

# Popularity Growth Patterns of YouTube Videos: A Category-based Study

Shaiful Alam Chowdhury and Dwight Makaroff

*Department of Computer Science, University of Saskatchewan, Saskatoon, SK, S7N 5C9, CANADA*  
{sbc882, makaroff}@cs.usask.ca

**Keywords:** Workload characterization: multimedia applications: content distribution; time-series clustering

**Abstract:** Understanding the growth pattern of content popularity has become a subject of immense interest to Internet service providers, content makers and on-line advertisers. This understanding is important for the sustainable deployment of content distribution systems. A significant amount of research has been done in analyzing the popularity growth patterns of YouTube videos. Unfortunately, little work has been done that investigates the popularity patterns of YouTube videos based on video object category. In this paper, we perform an in-depth analysis of the popularity pattern of YouTube videos, considering video categories. We find that the time varying popularity of different YouTube categories are different from each other. For some categories, views at early ages can be used to predict future popularity, whereas for some other categories, predicting future popularity is a challenging task and requires more sophisticated techniques (e.g. time-series clustering). The outcomes of these analyses can be instrumental towards designing a reliable workload generator, which can be further used to evaluate different caching policies and distribution mechanism for YouTube and similar sites.

## 1 INTRODUCTION

YouTube and other user generated content (UGC) sites have altered the way people watch video on the Internet. YouTube was the 4<sup>th</sup> most accessed Internet site in 2007 (Cheng et al., 2007), and its use was increasing over time in a power-law manner. Recent studies continue to support two central observations: 1) increasing number of videos and users (Ding et al., 2011; Siersdorfer et al., 2010) and 2) dissatisfying experiences of users in watching YouTube videos (Khemmarat et al., 2011). Other recent studies (Gember et al., 2011; Labovitz et al., 2010; Maier et al., 2010) suggest that YouTube is the most bandwidth intensive service of today's Internet, and it accounts for 20-35% of Internet traffic.

Much research has been done investigating request characteristics from both client (Gill et al., 2007; Zink et al., 2009) and server perspectives (Borghol et al., 2011; Cha et al., 2009; Ding et al., 2011; Figueiredo et al., 2011) in order to enable improved service. However, none of this earlier work considered the types of video objects. This aggregate data may not tell the whole story.

A proper understanding of YouTube's workload will aid in the design of new systems, as well as capacity planning, and network management for similar

types of systems. The methodology we have developed is useful for UGC sites that have a single cache for the region of requests captured.

In this paper, the time-varying global viewing patterns of a sample of YouTube videos from their upload time are analyzed, considering video category.<sup>1</sup> We present the results of one data collection period (5 months of views of videos uploaded in 2 consecutive days); a previous dataset showed similar characteristics and is not evaluated here. Our results show that different categories exhibit different viewing patterns in terms of overall popularity and detailed popularity over time. We confirmed that the number of views of the popular videos follows a Zipf distribution for most categories, whereas views of the unpopular videos follow a heavy tail distribution. We also show that time-series clustering can be successfully used to understand the growth patterns for the categories where early popularity cannot be used to predict future popularity of a video.

These observations contribute to a better understanding of the popularity dynamics of YouTube videos, enabling realistic testing scenarios for developing and evaluating various design parameters for UGC sites. While the request patterns for different

---

<sup>1</sup>as defined by the uploader

categories may vary around the world, our dataset and analysis provide a case study that shows that category differences persist in global access patterns, and therefore will exist in each region. Our analysis enables the development of category-specific workload generators which can be combined to form the input for simulators and prototype systems. While developing and evaluating a comprehensive workload generator remains as future work, we have a strategy for generating synthetic requests on a category basis and present preliminary results which match reasonably well for two categories: News and Music.

The remainder of the paper is organized as follows. Related work is described in Section 2. Section 3 explains the data collection methods of our study. We discuss the characterization of the request patterns in Section 4, and use the information from views over time to develop a workload generator for two categories in Section 5. Section 6 provides conclusions and future work.

## 2 RELATED WORK

Previous request characterization and video popularity analysis has been used to investigate the feasibility of different content delivery streaming techniques, and to design and evaluate caching policies/systems for UGC sites. Our work leverages the best practices in the previous literature to investigate *category* popularity over time.

YouTube video request traffic was captured at the packet level at the University of Calgary over a 4 month period (Gill et al., 2007). They investigated video popularity properties, usage patterns, and transfer behaviours as measured from the client edge of the distribution network. The traces examined contained data from both completed and incomplete requests. Their analysis suggests that appropriate caching decisions not only can improve end user experience, but also reduce network bandwidth usage.

Another study (Zink et al., 2009) observed the traffic of YouTube videos between a university campus and the YouTube server. Approximately 25% of the videos in the trace were requested more than once, leaving a long tail in the distribution. Three different content delivery techniques were analyzed: P2P based distribution, proxy caching and local caching. Proxy-caching outperformed the other techniques, and P2P based distribution sometimes exhibited worse performance than local caching.

These two results can be biased by the measurement locations which appropriately restrict the context of the studies and the solutions that are pro-

posed. For instance, it is claimed that video requests in YouTube follow a Zipf distribution (Gill et al., 2007), which is different from other works that consider global request patterns. For our purposes, global access patterns are essential.

