Revisiting Popularity Characterization and Modeling of User-generated Videos

M. Aminul Islam Derek Eager
University of Saskatchewan
Saskatoon, SK, Canada

Niklas Carlsson
Linköping University
Linköping, Sweden

Anirban Mahanti
NICTA
Alexandria, NSW, Australia

***Abstract*—This paper presents new results on characterization and modeling of user-generated video popularity evolution, based on a recent complementary data collection for videos that were previously the subject of an eight month data collection campaign during 2008/09. In particular, during 2011, we collected two contiguous months of weekly view counts for videos in two separate 2008/09 datasets, namely the "recently-uploaded" and the "keyword-search" datasets. These datasets contain statistics for videos that were uploaded within 7 days of the start of data collection in 2008 and videos that were discovered using a keyword search algorithm in 2008, respectively. Our analysis shows that the average weekly view count for the recently-uploaded videos had not decreased by the time of the second measurement period, in comparison to the middle and later portions of the first measurement period. The new data is used to evaluate the accuracy of a previously proposed model for synthetic view count generation for time periods that are substantially longer than previously considered. We find that the model yielded distributions of total (lifetime) video view counts that match the empirical distributions; however, significant differences between the model and empirical data were observed with respect to other metrics. These differences appear to arise because of particular popularity characteristics that change over time rather than being week-invariant as assumed in the model.**

I. Introduction

The Internet hosts a vast and rapidly-increasing quantity of user-generated content. User-generated content is a central feature of numerous online media sharing and social networking applications, and is a major contributor to Internet traffic volumes. The importance of user-generated content has motivated considerable research on its characteristics. Of particular interest has been popularity characteristics, including popularity distributions within specific catalogs of content items [5], [6], [8], content popularity evolution over time [2], [10], and popularity prediction [3], [12].

Studies of content popularity evolution have mostly considered only short time periods. The longest duration measurement study of which we are aware is that by Borghol *et al.* [2], in which view count statistics for two sets of YouTube videos were tracked for eight months. Prior work has not considered the question of how the popularity of user-generated content items evolves over longer, multi-year time periods.

In this paper, we address this question using new weekly view count statistics that we collected for the videos of Borghol *et al.*, over a two month measurement period from October 25th to December 26th 2011. These new measurements allow us to observe how the popularity of each video, as measured by number of weekly views, changed over the intervening time span from the first measurement period in 2008/09, to our second measurement period over two years later.

One of the sets of videos of Borghol *et al.* was obtained using a YouTube API call that returns details on videos that have been recently uploaded (within the past seven days), while the other was obtained using keyword search. Owing to space limitations, we omit our analyses of the new data we collected for the videos in the keyword-search dataset, as well as our examination of the popularity characteristics of the videos that were in the original datasets of Borghol *et al.*, but that had subsequently been removed from YouTube; see [9].

Although the keyword-search dataset is biased towards more popular videos (since keyword-search algorithms try to return the videos of most likely interest to the user), the recently-uploaded dataset is believed to contain a random sampling of such videos [2]. Our new measurements for these videos therefore allow us unbiased observations of how video popularity evolves from the first few months of video lifetime, to when a video has attained a relatively old, multi-year age. We are also able to use our new view count data to study the accuracy of a model for synthetic view count generation that was proposed by Borghol *et al.*, for time periods substantially longer than the eight month period considered in that work. Our main findings include the following:

- In some respects, popularity is surprisingly resilient. The average weekly view count for the recently-uploaded videos had not decreased by the time of the second measurement period, in comparison to the middle and later portions of the first measurement period. We also find that a significant number of videos attained their highest observed weekly view count (considering only the weeks in the two measurement periods) during the second measurement period, more than two years after video upload.
- The synthetic view count generation model of Borghol *et al.* can yield distributions of total (lifetime) video view counts matching the empirical distributions, even when evaluated for multi-year time periods. The model does, however, assume that particular popularity characteristics are time-invariant, an assumption that can break down over long time periods, leading to substantial differences in other metrics.

To appear in *Proc. 21st IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS 2013)*, San Francisco, CA, Aug. 2013. The final/official version will appear in the conference proceedings and the IEEE Xplore Digital Library.

