Characterizing Web-based Video Sharing Workloads

Siddharth Mitra, Mayank Agrawal, Amit Yadav
Indian Institute of Technology Delhi, New Delhi, India
and
Niklas Carlsson
Linköping University, Linköping, Sweden
and
Derek Eager
University of Saskatchewan, Saskatoon, SK, Canada
and
Anirban Mahanti
NICTA, Alexandria, NSW, Australia

Video sharing services that allow ordinary Web users to upload video clips of their choice and watch video clips uploaded by others have recently become very popular. This paper identifies *invariants* in video sharing workloads, through comparison of the workload characteristics of four popular video sharing services. Our traces contain meta-data on approximately 1.8 million videos which together have been viewed approximately 6 billion times. Using these traces, we study the similarities and differences in use of several Web 2.0 features such as ratings, comments, favorites, and propensity of uploading content. In general, we find that active contribution, such as video uploading and rating of videos, is much less prevalent than passive use. While uploaders in general are skewed with respect to the number of videos they upload, the fraction of multi-time uploaders is found to differ by a factor of two between two of the sites. The distributions of life-time measures of video popularity are found to have heavy-tailed forms that are similar across the four sites. Finally, we consider implications for system design of the identified invariants. To gain further insight into caching in video sharing systems, and the relevance to caching of life-time popularity measures, we gathered an additional data set tracking views to a set of approximately 1.3 million videos from one of the services, over a twelve week period. We find that life-time popularity measures have some relevance for large cache (hot set) sizes (i.e., a hot set defined according to one of these measures is indeed relatively "hot"), but that this relevance substantially decreases as cache size decreases, owing to churn in video popularity.

Categories and Subject Descriptors: C.2.0 [**Computer-Communications Networks**]: General

1. INTRODUCTION

Web-based video sharing services have recently become enormously popular and have revolutionized distribution of online video content. A video sharing service allows "user generated" video clips to be uploaded, and users of the service to view, rate, and comment on uploaded videos. One common aspect of these services is the ease of uploading, searching, and viewing videos. Typically, these sites allow content producers to upload content encoded using any commonly used codec. This uploaded video is converted by the service provider to a common format (in most cases Flash), thus enabling most Web users to view the content without searching for different codecs to view different videos. Availability of a large number of diverse videos, the ease of viewing at the click of a button, and in many cases the ability to form social groups with content uploaders and viewers alike also contributes towards the increased popularity of these services.

Video sharing services have only recently become popular, and thus it is not surprising that there has been only limited effort towards understanding the general characteristics of these workloads. To the best of our knowledge, prior work in this domain has focused mostly on the YouTube[1] video sharing service [Cha et al.; Gill et al.; Zink et al.; Cheng et al.; Halvey and Keane]. While YouTube is arguably the most popular video sharing service, there are many other popular services that offer similar services or offer services that target a niche audience (such as customers willing to pay for high quality videos, or a particular non-English speaking population). Studying the workload characteristics of other video sharing services, and in particular identifying invariant properties as well as significant differences among these services is an important step towards building a broader understanding of this new type of workload.

With the aforementioned objective, we collected traces from four video sharing services: Dailymotion[2], Yahoo! video[3], Veoh[4], and Metacafe[5].[6] Dailymotion is France's leading video sharing service and caters mostly to French-speaking demographics, while Yahoo! video, Veoh, and Metacafe are US-based services. (Metacafe was originally headquartered in Israel.) While all four host user generated video clips, Veoh, in addition, also serves content from major studios and independent production houses, and utilizes peer-to-peer technology to distribute videos longer than 20 minutes. Metacafe is distinctive among these services in its use of a revenue sharing model in which content creators are paid for videos that exceed a certain threshold of views. These services cover a spectrum of possibilities in the realm of video sharing; our principal contribution is their workload characterization and comparison.

Our key contributions are summarized below:

—We present and analyze workload data from *four* video sharing services. In aggregate, our traces contain meta-data on 1.8 million videos which together

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

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

[3] http://video.yahoo.com

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

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

[6] Our data sets are available at: http://www.cs.usask.ca/faculty/eager/TWeb10.html.

acquired more than 6 billion views.

- We identify seven key invariants of these workloads, concerning aspects such as the video popularity distribution, use of social and interactive features, and the uploading of new content. These invariants are summarized and discussed in Section 5.6.
- We also find some significant differences across these services. For example, while the number of video uploads by users follows the Pareto principle, the fraction of multi-time uploaders is almost two times larger with Veoh (65%) than with Yahoo! (33%).
- We also consider implications for system design of the identified invariants. Further insight into caching in video sharing systems, and the relevance to caching of life-time popularity measures are realized using an additional data set that tracks views to approximately 1.3 million videos from the Dailymotion service, over a twelve week period. We find that life-time popularity measures have some relevance for large cache (hot set) sizes (i.e., a hot set defined according to one of these measures is indeed relatively "hot"), but that this relevance substantially decreases as cache size decreases, owing to churn in video popularity. These systems implications are discussed in Section 6.

We believe that our work, together with recent complementary work characterizing YouTube workload, provides insights for the design of improved content distribution systems, new video search and recommendation systems, and appropriate workload models.

Our paper is structured as follows. Related work is discussed in Section 2. Section 3 presents a brief overview of the video sharing services considered in this paper. Our data collection methodology is described in Section 4, followed by the measurement results in Section 5. Section 6 presents systems implications and insights, in part using an additional data set. Conclusions are presented in Section 7.

2. RELATED WORK

Prior work on Web-based video sharing has primarily focused on YouTube, and is based on either crawling the site [Cha et al. 2007; Cheng et al. 2008; Halvey and Keane 2007a; 2007b] or on collection of YouTube specific traffic at university networks [Zink et al. 2008; Gill et al. 2007; 2008].

Various properties of YouTube flash video files have been examined. For example, the encoded bit rate of a large fraction of the videos is found to be between 300 and 400 Kbps [Gill et al. 2007; Cheng et al. 2008]. A typical video on YouTube is around 3 to 5 minutes long, illustrating that short videos are the norm [Gill et al. 2007; Cheng et al. 2008], and on average between 8 to 10 MB in size [Gill et al. 2007; Zink et al. 2008].

Video popularity and file referencing behavior have also been studied [Gill et al. 2007; Cheng et al. 2008; Zink et al. 2008; Cha et al. 2007]. For example, using meta-data on videos obtained by crawling YouTube's science and entertainment categories, Cha et al. [2007] find that the Pareto principle (cf. Section 5) applies to the total views since a video's upload. Gill et al. [2007] and Zink et al. [2008] find that the Pareto principle applies weakly to YouTube video accesses as seen at the gateway of their respective university networks. The applicability of Zipf's model to the number of video views (references) has also been considered. Crawling-based techniques show that the number of video views since upload follows Zipf-like behavior with cut off [Cha et al. 2007; Cheng et al. 2008], while edge-based analysis finds a Zipf model to be reasonable for video accesses [Gill et al. 2007]. The "fetch at most once" model [Gummadi et al. 2003] has been suggested as providing one possible explanation for deviation from Zipf-like behavior [Cha et al. 2007]. It has also been suggested that recommendation systems and common crawling approaches (that start from a list of the most popular videos and follow links to related videos) may skew results towards popular content and weed out the unpopular content [Cheng et al. 2008].

We believe that it is important to distinguish between popularity as measured by the total views since upload, and popularity as measured by the viewing rate; the former is considered by Cha et al. [2007] and Cheng et al. [2008], while both Gill et al. [2007] and Zink et al. [2008] implicitly consider the latter. Note that Zipf models for Web and media file popularity concern popularity as measured by referencing rate (as determined by the number of references over the fixed period of a trace), and not the total number of references to a file since its creation (e.g., see [Yu et al. 2006; Arlitt and Williamson 1997; Breslau et al. 1999]).

There has also been some effort towards understanding how users interact with YouTube. For example, it has been found that YouTube videos in the entertainment, comedy, and music categories are viewed the most [Gill et al. 2007; Cheng et al. 2008]. Halvey and Keane [2007b] explored the social dynamics in YouTube's video sharing service based on meta-data obtained by crawling. They found that most users do not form social networks and only a small number of users post comments, ratings, and use other interaction tools.

3. VIDEO SHARING SERVICES

The goal of our work is to provide a detailed characterization of video sharing workloads, identify workload invariants, and understand how similar or different these workloads are from traditional Web and media workloads. For achieving the objectives of this work, we collected data from four different video sharing services, where the services were selected to cover a wide range of content, user demographics, and popularity. (In addition, to draw insights with regards to an even more diverse set of services, our discussion leverages existing studies of YouTube.) This section presents a brief overview of these services along with some information on their popularity and user demographics.

Dailymotion is a leading video publishing and sharing Web service headquartered in Paris. This service was launched in March 2005, around the same time as YouTube was launched. As of this writing, the maximum allowed size and duration of uploaded videos are 150 MB and 20 minutes, respectively. Only users designated as "motion makers" are exceptions to the above-stated limits. Dailymotion's (November 2009) `Alexa`[7] global Internet traffic rank (based on the number of visits to the site) is 87, with about 50% of its users coming from France, the United States, and Japan.

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

Yahoo! video is Yahoo's video publishing and sharing service. It was initially launched as a video searching site, and in June 2006 launched its video sharing service. According to the service's home page, the maximum allowed size of uploaded videos is 150 MB; no information regarding any upper limit on video duration is available. Yahoo! has an overall Alexa Internet traffic rank of 3; however, its various services are not individually ranked. We also could not find any published information on the number of page views, unique visitors, and uploads per month.

