Characterizing Web-based Video Sharing Workloads

Siddharth Mitra¶ Mayank Agrawal¶ Amit Yadav¶
Niklas Carlsson‡ Derek Eager§ Anirban Mahanti†
¶Indian Institute of Technology Delhi, India
‡University of Calgary, Canada
§University of Saskatchewan, Canada
†NICTA, Australia

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

General Terms: Measurement, Modeling, Human Factors
Keywords: Workload Characterization, Video Sharing, UGC

1. INTRODUCTION

A video sharing service allows "user generated" video clips to be uploaded, and users of the service can view, rate, and comment on uploaded videos. Prior work has focused mostly on the YouTube video sharing service [1,2]. While YouTube is arguably the most popular video sharing service, studying the workload characteristics of other video sharing services, and identifying invariant properties as well as significant differences, is an important step towards building a broader understanding of this type of service.

With the aforementioned objective, we collected traces from four video sharing services: Dailymotion, Yahoo! video, Veoh, and Metacafe. 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. 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 longer videos. 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.

2. SUMMARY OF CONTRIBUTIONS

Our key contributions are summarized below:

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

- 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 show that video popularity can be measured in different ways and argue that one commonly used metric, specifically the number of views a video has received since it was uploaded, may not be appropriate when studying issues such as potential for caching. We define alternative metrics for quantifying video popularity that may be appropriate for such purposes.

For a complete description of our measurement methodology and a discussion of the above results we refer to our full paper [3]. Here we briefly describe invariants pertaining to video popularity.

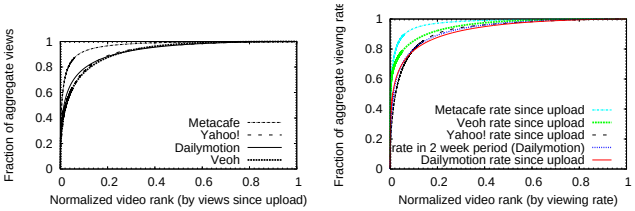
WWW2009, April 20-24, 2009, Madrid, Spain.

3. VIDEO POPULARITY

We distinguish between two different measures of popularity that have differing applications: 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*.

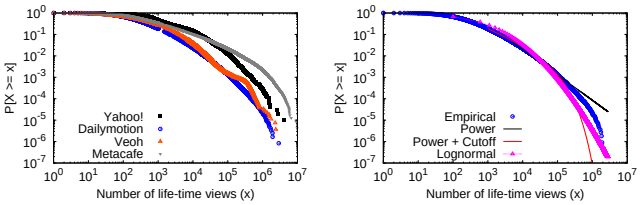
The total views popularity distribution is useful for understanding service features such as "all time" most popular listings, but does not provide an accurate picture of the distribution of the rates at which videos are viewed. The latter is very important when attempting to model the video reference 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.

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



(a) Total views popularity (b) Viewing rate popularity

Figure 1: Skewness in the video popularity.



(a) CCDF (b) Models (DM)

Figure 2: Total Views Popularity.

of the crawl. This latter measure removes, to some extent, the age bias in the total views popularity measure.

3.1 The 80-20 Rule for Video Popularity

We consider how skewed references are to the most popular videos. Figure 1(a) shows the cumulative distribution of the *total views popularity*. Figure 1(b) shows the cumulative distribution for the *viewing rate popularity* of the videos, using the two measures described above. For all three metrics, the Pareto principle holds for video views. Specifically:

- 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 [1] and the viewing rate [2] have been made for YouTube. We note, however, that the Metacafe service appears to exhibit significantly more skew than the other services we considered.

3.2 Zipf and Power Law Analysis

Power law is often 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 the scaling exponent of the distribution. 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 indicates power law scaling.

Figure 2(a) shows the CCDF plot for the total number of views since a video was uploaded (i.e., the *total views popularity*) for each of our data sets. 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 hotest videos.

Figure 2(b) shows the best fit power law, power law with exponential cut off, and lognormal distributions, for total views popularity as measured for the Dailymotion data set. It is often difficult to distinguish among the mathematical distributions that we consider in this graph, with respect to their goodness-of-fit to measured data. For example, the lognormal distribution can also exhibit a near straight line in the right tail of the CCDF plot when there is high variance in the distribution [4].

Table 1: Models for the total views popularity.

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

Using the likelihood ratio test [4], we compared power law fits with power law plus exponential cut off and lognormal fits. For Yahoo! video and Metacafe 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 1 presents the power law exponent α along with the minimum value of x for which power law or power law with cut off behavior holds for the data sets; the exponent ranges between 1.43 and 2.25, and is consistent with prior observations for YouTube's science and entertainment videos [1]. To summarize:

- 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 2.3. However, we note that neither a power law (or variants), nor a lognormal distribution, appear to fit the entire distribution well.

While the details of our analysis of the viewing rate popularity distributions are omitted we note that the average viewing rate since upload is generally more Zipf-like (or power law) than the total views popularity distribution. (While we have two snapshots only for the Dailymotion data set, we note that the average viewing rate for that two week period is even more Zipf-like than the average viewing rate since upload.) Furthermore, 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.

4. REFERENCES

[1] 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 IMC*, San Deigo, USA, October 2007.

[2] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube Traffic Characterization: A View from the Edge. In *Proc. ACM IMC*, San Deigo, USA, October 2007.

[3] S. Mitra, A. Yadav, M. Agrawal, N. Carlsson, D. Eager, and A. Mahanti. Characterizing Web-based Video Sharing Workloads. Technical Report, IIT Delhi, 2008.

[4] M. Newman. Power laws, Pareto distributions and Zipf's Law. *Contemporary Physics*, 46, 2005.