The remainder of the paper is organized as follows. Section II discusses related work. The recently-uploaded and keyword-search datasets, and our new measurements for these videos, are described in Section III. Section IV presents our analyses for the recently-uploaded videos. Section V uses our new view count data to further evaluate the accuracy of the Borghol *et al.* model. Conclusions are presented in Section VI.

II. Related Work

A number of studies have focussed on characterizing the popularity of user-generated videos [5]–[8], [10], [14]. Cha *et al.* [5] analyzed popularity characteristics of user-generated videos using traces collected from the YouTube and Daum services. Mitra *et al.* [10] analyzed popularity characteristics for four video sharing services, Dailymotion, Yahoo, Veoh and Metacafe, based on both total views popularity and viewing rate popularity. Gill *et al.* [8] and Zink *et al.* [14] studied the popularity of YouTube content within large edge networks by collecting network traces from the interconnection between their respective university campus access network and the Internet. Other studies have examined such issues as user interaction, and the impacts of search mechanisms and social networks on video views [1], [4], [7].

Some recent works have developed models for popularity evolution of online content [2], [11]–[13]. Borghol *et al.* [2] proposed and validated a model that can generate synthetic video view counts. The model assumes that video view counts in each of the three video lifetime phases (before-peak, at-peak, and after-peak) can be modelled using week-invariant distributions. The authors also extended their model to introduce additional churn in the synthetic view counts. Szabo *et al.* [12] and Tang *et al.* [13] proposed models to predict long term popularity characteristics using logarithmic transformation and k-transformation with a Zipf's law respectively. Ratkiewicz *et al.* [11] proposed a model that combines a classical "rich-get-richer" model with random popularity shifts.

III. Data Collection

We collected complementary view count data for two datasets that were previously the subject of an eight month long data collection campaign during 2008/09. The original data collection by Borghol *et al.* collected view count data from 27 July 2008 to 29 March 2009. The two datasets are referred to as "recently-uploaded" and "keyword-search", which correspond to a sample of recently uploaded videos (within a week of the time of sampling) and a sample of videos discovered through use of keywords over a one week period. A one week sampling frequency was chosen so that each video's data collection was always on the same day of the week at approximately the same time, so as to avoid potential day-of-the-week effects [2]. Meta-data for each video was collected once each week throughout those eight months. This procedure resulted in 35 "snapshots" for each video's meta-data, including the "seed" snapshot obtained during the first week of data collection. Each snapshot for a video contains the total number of views received by that video. Thus from the total view count at each snapshot i ($1 < i \leq 35$) one can determine the number of views each video received during the week between snapshot i and $i-1$. We will refer to this as the first period of data collection.

Our new, second period of data collection was of view count data for the same videos, excepting those videos no longer accessible from YouTube. Videos may become unavailable for reasons such as violation of YouTube terms and conditions, videos made private by uploader, or video deletion by the uploader of the content. The new data collection was done from 25th October to 26th December 2011. We obtained the total view count for each video once per week, on the same day of the week and at a similar time. This procedure resulted in 9 new snapshots for each video's view count.

TABLE I. Summary of data from 1st period of data collection (27 July 2008 to 29 March 2009; 35 snapshots)

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

TABLE II. Summary of data from 2nd period of data collection (25 October to 26 December 2011; 9 snapshots)

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

Tables I and II summarize the datasets. Throughout the second measurement period, 20,000 recently-uploaded and 627,002 keyword-search videos were available, representing 67.13% and 55.23% of the total numbers of videos in these datasets, respectively. We speculate that the higher proportion of removed keyword-search videos reflects the popularity bias in this dataset, and the greater likelihood for popular videos to be reported for copyright violations. During our new data collection, the 20,000 recently-uploaded videos received almost 5 million additional views, and the 627,002 available keyword-search videos received about 1.8 billion more views.