Veoh is a video sharing and Internet television service based in San Diego, California. This service hosts content from major studios, independent production houses, and ordinary Web users. Currently, Veoh offers two different services. The veoh.com site allows users to browse videos, create accounts, upload content, and preview videos within one's browser. The second service called VeohTV requires download of software that facilitates delivery of full-length, high-quality content, to desktops and portable devices. Veoh recommends upload of videos encoded for playback at 500 Kbps or higher. Videos of duration less than 20 minutes can be watched using a browser, while longer duration content requires use of the Veoh player application. As of November 2009, according to Alexa, a majority of the visits to this service were from Japan, and the service's global Internet traffic rank was 241.

Metacafe was founded in July 2003 in Israel, with the stated goal of promoting short videos that are specifically developed for entertaining the Internet audience. Every uploaded video is reviewed by a pool of volunteers to determine whether or not the video is suitable for the site. Metacafe attempts to ensure that copyrighted content is not uploaded, and has a preference for short duration clips. Metacafe also offers a unique revenue sharing model. Like all video sharing services, revenue is earned from advertisers. Uploaders of videos that achieve more than 20,000 views and an average rating of 3.0 or higher become eligible for a share of the revenue earned from advertisements. According to the site, some content producers have earned in excess of (US) $10,000 from their videos. Metacafe's current Alexa Internet traffic rank is 153; close to 40% of the visitors to this site are from India and the United States.

To summarize, the four services considered are popular, and are particularly strong in different geographic regions. Dailymotion is strong in France and other French-speaking countries, Veoh is popular in Japan, and Metacafe is popular in India. These services also differ with respect to their service models. While Dailymotion and Yahoo! video are free services, in contrast Veoh offers both free and premium services, and delivers full-length content using a peer-to-peer architecture. Metacafe is unique with respect to its revenue sharing program.

4. METHODOLOGY

4.1 Data Collection

Customized crawlers, written in Python, were designed to obtain meta-data associated with videos such as the number of views, number of ratings, and number of comments. Where necessary, the crawlers were used along with a customized Firefox browser. Requests issued by the crawlers were spaced in time to limit overloading the services. The crawlers did not download any videos and did not

Table I. High-level summary of data sets.

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

contribute to the view counts of the video clips; only textual information was downloaded. This approach reduced the overall network bandwidth consumption of our crawls, and also limited the load placed on the services. Only publicly available information is retrieved from the services. SQLite was used for all back-end needs. The remainder of this section describes how the sites were crawled.

4.1.1 *Dailymotion Data Collection.* At the time of data collection, videos on Dailymotion were divided into categories such as ‘Music’, ‘Action’, and ‘Humor’. Videos within a category could be found under several sub-categories such as “most popular”, “most recent”, “most viewed”, and “most commented”; these sub-categories sort the videos of a certain category according to the criterion chosen for listing. We determined that “most recent” and “most popular” gave us access to a large fraction of the videos in the other sub-categories. For this work, we crawled the ‘most recent’ and ‘most popular’ listings under the ‘music’ category, which consisted of approximately 90,000 and 50,000 pages, respectively, with each page listing 14 videos. (Dailymotion has since changed its site layout; currently, only a listing of 100 pages is accessible.) Our crawler downloaded the HTML source code of each video page and performed string matching on it to obtain per-video statistics. For each video, we collected the following information: video identifier, uploader identifier, number of views, time of upload, video duration, number of ratings, average rating, and number of comments.

We crawled Dailymotion twice, first on 8 March 2008, and again on 22 March 2008. Except for the analyses of video popularity, the aggregate from these two traces was used. For videos found in both crawls, meta-data from the latter crawl was used.

4.1.2 *Yahoo! Video Data Collection.* At the time of data collection, Yahoo! video had 20 categories. Each category list had at most 100 pages, with each page listing 20 videos. For each video on a page, the following four pieces of information were visible on the browser: video identifier, uploader identifier, video duration, and the number of views. Initially, we tried obtaining these fields using a modified version of our Dailymotion crawler; however, we noticed that the attribute fields of the videos were missing from the fetched pages' HTML source. These missing fields were being set after the page fetch, through a client-side Javascript, which our simple page crawler failed to replicate.

To overcome the above problem, we crawled each category by using a customized Firefox client that could automatically browse through pages by clicking on links and that could automatically save the contents of the page to disk. Specifically, our customized Firefox used a macro recorder called AutoHotKey[8] to browse pages of interest, and a Firefox extension called AutoSave[9] to automatically save the fetched page to a file on disk which we subsequently processed to obtain the relevant fields.

Our automated Firefox client provided us with data on approximately 50,000 videos. Our experience with Yahoo! video is that not all videos available on the site are displayed under categories. By using Yahoo! video's search engine, we discovered additional videos. Our customized client was used to discover additional videos by searching for English words from an online dictionary, discovering in this process approximately an additional 50,000 video identifiers. Once we had the list of video identifiers, the respective video pages were fetched for further details such as the number of ratings, average rating, and the time of upload.

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

4.1.3 *Veoh and Metacafe Data Collection.* Both Veoh and Metacafe associate uploaded videos with channels. We automated download of all channel pages, and extracted the meta-data of videos listed on each page. The number of ratings for each video was available only for Veoh, while the number of comments was available only for Metacafe. Some of the channels were skipped in our crawls because of a family filter scheme. Furthermore, duplicates of videos found to be listed on multiple channels were pruned. Our Veoh and Metacafe data collections were initiated on 18 March 2008 and 2 April 2008, respectively. Overall, we believe that a substantial portion of the videos available on these two sites, at time of data collection, were found by our crawlers.

4.2 Summary of Data Sets

Table I summarizes our data sets. Our data sets contain meta-data on approximately 1.8 million video clips. Together, these video clips have received close to 6 billion views. From the table, it appears that videos are rated much less frequently than they are viewed, that typically videos are significantly shorter than the typical full-length movie or television show, and that a small number of uploaders account for a large fraction of the video uploads.

[8] `http://swik.net/AutoHotkey`

[9] `https://addons.mozilla.org/en-US/firefox/addon/`

We note that Veoh serves significantly longer duration content than the other sites. Interestingly, as we will show later in this paper, Veoh appears similar to the other services, with respect to many other metrics, suggesting that there may be "invariants" that are not specific to services with YouTube-length videos, but that also may be applicable to services with longer content. Another noticeable difference between the workloads is the fraction of uploaders that upload content more than once. With Veoh 65% of the uploaders are multi-timers, while with Yahoo! only 33% upload more than once. With Dailymotion and Metacafe the corresponding percentages are 53% and 56%, respectively. We believe that these relatively large differences in multi-time uploaders are due to differences in the demographics of the uploaders. For example, Veoh, which has the most multi-time uploaders, hosts a significant amount of content from major studios and independent production houses, which are likely to be multi-time uploaders. These differences further support that we consider a wide range of video-sharing services.

4.3 Limitations and Comments on Data Sets

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

One potential cause for concern is the continually evolving nature of video sharing workloads. New videos are added every day, and viewing, rating, and commenting rates for each existing video change over time according to the current level of interest. This makes the task of workload characterization much more difficult than with static content. A single snapshot may provide representative data on the general characteristics of the service at the time of data collection. To understand video popularity, we obtained multiple snapshots from the Dailymotion service.

Second, how a site is crawled can also have an impact on the resulting analyses. In particular, biases may be introduced during data collection. A typical approach is to download pages that list videos. Often, a video sharing service offers multiple video listings, each under a different category (such as most recent videos, all time most popular, most popular this week, all time most rated, most rated this week, etc.). One strategy is to crawl all such pages and prune duplicate information. We applied this approach to crawl Dailymotion, Metacafe, and Veoh. We believe that we found a large fraction of the music videos that were available on Dailymotion, and believe that our data set is not biased towards popular videos. With the exception of channels that where skipped due to family filtering, we believe that we capture all channels of both Veoh and Metacafe. Furthermore, with the majority of channels captured, we believe these two data sets are relatively free from sampling biases of the files made available through these sites at the time of our crawls.

Some services severely limit the number of videos that can be found by browsing categories. For example, Yahoo! video limits the video listings per category to 100 pages. One can argue that content not easily accessible to the crawlers is also not readily visible to the users of the site, and thus the information gleaned by the crawler is representative enough. These considerations motivated us to attempt to augment our data by searching for videos using words from an online dictionary (cf. Section 4.1.2).

5. CHARACTERIZATION RESULTS

This section presents our characterization results, with particular emphasis on the invariants and differences among the four different services. Sections 5.1 and 5.2 discuss how users are interacting with the different video sharing services via ratings and comments, respectively. Section 5.3 analyzes characteristics of video uploaders. Video duration is studied in Section 5.4. Detailed analyses of video popularity are presented in Section 5.5. Section 5.6 summarizes our measurement results and identifies characteristics that may be considered to be invariant across video sharing services.

5.1 Ratings and Average Rating Score

The four services that we consider allow users to assign videos an integer rating between 0 and 5, with 0 indicating low quality or satisfaction and 5 indicating high quality or satisfaction. From the services considered, we were able to collect both the rating count and average rating score from all but the Metacafe service (from which we were only able to obtain the average rating score associated with each video).

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

Figure 1 shows the cumulative distribution of the number of times videos have been rated. We notice high variability in the number of times a video is rated. This figure reaffirms observations made above regarding the paucity of ratings. For example, from the figure we find that 90% of Yahoo!'s videos were rated 20 or fewer times; for Dailymotion and Veoh, the corresponding numbers of ratings are 8 and 12, respectively. Only a small fraction of views translate into ratings, and only a small fraction of the videos receive substantial (e.g., 50 or more) ratings. The number of ratings and views to a video can be expected to exhibit a strong positive linear correlation. Pearson's product-moment correlation between the number of views and ratings is 0.68, 0.66, and 0.24 for Dailymotion, Yahoo!, and Veoh, respectively. Clearly, the more a particular video is watched, the higher the expected number of ratings; however, the correlation is stronger for Dailymotion and Yahoo! then it is for Veoh.