2.5 million YouTube videos were obtained using related video links (Cheng et al., 2007). Access patterns of the popular videos did follow a Zipf-like distribution, in spite of having a heavy-tailed section in the distribution curve. Data collected indicated that the YouTube network is similar to small world networks, and P2P techniques could be successfully applied, contradicting earlier findings (Zink et al., 2009). Their dataset is likely to be biased to popular videos because of the crawling approach, and popularity over time is not investigated in detail.

A recent approach to investigate growth patterns in YouTube video requests was to use Google charts to collect views over time (Figueiredo et al., 2011). They analyzed the time-varying viewing patterns of popular videos, deleted videos and randomly selected videos. Popular videos usually experience a huge number of views on a single peak day or week. Unfortunately, using the Google charts API is not sufficient to have a proper, fine-grained understanding of the dynamics of video popularity as Google charts API always returns 100 data points, regardless of video age.

Recent work was done on nearly 30,000 videos, collected by using the recently uploaded standard feed provided by the YouTube API (Borghol et al., 2011). Their collection procedure claims to have an unbiased dataset; the *Most Recent* standard feed returns video information randomly that are uploaded recently. Most of these videos experienced their peak popularity within fewer than six weeks of their uploading time. Video collection based on keyword search is shown to be biased to popular videos. This observation suggests that in order to accurately characterize the viewing patterns of YouTube videos, the method of data collection is important.

## 3 DATA COLLECTION

No prior work measures the daily views of different categories of YouTube videos from the first day of their uploading time. We modified previous unbiased data collection methods (Borghol et al., 2011) since we speculate that the first week since uploading deserves more investigation, even though this may expose day-of-week effects. Moreover, similar numbers of videos from all the categories are needed for appropriate comparison between different categories. Multiple crawlers were deployed to obtain data used in

our analysis. Since the crawler obtained information from the API, the crawler location is irrelevant.

**(1) Most Recent crawlers.** 15 different crawlers were deployed on March 3<sup>rd</sup>, 2012 (a Saturday), collecting video IDs for 15 different categories,<sup>2</sup> by restricting the *Most Recent* queries to a specific category for each crawler. All crawlers collected video information for 24 hours, ensuring that subsequent video views began on the first day of their lifetimes. The *Most Recent* standard feed provides video information randomly, reducing bias to particular classes of videos. A similar procedure was followed on March 4<sup>th</sup>, 2012. After two days, a total of 71,208 videos’ information was obtained. Depending on the server load, the YouTube API returns at most 100 videos’ information for each request every one or two hours, limiting the dataset size.

**(2) Video view collection crawlers.** Video view collection using two separate crawlers was started from March 4<sup>th</sup>, 2012 and March 5<sup>th</sup>, 2012. This continued for 149 consecutive days (approximately 5 months). The crawlers ensured a 24-hour difference between view collections. Normalization was performed on the first day’s views. Due to network connection failures, some video views on days 20 and 58 of the measurement period were not captured. Fortunately, those days are not that important for most of the videos, as most of the significant events occur at the very early age of a video. After normalization, 147 day’s views are analyzed.

After 149 days, the number of videos in the dataset fell from 71,208 to 47,711 (an average deletion rate of 33%). Manually sampling of the data set revealed that a large percentage of the deleted videos had copyright infringement issues. Table 1 shows the summary of our dataset. Howto, Film, Entertainment and Tech videos experience the highest deletion rates. Analysis of deletion rates is left as future work though deletion rates for all categories decrease over time.

**(3) Uploading rate crawlers.** Another crawler was developed that collected category names of videos provided by YouTube’s *Most Recent* standard feed. The crawler ran for 5 months, starting from February 2<sup>nd</sup>, 2012 and collected approximately 365,000 unique videos’ information. This allows us to estimate the short-term current category-specific uploading rates. While not an accurate representation of the entirety of YouTube, it does give some insight.

<sup>2</sup><http://support.google.com/youtube/bin/answer.py?hl=en&answer=94328>

Table 1: Categories and Number of videos

Category	Number of videos (Day 1)	Number of videos (Day 149)	Deleted videos Pct
Howto	4773	1772	62.87
Film	4654	2346	49.59
Ent.	4991	2528	49.34
Tech	4942	2682	45.73
Games	4711	2966	37.04
People	4310	2730	36.65
Autos	4714	3245	31.16
Comedy	4744	3467	26.91
News	4623	3432	25.76
Travel	4918	3698	24.80
Sports	4812	3733	22.42
Music	4774	3477	21.93
Nonprofit	4624	3691	20.17
Education	4710	3801	19.29
Animals	4908	4143	15.58
Total	71208	47711	33.00

## 4 VIDEO REQUEST ANALYSIS

### 4.1 Time-Varying Category Popularity

Figure 1 shows the cumulative distribution functions (CDF) of time-to-peak for the videos from different categories with at least 100 views; a video with a very small number of views might contribute unfairly to the understanding of the actual growth pattern of a category. One consequence of this restriction is that the number of videos in each category is significantly reduced, down to 42% for News and Sports and 18% for Animals and Travel. We define *time-to-peak* as the day in which a video experienced the most views as in previous work (Borghol et al., 2011). Time to reach peak popularity is not the same for all categories.

News and Sports categories follow a similar distribution and the time to reach peak popularity for these two categories is the shortest. Approximately 85% of News and Sports videos reach peak popularity within the first 4-5 days of their lifetimes. As well, in every category, between 50% and 60% of the videos experience their peak viewing on Day 1. Other categories such as Music, Film, Howto, Tech and Education follow similar patterns and many videos in these categories reach peak popularity much later.

