Performance Evaluation 68 (2011) 1037–1055

Contents lists available at SciVerse ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Characterizing and modelling popularity of user-generated videos

Youmna Borghol [a,b], Siddharth Mitra [c], Sebastien Ardon [a,b], Niklas Carlsson [d], Derek Eager [e], Anirban Mahanti [a,b,*]

[a] NICTA, Locked Bag 9013, Alexandria, NSW 1435, Australia
[b] School of Electrical Engineering and Telecommunications, University of New South Wales, NSW 2030, Australia
[c] Department of Computer Science and Engineering, Indian Institute of Technology Delhi, New Delhi 110016, India
[d] Department of Computer and Information Science, Linköping University, Linköping, SE - 581 83, Sweden
[e] Department of Computer Science, University of Saskatchewan, Saskatoon, SK S7N 5C9, Canada

ARTICLE INFO

Article history:
Available online 3 August 2011

Keywords:
User-generated videos
Popularity dynamics
Video sharing
Workload modelling

ABSTRACT

This paper develops a framework for studying the popularity dynamics of user-generated videos, presents a characterization of the popularity dynamics, and proposes a model that captures the key properties of these dynamics. We illustrate the biases that may be introduced in the analysis for some choices of the sampling technique used for collecting data; however, sampling from recently-uploaded videos provides a dataset that is seemingly unbiased. Using a dataset that tracks the views to a sample of recently-uploaded YouTube videos over the first eight months of their lifetime, we study the popularity dynamics. We find that the relative popularities of the videos within our dataset are highly non-stationary, owing primarily to large differences in the required time since upload until peak popularity is finally achieved, and secondly to popularity oscillation. We propose a model that can accurately capture the popularity dynamics of collections of recently-uploaded videos as they age, including key measures such as hot set churn statistics, and the evolution of the viewing rate and total views distributions over time.

1. Introduction

The phenomenal success of online media sharing and social networking services has resulted in massive volumes of user-generated content being created and spawned new content consumption approaches. With the success of these services, there is interest in understanding the popularity characteristics of user-generated content as well as interest in understanding the characteristics and processes governing their popularity dynamics. Understanding the popularity characteristics of user-generated content can be helpful in identifying potential bottlenecks in discovering content [1]. New and efficient content distribution approaches can be developed using workload models developed from characterization of user-generated content usage [1–3]. From a theoretical point of view, the massive amount of data available from these online services provides an unprecedented opportunity to understand the underlying social behaviour and collective human dynamics governing content creation and consumption processes [4,5].

In this paper, we study and model how viewing rates of user-generated videos change over time, which we refer to as *popularity dynamics* or *popularity evolution*, using datasets collected from YouTube.[1] Studying the popularity dynamics, however, is challenging because of the extremely large and rapidly growing number of videos available from such services.

* Corresponding author at: NICTA, Locked Bag 9013, Alexandria, NSW 1435, Australia.
E-mail address: Anirban.Mahanti@nicta.com.au (A. Mahanti).

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

0166-5316/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.peva.2011.07.008

As an example, consider YouTube the most popular video-sharing service today. YouTube was launched in early 2005, and by mid-2006, had approximately 65,000 video uploads and 100 million video requests per day [6]. More recently, YouTube reported receiving 48 h of new videos each minute.[2] In this paper, we make several contributions which are discussed below.

Our first contribution concerns biases that may be introduced in the analysis of user-generated video popularity owing to use of sampling techniques. Sampling is necessary as popular services host millions of videos with restrictions on the rate at which data may be fetched from the service. Furthermore, sampling is not straightforward because services often restrict how videos may be discovered from these services. From YouTube, for example, videos may be sampled from various "most-popular" lists (such as most viewed today, this week, this month, or all time most popular), the "recently-uploaded" list, or by searching using keywords. Evidently, sampling from any of the most-popular lists provides a set of videos that are biased towards popular content. In this work, we used the YouTube developer's API to collect two datasets, one based on sampling from the *recently-uploaded* videos, and another based on *keyword searches*.[3] We tracked the views to these videos over an eight month period. Perhaps not surprisingly, we find that sampling based on keyword searches yields a dataset biased towards more popular content. Fortunately, however, our results suggest that the YouTube API call that returns details on recently-uploaded videos gives an unbiased sample of such videos.

Our second contribution is an examination of popularity dynamics and churn, for our sample of recently-uploaded videos, over the first eight months of their lifetime. An important observation is that the relative popularities of the videos are highly non-stationary. One cause of the observed non-stationarity is the presence of large differences in when videos peak in popularity. While a majority of the videos peak in popularity, as measured by weekly viewing rate, within the first six weeks of their lifetime, many others do not peak until much later. Another cause of non-stationarity is the presence of oscillations in video popularity.

Our third contribution is a three-phase characterization of popularity evolution for our sample of recently-uploaded videos. This characterization is motivated by the observed non-stationarity in the relative popularity of these videos and the differences in how long it takes video popularity to peak. For each week, the videos are partitioned into three disjoint sets, based on whether they are *before*, *at*, or *after* their observed popularity peak. Grouping the videos in this manner, we identify several interesting properties of video popularity evolution. First, we find that within each set of videos, the distribution of the number of weekly views is heavy tailed, where the tails may be approximated by a lognormal distribution. Second, we find that these distributions are approximately week-invariant. Third, as a specific consequence of the second property, we find that the viewing rate at peak popularity is approximately independent of how long it takes videos to attain their peak popularity.

Finally, based on our three-phase characterization, we develop a model that can capture the popularity evolution of newly-uploaded videos. In particular, using only a small number of distributions based on the three-phase characterization, our model is able to generate synthetic datasets in which key characteristics and consequences of the video popularity dynamics match those observed in the empirical data, including the distribution of the weekly viewing rate for videos at a particular age, the distribution of total accumulated views to videos at a particular age, and measures of churn in the relative popularity of videos. The model is developed in two stages. We first present a basic model in which popularity churn results *only* from the movement of videos (at varying times) between their before-peak, at-peak, and after-peak phases. We find that this model successfully captures the first-order dynamics of popularity evolution and yields results matching most of the characteristics of the empirical data. So as to better capture churn characteristics, in particular hot set evolution, we present an extended model that adds a tunable degree of additional popularity variation by shuffling the popularities of the videos within each phase. Our model can be considered a first step towards a synthetic workload generator for user-generated video services.

This paper is organized as follows. Section 2 presents related work within the context of the contributions of our work. Section 3 describes how we collected our datasets, and presents some initial analyses concerning possible biases in the datasets owing to use of sampling techniques. Section 4 examines popularity dynamics and churn for our dataset of recently-uploaded videos. Section 5 presents our three-phase characterization of popularity evolution, and provides the underpinnings for the model proposed in this work. Section 6 presents the basic model, its validation, and also insights drawn from the model. An extension of the basic model is described in Section 7, followed by conclusions and directions for future work in Section 8.

2. Related work

The success of video-sharing services has resulted in studies on various facets of YouTube and other similar services. Studies have examined the characteristics of user-generated video files [7,8,1,9,10], use of social networking features in video-sharing services [11,9], the structure of YouTube's "friend" network [12], the use of the "video response" feature of YouTube [13], the popularity characteristics of user-generated videos [1,10,7,9,14,15], and also models for user-generated video popularity prediction [16,17]. Here, we restrict attention mostly to related work on popularity characterization and modelling for user-generated videos.

[2] http://youtube-global.blogspot.com/2011/05/thanks-youtube-community-for-two-big.html.

[3] Our datasets are available at http://www.cs.usask.ca/faculty/eager/Performance11.html.

Prior work on user-generated video popularity characterization has typically relied either on network traffic traces from a network gateway [7,10] or meta-data sampled from video sharing services [1,9,15,14]. Both Gill et al. [7,18] and Zink et al. [10] analysed YouTube video requests from a campus network and observed that the video requests follow a Zipf-like distribution and network bandwidth savings may be feasible if large proxy caches are used.

Cha et al. [1] analysed meta-data collected by crawling YouTube and found that the lifetime views (i.e., total views since upload) for videos of various ages is heavy-tailed, and can be modelled as power law with exponential cut-off. A "limited fetch" model where some users do not repeatedly request the same videos was formulated by Cha et al. [1] to explain the exponential drop off observed in the tail of the total views distribution. Mitra et al. [9] studied four video sharing services other than YouTube and identified invariants in video sharing workload properties such as uploading characteristics, video ratings, life-time views, and average viewing rates. In addition, Mitra et al. distinguish between lifetime and short-term popularity measures, and evaluate their respective degrees of relevance for cache management. Both Cha et al. [1] and Mitra et al. [9] analysed meta-data collected at a few points close together in time, and did not focus on the long-term temporal aspects of the popularity dynamics.

In recent work, Figueiredo et al. [15] studied popularity dynamics using three different YouTube datasets, specifically, a sample of most popular videos, a sample of deleted videos, and a sample of videos obtained via keyword searches. The authors utilized a recently available feature of the YouTube API that provides high-level and limited statistics of how clicks to a particular video grows over time, along with some information on sources of these clicks. None of these prior works [1,9,15] considered the issue of sampling biases.

There has also been interest in classifying user-generated videos based on the diversity in changes to the viewing patterns [14,4]. For instance, heterogeneity in growth behaviour has been explained by differences in the "socialness" of a video [14], and by "viral", "junk", and "quality" descriptors [4]. We build upon this body of work [1,10,7,9,14,15] by considering possible sampling biases and developing a framework for studying the long-term popularity dynamics of user-generated videos.