The ability to rate videos is one of the features that enables visitors to these video sharing sites to express their degree of liking of the videos that they watch. The average of the rating scores can, therefore, be used as a metric to evaluate how satisfied users are with the sites' content.

A histogram of the average rating score for the videos in each of our data sets is shown in Figure 2. The "NR" column represents the fraction of videos that were not rated. Recall that we did not have rating counts in our Metacafe data set. Therefore, we were not able to distinguish between videos that have not been rated so far and videos that have been given a score of zero by the rater(s). In our Metacafe data set, 14,956 videos had an average rating of fewer than one out of which 14,944 had an average rating of zero; the latter 14,944 videos are included in the "$[0, 1)$" column but these could possibly be videos that have never been

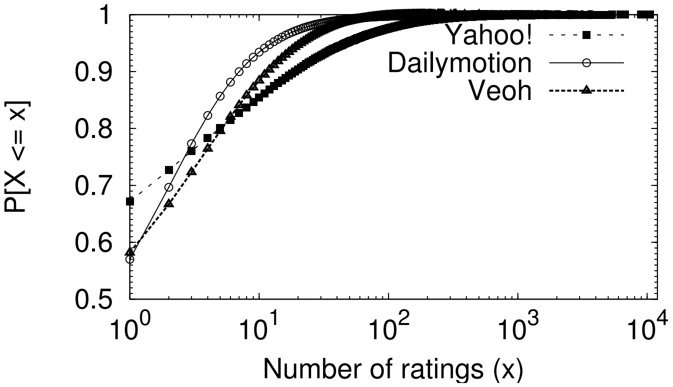


Fig. 1. Cumulative distribution of rating count.

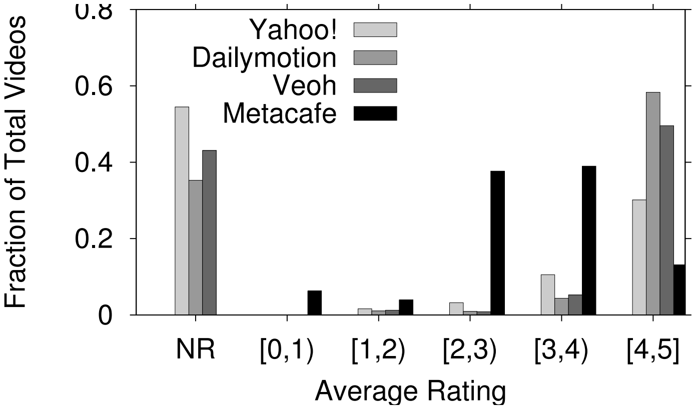


Fig. 2. Average rating score of videos.

rated (and thus should have been part of the "NR" column). Overall, our results may indicate that people tend to rate videos that they enjoyed watching. For Dailymotion, Yahoo! video, and Veoh, among the videos that have received ratings, we find that a majority have an average rating of 4 or higher. For Metacafe, we find that a majority of the videos have an average rating of 3 or higher.

5.2 Comments and Favourites

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

Unfortunately, we were able to obtain the number of comments for each video from only Dailymotion and Metacafe, and the number of favorites assigned to each video from only Dailymotion. Figure 3 presents the cumulative distribution of

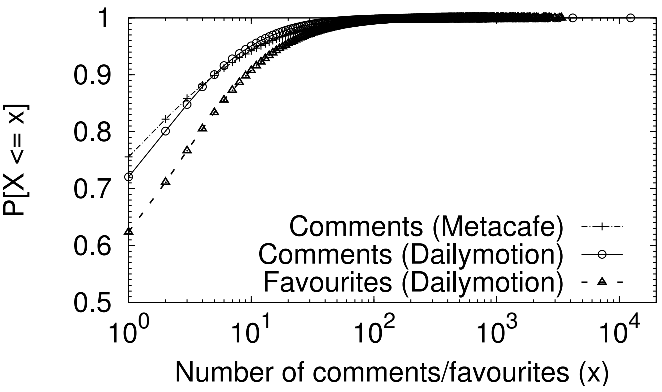


Fig. 3. Cumulative distribution of comments/favorites.

comments and favorites among the videos. In general, the number of comments (favorites) exhibits high variability with many videos receiving a small number of comments (favorites) and a handful of videos receiving many comments (favorites). In particular, a total of 624,885 videos, approximately 57% of the videos in the Dailymotion data set, have never been commented upon, and 95% of the videos have received 10 or fewer comments; however, the maximum number of comments observed for a video is 12,377. Qualitatively similar observations can be made for comments in the Metacafe data set.

The favorites feature is also sparsely used. Approximately 47% of Dailymotion videos have never been bookmarked as a favorite by any user, while 16% of the videos have been bookmarked as a favorite exactly once, and approximately 85% of the videos have been bookmarked as a favorite 10 or less times. Nevertheless, the video bookmarked as a favorite the most was bookmarked by 3,338 users.

5.3 Video Uploads and Uploaders

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

Figure 4 shows the cumulative distribution of the number of uploads per uploader. Here we find that a significant number of uploaders uploaded only one video. In the Yahoo! video data set, approximately 67% of the uploaders uploaded only once, whereas for Dailymotion, Metacafe, and Veoh the corresponding percentages are approximately 47%, 44%, and 35%, respectively. In general, from this figure we observe that most, approximately 95% or more, of the uploaders uploaded less than 50 videos.

We also analyzed whether or not the Pareto principle (cf. Section 5.5) applies to the distribution of the number of videos uploaded by each unique uploader. Our analysis suggests that the Pareto principle largely applies, with the top 20% of the uploaders accounting for close to 75-80% of the total videos in each data set.

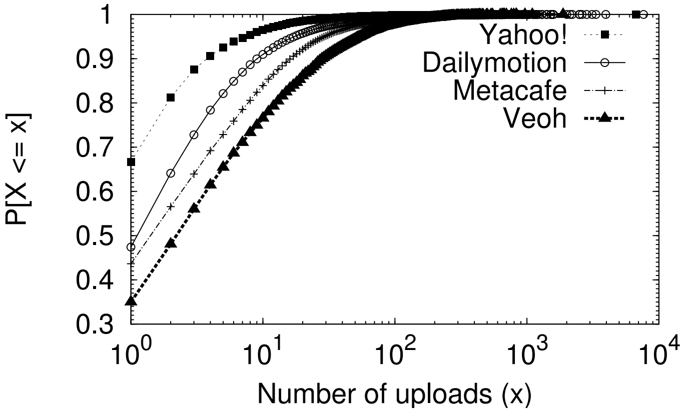


Fig. 4. Distribution of uploads by uploaders.

We manually analyzed the top 100 uploaders in each data set. For Metacafe and Veoh, most of the top uploaders appeared to be independent production houses. In the Yahoo! data set, the top contributors appeared to be Yahoo! applications such as Yahoo! music, news, and health; in fact, Yahoo! music is listed as the uploader of 191,263 videos in our data set. From the manual analysis of the Dailymotion data, we could not find any identifiable trend among the top uploaders as we found form the other data sets.

5.4 Video Duration

This section studies the duration of the videos found in our data sets. One problem with duration data, common across all data sets, was that a few video pages reported erroneous video durations. For example, in our Metacafe data set, we found one video for which a duration of 120 days was reported! A manual check of this video showed that it was, in fact, only a few minutes long. Similar issues were reported in the YouTube video duration analysis carried out by Gill et al. [2007]. To minimize the impact of erroneously reported video durations on our analysis, when analyzing average video durations (e.g., average duration reported in Table I) videos whose reported duration was longer than 6 hours were ignored. We note that videos with reported durations less than this threshold accounted for 99.99% of all videos in each of our data sets.

Figure 5 presents the cumulative distribution of video duration for each of our data sets. We draw several inferences from this figure and the results in Table I. In general, these Web-based video sharing systems are concerned with short duration videos. A typical video is between 2 and 4 minutes long for all but the Veoh service, which hosts some longer duration content from major production houses. We also find that only a very small fraction of the videos are very short; for example, less than 1 minute long. Metacafe, which is focused towards shorter-duration videos, is an exception as approximately 32% of the videos from this service are less than 1 minute long.

Note that the services we considered, with the exception of Veoh, either explicitly

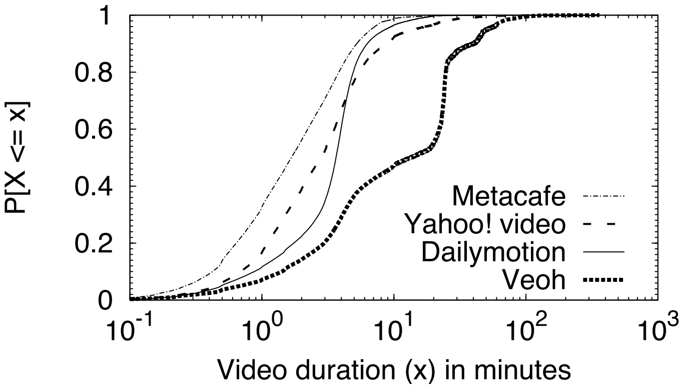


Fig. 5. Cumulative distribution of video duration.

