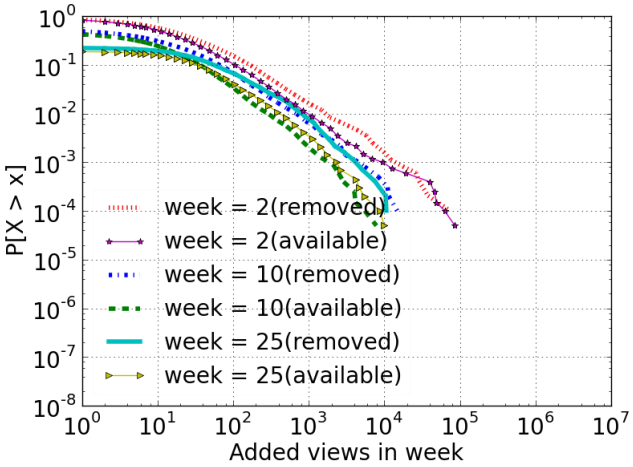
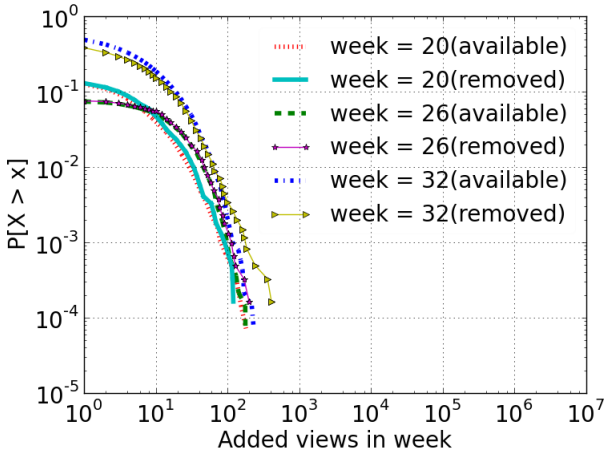


Figure 6.8: Distribution of added views for the available recently-uploaded and the removed recently-uploaded videos at different weeks

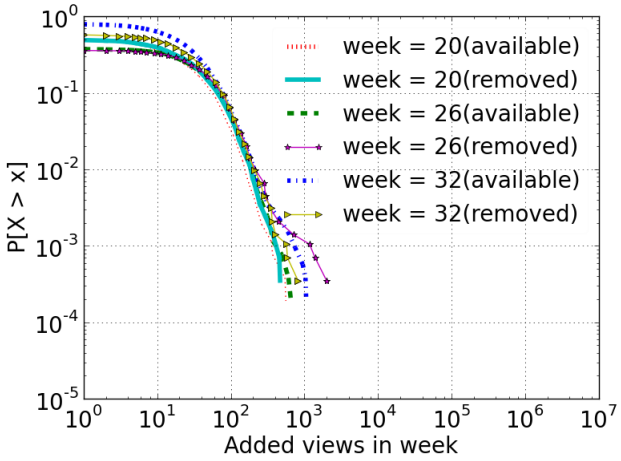
their added views in week 15. Figure 6.9 shows the CCDF of the added views for different weeks and various bins of the removed and available videos. Although there are some differences between the corresponding plots for the removed and available videos, in general the popularity distribution patterns are quite similar for both types of videos. Note, however, that the distributions of added views in week 15 for the removed and available videos are somewhat different (Figure 6.7), and therefore these videos have differing probabilities of being in a particular bin.

Figure 6.10 plots the weekly views in adjacent weeks for the removed recently-uploaded videos. The goal for such analysis is the observation of popularity churn patterns for these videos and how their popularity churn differs compared to that for the available recently-uploaded videos (Figure 4.4). Figure 6.10 and Figure 4.4 show similar probability churn patterns for the removed recently-uploaded videos and the available recently-uploaded videos. For example, for both the removed and available recently-uploaded videos, the points are more spread out for young ages (Figure 6.10(a)) than older ages, and as video age increases, the point spreads decrease gradually.