Perhaps most closely related to our work are the models by Szabo and Huberman [17] and Ratkiewicz et al. [19]. Szabo and Huberman [17] presented a model for predicting the total future view counts gathered by a video based on the total view count received at the time of prediction. They tracked approximately 7000 recently-uploaded YouTube videos for a period of one month and observed a strong linear correlation between the logarithmically transformed total views early in the lifetime and later in the lifetime of the video. The authors modelled future total view counts using the relationships between the logarithmically transformed total view counts. Ratkiewicz et al. [19] presented a model that combines the classical "rich-get-richer" model [20] with random popularity shifts, with the goal of capturing the influence of exogenous events on content popularity. The authors validated their model using click-through data for Wikipedia and the Chilean Web.

Predicting the future popularity of individual videos based on initial or current popularity is extremely difficult. It is perhaps not surprising that the absolute errors produced by the Szabo and Huberman [17] model are large [16]. Instead of trying to estimate the popularities of individual videos, Lee et al. [16] proposed an approach for estimating the probability that an online content item will reach a certain threshold popularity. Their method is based on survival analysis techniques, and it predicts the lifetime popularity based on early lifetime characteristics of the content. The authors validated their approach using data from online discussion forums. In contrast to the above works [17,16], we present a model that captures various dynamic characteristics observed in the empirical data, such as the weekly viewing rate and total views distributions, and the churn in video popularities as the videos age following their upload to the service.

3. Sampling approaches and bias

Studying the popularity dynamics of user-generated videos requires tracking a representative sample of videos over a period of time. Obtaining a random sample of user-generated videos from YouTube is, however, challenging because of the scale of the service, its continually-expanding catalogue of videos, and the service-specific limitations associated with discovering and tracking videos. Section 3.1 describes how we collected our datasets, including the two alternative sampling approaches used and the tracking of the sample videos over a period of eight months. Section 3.2 presents a high-level summary of our datasets. Section 3.3 describes the results of some initial analyses designed to identify possible biases in the datasets, as might result from the sampling approaches employed.

3.1. Data collection methods

We collected meta-data (such as number of views, ratings, and comments) on more than one million YouTube videos on a weekly basis for over eight months. The videos were selected by sampling, over a one-week period from 27 July to 2 August, 2008. A one-week sampling period was chosen to avoid potential day-of-the-week effects. We used two different sampling approaches, both based on functionality provided by the YouTube API,[4] as described below:

[4] http://code.google.com/apis/youtube/overview.html.

Table 1
Summary of datasets.

Dataset	Recently-uploaded	Keyword-search
Videos	29,791	1,135,253
Views (start)	1,203,755	40,094,514,507
Views (end)	39,089,184	64,019,907,026

- Sampling from the recently-uploaded videos: the API provides a call that returns details on 100 recently-uploaded videos. Using this API, we collected meta-data on approximately 29,500 videos during the one-week sampling period.
- Sampling using keyword search: the API also allows retrieval based on keyword searches; the API returns search results sorted by "relevance". We performed keyword searches using words chosen randomly from a dictionary. As search results for some words return a very large number of videos, for those returning more than 500 videos we selected only the first 500. We found approximately 1 million videos using this method during the one-week sampling period.

There are several other possible approaches to sampling videos. One approach is to sample from the "most-popular" lists. A closely-related variant is to start the sampling process from one or more videos in the "most-popular" list and subsequently follow "related videos". A detailed investigation of the biases introduced by other sampling approaches is left for future work.

During the remainder of our measurement period, specifically from 3 August 2008 to 29 March 2009, we collected meta-data for the videos identified in the sampling phase on a weekly basis. Using the timestamp at which the meta-data for a video was first captured, we ensured that subsequent measurements ("snapshots") were exactly one week apart. For example, if a video was sampled on Tuesday evening, then each weekly measurement for this video was performed as close to the same time of day as possible, on Tuesday evenings, in the following weeks. This form of staggering allowed us to track a large number of videos without exceeding YouTube's query rate limitations, while enabling easier management of our own measurement resources.

3.2. Summary of datasets

A summary of our datasets is presented in Table 1. In total, we have 35 snapshots for each sampled video's meta-data (counting also the "seed" snapshot collected during the sampling phase), with one-week spacings between consecutive snapshots. From the total view count at each snapshot i ($1 < i \leq 35$), we can determine how many times the video was viewed during the one-week period since snapshot $i - 1$, which we term the "added views" at snapshot i. The total view count at snapshot 1 (the seed snapshot) tells us the total views acquired by the video from its upload time until the start of our data collection for that video.

During the measurement period, the 29,791 recently-uploaded videos acquired about 38 million additional views, and the 1,135,253 keyword-search videos received about 24 billion additional views. Note that the keyword-search videos acquired additional views at a higher average rate than the recently-uploaded videos. This suggests possible bias in the keyword-search dataset towards more popular videos, which is investigated further in the next section.

3.3. Sampling bias in the datasets

One indicator of sampling bias is a skewed age distribution, where video age is defined as the time since upload of the video. Fig. 1 shows histograms for the age at seed time (i.e., when meta-data for the video was first collected) for the recently-uploaded videos, using 6 h bins (left plot), and for the keyword-search videos, using one-week bins (right plot). The age of the videos in the recently-uploaded dataset is approximately uniformly distributed within a week, which is consistent with the hypothesis that the YouTube API call used to obtain these videos returns randomly-selected videos at most one week old.[5] The age distribution of the keyword-search dataset videos, in contrast, shows that this dataset is far from being a random sample of (all ages of) YouTube videos. There is a strong skew towards younger videos, with a prominent spike in the distribution for the first bin corresponding to an age of at most one week. The age of the oldest video is about 38 months. One possible explanation of the observed skew is that the results returned from keyword searches are biased towards more popular videos. This hypothesis is supported by the popularity characteristics of the keyword-search and recently-uploaded videos, as described next.

Fig. 2 shows the complementary cumulative distribution function (CCDF) of the added views at snapshots $i = 2, 8, 32$, using logarithmic scales on both axes, for both datasets. Comparing the added views of the keyword-search and the recently-uploaded videos at the same snapshot, note that the keyword-search videos receive substantially more views than the recently-uploaded videos. This is reflected, for example, by a heavier right tail for the keyword-search video curves. At each snapshot, the most (currently) popular keyword-search videos (i.e., those with the most added views) have an order

[5] We notice a dip in the histogram for videos that are approximately 96–108 h old at time of collection. This dip is currently unexplained.

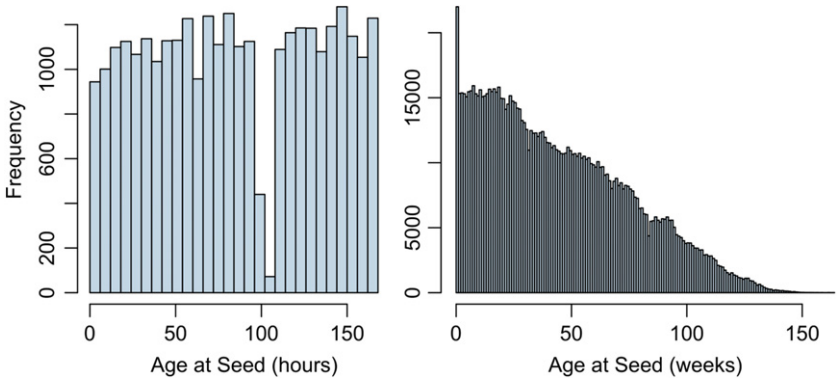


Fig. 1. Age distribution of the videos (left: recently-uploaded; right: keyword-search).

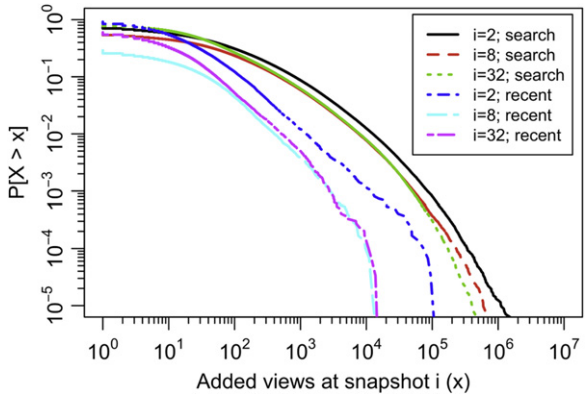


Fig. 2. Distribution of added views at snapshot *i*, for recently-uploaded and keyword-search videos.

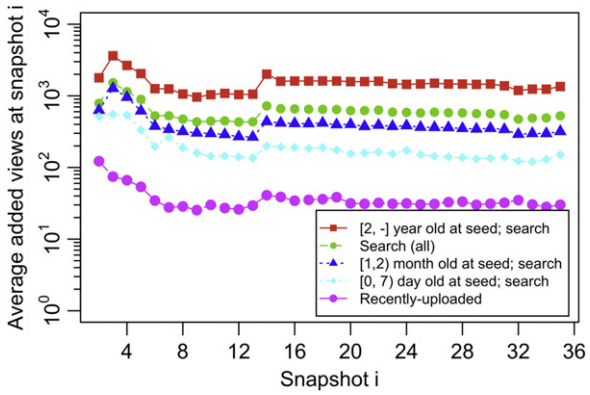


Fig. 3. Average added views at each snapshot.

of magnitude more new views than the most popular recently-uploaded videos. Further, for the recently-uploaded video dataset, as we look further into the measurement period the curves shift to the left, owing to a decreasing fraction of these videos that are currently popular. The corresponding shift for the keyword-search videos is less pronounced.