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

5.5 Video Popularity

Understanding the popularity characteristics of videos is important for managing video storage, designing content distribution architectures and caching strategies, developing marketing strategies, and workload modeling. This section presents results concerning video popularity distributions as observed for our data sets. We distinguish between two quite different measures of popularity, with differing applications and significance: the total number of views to videos since they were uploaded, referred to here as the *total views popularity*, and the rate with which videos accumulate new views, referred to here as the *viewing rate popularity* [Mitra et al. 2009].

5.5.1 *The 80-20 Rule for Video Popularity.* Often we are interested in understanding how skewed the references are to the most popular videos, because presence of such skewness can have immediate positive implications with respect to the potential effectiveness of content management strategies such as caching. When discussing skewness of distributions, the “80-20 rule”, also known as the Pareto principle, is often considered as it is found to be applicable in many diverse contexts. This rule, in its original context of wealth distribution, states that 20% of the wealthiest people account for 80% of the total wealth of the population [Newman 2005]. Skewness of references, and the potential applicability of this rule, has previously been discussed in the context of references to Web servers and proxies [Arlitt

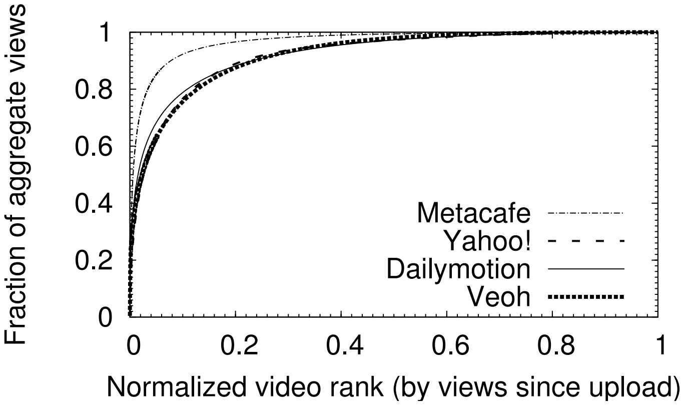


Fig. 6. Skewness in the total views popularity of videos.

and Williamson 1997; Mahanti et al. 2000], on-demand streaming systems [Yu et al. 2006], and more recently in the context of video views in YouTube [Gill et al. 2007; Cha et al. 2007; Zink et al. 2008].

Figure 6 shows the cumulative distribution of the *total views popularity*; i.e., the total number of views to a video since it was first uploaded, as measured at the time of our crawl. This life-time metric may be viewed as a measure of longer-term popularity of a video. With respect to this metric, we find that the Pareto principle generally holds for video views: 20% of the most popular videos accounted for approximately 85% or more of the total views, in the four data sets we analyzed. In general, our findings are similar to those obtained for YouTube videos [Cha et al. 2007]. However, we note that the Metacafe service appears to exhibit significantly more skew than the other services we considered.

The total views popularity distribution is useful for understanding service features such as "all time" most popular listings, but does not provide an accurate picture of the distribution of the rates at which videos are viewed. The latter is very important when attempting to model the video reference process, and in understanding the potential of different content distribution and caching architectures. For example, with the total views popularity metric, an older video with many views in the past may appear to be more popular than a recently uploaded video (and, erroneously, a better caching candidate) simply because the newer video has not been available for enough time to acquire more views. These issues are further discussed in Section 6.

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

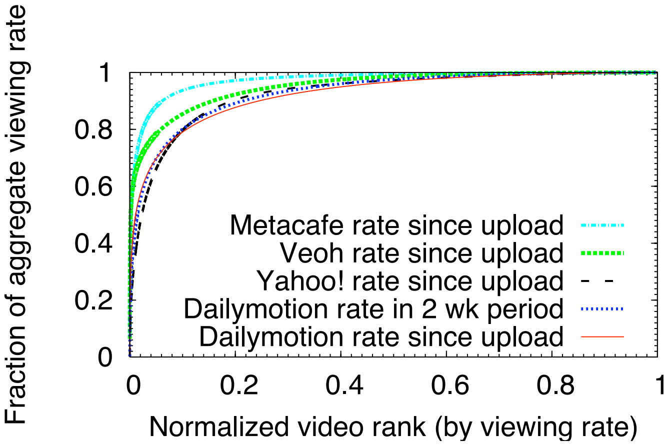
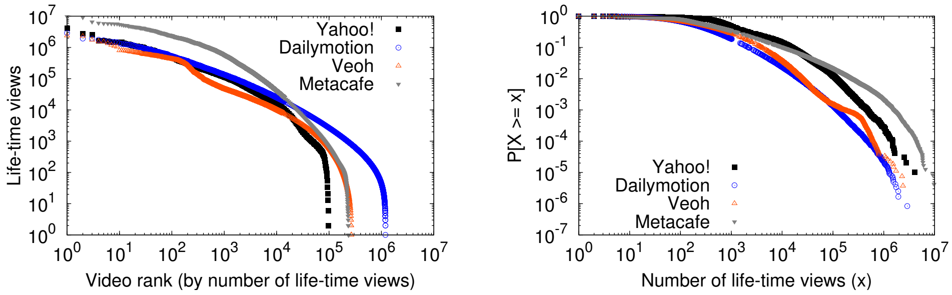


Fig. 7. Skewness in the average viewing rate over a two week period (Dailymotion), and the average viewing rate since upload.

the video at the time of the crawl. This latter measure removes, to some extent, the age bias in the total views popularity measure. Note that two closely spaced crawls allow us to measure short-term popularity of videos. In the case of viewing rate since upload, however, the measure captures short-term (long-term) characteristics depending on whether the video under consideration is relatively new (old).

Figure 7 shows the cumulative distribution for the viewing rate popularity of the videos. The two measures of the viewing rate popularity described above are used in this figure: the average viewing rate over a two week period, specifically the time span separating our two crawls of Dailymotion, and the average viewing rate since upload for each of the services. Our results show that, with respect to the average viewing rate since upload, videos exhibit skewness with 20% of the most popular videos by viewing rate accounting for 80% or more of the total viewing rate. Similar to results for the total views popularity metric, the videos in the Metacafe data set exhibit more skewness than videos in other data sets. We also note that with respect to viewing rate popularity, and unlike results for total views popularity, Yahoo! videos exhibit more skewness than Dailymotion and Veoh videos.

The results in Figure 7 also show that videos in the Dailymotion data set exhibit similar skewness properties with both measures of viewing rate popularity. With both of these measures, 20% of the most popular videos account for close to 88% of the total viewing rate. Interestingly, these results are similar to what we observed for the total views popularity for videos, with the average viewing rate popularity measures indicating only a slightly increased skewness in video popularity. The most popular videos according to viewing rate popularity, however, may be quite different than with total views popularity. In fact, plots of average popularity as a function of age for the four data sets show that popularity as measured by the average viewing rate since upload is generally lower for older videos (particularly for Dailymotion, Metacafe, and Veoh), while the total views popularity is generally higher for older videos. Other important differences between the total views and viewing rate popularity measures are discussed in the next section.



(a) Zipf plot

(b) CCDF plot

Fig. 8. Distribution of the total views popularity of videos.

When comparing popularity distributions for user generated content with those for traditional Web and media workloads, it should be noted that the measure of popularity used in the latter context is *number of accesses to (Web or media) files over the fixed time period of the trace* (see, for example, [Arlitt and Williamson 1997; Yu et al. 2006; Mahanti et al. 2000; Acharya et al. 2000; Almeida et al. 2001]), which in our context corresponds to the *viewing rate popularity*, and not to the total views popularity of the videos. Comparing our observations regarding viewing rate popularity with previous work on Web and media servers, we find that the popularities of videos on video sharing and publishing sites appear to be *more skewed* than object popularities in these other domains. In the traditional Web domain, for example, it has been reported that typically the top 20% of the most visited Web pages account for approximately 70% of all visits to a Web server [Arlitt and Williamson 1997; Mahanti et al. 2000].

5.5.2 *Zipf and Power Law Analysis.* Power laws can often be successfully used to describe phenomena in which "large" events are uncommon while "small" events occur frequently. A random variable X is said to follow a power law if $P[X \geq x]$ is approximately $Cx^{\alpha-1}$, where both C and α are constants; the parameter α is referred to as the exponent, shape, or scaling parameter of the distribution. A characteristic feature of power law distributions is the presence of a straight line on a complementary cumulative distribution function (CCDF) plot over several orders of magnitude when a logarithmic scale is used on both axes. In the literature on traditional Web and media workloads, for example, power laws have been found to apply to reference counts seen at Web proxies [Breslau et al. 1999; Mahanti et al. 2000], and media servers [Almeida et al. 2001]. The presence of power law behavior in these reference streams has important implications for the design of caching systems, which may store only a relatively small number of the most popular objects (i.e., Web or media files) in the cache with the goal of improving response times and saving bandwidth. This section considers the issue of whether or not power laws can be used to describe video popularity in the four measured video sharing services.