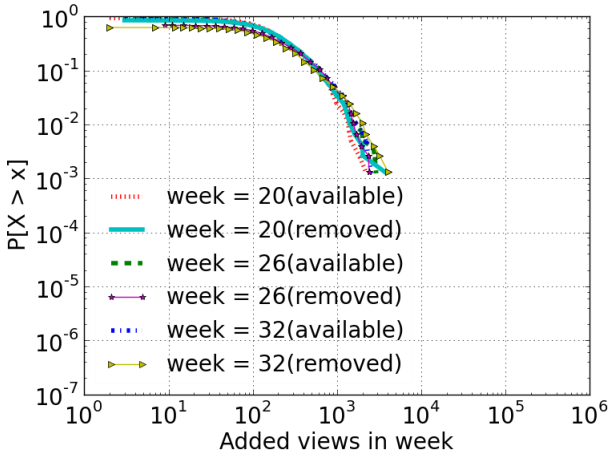
Non-stationarity in the relative popularities of the removed recently-uploaded videos is analyzed by looking at the popularity rank shifts in adjacent measurement weeks, in comparison with those for the available recently-uploaded videos. Figure 6.11(a) presents the CDF of the absolute value of the change in popularity rank for some example snapshots for both the available and the removed recently-uploaded videos. Example weeks are selected following the initial weeks, to avoid the high popularity churn observed for these early weeks. Significant differences in the rank change distribution are observed between the removed recently-uploaded videos and the available recently-uploaded videos. (When interpreting such figures, it should be noted that the total numbers of removed and available recently-uploaded videos are different.) For example, between snapshot ten and eleven approximately 25% of the available recently-uploaded videos experienced a change in rank of 8,000 or more, while below 5% of the removed recently-uploaded videos experienced such