The popularity characteristics of the recently-uploaded and keyword-search videos are investigated further in Fig. 3, which shows the average added views at each snapshot for both datasets. For the keyword-search videos, we also consider subgroups based on video age at the time of seeding. On average, the keyword-search videos (cf. the "Search (all)" line in the graph) attract more than 10 times as many views throughout the measurement period compared to the recently-uploaded videos, providing additional evidence that the keyword-search video dataset is biased towards more popular videos. (Note that the *y*-axis is on logarithmic scale.) The results for the keyword-search video subgroups based on age further support this conclusion. First, we note that the older videos in the keyword-search dataset appear to attract substantially more new

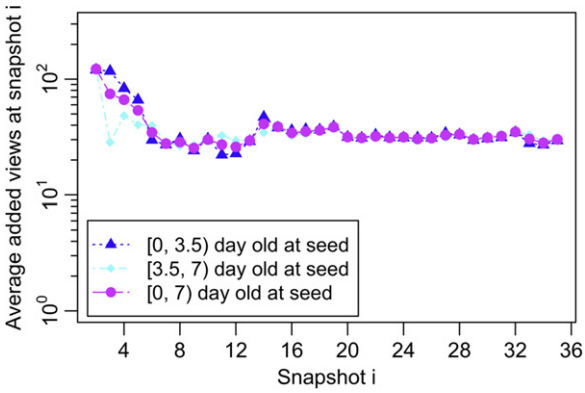


Fig. 4. Average added views at each snapshot for subgroups of the recently-uploaded videos.

views, on average, than their younger counterparts. Second, note that the week-or-less old keyword-search videos obtain new views at a higher rate than the recently-uploaded videos, which have the same age range, throughout the measurement period.

The fact that these popularity differences persist for the entire measurement period indicates that the keyword-search video dataset is biased towards videos with elevated long-term popularity. Fig. 3 also suggests bias based on elevated short-term popularity. In particular, consider the results for keyword-search videos that are 2 years or older at the time of seeding. For a randomly-selected set of videos of this age, one would expect to see a fairly stable average viewing rate over periods of a few weeks, whereas for this subgroup of keyword-search videos, as seen in Fig. 3 there is an initial period of significantly higher average viewing rate, reflecting elevated short-term popularity.

Next, we consider further the possibility of biases in the recently-uploaded video dataset. Recall from Fig. 1 that the age at seed time for these videos is approximately uniformly distributed, up to a maximum of one week. One indicator of bias towards more popular videos would be a correlation between the age at seed time, and the rate of accumulating new views, since it may be easier to predict (for the purposes of preferential selection) the future popularity of older videos. Fig. 4 shows the average number of added views at each snapshot for those videos in the recently-uploaded video dataset whose age at seed time is less than 3.5 days, for those videos whose age at seed time is at least 3.5 days, and for the entire dataset. Note that there are no observable longer-term differences in the viewing rate behaviour among these three groups of videos. There do exist some differences in the first few weeks of the measurement period, which is to be expected for recently-uploaded videos. We have also considered other properties such as the distribution of the total views at the end of the measurement period, and have similarly found no significant differences among these groups.

To summarize, our conclusions regarding sampling bias are:

- The keyword-search videos appear to be biased towards those videos that exhibit both higher short-term and long-term popularity.
- The recently-uploaded video dataset appears to exhibit no observable bias towards popular content.

Based on our analysis, we conjecture that the recently-uploaded video dataset is a random sample representative of the videos uploaded to the service. In the remainder of this paper, we characterize the popularity dynamics of these videos over the first eight months of their lifetime, and from this characterization develop a model for popularity evolution of newly-uploaded videos.

4. Popularity dynamics and churn

In this section, we dig deeper into the popularity dynamics of the videos in the recently-uploaded dataset, with the objective of developing insights for modelling the popularity evolution of these videos. One goal is to understand whether or not current popularity is a good predictor of future popularity of user-generated videos. If current popularity is indeed a good indicator of future popularity, then modelling the popularity evolution of individual videos is certainly feasible [17]. However, the popularity of an object may be influenced by many exogenous and endogenous factors [5,4], which may introduce some degree of inherent unpredictability [5,16]. In this section, we characterize the degree of (in)stability and (un)predictability of the popularity of individual videos and the extent of churn in the relative popularities of videos.

Fig. 5 shows scatter plots for the number of added views received by a video at adjacent snapshots for some example early and later snapshots. With our notion of added views at a snapshot (or the weekly viewing rate, as determined by our measurement granularity), this figure illustrates the change in viewing rate between consecutive snapshots. The scatter plots, especially for the first few snapshots since a video is uploaded, show substantial point spreads which indicates that a large number of videos experience significant variation in viewing rate from one week to another. We observe that a video

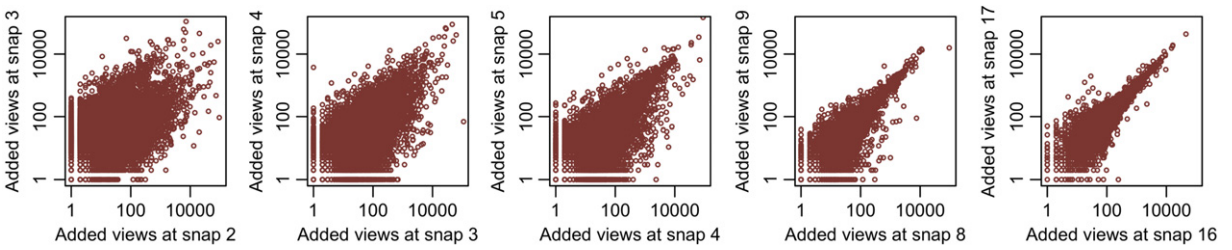
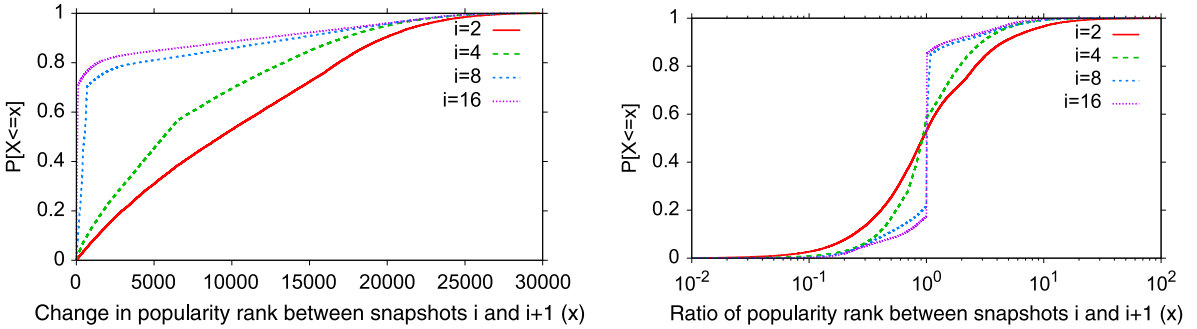


Fig. 5. Scatter plot of the number of added views at snapshots i versus $i + 1$.



(a) Absolute change in ranks.

(b) Ratio of new to old popularity ranks.

Fig. 6. Distribution of change in popularity ranks of videos.

that is mildly popular in the week prior to one snapshot can become highly popular before the following snapshot, and vice versa. Videos that have about 1000 views added at snapshot two, for example, could receive less than 100 additional views, or more than 10,000 additional views, at snapshot three. Overall, we observe substantial non-stationarity in the popularity of individual videos, especially within the first five to six weeks of their upload. Looking further in our measurement period, we see that there are fewer diverging points at the top right quadrant of the scatter plots, as the videos become older. Note that the scatter plots have fewer points for later snapshots owing to videos that received no views in one or both of the weeks of interest (and hence are not shown on the log–log plots).

We also computed the Pearson's correlation coefficient between the added views at adjacent snapshots. A correlation coefficient value of 0.8 or more is considered to reflect strong positive linear correlation [1]. The correlation coefficient between the added views at snapshots two and three is close to zero (0.09). Until week eight of our measurement, as may be expected from visual inspection of Fig. 5, at best, a weak positive linear correlation (less than 0.7) between added views at successive snapshots is observed. As videos become older, we observe a very strong positive linear correlation between the viewing rate of videos across adjacent snapshots. These observations, together with Fig. 5, indicate that the initial or current popularity of a random young video is likely not a reliable indicator of its future popularity; on the other hand, it appears that the current popularity of an older video may be indicative of its immediate future popularity.

The non-stationarity in the weekly views to videos impacts the relative popularity of videos. For any snapshot of our measurement period, we can rank the videos according to the number of views added to each video's view count at the considered snapshot. Ties are broken using an assigned video id. Based on the assigned ranks in any two snapshots, we can calculate how much each video's rank shifts. Fig. 6(a) shows the cumulative distribution of the absolute value of the rank shifts for some example snapshots. Early in the measurement period, and thus when the videos are young, videos experience significant rank changes. For example, between snapshots two and three more than 30% of the videos in our recently-uploaded dataset switch 10,000 or more rank positions. The changes in the relative popularities of videos stabilize after the initial weeks; however, there are still some videos that experience substantial rank shifts between consecutive weeks. This trend is consistent with the trend suggested by Fig. 5.

In Fig. 6(b), we present the cumulative distribution of the ratio of ranks for some example snapshots. This analysis complements the results presented in Fig. 6(a) by considering each video's popularity rank increase/decrease relative to its current rank. Significant changes in the relative ranks of videos are observed when videos are young. Between snapshots two and three, for example, about 30% of the videos gain a factor of two or more in popularity rank, whereas less than 10% of the videos experience similar increases in popularity rank in later snapshots. In fact, when videos become eight weeks or older, approximately 75% of them retain their popularity rank across (weekly) snapshots.