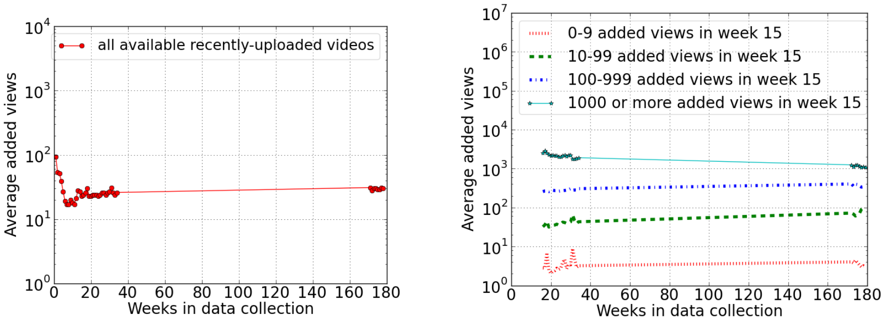
IV. Popularity Characteristics for the Recently-uploaded Dataset

Some basic popularity characteristics for the recently-uploaded dataset are already observed by Borghol *et al.* [2]. However, those analyses are done on the data collected in the first measurement period only. We study those characteristics again to observe whether or not they remain similar after a long period has passed since the videos were uploaded.

A. Average View Count for Each Week

Figure 1(a) shows the average added views in different weeks for the recently-uploaded videos. Since data points exist only for the weeks within the first and second measurement periods, this plot (as in similar later figures) displays a straight line across the weeks between these two periods. Note that the average viewing rate for these videos has not decreased by the time of the second measurement period. In fact, it slightly increased. This is somewhat surprising; one might expect that the average viewing rate of these videos would decrease as the videos age, considering the vast amount of new content uploaded over this time period. A possible contributing factor is growth in the YouTube user population.

To gain a better understanding of the evolution of the average viewing rate, videos are separated into bins according



(a) Average added views in each week

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

Fig. 1. Average view count for the recently-uploaded videos

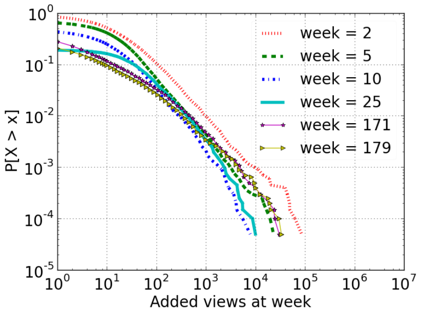


Fig. 2. Distribution of added views for the recently-uploaded videos at different ages

to their view count during week 15. We chose week 15 to avoid the higher popularity churn at earlier video ages. The results shown in Figure 1(b) are again somewhat surprising. It appears that the average viewing rate of videos with intermediate popularity increases, evidently resulting in the increased overall average viewing rate. The average viewing rate of the highly popular videos, however, decreases.

B. View Count Distribution

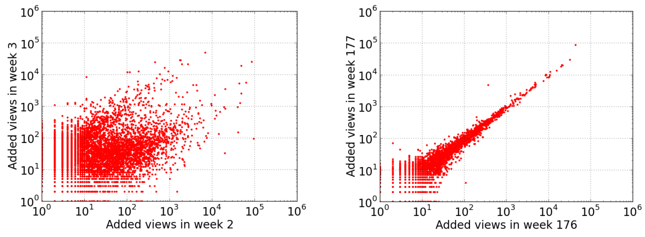
Figure 2 shows the complementary cumulative distribution function (CCDF) of the added views in various weeks for the recently-uploaded videos. As in subsequent figures, representative results are shown for selected weeks only. Note that during the first measurement period, videos tend to receive more views at an early age than when they are older. The heavier right tail of the curve shows an order of magnitude more new views in week 2 than in weeks 10 and 25. However, when the video age is very high (weeks 171, 179) the heavier right tails of the curves again show substantially more new views for highly popular videos than in weeks 10 and 25.

C. Popularity Dynamics and Churn

Figure 3 shows scatter plots of added views at different pairs of consecutive weeks. The points are spread out more for early video ages than for the later ages. For example, highly unpopular videos during week 2 could be highly popular in week 3, and vice versa. As the videos age, churn decreases; i.e., view counts are highly variable from one week to the next early in a video's lifetime, but become more stable with age. Interestingly, Figure 3(a) shows two distinct clusterings of points. This may reflect differences in popularity phase (i.e., before-peak vs. after-peak, as described in Section V).