The other categories follow similar distributions, and peak distributions of these categories lie within the previous two groups. The significance of time-to-peak can be enhanced by Figure 2 which depicts the CDF of percent of total views over time for all the videos in a subset of categories. Music and Film

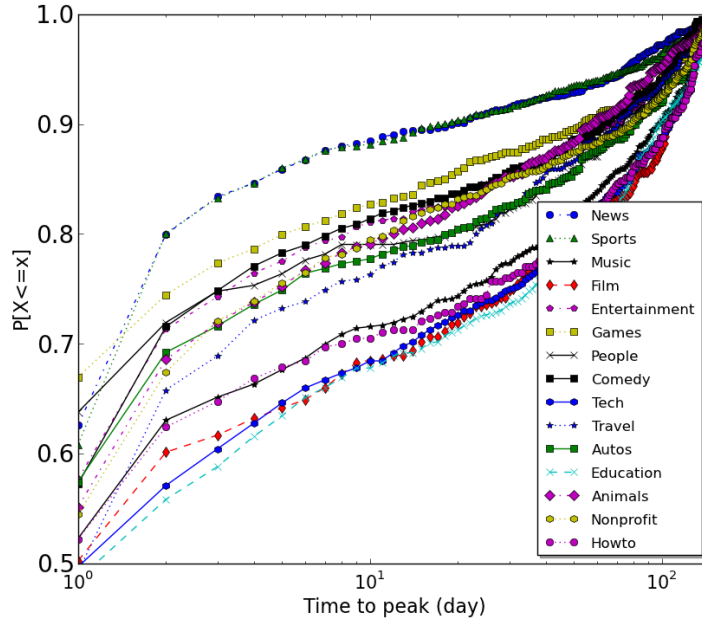


Figure 1: CDF of time-to-peak

videos experience relatively fewer views early in their lifetime. Film videos follow an almost constant viewing rate for the entire measurement period. News and Sports videos, however, experience a significant portion of the total views early.

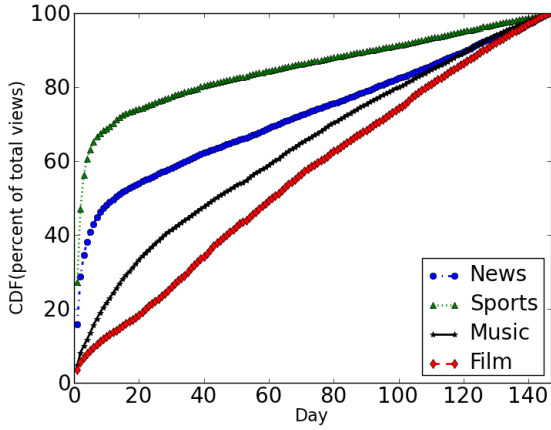


Figure 2: Percent of total views over time

It is important to understand if the peak day differs significantly from other days of a video's lifetime in order to determine if our previous statistic is helpful. Figure 3 shows the complementary cumulative distribution function (CCDF) of the most distant day  $x$  after the peak such that the views on day  $x$  is at least 50%

of the peak views, defined as follows:

$$x = \max(i) : \text{view}(i) \geq 50\% \times \text{view}(\text{peak}) \ \& \ i > \text{peak} \quad (1)$$

where  $\text{view}(i)$  is the views on day  $i$  and  $\text{view}(\text{peak})$  is the number of views on the peak day. Only videos with more than 100 views are considered. Figure 3 shows the peak day as a unique point in the lifetime of videos for faster-growing categories (e.g., News and Sports). These categories experience a popularity burst, and quickly decline to a lower viewing rate.

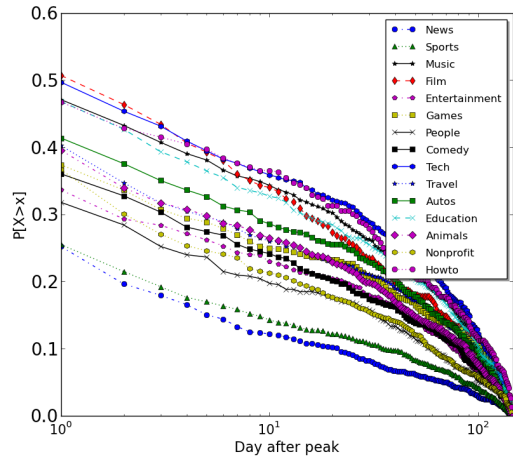


Figure 3: CCDF of time-after-peak

Many Music, Film, Howto, Education and Tech videos that reach peak popularity comparatively lately do not have that drop in their popularity (Figures 1 and 3), so time to reach peak popularity is proportional to the active lifespan of a video. For example, over 75% of the News and Sports videos *never* experience half of their peak days' views after the peak day (Figure 3), but fewer than 50% for Film and Tech videos have this characteristic. The stability of Film and Tech videos suggests that a longer measurement period would increase the difference between these categories and News/Sports.

We are also interested to know if the categories that reach peak popularity faster than others also experience differing numbers of views. Figure 4 depicts the 95<sup>th</sup> percentile of views of all categories over time. We show the 95<sup>th</sup> percentile to remove the potential effect of outliers. This shows which categories have a minimum percentage of popular videos (5%) during the first 100 days of the data collection and the relative popularity of the categories for those popular videos. The last 49 days of the collection period are virtually identical to days 50-99 in terms of this measure.