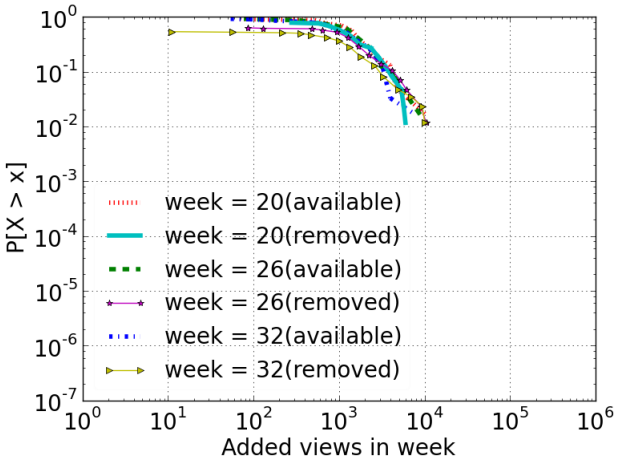
(a) 0 to 9 added views in week 15



(b) 10 to 99 added views in week 15

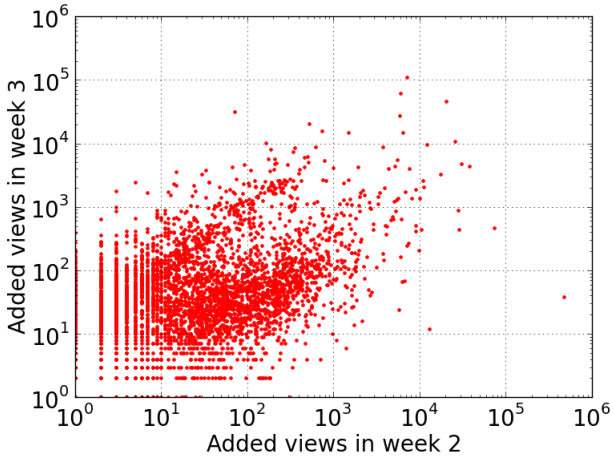


(c) 100 to 999 added views in week 15

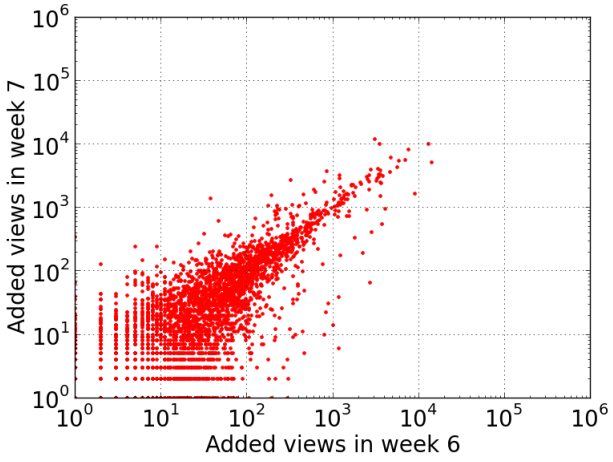


(d) 1000 to 9999 added views in week 15

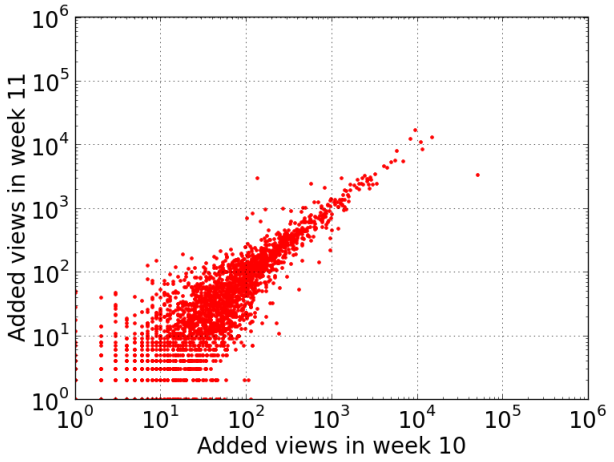
Figure 6.9: Distribution of added views for the available recently-uploaded and the removed recently-uploaded videos binned by added views in week 15