D. Time-to-peak Distribution

Widely differing rates at which videos achieve their peak popularity is a major cause of popularity churn. We calculated the time-to-peak for each video by comparing weekly view counts and keeping track of the week with the maximum weekly view count. Ties are broken by randomly picking one of these weeks. After determining a week in which the video attains its maximum weekly view count, the time-to-peak value is calculated using the video age, in the same way as described by Borghol *et al.* [2]. We find that most videos (almost 80%) reach their peak popularity (over the weeks in the first and second measurement period) within their first six weeks since upload. The time to achieve peak popularity for the remaining 20% of the videos is approximately uniformly distributed throughout the rest of the observed weeks. Note that some videos received their peak weekly number of views long after upload, during the second measurement period (from week 171 to 179). It should be emphasized that only the weekly view counts in the measurement periods are considered; any of the videos could have its actual peak weekly view count during a week outside of these measurement periods.



(a) Week 2 vs 3

(b) Week 176 vs 177

Fig. 3. Scatter plot of added views for the recently-uploaded videos in week i vs week i+1

V. Modelling Popularity Evolution

A. Background on the Basic Model of Borghol et al.

Based on observations from the recently-uploaded dataset, Borghol *et al.* [2] developed a model that generates synthetic weekly view counts with characteristics similar to those observed for newly-uploaded videos as they age. The model generates the weekly view counts for each video within a collection of synthetic newly-uploaded videos using a three-phase characterization of popularity evolution, in which each video is either "before-peak" (i.e., has not yet attained its highest weekly view count), "at-peak", or "after-peak". The number of synthetic videos whose popularity peaks in any particular week after video upload is determined using a time-to-peak distribution parameterized from the empirical data.

The model uses three view count distributions, one for each of the "before-peak", "at-peak", and "after-peak" phases. For each modelled week after upload, view counts sampled from the before-peak and at-peak distributions will be assigned to videos that were in their before-peak phase during the previous week, and views sampled from the after-peak distribution will be assigned to videos that were in their at-peak or after-peak phase during the previous week. The view counts sampled from the before-peak, at-peak, and after-peak distributions are assigned to the synthetic videos in the respective phases so as to preserve the relative popularities of videos in the same category. Borghol *et al.* fit analytic distributions to the empirical before-peak, at-peak, and after-peak view count distributions, making the approximation that these distributions do not depend on which week is considered.

B. Three-phase Characteristics

Figure 4 presents the CCDF of weekly views in each phase, for the videos in the recently-uploaded dataset. The distribution of weekly views in each phase is heavy-tailed even in week 172, which is long after the videos were uploaded. It is also observed that the view count distribution for before-peak videos in week 172 (Figure 4(a)) as well as for at-peak videos (Figure 4(b)) is quite distinct from that for the other weeks. In contrast, except for unimportant differences for videos with lower view counts, the after-peak distribution (Figure 4(c)) appears week-invariant. A topic of future work would be to modify the Borghol *et al.* model to make the before-peak and at-peak analytic view count distributions week-dependent.

Assuming week invariance as in the Borghol *et al.* model, Figure 5 presents the CCDF of weekly views in each phase when data is aggregated across all weeks. Although the at-peak and before-peak distributions show significant differences in the second measurement period, relatively few videos are in these phases in the second period, and so the impact on the distributions in Figure 5 is quite small. The same analytic distribution fits as used by Borghol *et al.* are used in this work.

C. Model Evaluation

In our evaluation of model accuracy, we generated a set of synthetic weekly views for 20,000 synthetic videos (the number of videos in the recently-uploaded dataset) and 179 weeks (the time span from the beginning of data collection for the recently-uploaded dataset until the end of the second measurement period). Note that Borghol *et al.* used their model to generate view counts for 29,791 videos (the same number of videos as in the first measurement period for the recently-uploaded dataset) and 34 weeks.

Figure 6(a) presents the CCDF of total views acquired by weeks 2, 32 and 172 both for the videos in the recently-uploaded dataset and for the synthetic videos. The figure shows excellent matches between the distributions for the synthetic and empirical videos. Note that the model is not parameterized using empirical total view count statistics, but instead the synthetic total view counts result from the view generation algorithm and modelling parameters derived from the model's three-phase characterization of video popularity evolution.