These graphs illustrate how viewing patterns of different categories change throughout the early part of their lifetimes. Although the most similar dataset collected (Borghol et al., 2011) shows that the views of Music category exceeds all other categories within their 8-month measurement period,<sup>3</sup> our dataset shows that popular News, and Sports videos enjoy higher viewing rates than any other types of videos for the first couple of days since publication. Figure 4 suggests that almost all categories have at least 5% of their videos that experience a high initial viewing rate; the difference is that after these few peak days, views for most of the categories become very low, except Music and to a lesser extent, Film and Tech videos. The results indicate the variations in active life spans of different categories.

Although similar results can be observed from the average views per day (Figure 5), this can be misleading because of the high variance of views. The higher early average views of Sports videos than News videos is due to the most popular video in the entire dataset, which happens to be a single enormously popular Sports video with almost 24 times that of the second most popular Sports video.

## 4.2 Fractions of Popular Videos

The percent of videos with different views of the YouTube categories are shown in Table 2. Only ap-

<sup>3</sup>We collected category names of the videos which had not been deleted by running another crawler.

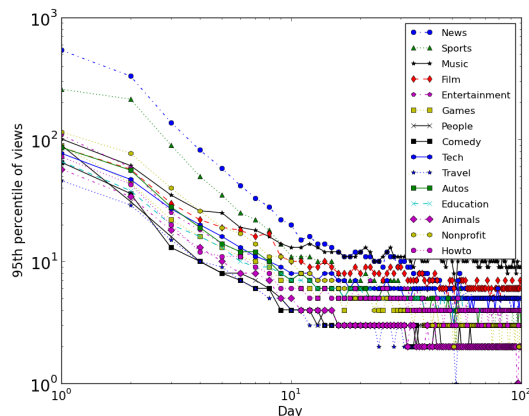


Figure 4: 95<sup>th</sup> percentile of views per day

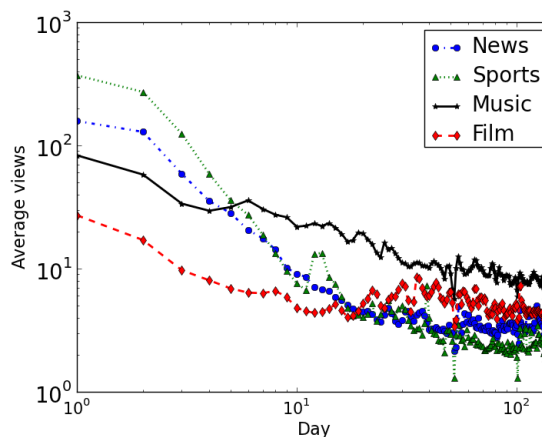


Figure 5: Time varying average added views

proximately 10% of the Music videos enjoy *fewer* than 10 views; this value is over 30% for Howto, People, Autos, Comedy, and Travel. Music, News, Sports, and Film contain most of the popular videos in our dataset (> 1.11% with over 10,000 views). The most unpopular videos are in the Travel category, followed by Comedy and Animals. Only 0.44% of the People videos had more than 10,000 views, in spite of the highest uploading rate (shown later). Although uploaders currently upload more UGC videos, users are still not attracted to UGC videos compared to UCC (user copied content) videos.

## 4.3 Current Uploading Rate

In order to design a request generator for YouTube, it is important to know the category uploading rate. In 2007, Music was in the top position in number of uploaded videos followed by Entertainment, Comedy, Sports and Film (Cheng et al., 2007). Manual

Table 2: Percent of popular videos

Category	≤10 views		11 to 100		101 to 1000		1001 to 10000		10001 to 100000		> 100000	
	Pct	Num	Pct	Num	Pct	Num	Pct	Num	Pct	Num	Pct	Num
Music	10.44	363	48.72	1694	32.87	1143	6.38	222	1.29	45	0.29	10
News	18.85	647	39.57	1358	31.61	1085	8.42	289	1.4	48	0.15	5
Sports	20.79	776	46.0	1717	26.12	975	5.97	223	1.04	39	0.08	3
Tech	22.56	605	47.28	1268	24.61	660	4.85	130	0.63	17	0.07	2
Film	23.06	541	49.53	1162	20.84	489	5.46	128	1.07	25	0.04	1
Entertainment	27.77	702	46.88	1185	20.61	521	3.88	98	0.75	19	0.12	3
Howto	43.79	776	34.59	613	17.04	302	4.01	71	0.45	8	0.11	2
Nonprofit	24.11	890	48.04	1773	23.49	867	3.85	142	0.46	17	0.05	2
Education	24.73	940	48.83	1856	21.7	825	4.34	165	0.37	14	0.03	1
Animals	25.59	1060	56.48	2340	15.52	643	2.05	85	0.34	14	0.02	1
Games	27.51	816	49.36	1464	19.08	566	3.44	102	0.51	15	0.1	3
People	29.52	806	49.93	1363	17.69	483	2.42	66	0.4	11	0.04	1
Autos	30.57	992	41.45	1345	23.17	752	4.07	132	0.68	22	0.06	2
Comedy	32.33	1121	51.08	1771	14.08	488	2.08	72	0.35	12	0.09	3
Travel	33.75	1248	48.89	1808	15.44	571	1.76	65	0.14	5	0.03	1

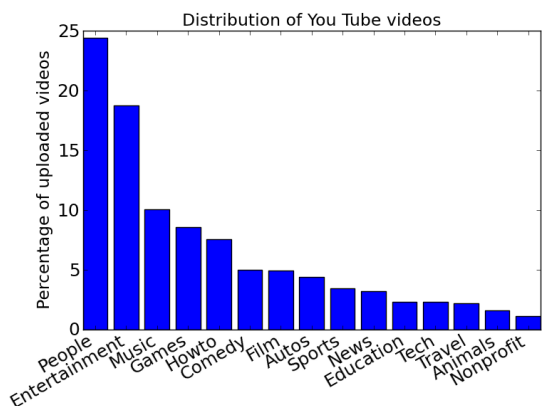


Figure 6: Category Uploading Rate (365,000 videos)

sampling revealed that these categories are dominated by UCC rather than UGC, so most of the videos in YouTube were actually UCC.