Differences in how rapidly videos attain their peak popularity can be a major cause of churn in relative popularities. Fig. 7 shows the cumulative distribution of time-to-peak for the videos in the recently-uploaded dataset, where we define time-to-peak for a video as its age (time since upload) at which its weekly viewing rate is the highest within our measurement

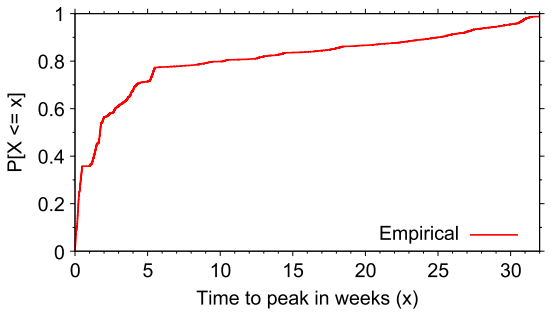


Fig. 7. Time-to-peak distribution for videos.

period.[6] The time-to-peak distribution shows that a large fraction of the videos, approximately three-quarters of them, peak within the first six weeks since their upload. The remainder peak at times approximately uniformly distributed between week six and the end of our measurement period. For those videos that peak within the first six weeks after upload, we find the time-to-peak to be approximately exponentially distributed. As we show later in the paper, the fact that many videos reach their peak popularity quickly plays an important role in explaining the high churn observed in the relative popularity of the videos over the first few weeks of the measurement period.

5. Three-phase characterization

The results of Section 4 suggest the futility of attempting to reliably model the popularity evolution of individual videos. We can, however, attempt to model the popularity dynamics of a collection of videos. In this section, we develop a characterization of the popularity evolution observed for our dataset of recently-uploaded videos. This characterization is applied to develop a popularity evolution model in Section 6.

Perhaps the biggest challenge in developing such a characterization is that of capturing the churn in the relative popularities of videos that is observed in the empirical data. As noted in Section 4, variations in time-to-peak may be an important factor in this churn. This motivates us to develop a *three-phase* characterization of popularity evolution, in which videos are grouped according to whether they are *before*, *at*, or *after* the age at which they attain their peak popularity.

Of particular interest in this characterization are: (a) the movement of videos among these phases (i.e., the time-to-peak distribution, as examined in Section 4), (b) the distribution of the viewing rate for the videos belonging to each group, and (c) the dependence of these distributions on video age.

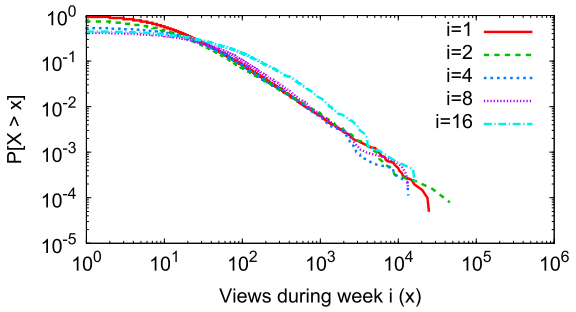
First, consider the distribution of weekly views to videos in each phase. Fig. 8 shows the complementary cumulative distribution of views during a week within each phase, using a logarithmic scale on each axis. Note that, by definition, none of the videos are past their peak (i.e., in the after-peak phase) in the first week. Similar to the distribution of views during a week for all videos, the distribution of views for the videos within each phase is also heavy-tailed. It also appears that the skew towards larger view counts is the largest when the videos are at their peak, and the least when the videos are past their peak. By inspection of Fig. 8, we notice that, within each phase the distribution of views each week is approximately similar and suggests the possibility of modelling the distributions as week-invariant.

We now investigate the efficacy of assuming the weekly viewing rate within each phase to be approximately week-invariant. Fig. 9 shows for each week and phase (of the lifetime of the videos) the average number of weekly views. The average viewing rate at peak exhibits a fair degree of variability. The observed variability may be expected as videos peak with varying (and occasionally very large) numbers of views during a week. An interesting observation is that there is no discernible trend in the average viewing rate in the at-peak phase; this provides additional evidence in support of a modelling approach in which the at-peak views distribution is modelled as week-invariant.

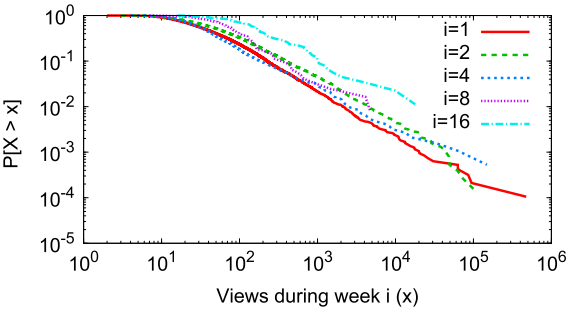
Unlike the high variability observed for the average number of weekly views to videos that are in their at-peak phase, the average views for after-peak videos appears to be quite stable throughout the measurement period. The average views for before-peak videos also lacks the high variability that is observed for at-peak videos, but appears to exhibit an increasing trend. This increasing trend may be an artefact of the finite measurement period, however; note that as the end of the measurement period grows closer, the maximum time period until each of the before-peak videos peaks corresponding shrinks.

Working further with our week-invariant assumption, we take a closer look at the distribution of views during a week, or equivalently the distribution of the viewing rates, for each of the three phases. Our goal was to get an understanding of the distribution of the viewing rate when videos are grouped by phases, ignoring week-specific behaviour. Fig. 10 shows that the distribution of weekly views to videos within each phase is heavy-tailed. As expected from our earlier discussion, the skew towards larger views is greatest when videos are at their peak, and least when they are past their peak.

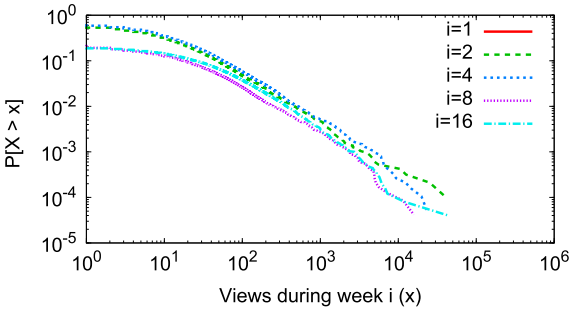
[6] Appendix A describes the details of how we determine this age given the fairly coarse granularity of our measurements. In general, we have found our results to be insensitive to alternative choices for these details.



(a) Before-peak phase.



(b) At-peak phase.



(c) After-peak phase.

Fig. 8. Distribution of weekly views to videos in the before-peak, at-peak, and after-peak phases for example weeks i ($i = 1, 2, 4, 8, 16$).

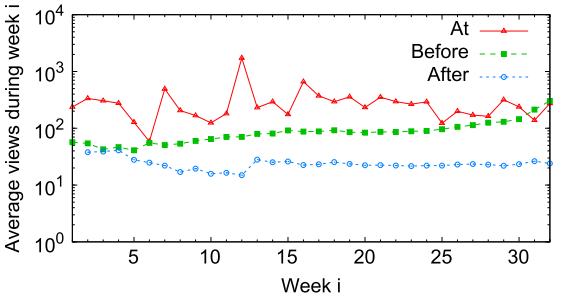


Fig. 9. The average weekly viewing rate of videos in the before-peak, at-peak, and after-peak phases.

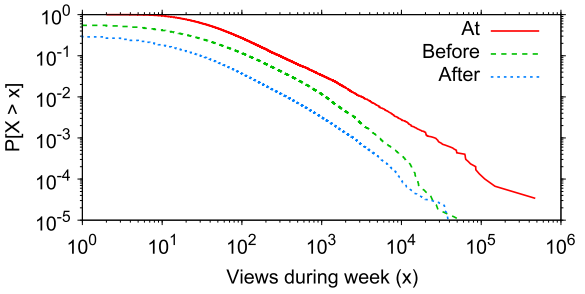


Fig. 10. Distribution of views during a week for videos that are in their before-peak, at-peak, and after-peak phases.

Because of the heavy-tailed nature of the views distribution in each phase, we took a closer look at the tail of each distribution where we define the tail to consist of only the largest ten percent of the views within each phase. Using our definition of the tail, we determine thresholds of 116, 296, and 31 weekly views for a video to be considered in the tail of the before-peak, at-peak, and after-peak view distributions, respectively. Fig. 11 shows, for each week of our measurement period, the average number of weekly views for those videos that acquired greater than or equal to these threshold weekly views. The average viewing rate is quite steady for each phase, suggesting that the week-invariant assumption is a reasonable approximation for the distribution tails.

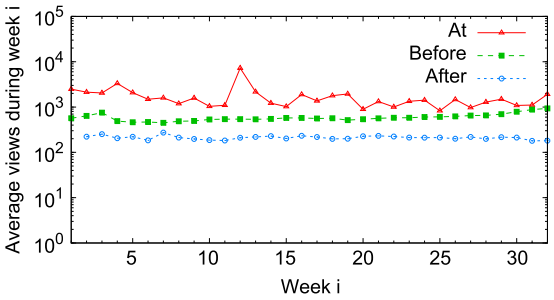


Fig. 11. The average weekly viewing rate of videos in the *tail* of the before-peak, at-peak, and after-peak distributions.