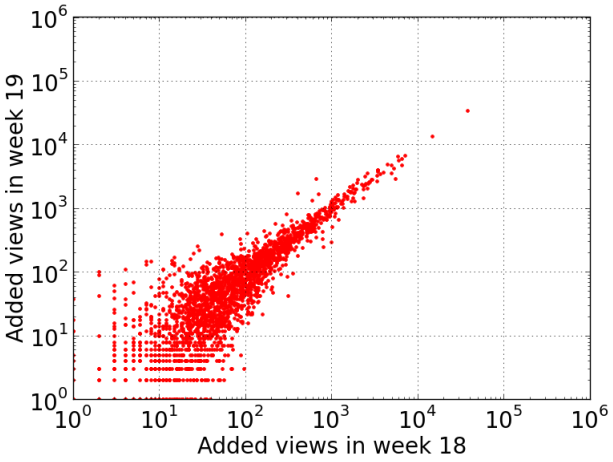
(a) Week 2 vs 3



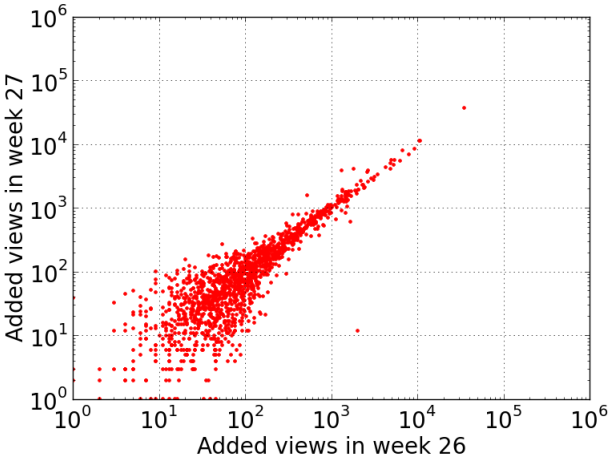
(b) Week 6 vs 7



(c) Week 10 vs 11

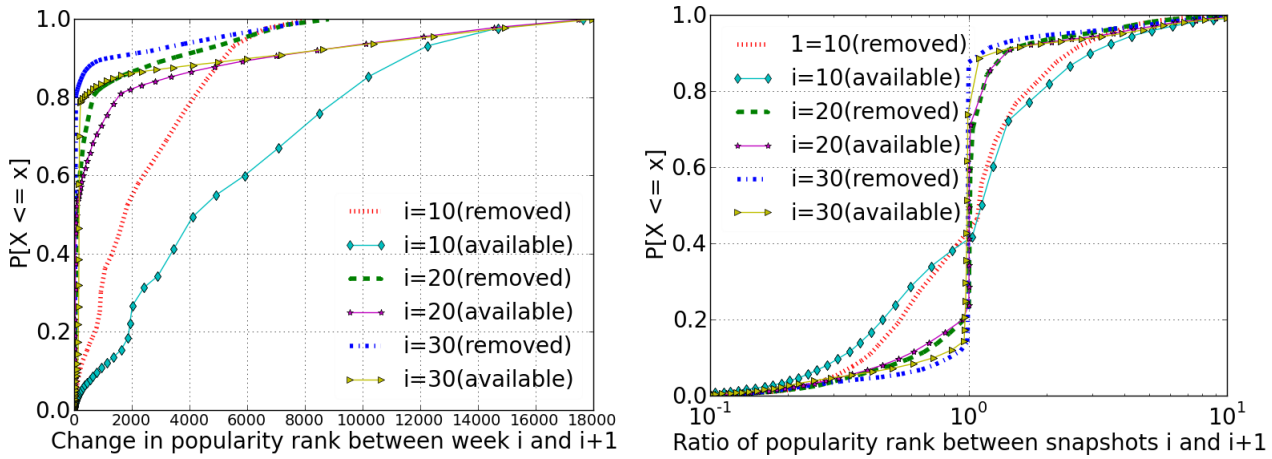


(d) Week 18 vs 19



(e) Week 26 vs 27

Figure 6.10: Scatter plot of added views for the removed recently-uploaded videos in week i vs week i+1



(a) Absolute change in popularity rank

(b) Ratio of new to old popularity rank

Figure 6.11: Distribution of change in popularity rank for the removed recently-uploaded videos

a rank change. Figure 6.11(b) complements Figure 6.11(a) by plotting the CDF of the ratio of new to old popularity ranks across adjacent weeks. There are no significant differences between the removed recently-uploaded videos and the available recently-uploaded videos observed in the figure. The recently-uploaded videos that experience higher rank change (Figure 6.11(a)) tend to be relatively unpopular videos with large rank values (and so even large absolute changes give small change ratios).

Chapter 7

Conclusions and Future Work

An efficient content distribution system can have technical, economical and social impact. Inefficient systems for sharing and distributing content can have service bottlenecks and high transmission latency. Knowledge of content popularity characteristics can provide insight into significant system design trade-offs for efficient content distribution policy design. Workload models based on empirically observed content popularity characteristics can be usefully applied in evaluation of various content distribution policies and architectures. One goal of this thesis is the observation of popularity dynamics and churn for user-generated content, specifically for YouTube videos. The assessment of the accuracy of the model of Borghol *et al.* [5] when used to generate synthetic video view counts for long time periods, and the observation of the popularity characteristics of videos that are removed from YouTube, are also goals of this thesis. This chapter presents a summary of the work that has been done throughout the thesis, its main contributions, and some guidelines for future work. Section 7.1 presents the summary of the thesis. The main contributions of the thesis are presented in Section 7.2. Section 7.3 describes some possible directions for future work.

7.1 Thesis Summary

Borghol *et al.* [5] collected two YouTube empirical datasets in 2008/09, namely the "recently-uploaded" and "keyword-search" datasets, for use in characterization and modelling of popularity evolution. Based on the empirical behaviour observed in the recently-uploaded dataset, they developed a workload model that can generate synthetic view counts with popularity dynamics similar to those of collections of recently-uploaded videos. The distribution of the total view counts and that of the weekly view counts generated by the model accurately match the corresponding distributions for the recently-uploaded empirical dataset. Borghol *et al.* [5] observed the popularity evolution of the videos in the recently-uploaded dataset for only the first 8 months of the video lifetime, and this was the time period for which their workload model was tested. This thesis was motivated by a wish to observe and characterize popularity evolution for older video ages, and to assess the accuracy of the Borghol *et al.* model for longer time periods.

Chapter 2 presents an overview of some previous work on popularity characteristics for traditional content such as web pages, videos in video on demand systems, files in peer-to-peer file sharing systems and one-click hosting systems, and TV channels in IPTV systems. Researchers have also done significant work

on user-generated content popularity and popularity modelling. Chapter 2 also contains a review of previous work that has examined the popularity characteristics of user-generated videos and pictures, as well as work concerned with modelling popularity evolution.