Figure 6 shows the current uploading trend of YouTube videos obtained by crawler 3. We see that the uploading trend in YouTube has changed over time. The People category is at the top position with approximately 24% of all the new videos, which was at the 6<sup>th</sup> position in 2007, only 8% of all the videos. Samples from the People category contain comparatively more UGC objects than other categories.

#### 4.4 Category Popularity Distributions

Figure 7 shows the Rank-frequency distribution for the 6 categories that showed the most interesting patterns. Previous studies (Abhari and Soraya, 2010; Cheng et al., 2007) showed that although requests for

popular YouTube videos follow a Zipf-like distribution, a Weibull distribution fits better because of the heavy tail section, which indicates a large number of very unpopular videos in YouTube. After considering video categories, only News videos follow a Weibull distribution for the first 80% of the videos, because of the comparatively flatter head section of News access pattern. This is consistent with *fetch-at-most-once* behaviour (Gummadi et al., 2003), as would be expected in watching news videos. For all other categories, request distributions of popular videos follow Zipf distributions and the heavy tail sections of the categories can be fit with a Weibull cutoff, as can be seen with the high goodness of fit statistic ( $R^2$ ). The number of videos that exhibit Zipf behaviour differs between the categories, showing different-sized tails.

Another measure that we calculated was the CCDF of total views over the measurement period. There were a substantial number of videos in certain categories that had at most 1 view. This can skew the popularity measures. The HowTo and Autos category had 17% and 12.6% of videos with at most 1 view, respectively, while 9% of HowTo videos had 0 views. There is a section of completely unpopular videos that get published, but never viewed. Figure 8 shows the CCDF of the total views for a selected number of categories. We truncate the x-axis to see the behaviour of views for unpopular videos more clearly. Entertainment is used as an example of a group of categories that had very similar CCDFs (Entertainment, Games, People, Education and Tech). The shape of the distribution of total views is very similar in these categories, but that of views over time is not. Music has very few videos below 20 views, but HowTo has almost 50% of the videos below 20 views.