Figure 6(b) shows the CCDF of weekly views during weeks 2, 32, and 172 for both the videos in the recently-uploaded dataset and the synthetic videos. Although there is a good match between the general forms of the corresponding distributions for the synthetic and empirical videos, there are some significant differences that are apparent in the figure. The model's approximation that the distributions used in its three-phase characterization are week-invariant may be a cause of these differences. The substantial growth in the YouTube user population between the first and second measurement periods (not accounted for in the model) may also be a factor.

D. Extended Model

Borghol *et al.* found that their basic model could not accurately match the empirical data with respect to popularity churn metrics, which led them to develop an extended model. Borghol *et al.* considered specifically hot set overlap metrics, which we also consider here. For this purpose, the most popular 10% and 1% of the videos in a week (two differently-sized "hot sets" for that week, where the notion of "most popular" is based on the new views acquired in that week only) are compared to the correspondingly-sized hot sets for some other week, and the overlap in videos is determined. The basic model achieves popularity churn only by moving videos among the three phases. The extended model introduces additional churn in video popularity by repeatedly exchanging the added view counts of selected videos. For each exchange, a random week i and two videos u and v are selected, both of which are in the same phase. Furthermore, the exchange of the added view counts is only carried out if the two currently assigned added view counts in that week i are within each other's respective exchange window:

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

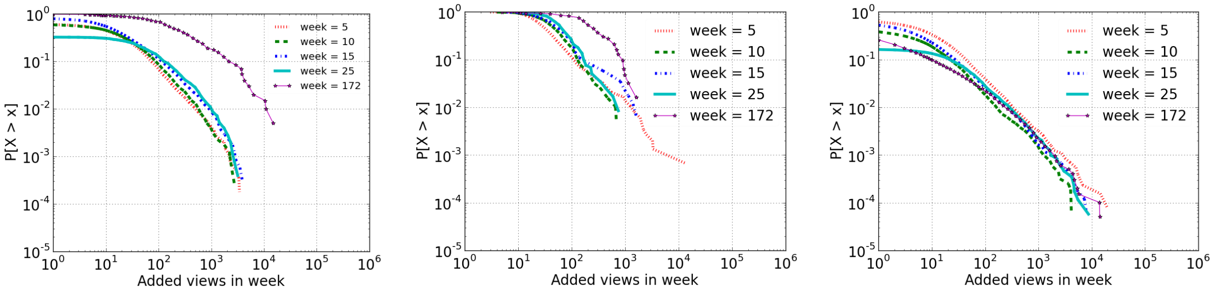
where x_i^v is the added views assigned to video v for week i by the basic model and x_{max}^v is the maximum weekly views assigned to the video v during its lifetime by the basic model. Finally, g ($1 \leq g \leq \infty$) is a model parameter that determines the amount of possible churn.

Note that the weekly views generated by the extended model for $g = 1$ are the same as the weekly views generated by the basic model. For the results in this paper, a random week and a potentially eligible pair of videos are picked 5 million times for each of several choices of the model parameter g.

Figure 7 compares the hot set overlap for both the videos in the recently-uploaded dataset and the synthetic videos. For the hot set overlap between adjacent weeks (Figure 7(a) and Figure 7(c)), a better match is observed with the extended model for a relatively high value of g, for both 10% and 1% hot sets, for young video ages, while as the video age increases (i.e, for later weeks) the best value of g decreases. For the time period and values of g considered, the best value of g changes from $g = 16$ for the initial weeks to $g = 2$ for the weeks corresponding to the second measurement period. The hot set overlap with week 2 (Figure 7(b) and Figure 7(d)) shows approximately similar churn for different g values, and similar behaviour as the empirical results.