To summarize, our three-phase characterization suggests that the viewing rate distribution within each phase could be modelled as week-invariant. We find that the tail of the viewing rate distribution can be modelled separately using heavy-tailed distributions. Appendix B presents the specific distribution fits that are found to best capture the characteristics of the empirical data. Overall, we find that the distribution of weekly views, for each of the three phases, can be modelled using an appropriately parameterized lognormal distribution for the tail and a beta distribution for the views that are not in the tail.

6. Basic model

Guided by the observations made in the foregoing sections, we develop a basic model for generating weekly views to *individual* videos in a *collection* of newly-uploaded videos. The model is developed using the observations pertaining to the before-peak, at-peak, and after-peak phases. The distribution of weekly views to videos within each phase is modelled to be week-invariant. From a modelling point of view, this is an attractive property as the distribution of weekly views can be succinctly represented using only three distributions, one for each phase. Transitions of videos between phases, specifically from being in their before-peak phase, to their at-peak phase, and then to their after-peak phase, are modelled using a time-to-peak distribution (such as the one shown in Fig. 7).

The basic approach consists of sampling views from the before-peak, at-peak, and after-peak distributions (cf. Fig. 10), and assigning them to videos. For each modelled week, we sample views from the before-peak, at-peak, and after-peak distributions based on how many videos are in each of these phases. Note that the number of videos that peak in any week is determined using a time-to-peak distribution (cf. Fig. 7). At the start of an arbitrary week, from among the videos that have not peaked thus far, some videos transition to being at their peak; subsequently, at the end of this week these videos will move into the after-peak phase.

For this approach to yield weekly views for individual videos, a framework for assigning the sampled views to individual videos is required. A straightforward approach for assigning weekly views to videos is based on an assumption that the relative popularities of videos in the same phase, or that were in the same phase during the previous week, are unchanged from the previous week, and precedes as follows. Assign the views sampled from the before-peak and at-peak distributions to those videos that were in their before-peak phase during the previous week; similarly, assign the sampled views from the after-peak distribution to those videos that were in their at-peak or after-peak phases during the previous week. In both cases, the assignment is made such that the relative popularities of the respective videos are preserved. Now, among those videos that were in the before-peak phase during the previous week, those that are assigned views from the at-peak distribution are assumed to peak in this week (and will be in their after-peak phase for all subsequent weeks). With this approach for assignment of views, churn with respect to the relative popularity of videos is introduced by videos moving between the three phases.

6.1. Views generation algorithm

Our algorithm requires the following input: the total number of newly-uploaded videos N, the total number of weeks d, a time-to-peak distribution, a distribution for the weekly views for videos in the before-peak phase, a distribution for the weekly views for videos in the at-peak phase, and a distribution for the weekly views for videos in the after-peak phase. The main steps of the algorithm are as follows:

1. **Determine the number of videos in the before-peak, at-peak, and after-peak phases.**
   Sample N values from the time-to-peak distribution and determine the number of videos n_i^{at} that peak at week i, for all $i \leq d$. Note that $n_i^{before} = n_{i-1}^{before} - n_i^{at}$, $n_i^{after} = n_{i-1}^{after} + n_{i-1}^{at}$, for $i > 1$. Also note that $N = n_i^{before} + n_i^{at} + n_i^{after}$ and $n_1^{after} = 0$. Therefore, following this step, we know the number of videos n_i^{before}, n_i^{at}, and n_i^{after} that are in the at, before, and after-peak phases, respectively, during week i. In our experiments, as described in Appendix B the time-to-peak distribution is chosen as a mix of an exponential and a uniform distribution.
   For $i = 1, 2, \ldots, d$:

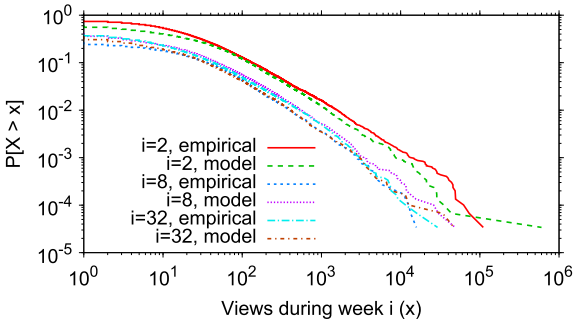


Fig. 12. Distribution of the views during week i in the recently-uploaded dataset and the basic model ($i = 2, 8, 32$).

2. **Sample views from the before-peak, at-peak, and after-peak distributions.**
   Sample n_i^{before}, n_i^{at}, and n_i^{after} times from the before-peak, at-peak, and after-peak distributions. In our experiments, we use a mixture of beta and lognormal distributions for each of these three phases.
3. **Assign views to the videos.**
   (a) if $i = 1$: Note that $n_1^{after} = 0$, i.e., there are no videos in week one that are after their peak. Assign the N ($= n_1^{before} + n_1^{at}$) sampled views to the videos.
   (b) if $i > 1$: Sort the sampled n_i^{at} "at-peak" views and the n_i^{before} "before-peak" views and assign them to those videos that are in the "before-peak" phase during week $i - 1$ such that the video with the highest view during week $i - 1$ is assigned the highest view in week i, the video with the second highest view during week $i - 1$ is assigned the second highest view during week i, and so on. Similarly, assign the sampled n_i^{after} after-peak views to those videos that were either at or after their peak in week $i - 1$.
4. **Determine the videos that peak in this week.**
   The videos that were assigned views sampled from the "at-peak" distribution are assumed to peak this week; for all subsequent weeks these videos will be in their "after-peak" phase.

6.2. Results and discussion

This section presents results from our basic model. We used our implementation of the basic model to generate synthetic views for $N = 29{,}791$ videos for a total of $d = 32$ weeks. The parameterization of the distributions (i.e., time-to-peak, weekly views during each phase) was done as specified in Appendix B.

Simple tests of our model include comparisons of the time-to-peak distribution, and the viewing rate distributions for videos in each of the three phases, for the synthetic data versus the corresponding empirical distributions for the recently-uploaded video dataset. Good matches are obtained but this is not surprising since the model was parameterized from these empirical distributions. Such tests do not show that our simple three-phase characterization (on which our model is based) captures enough of the detail of popularity evolution, to ensure that our synthetic data matches the empirical data on the metrics of practical interest concerning popularity and its evolution.

For such an evaluation, we test whether the synthetic data matches the empirical data from the recently-uploaded video dataset with respect to: (a) the distribution (over all videos) of views received during each week (e.g., the skewness in popularity among the videos, and the evolution of this skewness over time), (b) the distribution of the total views since upload at the end of each week (e.g., the skewness in total accumulated video views, or "long term average popularity", and the evolution of this skewness over time), and (c) hot set dynamics (e.g., how much churn is experienced in hot sets of various sizes from week to week). Note that the evaluation metrics considered above are not directly fitted from the empirical data. Instead, for the synthetic data, these evaluation metrics are consequences of the views generation algorithm and modelling parameters derived from the three-phase characterization of the recently-uploaded dataset.

6.2.1. Distribution of weekly views

Fig. 12 shows the CCDF of the views received during week i, for $i = 2, 8, 32$, for both the recently-uploaded dataset and the synthetic views generated by our basic model. We first observe that the weekly views distributions exhibit heavy tails, with videos receiving fewer large views in later weeks than earlier weeks. Further, we observe that the average weekly views to videos do not significantly change after the initial six weeks.

Overall, the match between the model generated views and the views in the recently-uploaded dataset is good, except for some differences for the least popular videos during a week (which is to be expected owing to our simplifying assumption of week-invariant distributions for the phases). Quantile-to-quantile (Q–Q) plots were used to evaluate the match for the body and tail of the distribution (cf. Appendix C). The observation pertaining to the average weekly views changing substantially only in the first six weeks is explained by the fact that a large majority, approximately 80% of the videos, peak within the

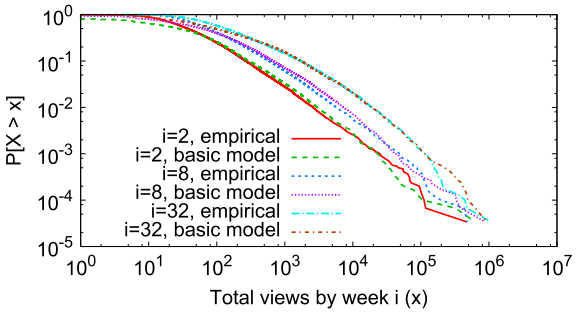
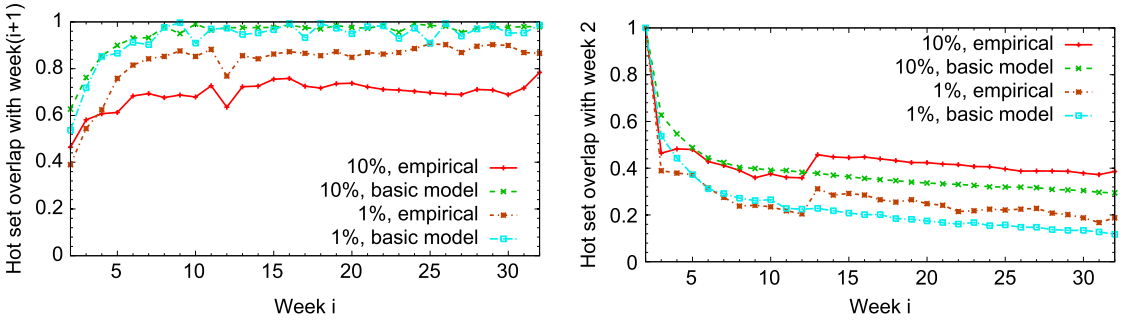


Fig. 13. Distribution of the total views by week i in the recently-uploaded dataset and the basic model ($i = 2, 8, 32$).