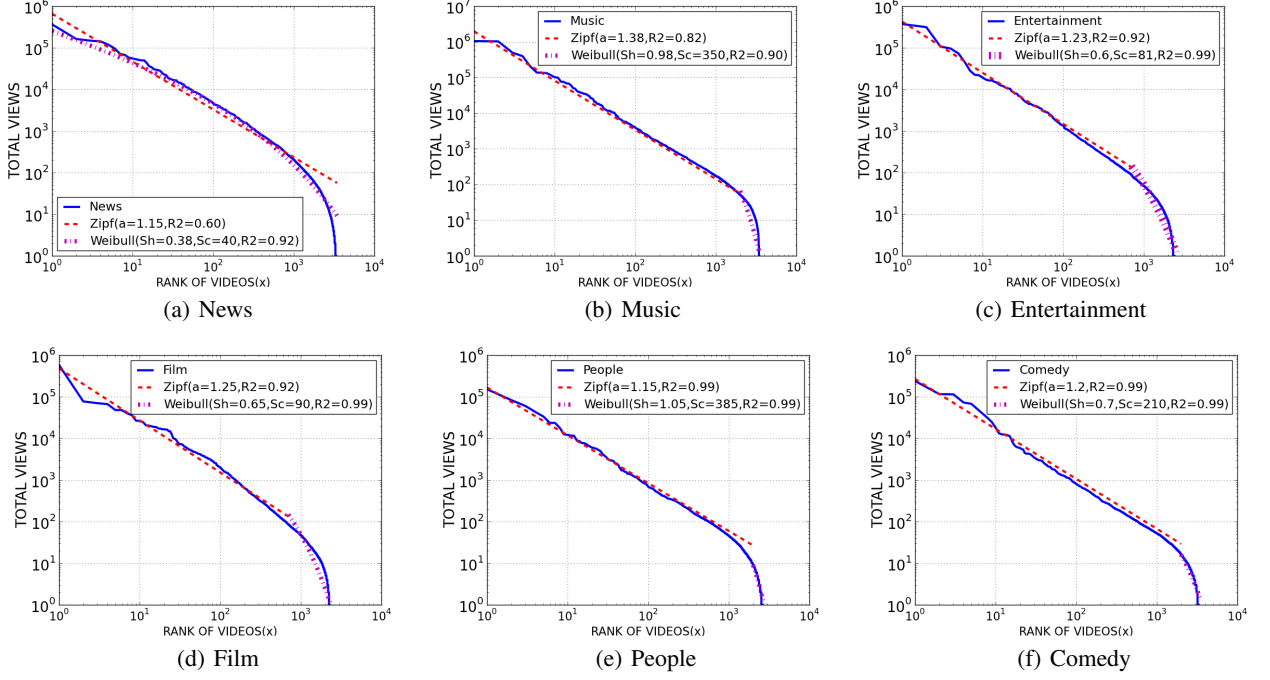


Figure 7: Number of views against rank for categories

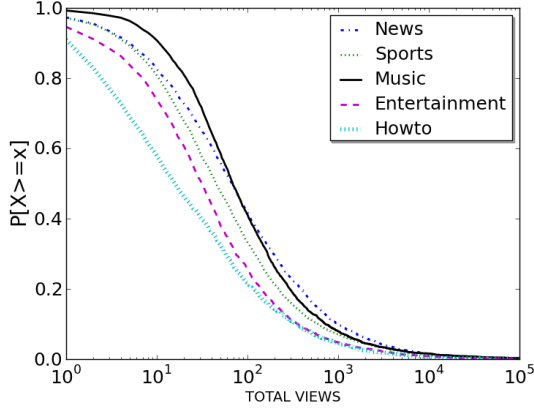


Figure 8: Selected CCDF of total views

## 5 TOWARDS A WORKLOAD GENERATOR

### 5.1 Predicting Popularity

As an approach to predict future popularity of videos, Pearson’s correlation coefficient (Equation 2) is calculated between the added views<sup>4</sup> at different snap-

<sup>4</sup>Added views is the number of views on a particular day

shots of the measurement period.

$$r_{xy} = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}} \quad (2)$$

A high correlation coefficient between early views and the rest of the period implies that prediction of future views of individual videos is achievable (Szabo and Huberman, 2010). We got very encouraging results for some of the categories including Sports, Travel, Howto, Tech and Games.<sup>5</sup>

However, for other categories like Film, News, Entertainment the coefficients are very poor, indicating the significant changes in the set of popular videos. Music shows a bit different characteristics though, if we take first 10 days as our first snapshot.

### 5.2 Time-Series Clustering

This category variation led us to model the growth patterns differently. Three-phase characterization (Borghol et al., 2011), does not work for the category specific modeling, as the number of videos that are at or before their peak phases in a particular day are very different between first few days and last few days in our measurement period. We thus decided to investigate whether the growth patterns of videos in

<sup>5</sup>Sports is 0.99 for the first day’s views and the rest of the measurement period



a specific category follow similar shapes. This approach can be considered as a time-series clustering problem and becomes a challenging problem as different videos reach peak popularity at different times. Inspired by a study on viral videos (Broxton et al., 2010), we translate all the time-series so that the x-axis is centred on the peak day, since most of the significant events happen around the peak periods.

Another challenging issue is to select the appropriate time-series clustering algorithm. We are particularly interested to identify similar shapes of the views per day, regardless of the time to peak. Moreover, the algorithm should not be affected much by outliers. We selected K-SC clustering (Yang and Leskovec, 2011), which has been found to be accurate in identifying the growth patterns of other Web content. Unlike K-means clustering, K-SC cluster centroids are not distorted by outliers. Instead of considering Euclidean distance between the curves, K-SC applies a scale and shift invariant distance metric (Chu and Wong, 1999). We evaluated the performance of K-SC algorithm for only two categories: Music and News. The clustering was performed only for the top 2000 videos in order to present more accurate results.

Figure 9 shows the six clusters for Music videos found by K-SC. Forcing K-SC to select fewer than six clusters drops the accuracy significantly, as we lose some of the interesting patterns. However, more than six clusters does not significantly improve the accuracy as we observe the repetition of similar clusters.

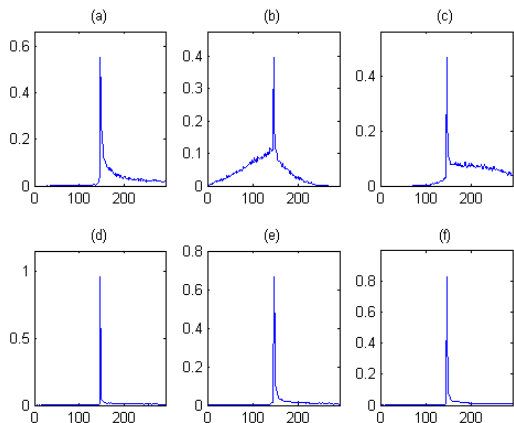


Figure 9: Music-clusters

The cluster shapes for News videos (Figure 10) are very similar to Music (except very little difference between cluster (a) and (e) in Figures 9 and 10 respectively).<sup>6</sup> However, the numbers of videos

<sup>6</sup>matching clusters are not in the same position in the graphs

in each cluster differ between these two categories, complementing our earlier findings. 46% of Music videos are contained within the slower-decaying clusters; this drops to 15% for News videos.

An important question that must be answered is whether a particular cluster is more biased to popular videos than others. This can be answered by taking the average of the rank values of all the videos in a cluster. The central limit theorem suggests that the

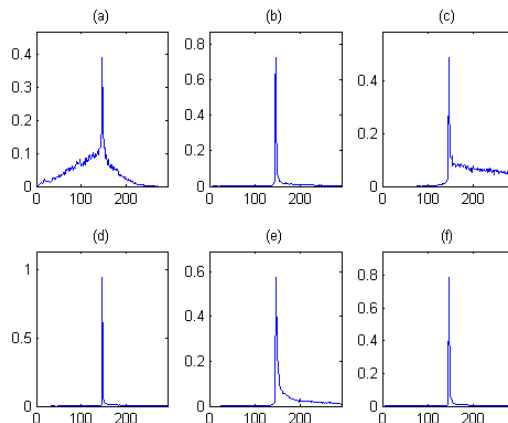


Figure 10: News-clusters

average rank of each cluster of videos should be 1000 if it is not popularity-biased. For News videos, the average rank values are very similar for each cluster (near 1000). For Music videos, the clusters with comparatively slower decay contain more popular videos, with average rank values of approximately 700. Popular Music videos observed a sharp decay with less frequency than popular News videos.