Zipf's law, an alternative characterization of power law behavior [Adamic ; Newman 2005], is frequently used in the literature on traditional Web workloads. Zipf's law states that if objects are ranked in order of their frequency of occurrence, with

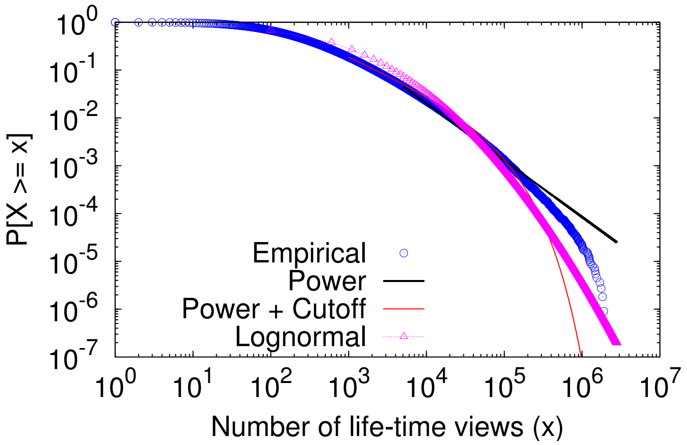


Fig. 9. Models of the total views popularity of videos for Dailymotion.

the most frequently occurring object assigned rank one, the second most frequently occurring object assigned rank two, and so on, then the number of occurrences y relates to the rank of the object r as $y \sim r^{-\theta}$, where θ is the exponent of the Zipf distribution. A rank-frequency plot is often used to study Zipf-like behavior, with the presence of an approximate straight line when a logarithmic scale is used on both axes indicating the likely presence of this behavior. While the rank-frequency plot shows most clearly the distribution of the "lukewarm" and "cold" objects, the CCDF plot shows most clearly the distribution of the "hot" and "lukewarm" objects. We use both types of plots here.

Figure 8 shows rank-frequency and CCDF plots for the total number of views since a video was uploaded (i.e., the *total views popularity*) for each of our data sets. A number of inferences can be drawn from Figure 8(a). We observe that the total views popularity appears to be Zipf-like for a substantial range of video ranks for all data sets, with a pronounced exponential cut off for the least popular videos. The presence of an exponential cut off suggests that there are not that many videos that receive a very small number of views since they are uploaded, a characteristic that would be necessary for the total views popularity to exhibit Zipf-like behavior for the coldest videos.

For each of the four data sets, the CCDF plot for the total views popularity, shown in Figure 8(b), has a right tail that spans four to five orders of magnitude. This is indicative of high variability in the total views achieved by the videos. The presence of skewness (i.e., a small number of videos account for a large fraction of the aggregate total views) combined with the long right tail indicates that the total views popularity distribution is heavy-tailed. Visual inspection of the graph suggests that the total views popularity distribution may have power law behavior over a portion of its range. For Metacafe, for example, power law behavior appears to exist for life-time views in excess of 100, with a drop-off for the hottest videos.

Figure 9 shows the best fit power law (key: "Power"), power law with exponential cut off (key: "Power + Cutoff"), and lognormal (key: "lognormal") distributions,

Table II. Summary of Distributions.

Distribution	$f(x)$	Shape parameter(s)
Power law	$x^{-\alpha}$	α
Power + Cutoff	$x^{-\alpha}e^{-\lambda x}$	α, λ
Lognormal	$\frac{1}{x}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$	μ, σ

Table III. Models for the total views popularity distribution.

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

for total views popularity as measured for the Dailymotion data set. (Table II summarizes the distributions used.) It is often difficult to distinguish among the mathematical distributions that we consider in this graph, with respect to their goodness-of-fit to measured data. For example, the lognormal distribution can also exhibit a near straight line in the right tail of the CCDF plot when there is high variance in the distribution [Newman 2005; Mitzenmacher 2004]. From the best fit curves, it appears that no fit is qualitatively significantly better than the other fits. For example, while the middle region of the curve (e.g., life-time views between 1000 and 100,000) appears to be best modelled by a power law, the left-most region and part of the middle region (e.g., life-time views between 10 and 10,000) appears to be better modelled by a lognormal distribution or by a power law with cut off. In general, however, the total views popularity appears to be heavy-tailed. Qualitatively similar results hold for the other data sets as well.

Using the likelihood ratio test [Clauset et al. 2009], we compared power law fits with power law plus exponential cut off and lognormal fits. Note that this test only tells us which of the competing distributions is a better model for the data but does not tell us whether or not the winner is a good model for the underlying empirical distribution. For the Yahoo! video and Metacafe empirical video popularity distributions, the comparison indicates that a pure power law is a better model, while for Dailymotion and Veoh we find that both power law with exponential cut off and lognormal distributions provide better fits. Table III presents the power law exponent α along with the minimum value of x for which power law or power law with cut off behavior holds for the data sets; the exponent ranges between 1.43 and 2.25, and is consistent with prior observations for YouTube's science and entertainment videos [Cha et al. 2007].

Figure 10 shows a rank-frequency plot of the *average viewing rate since upload* for each of the services, as well as the *average viewing rate over a two week period* for the Dailymotion data set. Figure 11 shows the corresponding CCDF plots along with the corresponding best fit models for both the Dailymotion measurements. Several inferences can be drawn from the rank-frequency plots in Figure 10.

First, Zipf-like behavior is considerably more apparent in these plots than in those for the total views popularity in Figure 8(a). In particular, although the rank-frequency plots in Figure 10 do show a drop for the least popular videos, this

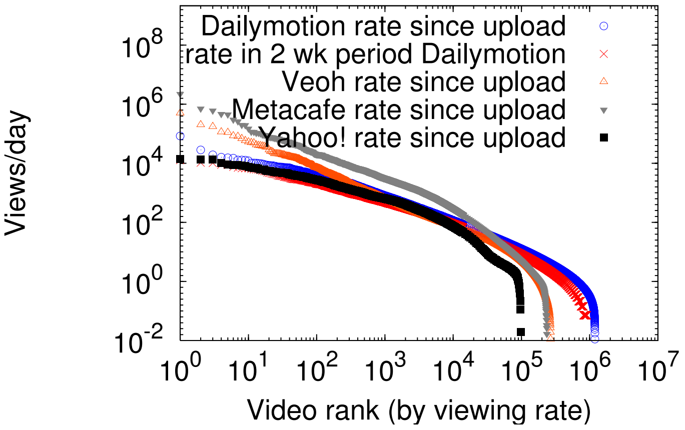
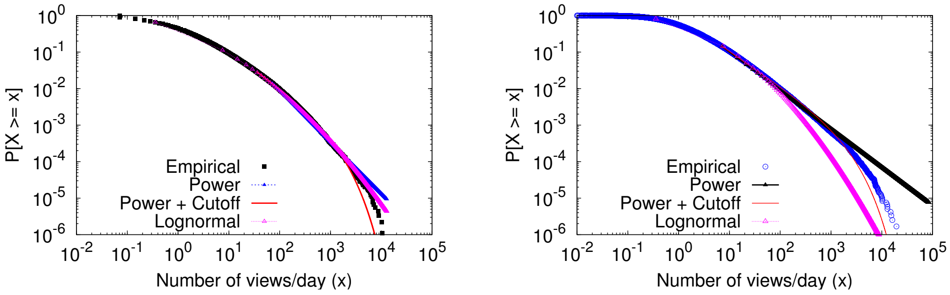


Fig. 10. Distribution of the viewing rate popularity.



(a) Two-week average rate CCDF (b) Average rate since upload CCDF

Fig. 11. Models of the viewing rate popularity for Dailymotion.

drop is not as pronounced as those in Figure 8(a), indicating that, with respect to viewing rate, there are more unpopular videos than one would conclude from data on total views popularity. The difference in drop off between Figures 10 and 8(a) may be related to videos tending to have a minimum number of life-time views, as is discussed in Section 6.2.

Second, although the rank-frequency plot for the average viewing rate since upload is quite similar to that for the average viewing rate over a two week period, the latter plot shows somewhat less of a drop for the least popular videos. This difference in the drop of the popularity is likely due to, again, videos tending to have a minimum number of life-time views (which may not occur within the considered two week period). We conjecture that rank-frequency plots for average viewing rate over an even shorter time scale such as a day, may show even more purely Zipf-like behavior than the two week average viewing rate plots. This conjecture is a topic of future work.

Third, while there are very few, approximately 0.01%, Dailymotion videos (for example) with only one view since they were uploaded, close to 10% of the videos had only a single view within the two week period between our crawls. The percentage of these so called one-timers, i.e., videos that have been viewed only once in a trace period, is somewhat comparable to that observed in the context of traditional Web and media servers. In the latter context, studies have reported between 15 and 75% of the total referenced objects to be one-timers [Arlitt and Williamson 1997; Mahanti et al. 2000].

Figures 11(a) and (b) present the empirical CCDF of the average viewing rate over a two week period and average viewing rate since upload, respectively, for the Dailymotion data set, along with the best fits for power law, power law with exponential cut off, and lognormal distributions. These graphs can be used to explore the differences in these two average viewing rate metrics with respect to the most popular videos. More formally, using the likelihood ratio test, we find that a power law ($\alpha = 2$) is the best candidate when modelling the distribution of the average viewing rate over a two week period for the Dailymotion data set (as it beats both a power law with cut off and lognormal fitting). A similar analysis on the distribution of the average viewing rate since upload for this data set suggests that power law ($\alpha = 1.93$) with cut off is the best candidate for that metric; similar analysis on Metacafe also suggests power law ($\alpha = 1.46$) with cut off, while for Yahoo! and Veoh it appears that both lognormal and power law with cut off are good contenders. Our analyses show that power law and related variant distributions might be useful for modeling the tail of the viewing rate popularity distribution. In addition, we also find that for most of the tail, the two metrics for viewing rate exhibit similar behavior although they diverge for the values on the extreme right (i.e., at the most popular end) of the distribution. Comparing Figures 9 and 11, we note that there appears to be less of a drop off with the short-term (rate) metric. We conjecture that the increased drop off in the long-term case, in comparison to the short-term case, is due to churn in video popularity. (See Section 6 for a discussion about the impact of churn.)