VI. Conclusions

Prior work concerning the popularity evolution of online content, and user-generated videos in particular, has considered only relatively short time periods. In this paper we address this limitation, using new view count statistics that we collected for two previous YouTube datasets. Our new data allows



(a) Before-peak phase

(b) At-peak phase

(c) After-peak phase

Fig. 4. Distribution of weekly views for the empirical recently-uploaded videos in the before-peak, at-peak and after-peak phases

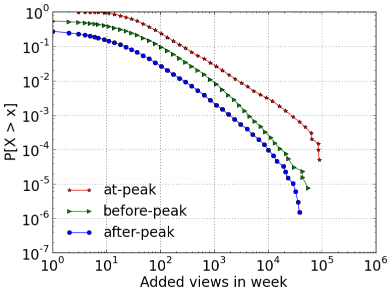
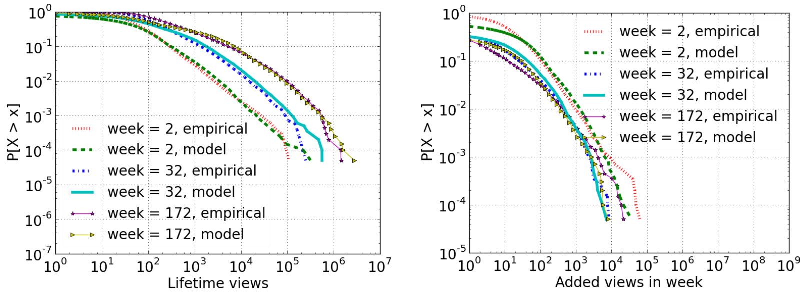


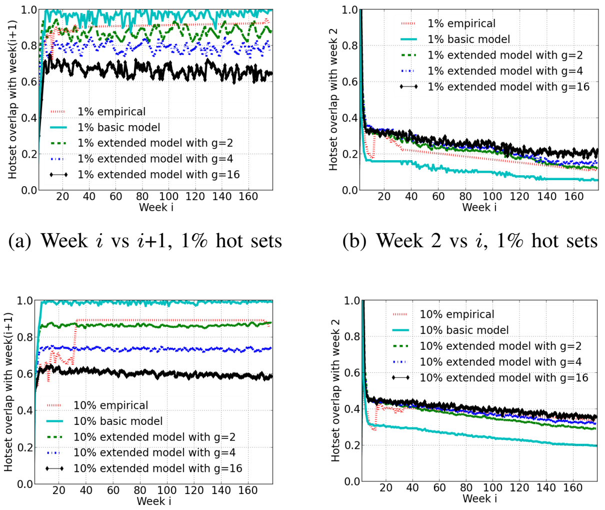
Fig. 5. Distribution of weekly views in each phase, when data is aggregated across all weeks



(a) Total views

(b) Added views

Fig. 6. Distribution of views for the empirical and synthetic recently-uploaded videos



(a) Week i vs i+1, 1% hot sets

(b) Week 2 vs i, 1% hot sets

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

(d) Week 2 vs i, 10% hot sets

Fig. 7. Churn for the empirical and synthetic recently-uploaded videos

observation of how video popularity has evolved over a multi-year time period. We are also able to assess the accuracy of a previously proposed model for synthetic view count generation for a much longer time period than was previously possible.

References

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

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

[3] Y. Borghol, S. G. Ardon, N. Carlsson, D. L. Eager, and A. Mahanti. The untold story of the clones: content-agnostic factors that impact YouTube video popularity. In *Proc. ACM KDD*, Beijing, China, Aug. 2012.

[4] T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer. Catching a viral video. In *Proc. IEEE ICDMW*, Sydney, Australia, Dec. 2010.

[5] 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 Diego, CA, Oct. 2007.

[6] X. Cheng, C. Dale, and J. Liu. Understanding the characteristics of Internet short video sharing: YouTube as a case study. Technical Report arXiv:0707.3670v1 [cs.NI], Cornell University, arXiv e-prints, July 2007.

[7] F. Figueiredo, F. Benevenuto, and J. M. Almeida. The tube over time: characterizing popularity growth of YouTube videos. In *Proc. ACM WSDM*, Hong Kong, China, Feb. 2011.

[8] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube workload characterization: a view from the edge. In *Proc. ACM IMC*, San Diego, CA, Oct. 2007.

[9] M.A. Islam. *Popularity Characterization and Modelling for User-generated Videos*. M.Sc. Thesis, Univ. of Saskatchewan, January 2013.

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

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

[12] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80-88, Aug. 2010.

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

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