The datasets that are used in the thesis research are described in Chapter 3. These datasets (recently-uploaded and keyword-search) were collected in two phases. The first phase of data collection was carried out by Borghol *et al.* from 27 July 2008 to 29 March 2009 using the YouTube API. They collected meta-data for 29,791 recently-uploaded videos and 1,135,253 keyword-search videos. The second phase of data collection was carried out as a part of the thesis research for the same videos from 25 October 2011 to 27 December 2011 by a crawler developed using the YouTube API. Before that a crawler separated the videos that were still available, from those that had been removed since the first period of data collection. The second phase of data collection was carried out for the meta data of 20,000 recently-uploaded and 627,002 keyword-search videos.

The popularity characteristics for both recently-uploaded and keyword-search datasets and both phases of data collections are presented in Chapter 4. Popularity characteristics analysis in this chapter include average viewing rate and view count distribution, including when videos are binned by age and/or popularity in some prior week. Popularity dynamics and churn are analyzed by showing the scatter plot of added views in adjacent weeks, the distribution of absolute change in popularity rank and the distribution of the ratio of new to old popularity rank. The analysis in this chapter shows that the popularity evolution pattern is similar for keyword-search videos of the same age even at different measurement weeks.

Chapter 5 presents the popularity modelling using both the "basic" and the "extended" variants of the model defined by Borghol *et al.* [5]. Development of the model requires the distribution of the time-to-peak. Therefore, the time-to-peak distribution for the recently-uploaded videos for both the first measurement period and the second measurement period is analyzed in this chapter, and an analytic time-to-peak distribution is determined. The three-phase viewing rate distribution characteristics for the empirical dataset and their comparison with those of the synthetic dataset resulting from use of the model are also analyzed in this chapter. Distribution of the weekly view counts and the total view counts of the synthetic dataset are also analyzed and the model accuracy is analyzed by comparing with the empirical dataset. Popularity dynamics and churn in the synthetic dataset generated by the basic model is also analyzed and for achieving additional accuracy, the extended model is developed by exchanging view counts for videos within the same phase (before-peak or after-peak). Popularity churn in the dataset generated by the extended model is compared with that in the empirical dataset by analyzing the hot set overlap between adjacent weeks and with week 2 of the measurement period.

Popularity characteristics for the videos (removed videos) that were available in the first measurement period but absent in the second measurement period are analyzed in Chapter 6. The average viewing rate and the view count distribution for both the removed recently-uploaded videos and the removed keyword-search videos are analyzed and compared to the corresponding characteristics for the videos that were still available

in the second measurement period. The comparison of popularity dynamics and churn between the available and the removed datasets is performed using observations from scatter plots, the distribution of the absolute change in popularity rank and the distribution of the ratio of new to old popularity rank.

7.2 Contributions

There are three main contributions in this thesis. Firstly, analysis of how popularity characteristics change as videos age beyond the 8 months covered in the recently-uploaded dataset collected by Borghol *et al.* [5], and analysis of video popularity characteristics and biases in the keyword-search dataset for both the initial measurement period and the second phase that is carried out as a part of this thesis research. Secondly, assessment of the accuracy of the Borghol *et al.* model when applied for periods of time longer than the 8 month period considered by Borghol *et al.* [5]. Thirdly, analysis of the popularity characteristics for the videos that were available in the first phase of the data collection but unavailable in the second phase of data collection. In particular, the contributions are as follows:

- The first contribution is the development of two crawlers using the YouTube API. The first crawler is used to separate available and removed videos in the recently-uploaded and keyword-search datasets collected by Borghot *et al.* [5]. It was observed that 67.13% of the recently-uploaded videos and 55.23% of the keyword-search videos were still available. The second crawler collected the weekly view counts for the available videos for over two months.

- The analysis of view count distributions for both the recently-uploaded and keyword-search datasets is the second contribution of this thesis. The analysis showed that the average weekly view count for recently-uploaded videos is approximately the same in the second measurement period in comparison to the middle and later portions of the first measurement period. However, analysis of the view count distribution for the recently-uploaded videos showed that the videos that are most popular in the weeks of the second measurement period, have substantially higher weekly view counts than those that are most popular in the middle and later portions of the first measurement period. For the keyword-search videos, the analysis suggests that even though this dataset is biased towards more popular videos, the observed popularity evolution when videos are binned according to both age and previous popularity may be representative of what one would see with randomly selected videos in the same age and popularity bins.