### 5.3 Performance of K-SC

In order to evaluate the performance of K-SC, we designed a synthetic workload generator for News and Music videos. The synthetic data should show similar characteristics to the empirical YouTube data if the clustering of K-SC is accurate.

The workload generator can be described as follows. A rank value is assigned to each of the 2000 videos as suggested by the chosen distributions for Music and News respectively. Then centroid/cluster is assigned to the videos based on the distribution we observed earlier. We also imposed a little bias for the popular videos before selecting the appropriate cluster in order to match our observed average rank value. Although for News videos the time-to-peak distributions are very similar for each of the clusters, we found very different results for Music videos, as



shown in Figure 11. We consider these peak distributions separately in our request generator.

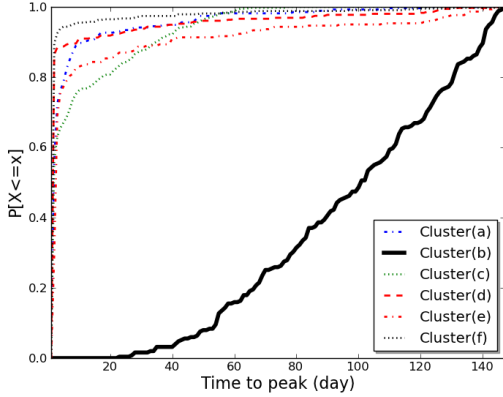


Figure 11: Peak distribution for music videos

We test similarity between the synthetic and empirical data from four different perspectives: 1) The total view distribution, 2) time-to-peak distribution, 3) Average daily views over time, and 4) 95<sup>th</sup> percentile of views over time.

Figures 12 and 13 indicate very good matches between synthetic and empirical data for metrics 1 and 2, which does not in itself indicate high accuracy of K-SC. We imposed the distributions for these two cases from our observations, i.e., fixed peak and Zipf/weibull distributions. Metrics 3 and 4 show, however, that the clusters found by K-SC algorithm for both categories represent most of the videos growth patterns (Figure 14 and 15, respectively).

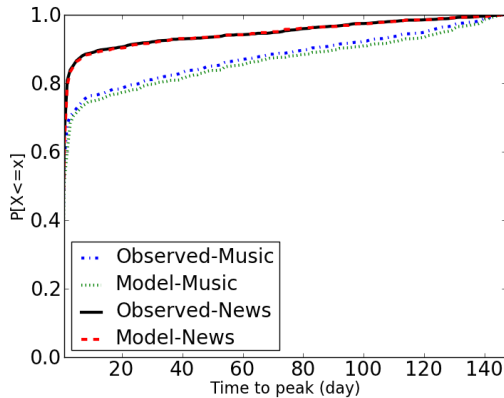


Figure 12: Time-to-peak distributions

Similar daily average views on a particular day indicates that view distribution among videos on that particular day are similar both in the empirical and synthetic data. These results show a smaller number of outliers in both of the categories.

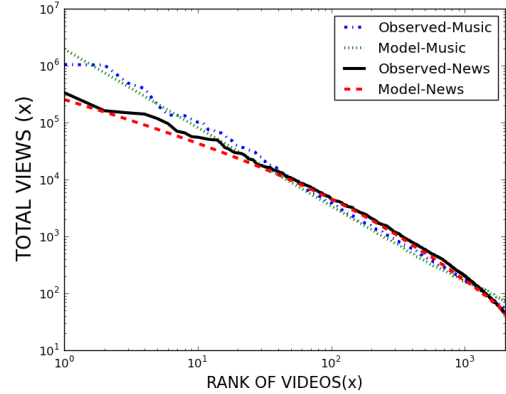


Figure 13: Popularity distributions

## 6 CONCLUSIONS AND FUTURE WORK

In this paper, we analyzed global daily viewing patterns of a representative subset of YouTube videos from their time of publication until they were 5 months old. We discovered significant time-varying popularity differences between categories.

Most videos exhibit their peak viewing day very soon after publication and then there is a decay; relatively few videos ever return to near their peak popularity. We determined that video categories which reached their peaks later were more stable. This is expected and matches our intuitions. We developed an analysis method that permits quantification of these differences on a particular dataset. The confirmation of Zipf distributions for the total views of popular videos in nearly every category indicates that caching would be effective.

We were also able to determine the relative trends of viewing patterns of videos within categories over the first few months of their lifetimes. Some categories contain a non-trivial number of videos which are still popular 5 months after upload date, whereas other categories have viewing patterns which dwindle to nothing. Some categories have videos which exhibit stationary behaviour that allows prediction of which videos will remain relatively popular based on their early views. Popularity changes around peak time can be captured by appropriate time-series clustering. While we use a dataset from YouTube, issues regarding the scale and deployment make direct applicability to YouTube impractical. Multiple regional caches are needed to satisfy the demand and regional differences (Brodersen et al., 2012). Our methodology and analysis could be used to help design, configure, and deploy any category specific UGC site.