(a) Week i and $i + 1$ hot sets.

(b) Week 2 and week i hot sets.

Fig. 14. Churn in video popularity measured by changes to the hot set for the recently-uploaded dataset and the basic model.

first six weeks since their upload. Once a video is past its peak, it is in the after-peak phase where the viewing rate is approximately week-invariant. With a majority of the videos in the after-peak phase, the average viewing rate remains approximately constant.

6.2.2. Distribution of total views

The next property we consider is the distribution of total views as a function of weeks since upload (or equivalently the age of the videos). Fig. 13 shows the CCDF of the total views received by week i, for $i = 2, 8, 32$, for both the recently-uploaded dataset and the synthetic views generated by our basic model. We observe an excellent match between the empirical and synthetic datasets. Again, Q–Q plots were used to evaluate the match for the body and tail of the distributions (cf. Appendix C).

The general shape of the total views distribution from the recently-uploaded dataset and the model provide insights into the popularity dynamics of the videos. Notice that the distribution of views during a single week, shown in Fig. 12 for several representative weeks, appears to have a fairly "straight" tail. The total views distribution, for both the empirical dataset and the model, is fairly straight for the first few weeks, but transitions to a more "curved" tail as the videos become older. In the figure presented, this change can be seen by comparing the curves for weeks two and eight. If videos that are currently popular continue to be popular in the future, as one expects in simple rich-get-richer models, then we expect the distribution of the sum of the views to also exhibit a "straight-ish" tail. We believe that this change in the characteristics of total views can be explained by the presence of churn in video popularity. Our basic model, which retains strong correlation between current viewing rate and future viewing rate except when videos change state (e.g., move from being in the before-peak phase to at-peak phase to after-peak phase), exhibits a very similar change in the shape of the distribution.

6.2.3. Churn and hot set dynamics

The skew observed in the popularity of videos can aid in caching [7,1]. However, caching decisions become difficult with increase in churn among the videos [9]. This section quantifies the amount of churn among the most popular videos, its potential impact on caching decisions, and studies how well our model captures the churn observed in the recently-uploaded dataset.

For studying churn, we define the hot set at week i to consist of the most popular x% of the videos with respect to the views received during week i. Fig. 14(a) shows the overlap between hot sets of successive weeks for both the recently-uploaded dataset and our basic model, for hot sets of size $x = 1\%$ and $x = 10\%$. If caching decisions are made based on the hot sets of the week immediately preceding the current week, then these graphs give us an indication of the amount of cache replacement traffic. Effectively, these results tell us how good an indicator the immediate past is with respect to the immediate future.

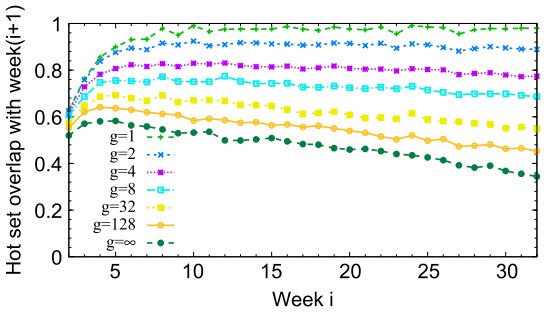


Fig. 15. Impact of the churn modelling parameter, with respect to the weekly churn in video popularity, as measured by weekly changes to the hot set using the extended model.

The results indicate presence of substantial churn. With a hot set of size $x = 10\%$, for example, we observe between 20%–60% change in the constitution of the hot sets between two consecutive weeks, with significantly higher churn observed in the first eight weeks. Comparing the results for the smaller and larger hot sets, the percentage change is more for larger hot sets, because of replacement of videos in the hot set with videos of similar popularity. Our model captures the trend of increased churn early in the lifetime of videos. However, our model suggests relatively smaller week-to-week changes than the corresponding empirical dataset.

The results presented in Fig. 14(a) show the relative change in the hot set from week-to-week. Fig. 14(b) presents an example result for the absolute change in the hot set. Here, we measure the number of common videos between the actual hot set at week i and the hot set of week two.[7] Both the model and the empirical hot set analyses show that there is substantial non-stationarity in the hot sets, with the model capturing the trend exhibited by the recently-uploaded dataset.

It is not surprising that our basic model does not exhibit as much churn as seen in the recently-uploaded dataset. Our basic model introduces churn by transitioning videos between phases. The model captures the large change in position caused by videos moving from being before their peak to being at their peak, and subsequently being after their peak. As we capture churn caused only by movement between phases, we see a better match for the smaller hot set then the larger hot set. The next section extends the model to introduce *second-order churn* effects with respect to the relative popularity of videos within each phase of their lifetime.

7. Model extension: perturbations

Our basic model introduces churn in relative popularity of videos only owing to the videos moving between the three phases during their lifetime. We now extend the model to introduce *second-order churn* by shuffling the popularity of videos within each phase, while preserving each video's phase.

The model extension is as follows. We first generate weekly views for videos conforming to the basic model. Then, we introduce perturbations in the relative popularity of videos during a week by exchanging the views assigned to selected videos. The views are exchanged such that none of the key characteristics of the basic model, specifically the distribution of weekly views for the before-peak, at-peak, and after-peak phases, as well as the distribution of time-to-peak are affected.

The algorithm for introducing additional churn is as follows. First, we assign weekly views to N videos for a period of d weeks according to our basic model. Then, for each week i and video v that does not peak in that week, we define a window W_i^v that specifies the bounds on views for a possible exchange:

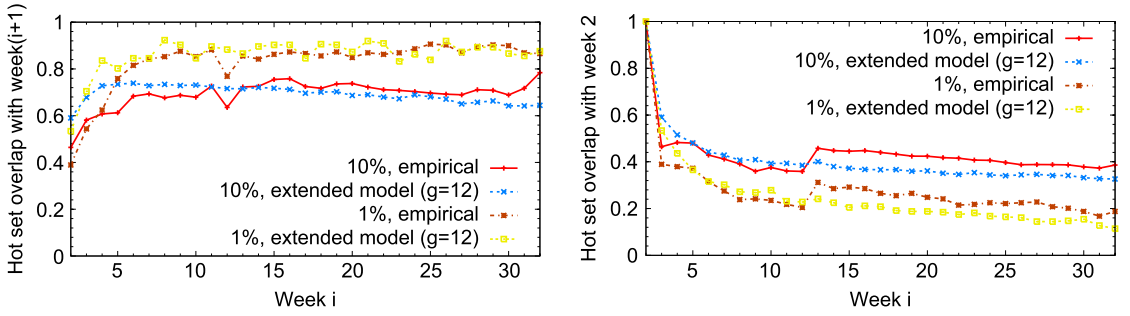
$$W_i^v = \left[\frac{x_i^v}{g}, \min(x_i^v \times g, x_{max}^v)\right], \quad g \in [1, \infty]$$

where x_i^v is the view assigned to video v during week i, x_{max}^v is video v's peak weekly viewing rate (i.e, $x_{max}^v = \max_j x_j^v$), and g is a modelling parameter that controls the maximum distance a video would be shifted with respect to its view count in a week. Specifically, we repeat the following step a sufficiently large number of times:

- Randomly pick a week i and two videos u and v such that either both videos peaked *before* week i, or both peak *after* week i. If the views during week i to videos u and v can be exchanged without causing either video's views for week i to move outside their respective windows W_i^v and W_i^u, switch the views.

Following a large number of iterations, views would be "uniformly mixed" subject to the constraint specified by the tunable parameter g; $g = 1$ conforms to the basic model outlined in the preceding section, and $g = \infty$ incorporates the maximum possible churn while still preserving the per-phase properties.

[7] We present comparisons with the hot set for week two because the empirical dataset has only a partial first week (i.e., a snapshot at seed); refer to Section 3 and Appendix A for details on data collection and sampling granularity.



(a) Week i and $i+1$ hot sets.

(b) Week 2 and week i hot sets.

Fig. 16. Churn in video popularity measured by changes to the hot set for the recently-uploaded dataset and the extended model.

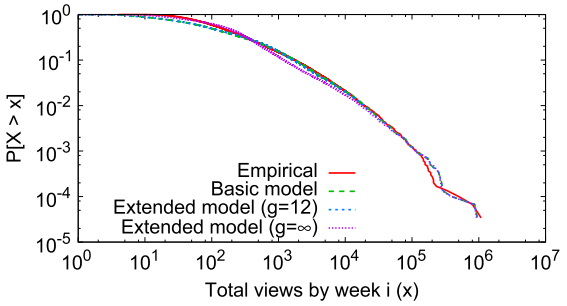


Fig. 17. The total views distribution after 32 weeks, in the recently-uploaded dataset and the extended model.

Fig. 15 presents results for a hot set evolution for hot set of size $x = 10\%$ for various values of g. As expected, with increasing g there is increased churn. This allows the model to capture a wide range of churn activity. With no additional churn, as described by the $g = 1$ (basic model), the constitution of hot sets between adjacent weeks changes by less than 10% once videos are seven weeks old. With $g = \infty$, the composition of videos in the hot sets change by as much as 50%–60% during the course of the 32 weeks.

Through experimentation, we found that g in the range $8 \leq g \leq 16$ yields a close match to the churn observed in the recently-uploaded dataset. Fig. 16 compares the hot set churn in the recently-uploaded dataset with the hot set churn using our extended model, with $g = 12$. Results are shown for both hot set size $x = 1\%$ and $x = 10\%$. We obtain a much better match between the curves than when only using the basic model.