- The popularity dynamics and churn analysis using the recently-uploaded dataset showed that the popularity churn decreases with age. One of the reasons for such popularity churn is differences in the rate at which videos attain their peak popularity. Approximately three quarters reach their peak popularity during the first six weeks since upload. Some videos received their peak popularity in the

second measurement period. Therefore, videos showed higher churn at their early age and popularity churn decreased with the increase of age.

- The analysis of the accuracy of the both basic and extended model of Borghol *et al.* [5] for the second measurement period is the most significant contribution of the thesis. The model can exhibit the empirical view count distribution and showed excellent match in total view count distribution, but due to the week-invariant assumptions in the model development and week-dependent behaviour in the empirical data, the model shows difference in the weekly binned view distribution and the popularity churn. Therefore to observe good matches in all metrics, it appears that the model needs to change the view count distribution for each phase to make it dependent on the week.

- A significant number of the videos in the recently-uploaded and keyword-search datasets were removed from YouTube during the intervening time between the first and second measurement periods. The analysis of the popularity characteristics of those removed videos is carried out in the thesis. The analysis shows that the removed recently-uploaded videos have a substantially higher average weekly view count in the first measurement period than the recently-uploaded videos that were still available in the second period. Such a difference is not observed for the keyword-search videos. Also, for both datasets, the removed videos tended to experience lower popularity churn during the first measurement period than the videos that were still available in the second measurement period.

7.3 Future Work

The popularity characteristics and the model accuracy analyzed in this thesis could be further improved in future by some additional research work. Some directions for future work are as follows.

- The accuracy analysis of the model shows that the model data can reveal the same distribution for the total view count and the added view count as observed in the empirical dataset. However, the model could not significantly reveal the empirical characteristics of view count distribution for different age and popularity bins. Similarly, popularity churn in the scatter plot of added views and in the number of videos peaking in different weeks could not exactly match due to the use of uniform distribution policy after some initial weeks. Therefore, a complete model could be designed in the future that can provide exactly same characteristics as observed in the empirical recently-uploaded dataset and also that will be applicable not only for the recently-uploaded YouTube datasets but also for other content delivery datasets.

- The workload model is compared with the empirical recently-uploaded dataset which is quite small, even smaller for the second measurement period. Therefore, the view count distribution for different popularity bins may not reveal the characteristics accurately due to the limited number of videos in

each bin. A large empirical recently-uploaded dataset (for example, a dataset ten times bigger) can provide a more accurate assessment of the quality of the match between the empirical and the synthetic dataset.

- The extended model introduces additional churn and the level of churn depends on the value of the model parameter g. In the model accuracy evaluation, it was observed that the model can provide approximately similar churn in both the first and second measurement period but with different g values. Therefore the future work can develop an efficient model that will also provide the trade-off between the video age and g value for the achievement of more accurate churn.

- The model is evaluated with only one dataset (recently-uploaded YouTube dataset). In future work the accuracy of the model could be evaluated for different sets of empirical data perhaps including data collected from different types of content delivery sites (e.g. picture sharing sites instead of video sharing sites).

- YouTube receives a huge number of new videos, as well as new users, every day. In particular, YouTube received a large number of such new users between the first and the second measurement periods over which the recently-uploaded and keyword-search datasets were collected. The model proposed by Borghol *et al.* [5] can not reveal the impact of the new users on popularity characteristics. Therefore as future work, a more suitable model could be developed that can reveal the impact of the activities of new users, as well as modelling the addition of new videos to the system.

References

[1] S. Acharya, B. C. Smith, and P. Parnes. Characterizing user access to videos on the World Wide Web. In *Proc. SPIE Int Soc Opt Eng '99*, pages 130–141, Porsn Lule, Sweden, Dec. 1999.

[2] J. M. Almeida, J. Krueger, D. L. Eager, and M. K. Vernon. Analysis of educational media server workloads. In *Proc. ACM NOSSDAV '01*, pages 21–30, Port Jefferson, NY, USA, Jan. 2001.