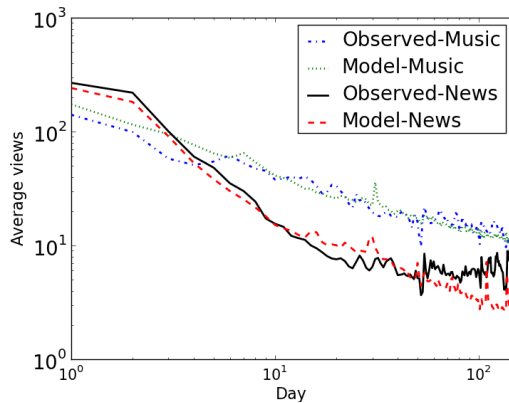


Figure 14: Daily views over time

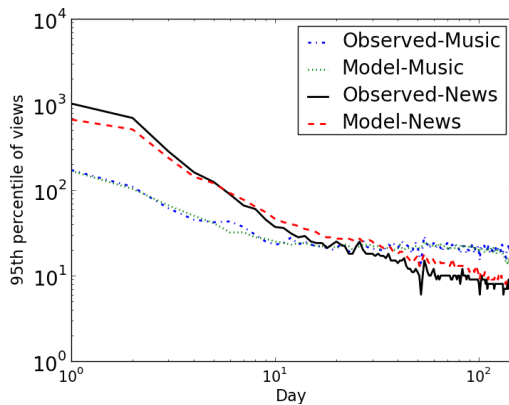


Figure 15: 95th percentile of views over time

As future work, we are in the process of building a complete workload generator that encompasses more aspects of user-generated content video requests. In particular, we will incorporate category-specific introduction of new content over time to drive simulations and/or prototype content distribution networks to evaluate different design policies for storing and delivering videos.

## REFERENCES

- Abhari, A. and Soraya, M. (2010). Workload Generation for YouTube. *Multimedia Tools and Applications*, 46(1):91–118.
- Borghol, Y., Mitra, S., Ardon, S., Carlsson, N., Eager, D., and Mahanti, A. (2011). Characterizing and Modelling Popularity of User-Generated Videos. *Performance Evaluation*, 68:1037–1055.
- Brodersen, A., Scellato, S., and Wattenhofer, M. (2012). YouTube Around the World: Geographic Popularity of Videos. In *World-Wide Web 2012*, pages 241–250, Lyon, France.
- Broxton, T., Interian, Y., Vaver, J., and Wattenhofer, M. (2010). Catching a viral video. In *IEEE Data Mining Workshops*, pages 296–304, Sydney, Australia.
- Cha, M., Kwok, H., Rodriguez, P., Ahn, Y., and Moon, S. (2009). Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems. *IEEE/ACM Trans. Netw.*, 17(5):1357–1370.
- Cheng, X., Dale, C., and Liu, J. (2007). Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study. Technical report, Cornell University, arXiv e-prints.
- Chu, K. K. W. and Wong, M. H. (1999). Fast time-series searching with scaling and shifting. In *ACM PODS 1999*, pages 237–248, Philadelphia, PA.
- Ding, Y., Du, Y., Hu, Y., Liu, Z., Wang, L., Ross, K., and Ghose, A. (2011). Broadcast Yourself: Understanding YouTube Uploaders. In *ACM IMC 2011*, pages 361–370, Berlin, Germany.
- Figueiredo, F., Benevenuto, F., and Almeida, J. (2011). The Tube over Time: Characterizing Popularity Growth of Youtube Videos. In *ACM WSDM 2011*, pages 745–754, Hong Kong, China.
- Gember, A., Anand, A., and Akella, A. (2011). A Comparative Study of Handheld and Non-handheld Traffic in Campus Wi-Fi Networks. In *PAM 2011*, pages 173–183, Atlanta, GA.
- Gill, P., Arlitt, M., Li, Z., and Mahanti, A. (2007). Youtube Traffic Characterization: A View From the Edge. In *ACM IMC 2007*, pages 15–28, San Diego, CA.
- Gummadi, K. P., Dunn, R. J., Saroiu, S., Gribble, S. D., Levy, H. M., and Zahorjan, J. (2003). Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In *ACM SOSP 2003*, pages 314–329, Bolton Landing, NY.
- Khemmarat, S., Zhou, R., Gao, L., and Zink, M. (2011). Watching User Generated Videos with Prefetching. In *ACM MMSYS 2011*, pages 187–198, San Jose, CA.
- Labovitz, C., Iekel-Johnson, S., McPherson, D., Oberheide, J., and Jahanian, F. (2010). Internet Inter-Domain Traffic. In *ACM SIGCOMM 2010*, pages 75–86, New Delhi, India.
- Maier, G., Schneider, F., and Feldmann, A. (2010). A First Look at Mobile Hand-held Device Traffic. In *PAM 2010*, pages 161–170, Zurich, Switzerland.
- Siersdorfer, S., Chelaru, S., Nejd, W., and Pedro, J. S. (2010). How Useful are Your Comments?: Analyzing and Predicting YouTube Comments and Comment Ratings. In *World-Wide Web 2010*, pages 891–900, Raleigh, NC.
- Szabo, G. and Huberman, B. (2010). Predicting the popularity of online content. *CACM*, 53(8):80–88.
- Yang, J. and Leskovec, J. (2011). Patterns of temporal variation in online media. In *ACM WSDM 2011*, pages 177–186, Hong Kong, China.
- Zink, M., Suh, K., Gu, Y., and Kurose, J. (2009). Characteristics of YouTube Network Traffic at a Campus Network - Measurements, Models, and Implications. *Computer Networks*, 53(4):501–514.