Finally, when discussing our fitting results, we should note that we we limited ourselves to power law distributions, their variants, and the lognormal distribution because these distributions have been used in the recent work on YouTube workloads [Cha et al. 2007; Gill et al. 2007], have been widely used in traditional Web workload studies [Mitzenmacher 2004; Breslau et al. 1999; Mahanti et al. 2000; Newman 2005; Barford and Crovella 1998], and have also generated discussion on how these distributions may be applied to empirical data [Downey 2005; Clauset et al. 2009]. Other distributions such as the Stretched Exponential [Guo et al. 2008], Zipf-Mandelbrot [Hefeeda and Saleh 2008], and Double Pareto [Mitzenmacher 2004] could also prove useful. We leave comparison with these and other distributions for future work.

5.6 Summary of Invariants

The preceding sections presented multi-dimensional analyses of video sharing workloads. Our analyses found several characteristics that appear to be common across video sharing services. We note that the video sharing services considered here span a wide range, including Veoh which serves longer content and Metacafe that uses a revenue sharing model to give uploaders incentives. Furthermore, many of the observations that hold for the services studied here also hold for YouTube [Gill et al. 2007; Cha et al. 2007; Zink et al. 2008; Halvey and Keane 2007b; 2007a] which is by far the largest and most popular video sharing service. Characteristics that appear to be *invariants* across these diverse video sharing services are summarized below:

(1) Users are primarily interested in watching videos; social interaction tools for rating, bookmarking, and commenting on videos are relatively infrequently used. Similar observations hold for YouTube [Halvey and Keane 2007a; 2007b].

(2) The number of uploaders to a service are an order of magnitude smaller than the number of uploaded videos, and several orders of magnitude smaller than the number of views to these videos. We found that the Pareto rule applies, with the top 20% of the uploaders contributing between 75-80% of the total videos.

(3) The typical video available from these services is of short duration. This observation, based on data from the Dailymotion, Veoh, Yahoo! video, and Metacafe services, is similar to that made for YouTube [Gill et al. 2007; Zink et al. 2008]. Some services, such as Veoh, however, are starting to feature longer duration videos, using peer-to-peer technology to address the problem of efficiently distributing such videos.

(4) Both the total number of views to videos, and the rate with which new views occur, follow the Pareto rule, with 20% of the most popular videos accounting for 80% or more of the views. Similar observations regarding the number of views since upload [Cha et al. 2007] and the viewing rate [Gill et al. 2007] have been made for YouTube.

(5) The total views popularity distribution is heavy-tailed and may be modelled as power law or power law with exponential cut off, with power law exponent between 1.4 and 2.5. However, we note that neither a power law (or variants), nor a lognormal distribution, appear to fit the entire distribution well. These results are consistent with those reported for YouTube's videos in the science and entertainment categories [Cha et al. 2007]. Similarly, the total views to videos may be modelled as Zipf with cut off.

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

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

Finally, we note that the fact that these invariants appear to hold across a wide range of services (selected to cover a wide range of content, user demographics, and popularity), provides some confidence that these observations hold true for other sites than only YouTube. While YouTube is the most popular video sharing service, we note that these invariants can be valuable to a much larger set of site developers and services. In the following section we discuss some systems insights that we believe are of value to such developers.

6. SYSTEMS IMPLICATIONS

In the ensuing discussion, we focus on systems implications. Our discussion is based on our observed invariants. In addition, to gain further insights into temporal aspects of these system implications, Section 6.3 considers a complementary data set. This data set allows us to draw insights with regards to how popularity evolution and churn in file popularity affect caching efficiencies.

6.1 Social Interactivity Tools

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

6.2 Content Popularity and Caching

Our next set of comments is about the popularity characteristics of the videos found on video sharing services. We believe that the very nature of user generated videos, and the sociological behavior of video sharers, precludes the possibility of a large percentage of the video clips having few (e.g., less than 5) or no total views since upload. Typically, video sharing Web sites have a link to videos that have been recently uploaded. A typical surfer of such sites may navigate through this listing and perhaps view some videos that are thought to be of possible interest.

Another potential cause of few one-timer videos may be publicity by the creator or uploader of a video. These sites are often venues for sharing interesting videos among friends and family members. A common practice of uploaders is to send friends and family the link of the uploaded video (or post a link to the newly uploaded video in one's own social network profile page or blog space), and people in the social network of the uploader may be expected to oblige. Finally, it is possible that the uploaders themselves view their uploaded videos.

Overall, we observed that video sharing systems exhibit high variability, both with respect to how videos are viewed and with respect to the actions taken on meta-data associated with videos (e.g., ratings, comments, bookmarking). Our measurements showed, similar to earlier results for YouTube video popularity [Cha et al. 2007; Gill et al. 2007], that a large number of videos on these services have only lukewarm popularity. Referring to Figure 10, for example, we can see that approximately 10,000 videos out of the 1 million videos sampled from Dailymotion receive more than 100 views/day. Yet another view can be obtained by looking at Figure 7, specifically by considering the Dailymotion data with two snapshots. Our analysis shows that the top 1% and top 10% of the videos (in terms of views received in the 2 week period) account for roughly 40% and 80%, respectively, of the total views in this measurement interval for the videos in the data set. Similarly, we observed high skews in ratings and comments to videos. We also observed a positive correlation between view counts and numbers of ratings. Such highly skewed access patterns can be exploited for improving system performance by caching popular content and their associated meta-data, in-memory and at content distribution servers.

While the skew in popularity of videos can aid in caching, the long-tail of lukewarm videos that each have modest popularity exacerbates the cost of distributing video files. The hot content can be cached and distributed efficiently but the large number of lukewarm video files is not readily amenable to scalable content distribution and requires careful resource (e.g., hardware, network capacity) management and much engineering effort.[10] Design of content distribution approaches that can more efficiently serve the long-tail is an open problem that requires further attention from the systems community.

6.3 Caching Efficiency and Popularity Evolution

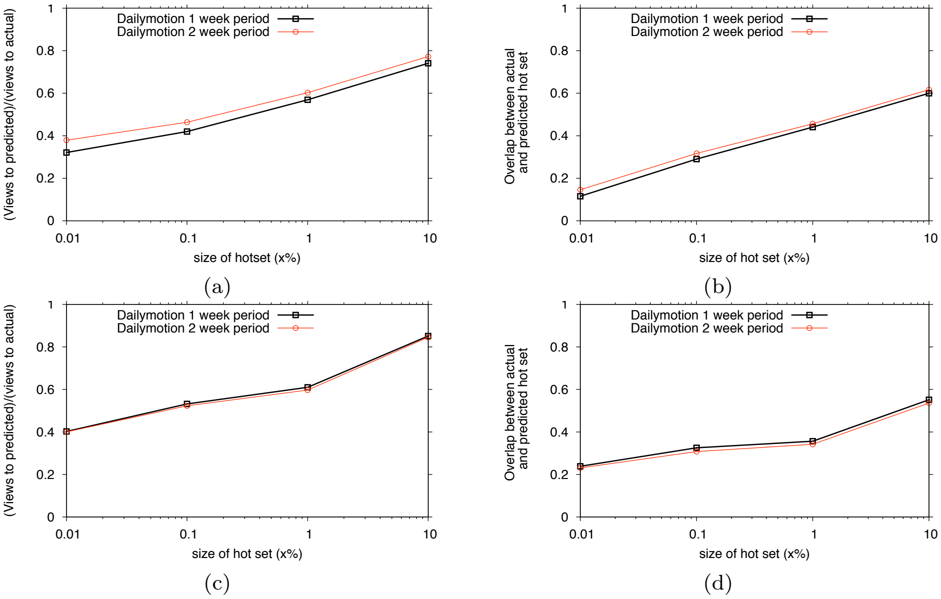
The remainder of this section provides further insights to caching effectiveness and the popularity evolution of video content. We also discuss the efficacy of the popularity metrics introduced earlier.

Our work here utilizes complementary Dailymotion data sets which were collected as follows. In particular, we crawled Dailymotion every day for seven days looking for recently added videos (i.e., videos added in the past 24 hours). Our crawls for new videos returned approximately 15,000 unique videos per day. In parallel to collecting new videos, we also used keyword search, similar to as described for Yahoo! videos, to obtain additional videos. The process of gathering these videos lasted for a week from July 13 to July 20, 2008. We collected the view count for each video twelve additional times, at exactly one week spacings. Overall, we collected statistics on 1,306,931 videos. Together, these videos had acquired approximately 3 and 3.6 billion views at the start and end of our data collection, respectively.

Figure 12 compares how the different measures of popularity, in particular the life-time views and life-time viewing rate popularity metrics, predict shorter-term popularity of videos. In all cases the *hot set* is defined as the top x% of videos with respect to either actual views during the indicated time period (*actual hot set*) or one of the life-time measures (*predicted hot set*). The predicted hot set is computed using one of the life-time popularity metrics, which in turn is computed using the data at the point of the initial data collection. The figures on the left-hand side show the ratio of the views to the *predicted* hot set to that of the views to the *actual* hot set, for varying sizes of the hot set; the figures on the right-hand side show the overlap in videos between the predicted and actual hot sets. In the figure legends, "1 week period" refers to the hot set defined by the views to the videos in the first week of data collection; "2 week period" refers to the hot set defined by the views to the videos in the first two weeks of data collection.

From Figure 12, we observe that the "most popular" 0.01% of the videos ac-

[10] It has been reported that YouTube expends substantial effort in managing the request load owing to lukewarm videos [C. D. Cuong 2007].