[3] D. Antoniades, E. P. Markatos, and C. Dovrolis. One-click hosting services: a file-sharing hideout. In *Proc. ACM SIGCOMM '09*, pages 223–234, Chicago, IL, USA, Nov. 2009.

[4] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and K. Ross. Video interactions in online video social networks. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 5(4):30:1–30:25, Nov. 2009.

[5] Y. Borghol, S. Mitra, S. G. Ardon, N. Carlsson, D. L. Eager, and A. Mahanti. Characterizing and modeling popularity of user-generated videos. *Performance Evaluation*, 68(11):1037–1055, Nov. 2011.

[6] L. Breslau, p. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and Zipf-like distributions: evidence and implications. In *Proc. IEEE INFOCOM '99*, pages 126 –134, New York, NY, USA, Mar. 1999.

[7] T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer. Catching a viral video. In *Proc. IEEE ICDMW '10*, pages 296–304, Sydney, Australia, Dec. 2010.

[8] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proc. ACM SIGCOMM '07*, pages 1–14, San Diego, CA, USA, Oct. 2007.

[9] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Networking*, 17(5):1357–1370, Oct. 2009.

[10] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proc. ACM WWW '09*, pages 721–730, Madrid, Spain, Apr. 2009.

[11] M. Cha, P. Rodriguez, J. Crowcroft, S. Moon, and X. Amatriain. Watching television over an IP network. In *Proc. ACM SIGCOMM '08*, pages 71–84, Vouliagmeni, Greece, Oct. 2008.

[12] M. Cha, P. Rodriguez, S. Moon, and J. Crowcroft. On next-generation telco-managed P2P TV architectures. In *Proc. IPTPS '08*, pages 5–5, Tampa Bay, FL, USA, Feb. 2008.

[13] X. Cheng. Understanding the characteristics of Internet short video sharing: YouTube as a case study. In *Proc. ACM SIGCOMM '07*, pages 28–28, San Diego, CA, USA, Oct. 2007.

[14] M. Chesire, A. Wolman, G. M. Voelker, and H. M. Levy. Measurement and analysis of a streaming-media workload. In *Proc. USITS '01*, pages 1–12, San Francisco, CA, USA, Mar. 2001.

[15] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, Nov. 2009.

[16] R. P. Doyle, J. S. Chase, Gadde S., and A. M. Vahdat. The trickle-down effect: Web caching and server request distribution. *Computer Communications*, 25(4):345 – 356, June 2000.

[17] F. Figueiredo, F. Benevenuto, and J. M. Almeida. The tube over time: characterizing popularity growth of YouTube videos. In *Proc. WSDM '11*, pages 745–754, Hong Kong, China, Feb. 2011.

[18] S. Fortunato, A. Flammini, and F. Menczer. Scale-free network growth by ranking. *Physics Review Letters*, 96(21):218701, May 2006.

[19] C. Griwodz, M. Bär, and L. C. Wolf. Long-term movie popularity models in video-on-demand systems:

or the life of an on-demand movie. In *Proc. ACM MULTIMEDIA '97*, pages 349–357, Seattle, WA, USA, Nov. 1997.

[20] 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 *Proc. ACM SOSP '03*, pages 314–329, Bolton Landing, NY, USA, Dec. 2003.

[21] L. Guo, S. Chen, Z. Xiao, and X. Zhang. Analysis of multimedia workloads with implications for Internet streaming. In *Proc. ACM WWW '05*, pages 519–528, Chiba, Japan, May 2005.

[22] X. Hei, C. Liang, Liang J., Y. Liu, and K. W. Ross. A measurement study of a large-scale P2P IPTV system. *IEEE Transactions on Multimedia*, 9(8):1672 –1687, Dec. 2007.

[23] C. Huang, J. Li, and K. W. Ross. Can Internet video-on-demand be profitable? In *In Proc. ACM SIGCOMM '07*, volume 37(4), pages 133–144, Kyoto, Japan, Aug. 2007.

[24] S. Ihm and V. S. Pai. Towards understanding modern web traffic. In *Proc. ACM SIGCOMM '11*, pages 295–312, Berlin, Germany, Nov. 2011.

[25] D. L. Johnson, E. M. Belding, K. Almeroth, and V. G. Stam. Internet usage and performance analysis of a rural wireless network in Mache, Zambia. In *Proc. ACM NSDR '10*, pages 7:1–7:6, San Francisco, CA, USA, June 2010.