Note that the model extension is constructed such that the distribution of weekly views is not impacted by the introduction of additional churn. We close our discussion on model extensions by investigating how the additional churn influences the total views distribution of videos. The results are presented in Fig. 17. Also here, we note that introduction of additional churn does not affect the tail of this distribution. Addition of further churn, by tuning the parameter g, however, has some (mostly negligible) impact at the head of the distribution.

8. Conclusions

Content popularity dynamics are governed by complex processes dependent on several endogenous and exogenous factors. A good understanding of these dynamics can have technological, economical, and societal implications. In this work, we focus on the popularity dynamics of user-generated content, specifically user-generated video, given its widespread appeal. The volume of such content, as provided by popular services such as YouTube, however, makes it challenging for researchers to study these dynamics. In this paper we make several contributions that address this challenge.

Our first contribution concerns the use of sampling. Sampling is necessary given the huge volume of content available from popular services, but sampling may yield datasets biased towards content with elevated short-term and/or long-term popularity. We find that sampling from recently-uploaded videos, as provided by the YouTube API, appears to yield a dataset that is seemingly unbiased, unlike sampling based on keyword searches.

We next show that there is substantial churn in the relative popularities of videos, particularly young videos, and that the current popularity of a video is not a reliable predictor of its future popularity. This finding motivates models that attempt to capture the popularity dynamics of collections of videos, rather than attempting to predict the popularity evolution of individual videos. To this end, we develop a novel three-phase characterization of the popularity evolution of a dataset of recently-uploaded YouTube videos. This characterization provides a basis for a model that, using a small number of distributions as input, is able to generate synthetic data matching empirically observed characteristics with respect to key metrics concerning popularity and its evolution, such as hot set churn statistics, and the evolution of the viewing rate and

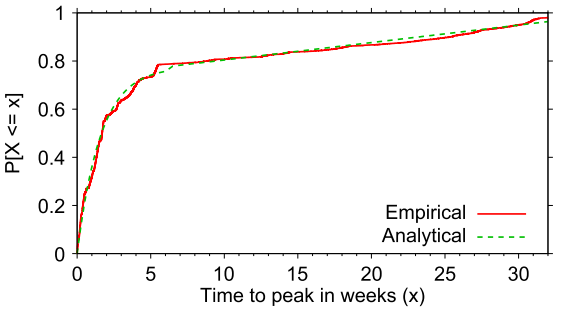


Fig. 18. Time-to-peak distribution of videos.

total views distributions over time. Future work involves large-scale tests of the model, study of the model's applicability for content from other user-generated content services, and study of the factors that influence the key evolution properties observed in our three-phase characterization.

Acknowledgements

The authors are grateful to Martin Arlitt, Guillaume Jourjon, Thierry Rakotoarivelo, Vinay Ribeiro, Aaditeshwar Seth, and the Performance 2011 reviewers for their constructive suggestions, which helped to improve the clarity of our original paper. This work was supported by the Commonwealth of Australia under the Australia–India Strategic Research Fund, the Natural Sciences and Engineering Research Council (NSERC) of Canada, and CENIIT at Linköping University.

Appendix A. Sampling granularity

The recently-uploaded dataset was obtained by sampling the video popularity at a weekly time granularity. By taking the difference of the total view counts between consecutive snapshots, we can measure the weekly viewing rate (i.e., the number of views in a week). For simplicity, and for the purpose of our analysis, we say that a video's peak viewing rate occurs at the midpoint between the two snapshots, between which the highest weekly viewing rate was observed. (As we sample videos of different ages at the time that we first start tracking them, we can still have videos peaking at an arbitrary age less than the total measurement period.)

To obtain a weekly viewing rate associated with the first snapshot (at which the videos may be of any age between 0 and 7 days), we inflate the view count at the first snapshot using a fraction of the added views during the following week to account for the missing days needed to get a weekly view count. Note that alternative ways of calculating a weekly viewing rate, such as dividing the views at the first measurement point by the time since upload, may result in extremely large viewing rates, if the time since upload is small, for example. Finally, in the case that the initial viewing rate is higher than for any other measurement point, we say that the video peaked at the halfway point between its upload time and the initial measurement point (as the rate in this interval, in such a case, is higher than the average rate during following weeks).

Appendix B. Model parameterization

As discussed in Section 4, a large fraction of the videos, approximately three-quarter of our sample, peak within the first six weeks since their upload. The remaining peak at times uniformly distributed between week six and the end of our measurement period. To estimate the rate parameter λ of the exponential part, we use the *Maximum Likelihood Estimation* (MLE) method. For the recently-uploaded dataset, we determine $\lambda = 0.598$. Time-to-peak values greater than six weeks are then drawn from a uniform distribution $U(6, d)$, where d is the duration of the measurement period. Fig. 18 shows the cumulative distribution function (CDF) of the empirical time-to-peak and the analytical fitting.

We parametrize the weekly views of videos belonging to the before-, at-, and after-peak phases. As our model assumes week-invariant distributions for the three phases, we only consider the aggregated before-peak, at-peak, and after-peak weekly views. The body and tail are modelled separately, with the tail assumed to consist of all videos with weekly views greater or equal to a threshold x_{thresh} views, selected such that the tail of each phase contains the 10% videos with the largest weekly view counts.

The distribution of weekly views within each phase is heavy-tailed (cf. Fig. 10). Using the approach developed by Clauset et al. [21], we investigated whether the tail of each phase could be modelled using a power-law distribution or a lognormal distribution.

The MLE method is used to estimate the respective distribution parameters. The power law scaling parameter α (for the continuous case) can be estimated using:

$$\alpha = 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{thresh}}\right]^{-1},$$

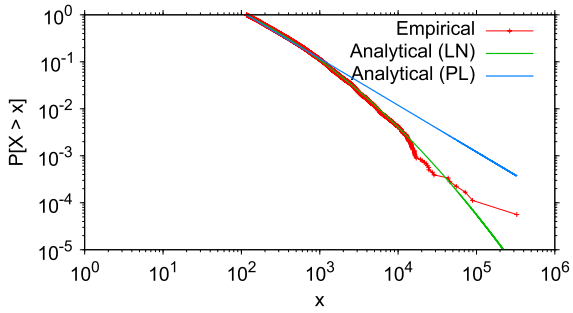


Fig. 19. Power law and lognormal fits for the before-peak phase.

Table 2
Power law and lognormal fits for the tails of the distributions.

Phase	Parameters			Power law fits	Lognormal fits		$R = \frac{L_p}{L_{ln}}$
	n_{tail}	x_{thresh}	x_{max}	α	μ	σ	
Before-peak	16,829	119	89,090	1.996	2.000	2.135	−186.947
At-peak	2986	297	476,100	1.950	−3.826	3.477	−8.164
After-peak	83,148	30	94,930	1.895	−0.356	2.533	−762.172

where n is the number of unique view count observations that fall into the tail distribution. To estimate the parameters μ and σ of the lognormal model, direct maximization of tail-conditional log-likelihood was applied [21]. Note that to model the empirical data using a lognormal distribution above the specified lower threshold x_{thresh}, we use the *tail-method*, where we consider that the right tail exhibits the same shape as the right tail of a lognormal distribution, without essentially having an equal probability of being in the tail.

Table 2 presents the parameter estimation results for the tail of the before-peak, at-peak, and after-peak distributions. Each group has n_{tail} observations $x \in [x_{thresh}, x_{max}]$, α is the scaling parameter of the power law model, and μ and σ are the parameters for the lognormal model.

In order to compare the power-law and lognormal models, we apply the *Log Likelihood Ratio* (LLR) test to determine which distribution best fit the empirical data [21]. This test computes the ratio of the logarithm of the likelihood of our empirical data in the two candidate distributions, and find which distribution best fits the data depending on the sign of LLR. In our case, we calculate the log-likelihood ratio, $R = \log(\frac{L_{PL}}{L_{LN}})$ where, L_{PL} is the likelihood of the power law distribution, and L_{LN} is the likelihood of the lognormal distribution. The values of R in Table 2 suggest that the lognormal hypothesis is more suitable to model the distribution of weekly views for each of the three phases. To verify that the sign of R can be reliably used to make a quantitative judgement about which model is a better fit, we computed the standard deviation of R using the *Vuong* method [21]. Fig. 19 shows the CCDF of the weekly views in the tail of the before-peak distribution, and the analytical lognormal (LN) and power law (PL) fittings.

To determine a distribution that best models the body of the weekly views distribution for each phase, we tried several probability distributions and found the beta distribution provides the best approximation to the empirical data. Since there is no closed-form of the maximum likelihood estimates for the parameters of the beta distribution, we estimate the shape parameters α and β, over an interval $[x_{min}, x_{thresh}]$, using the method-of-moments, where:

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

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

with

$$\tilde{x} = \frac{E[x] - x_{min}}{x_{thresh} - x_{min}}$$

and

$$v = \frac{V[x]}{(x_{thresh} - x_{min})^2}.$$

Here, the $E[x]$ is the sample mean and $V[x]$ is the sample variance. The estimated α and β parameters results are shown in Table 3. Each group has n_{body} observations, $x \in [x_{min}, x_{thresh}]$, where x_{min} is the smallest observation in the dataset and x_{thresh} is equal to the threshold separating the body from the tail.