(a) (b) (c) (d)

Fig. 12. Comparing effectiveness of predicting hot sets using: (a, b) life-time total views; (c, d) life-time viewing rate.

cording to either of the life-term popularity metrics has substantially lower current popularity than the actual hot set although it appears that the life-time viewing rate may be slightly more indicative of current popularity than the life-time total views. Even for large hot sets (e.g., top 10%), the overlap between the hot set defined according to either of the life-time popularity metrics, and the actual hot set, is only 50-60%, although in this case, the hot set according to life-time popularity is indeed relatively "hot" (with about 80% as many views as to the actual hot set). The latter phenomenon can be explained by the long-tailed Zipf-like distribution for video popularity. Although there are substantial differences in membership between the 10% hot set based on life-time popularity, and the actual 10% hot set, many of the differing videos may be among the large set of lukewarm videos with similar popularities.

Figure 13 provides further insights to the results in the previous graphs. In this scatter plot, there is a point plotted for each video, giving its "short-term" popularity as the point's y value, and its "long-term" popularity as the point's x value. Here short-term popularity is defined by the video's view count during the "current" week (in this case, the first week of data collection), while long-term popularity is defined by the average number of weekly views to the video since it was uploaded (i.e., the life-time viewing rate, measured in units of views per week). The graph also shows, using solid squares, the 5th and 95th percentile actual views (in the measurement period) for videos binned logarithmically by their life-time viewing rate. Specifically, life-time viewing rates were binned using a multiplicative factor of 2. Thus, videos were binned according to life-time viewing rates in the intervals [1, 2), [2, 4), [4, 8), [8, 16), and so on, and the 5th and 95th percentile of

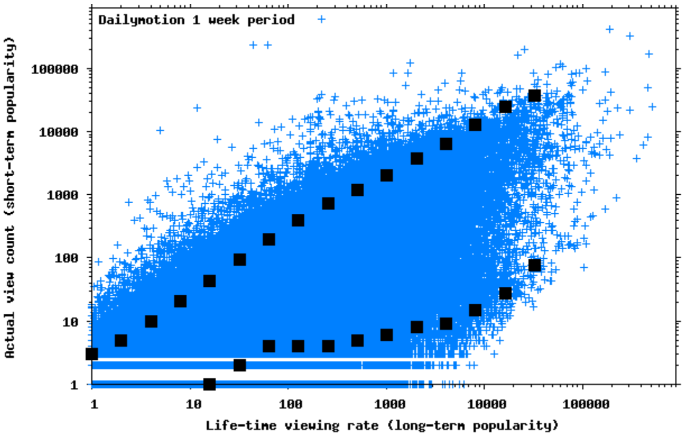
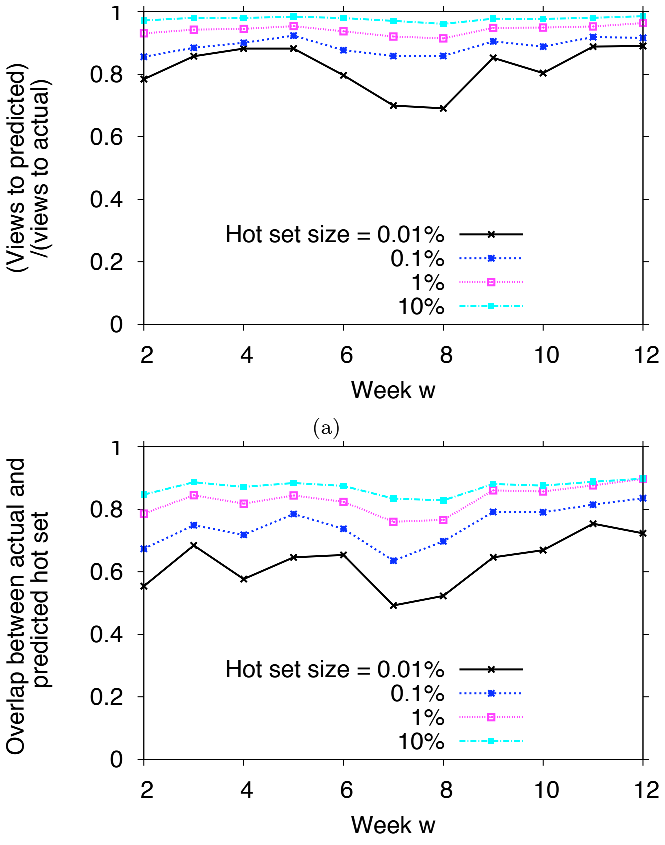


Fig. 13. Scatter plot for life-time viewing rate versus short-term popularity of videos.

the actual view counts for each bin with at least 200 videos was calculated. The graph shows that a video of certain popularity, as measured by the life-time viewing rate metric, can result in widely varying short-term popularity (illustrated by the wide spread between the 5-95 percentile markings, for example) and is indicative of the churn in popularity of videos. The churn in popularity of videos is investigated further using the multiple snapshots we collected for the Dailymotion videos.

We first study churn through investigation of how well a video hot set, as determined according to short-term video popularities (i.e., views during the current week), could be predicted from the hot set of the preceding week. Figure 14(a) shows the ratio of the total views to the videos in a predicted hot set, to the total views to the videos in the actual hot set, for various weeks w of data collection. The actual hot set is composed of the videos with the highest short-term popularity during week w. The predicted hot set is determined according to the short-term video popularities during week w-1. We show curves for four different hot set sizes: 0.01%, 0.1%, 1%, and 10%. Figure 14(b) shows the corresponding overlap in videos between the predicted and actual hot sets. We note that the overlap between videos in the predicted and actual hot set is modest, between 50-70%, for the smallest hot set of size 0.01%. This indicates presence of churn in the relative popularity of videos. While predicting videos that may be popular later based on current views appears hard, it is interesting to note that for the smallest hot set of 0.01%, the total views to videos in the predicted hot set is between 70-85% of the total views to videos in the actual hot set. Thus, it appears that caching decisions could be made on the basis of shorter-term measures of popularity and be fairly successful in terms of cache hit rates.

Figure 15 shows a scatter plot similar to the one shown in Figure 13. Here we plot the number of views in the first week interval on the x-axis and the number of views in the second week interval on the y-axis. Using an approach similar to Figure 13, we show using solid squares, the 5^{th} and 95^{th} percentile of views for videos in the second week binned logarithmically by their corresponding views in the previous week. Clearly, there is a strong correlation between the number of



(a)

(b)

Fig. 14. Effectiveness of hot set prediction using short-term popularity metric (i.e., views to videos within one week): (a) measured by views to files in hot set; (b) measured by number of files in hot set.

views across weeks. However, we also observe that there are videos that go from being cold to being wildly popular and vice versa. This churn in popularity of videos may impact caching decisions as objects may have to be pulled in and out of the cache with changes in their popularity.

With the type of data we have, it is not possible to directly study caching algorithms. We can, however, gain insights into the quantity of cache replacement traffic that would be required to track, at some level of precision, hot set evolution as observed in our data. Figure 14 (b) shows the overlap between hot sets of successive weeks, and therefore quantifies the cache replacement traffic that would be required for a cache configured to store the weekly most popular videos. For a hot set size of 0.01%, for example, Figure 14 (b) shows that over 40% of the videos in the hot set change from week 1 to week 2, implying substantial cache replacement traffic. This traffic could be reduced, however, if the hot set was tracked less precisely. Figure 14 (a) shows that a cache storing the hot set for week 1 could, without any cache replacements, serve roughly 80% of the week 2 views that could

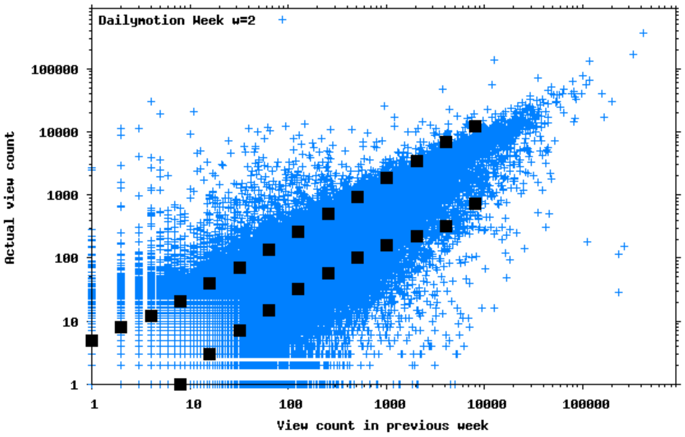


Fig. 15. Scatter plot of views to videos in the first week versus that in the second week.

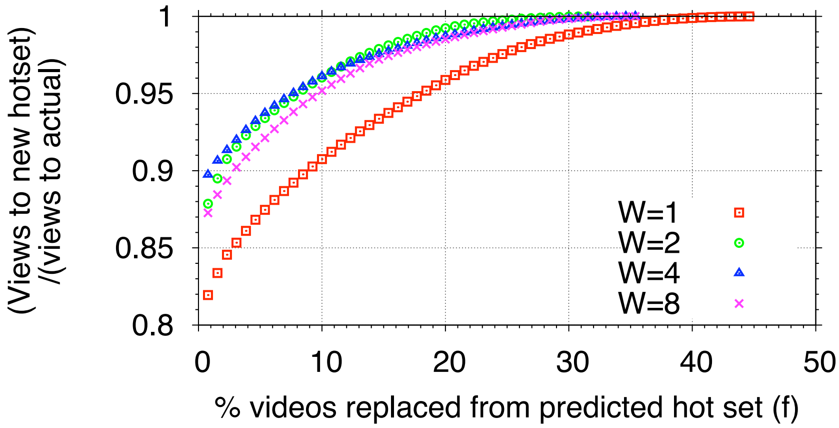


Fig. 16. Impact of replacing cold videos from the predicted hot set as measured by the views to the modified hot set (x = 0.01%).

be served by a cache storing the new hot set for that week.