[26] A. Kaltenbrunner, V. Gomez, and V. Lopez. Description and prediction of slashdot activity. In *Proc. LA-WEB '07*, pages 57 –66, Santiago de Chile, Chile, Nov. 2007.

[27] J. G. Lee, S. Moon, and K. Salamatian. An approach to model and predict the popularity of online contents with explanatory factors. In *Proc. IEEE WI-IAT '10*, pages 623–630, Toronto, ON, Canada, Aug. 2010.

[28] N. Leibowitz, A. Bergman, R. Ben-shaul, and A. Shavit. Are file swapping networks cacheable? characterizing p2p traffic. In *Proc. Workshop on Web Content Caching and Distribution '02*, Boulder, Colorado, USA, Aug. 2002.

[29] H. Ma and K. G. Shin. Multicast video-on-demand services. *ACM SIGCOMM Computer Communication Review*, 32(1):31–43, Jan. 2002.

[30] A. Mahanti, N. Carlsson, and C. Williamson. Content sharing dynamics in the global file hosting landscape. In *Proc. IEEE MASCOTS '12*, pages 219–228, Arlington, VA, USA, Aug. 2012.

[31] A. Mahanti, C. Williamson, N. Carlsson, M. Arlitt, and A. Mahanti. Characterizing the file hosting ecosystem: A view from the edge. *Performance Evaluation*, 68(11):1085 – 1102, Nov. 2011.

[32] M. Marcon, B. Viswanath, M. Cha, and K. P. Gummadi. Sharing social content from home: a measurement-driven feasibility study. In *Proc. ACM NOSSDAV '11*, pages 45–50, Vancouver, BC, Canada, June 2011.

[33] S. Mitra, M. Agrawal, A. Yadav, N. Carlsson, D. Eager, and A. Mahanti. Characterizing web-based video sharing workloads. *ACM Transactions on the Web*, 5(2):8:1–8:27, May 2011.

[34] R. D. Oliveira, M. Cherubini, and N. Oliver. Looking at near-duplicate videos from a human-centric perspective. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 6(3):15:1–15:22, Aug. 2010.

[35] V. N. Padmanabhan and L. Qiu. The content and access dynamics of a busy web site: findings and implications. In *Proc. SIGCOMM '00*, pages 111–123, Stockholm, Sweden, Aug. 2000.

[36] T. S. Perry. The trials and travails of interactive tv. *IEEE Spectrum Magazine*, 33(4):22 –28, Apr. 1996.

[37] T. Qiu, Z. Ge, S. Lee, J. Wang, Q. Zhao, and J. Xu. Modeling channel popularity dynamics in a large iptv system. In *Proc. ACM SIGMETRICS '09*, pages 275–286, Seattle, WA, USA, June 2009.

[38] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Characterizing and modeling the dynamics of online popularity. *Physic Review Letters*, 105(15):158701, Oct. 2010.

[39] K. Salah, J. Hamodi, Z. A. Baig, and F. Al-Haidari. Video-on-demand (vod) deployment over hospitality networks. *International Journal of Network Management*, 21(4):65–80, July 2011.

[40] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3-4):425–440, Dec. 1955.

[41] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the*

ACM, 53(8):80–88, Aug. 2010.

[42] W. Tang, Y. Fu, L. Cherkasova, and A. Vahdat. Long-term streaming media server workload analysis and modeling. *HP Technical Report, HP Laboratories*, Jan. 2003.

[43] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng. Understanding user behaviour in large-scale video-on-demand systems. *SIGOPS Operating Systems Review*, 40(4):333–344, Apr. 2006.

[44] J. Zhou, Y. Li, V. K. Adhikari, and Z. Zhang. Counting youtube videos via random prefix sampling. In *Proc. ACM SIGCOMM '11*, pages 371–380, Berlin, Germany, Aug. 2011.

[45] M. Zink, K. Suh, Y. Gu, and J. Kurose. Watch global, cache local: Youtube network traffic at a campus network - measurements and implications. In *Proc. IEEE MMCN '08*, pages 1–13, San Jose, CA, USA, Jan. 2008.