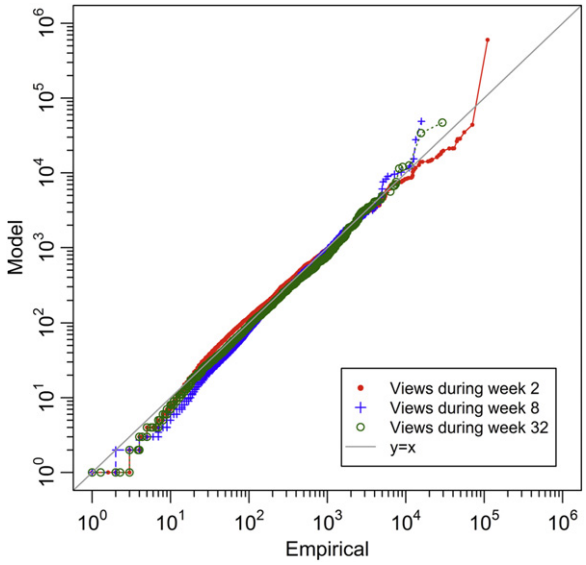


Fig. 20. Q–Q plot for the views during a week from the model and the recently-uploaded dataset.

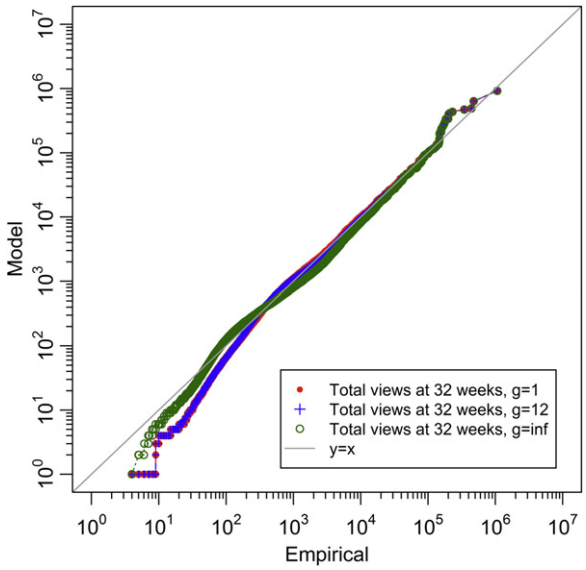


Fig. 21. Q–Q plot for the total views from the model and the recently-uploaded dataset.

Table 3
Beta fits for the body of the distributions.

Phase	Parameters			Beta fits	
	n_{body}	x_{min}	x_{thresh}	α	β
Before-peak	151,051	0	119	0.191	1.330
At-peak	26,805	4	297	0.543	2.259
After-peak	732,075	0	30	0.077	0.968

Appendix C. Model validation

To evaluate the goodness of fit of the synthetic data obtained from our models, in addition to the graphical illustrations presented in Sections 6 and 7, we present here quantile–quantile (Q–Q) plots. Fig. 20 shows the Q–Q plot for the views during a week. Recall that both the basic model and the extended model generate identical views during a week, and thus this plot is representative of both models. Fig. 21 shows the Q–Q plots for the total views by week 32, for three different values of g, including $g = 1$ (basic model), $g = 12$ (extended model), and $g = \infty$ (extended model).

In general, the Q–Q plots show that the models are able to generate synthetic data matching the empirical views distributions. A good match is observed for the body and tail of the distributions. We observe some biases in the head of the distributions; however, these are for the less popular videos and we did not focus on accurately modelling these videos.

References

[1] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, S. Moon, I Tube, You Tube, Everybody Tubes: analyzing the world's largest user generated content video system, in: Proc. ACM Internet Measurement Conference, IMC, San Diego, CA, October 2007, pp. 1–14.
[2] D. Applegate, A. Archer, V. Gopalakrishnan, S. Lee, K.K. Ramakrishnan, Optimal content placement for a large-scale VoD system, in: Proc. ACM International Conference on emerging Networking EXperiments and Technologies, CoNEXT, Philadelphia, PA, December 2010, pp. 4:1–4:12.
[3] D. Beaver, S. Kumar, H.C. Li, J. Sobel, P. Vajgel, Finding a needle in haystack: Facebook's photo storage, in: Proc. USENIX Symposium on Operating Systems Design and Implementation, OSDI, Vancouver, BC, Canada, October 2010, pp. 47–60.
[4] R. Crane, D. Sornette, Robust dynamic classes revealed by measuring the response function of a social system, Proc. Natl. Acad. Sci. 105 (41) (2008) 15649–15653.
[5] M.J. Salganik, P.S. Dodds, D. Watts, Experimental study of inequality and unpredictability in an artificial cultural market, Science 311 (5762) (2006) 854–856.
[6] USA Today, YouTube Serves up 100 million Videos a Day Online, July 2006.
[7] P. Gill, M. Arlitt, Z. Li, A. Mahanti, YouTube traffic characterization: a view from the edge, in: Proc. ACM Internet Measurement Conference, IMC, San Diego, CA, October 2007, pp. 15–28.
[8] X. Cheng, C. Dale, J. Lui, Statistics and social network of YouTube videos, in: Proc. International Workshop on Quality of Service, IWQoS, Enskede, The Netherlands, June 2008, pp. 229–238.
[9] S. Mitra, M. Agrawal, A. Yadav, N. Carlsson, D. Eager, A. Mahanti, Characterizing web-based video sharing workloads, ACM Trans. Web 5 (2) (2011) 8:1–8:27.
[10] M. Zink, K. Suh, J. Kurose, Watch global, cache local: YouTube network traffic at a campus network — measurements and implications, in: Proc. SPIE Multimedia Computing and Networking Conference, MMCN, San Jose, CA, January 2008, pp. 681805-1–13.
[11] M. Halvey, M. Keane, Exploring social dynamics in online media sharing, in: Proc. International Conference on World Wide Web, WWW, Banff, AB, Canada, May 2007, pp. 1273–1274.
[12] A. Mislove, M. Marcon, K.P. Gummandi, P. Druschel, B. Bhattacharjee, Measurement and analysis of online social networks, in: Proc. ACM Internet Measurement Conference, IMC, San Diego, CA, October 2007, pp. 29–42.
[13] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, K. Ross, Video interactions in online video social networks, ACM Trans. Multimedia Comput. Commun. Appl. 5 (4) (2009) 30:1–30:25.
[14] T. Broxton, Y. Interian, J. Vaver, M. Wattenhofer, Catching a viral video, in: ICDM Workshops, Sydney, Australia, December 2010, pp. 296–304.
[15] F. Figueiredo, F. Benevenuto, J.M. Almeida, The tube over time: characterizing popularity growth of YouTube videos, in: Proc. ACM International Conference on Web Search and Data Mining, WSDM, Hong Kong, China, February 2011, pp. 745–754.
[16] J. Lee, S. Moon, K. Salamatian, An approach to model and predict the popularity of online contents with explanatory factors, in: Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence, WI, London, ON, Canada, August 2010, pp. 623–630.
[17] G. Szabo, B.A. Huberman, Predicting the popularity of online content, Commun. ACM 53 (8) (2010) 80–88.
[18] P. Gill, M. Arlitt, Z. Li, A. Mahanti, Characterizing YouTube user sessions, in: Proc. SPIE Multimedia Computing and Networking Conference, MMCN, San Jose, CA, January 2008, pp. 681806-1–8.
[19] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, A. Vespignani, Characterizing and modeling the dynamics of online popularity, Phys. Rev. Lett. 105 (15) (2010) 158701.
[20] A. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[21] A. Clauset, C. Shalizi, M. Newman, Power-law distributions in empirical data, SIAM Rev. 51 (4) (2009) 661–703.

Youmna Borghol is a NICTA graduate researcher and a doctoral candidate in the Department of Electrical Engineering and Telecommunications at the University of New South Wales (UNSW). She has a B.Sc. in Computer Engineering from the University of Balamand, an M.Sc. in Telecommunications from UNSW, and an M.Sc. in Engineering Management from the University of Technology, Sydney. Her research interests include peer-to-peer networking and distributed algorithms, models for the statistical analysis of user-generated content, and social networks.

Siddharth Mitra has a B.Tech. in Information Technology from Vellore Institute of Technology and an M.Tech in Computer Science from Indian Institute of Technology, Delhi. He is currently the Technology Lead at Teaspiller LLC, and the founder of a Design and Development Studio, Cloudshuffle. He has previously worked at Honeywell. His general research interests are social network analysis, recommendation systems, and natural language processing. Current research encompasses study and characterization of the online social networks, workload study of online video sharing sites, and content delivery systems.

Sebastien Ardon obtained his Ph.D. from LIP6, at University Pierre et Marie Curie in Paris. He is now a researcher at NICTA, Australia, with research interests in content distribution, and network performance analysis.

Niklas Carlsson is an Assistant Professor at Linköping University, Sweden. He received his M.Sc. Degree in Engineering Physics from Umeå University, Umeå, Sweden, and his Ph.D. in Computer Science from the University of Saskatchewan, Canada. He has also been working as a Postdoctoral Fellow at the University of Saskatchewan, Canada, and as a Research Associate at the University of Calgary, Canada. His research interests are in the areas of design, modelling, characterization, and performance evaluation of distributed systems and networks.

Derek Eager received the B.Sc. Degree in Computer Science from the University of Regina, Regina, SK, Canada, in 1979, and the M.Sc. and Ph.D. Degrees in Computer Science from the University of Toronto, Toronto, ON, Canada, in 1981 and 1984, respectively. He is currently a Professor in the Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada. His research interests are in the areas of performance evaluation, streaming media and bulk data content distribution, and distributed systems.

Anirban Mahanti is a Senior Researcher at NICTA. He received the B.E. Degree in Computer Science and Engineering from the Birla Institute of Technology (at Mesra), India, and the M.Sc. and Ph.D. Degrees in Computer Science from the University of Saskatchewan, Saskatoon, SK, Canada. His research interests are in the areas of network measurements, network protocols, performance evaluation, and distributed systems.