Figure 16 gives insight into potential intermediate policies. The figure shows, for example, that if only 10% of the videos in the 0.01% hot set for week 1 are replaced (out of the more than 40% of these videos that are not in the hot set for week 2), then the cache could achieve more than 90% of the hit rate of a cache storing the exact week 2 hot set. Here it is assumed that the videos removed from the cache are replaced by the hottest videos for the current week that are not already present in the cache. The figure shows diminishing returns with respect to more precisely tracking the hot set. In general, it appears that the quantity of cache replacement traffic required to achieve most of the benefits of caching the hot set may be substantially lower than would be suggested by Figure 14 (b). Further, our analysis also suggests that simple measures of short-term popularity may perform

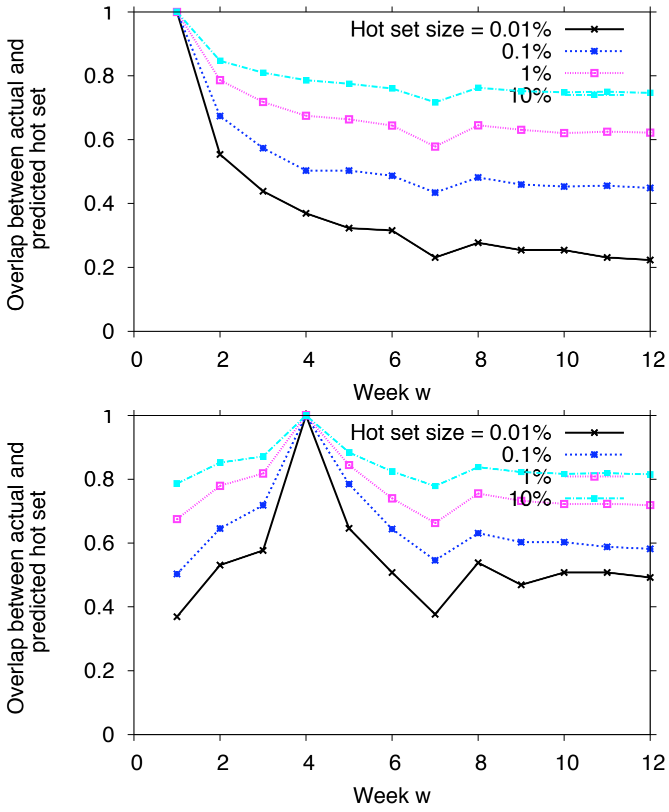


Fig. 17. Churn in video popularity measured by the overlap between actual and predicted hot sets: (a) prediction assumes hot set fixed to that for week 1; (b) prediction assumes hot set fixed to that for week 4.

reasonably well for achieving high hit rates to preloaded caches.

The final set of graphs, shown in Figure 17, further explore how the hot set evolves. We show two example cases. One considers the impact of fixing the hot set to the videos that received the most views in the first week (top sub-figure). Another example fixes the hot set to the videos that received the most views in the forth week (bottom sub-figure). For these two example hot sets, we show the fraction overlap in the videos between the predicted and actual hot sets. Referring to the percentage overlap between hot sets, we see that the set of most popular videos exhibits significant churn. For example, only about 20% of the videos in the hot set in week 1 are still in the hot set at the end of week 12 when considering the smallest hot set x = 0.01% (cf. top sub-figure). As the hot set increases in size, we observe less change in the hot set across weeks. The sub-figure on the bottom illustrates the formation, as well as the decay, of a hot set (specifically, the hot set for week 4). Note that there is substantial change in the relative popularity of videos between weeks 3, 4, and 5. We also note that the rate with which content moves in and out of the hot set before and after week 4 is roughly symmetrical. Both figures suggest that the substantial churn in popularity would require videos to move in and out of caches rather frequently, if the weekly hot set was tracked precisely.

7. CONCLUSIONS

This paper presented detailed analyses of the workload characteristics of four video sharing services, namely Dailymotion, Yahoo! video, Veoh, and Metacafe. Metadata was collected on approximately 1.8 million videos which together have been viewed approximately 6 billion times. The four services that we considered cover a range of user communities and service models. Our study of four different video sharing services, along with prior studies of YouTube, allowed us to postulate various video sharing workload invariants.

In addition, we identified differences in how video popularity is measured that may be important in workload modelling. For example, we found that popularity as measured by the number of views within a fixed, relatively short, time period exhibits Zipf-like behavior, whereas popularity as measured by the total number of views to videos since their upload exhibits Zipf-like behavior with cut off.

Finally, we considered implications for system design. In addition to system insights gained based on the identified invariants, we collected an additional data set tracking the views to a set of videos from Dailymotion, over a twelve week period, and used it to draw further insights into caching in video sharing systems, and the relevance to caching of life-time popularity measures. We found that lifetime popularity measures have some relevance for large cache (hot set) sizes but that this relevance substantially decreases as cache size decreases, owing to churn in video popularity.

Many avenues remain for future work. Our ongoing work is concerned with developing models for video reference streams, and studying novel content distribution mechanisms for user generated content. Other open problems include determining general characteristics of user generated content sharing services (including photo sharing rather than just video sharing, for example), developing efficient storage management solutions for large scale user generated content sharing services, and understanding the impact geographic diversity has on file popularity.

8. ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Informatics Circle of Research Excellence (iCORE) in the Province of Alberta, the National Information and Communications Technology Australia (NICTA), and Network Appliance (NetApp).

REFERENCES

Acharya, S., Smith, B., and Parnes, P. 2000. Characterizing User Access to Videos on the World Wide Web. In *Proc. of SPIE Multimedia Computing and Networking (MMCN) Conference*. San Jose, USA.

Adamic, L. Zipf, Power-laws, and Pareto - A Ranking Tutorial. http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html.

Almeida, J., Krueger, J., Eager, D., and Vernon, M. 2001. Analysis of Educational Media Server Workloads. In *Proc. of International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV)*. Port Jefferson, USA, 21–30.

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

Barford, P. and Crovella, M. 1998. Generating Representative Web Workloads for Network and Server Performance Evaluation. *SIGMETRICS Perform. Eval. Rev. 26,* 1 (June), 151–160.

Breslau, L., Cao, P., Fan, L., Phillips, G., and Shenker, S. 1999. Web Caching and Zipf-like Distributions: Evidence and Implications. In *Proc. IEEE INFOCOM*. New York, USA, 126 – 134.

C. D. Cuong. 2007. YouTube Scalability, `http://www.techpresentations.org/YouTube_Scalability`. Google Seattle Conference on Scalability.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y., and Moon, S. 2007. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *Proc. ACM Internet Measurement Conference (IMC)*. San Deigo, USA, 1–14.

Cheng, X., Dale, C., and Lui, J. 2008. Statistics and Social Network of YouTube Videos. In *Proc. International Workshop on Quality of Service (IWQoS)*. Enskede, The Netherlands, 229 – 238.

Clauset, A., Shalizi, C., and Newman, M. 2009. Power-law Distributions in Empirical Data. *SIAM Review 51,* 4 (November), 661–703.

Downey, A. B. 2005. Lognormal and Pareto Distributions in the Internet. *Computer Communications 28,* 7 (May), 790–801.

Gill, P., Arlitt, M., Li, Z., and Mahanti, A. 2007. YouTube Traffic Characterization: A View from the Edge. In *Proc. ACM Internet Measurement Conference (IMC)*. San Deigo, USA, 15–28.

Gill, P., Arlitt, M., Li, Z., and Mahanti, A. 2008. Characterizing YouTube User Sessions. In *Proc. SPIE Multimedia Computing and Networking (MMCN) Conference*. San Jose, USA, 1–8.

Gummadi, K., Dunn, R., Saroiu, S., Gribble, S., Levy, H., and Zahorjan, J. 2003. Measurement, Modeling and Analysis of a Peer-to-Peer File-Sharing Workload. *SIGOPS Oper. Syst. Rev. 37,* 5 (December), 314–329.

Guo, L., Tan, E., Chen, S., Xiao, Z., and Zhang, X. 2008. The Stretched Exponential Distribution of Internet Media Access Patterns. In *Proc. ACM Symposium on Principles of Distributed Computing (PODC)*. Toronto, Canada, 283–294.

Halvey, M. and Keane, M. 2007a. Analysis of Online Video Search and Sharing. In *Proc. ACM Hypertext and Hypermedia Conference*. Manchester, UK, 217–226.

Halvey, M. and Keane, M. 2007b. Exploring Social Dynamics in Online Media Sharing. In *Proc. International Conference on World Wide Web (WWW)*. Banff, Canada, 1273–1274.

Hefeeda, M. and Saleh, O. 2008. Traffic Modeling and Proportional Partial Caching for Peer-to-Peer Systems. *IEEE/ACM Trans. Netw. 16,* 6 (December), 1447–1460.

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

Mitra, S., Agrawal, M., Yadav, A., Carlsson, N., Eager, D., and Mahanti, A. 2009. Characterizing Web-based Video Sharing Workloads. In *Proc. International Conference on World Wide Web (WWW)*. Madrid, Spain, 1191–1192.

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

Newman, M. 2005. Power Laws, Pareto Distributions and Zipf's Law. *Contemporary Physics 46,* 5 (September), 323–351.

Yu, H., Zheng, D., Zhao, B., and Zheng, W. 2006. Understanding User Behavior in Large-Scale Video-on-Demand Systems. In *Proc. ACM SIGOPS/EuroSys European Conference on Computer Systems (EuroSys)*. Leuven, Belgium, 333–344.

Zink, M., Suh, K., and Kurose, J. 2008. Watch Global, Cache Local: YouTube Network Traffic at a Campus Network - Measurements and Implications. In *Proc. SPIE Multimedia Computing and Networking (MMCN) Conference*. San Jose, USA.