What Drives the Growth of YouTube? Measuring and Analyzing the Evolution Dynamics of YouTube Video Uploads

Golshan Golnari[1], Yanhua Li[2], Zhi-Li Zhang[1]
[1]University of Minnesota, Twin Cities
[2]HUAWEI Noah's Ark Lab, China
golna001@umn.edu, li.yanhua1@huawei.com, zhzhang@cs.umn.edu

Abstract

We make the first attempt to study *the evolution dynamics of YouTube, from the perspectives of uploaded videos and uploaders*. Using *unbiasedly estimated* video statistics, we study how YouTube grows over time, from the inception of YouTube in 2005 up until now. We show that the growth of YouTube videos undergoes several phases: i) an initial growth phase best fitted by a quadratic curve, ii) an exponential growth phase that starts circa late 2009, interrupted by iii) a sudden drop that lasts a few months in early 2012, and followed by iv) a (resumed) rapid growth phase again. To further understand what drives the growth in YouTube uploaded videos, we examine several factors/questions related to the growth of YouTube videos, and develop models to predict the growth in the video uploads. To the best of our knowledge, our study is the first of its kind in studying the evolution dynamics of YouTube; it may shed useful light on its future growth.

1 Introduction

Since its inception in 2005, YouTube has seen explosive growth in its popularity. It has become a unique cultural and social media phenomenon, attracting millions of users all over the world every day. As a primarily "user-generated" video sharing service and a quintessential social media site, clearly the success of YouTube hinges on two interacting factors: continually attracting both users to upload videos (*uploaders*) and users to watch uploaded videos (*viewers*); without either, YouTube would not have attained the popularity and explosive growth that we have witnessed. So what drives the growth of YouTube over time, making it such a great success? Will YouTube be able to sustain its phenomenal growth rate in the future as it has achieved so far? Gaining a deeper understanding of the evolution dynamics of YouTube can explicate major commercial and technical implications to user generated content service site administrators (e.g., capacity planning), content creators and owners as well as advertisers. It may also shed valuable light on the future evolution of video sharing services and other social media sites, especially in terms of the roles and behaviors of *contributors* (e.g., video uploaders in the case of YouTube) to such social media sites.

Much of research attention has been devoted to studying YouTube video *popularity* and *viewer behavior*, with applications in online advertising, video recommendation, and so forth (say, e.g., [1, 2, 3, 4, 5, 6]). Some studies [7, 8] have examined the social relations among YouTube users to explore their impact on YouTube uploading behaviors. Most of these existing studies are based on snapshots of the statistics collected from the YouTube site or based on YouTube related data extracted from passively collected network traces at certain vantage points (e.g., a campus network). Few have investigated the evolution dynamics of YouTube over a longer period of time. While videos' popularity (view) statistics and dynamics can be crawled directly from YouTube site, it is not a trivial task to tease out the historical growth process of YouTube (uploads), since YouTube has not made such statistics publicly available.

In this paper we make the first attempt to study *the evolution dynamics of YouTube from the perspectives of uploaded videos and uploaders*, with the goal to illuminate the roles and behaviors of social media contributors in the evolution of social media sites such as YouTube. Our study is made feasible by a novel random prefix sampling technique developed in [9] which we have generalized and extended to estimate the number of daily video uploads, the number of uploaders and other statistics associated with the sampled videos in an *unbiased* manner (see Section 3). Using this technique we are able to sample large (unbiased) collections of user uploaded videos at several different points in time, and use the samples to estimate the number of YouTube videos uploaded by users as well as various statistics (e.g., video uploaders, categories, etc.) associated with the uplodaded videos*from its inception in 2005 up until now.* In studying the evolution dynamics of YouTube from the perspectives of uploaders and uploading behavior, we are particularly interested in answering the following questions: i) How does the number of YouTube uploaded videos grow over time? Does it have an exponential growth rate? ii) Are there certain changes in the growth rates over time? What may have caused it? iii) What contributes to the growth in the number of uploaded videos? Is it due to videos uploaded by *new* uploaders who join the site, or is it due to existing uploaders contribute more videos? iv) Is the growth contributed primarily by videos belonging to certain categories or countries, or by uploaders from certain countries, or due to other factors (e.g., "major" events such as emergence of a new technology, promulgation of a new policy)? We summarize our contributions and key findings as follows.

• By analyzing the dataset, we estimate the total number of YouTube videos uploaded per day and the growth of (new) uploaders over time since its inception in 2005 up until now. We show that the growth of YouTube videos undergoes several phases: i) an initial growth phase best fitted by a quadratic curve, ii) an exponential growth phase starting circa late 2009, interrupted by iii) a sudden drop that lasts a few months in early 2012, and followed by iv) a (resumed) exponential growth phase. The growth rate of daily new *uploaders* undergoes similar phases, with a key exception – after the disruptions in early 2012, the growth rate of daily new uploaders never fully recovers (see Sec 4). We also analyze the growth of YouTube videos over time in terms of various categories, which provides hints regarding the interests of both uploaders and viewers (see Sec 4).

- Analyzing the contributors to the distinct growth phases in the YouTube video evolution dynamics, we find that the exponential growth phase of YouTube coincides and is strongly correlated with the emergence and exponential growth of videos uploaded via mobile devices. On the other hand, the sudden drop in the growth of YouTube is plausibly caused by Google's new privacy policy, which is evidenced by the same sudden drop in the growth of daily uploaders (as measured by the number of user id's seen in the datasets), which in fact never fully recovers (see Sec 5).
- To further analyze what contributes to the YouTube growth, we conduct an in-depth study of the uploading behavior of users. We show that "new" uploaders who joined the system in the later years tend to show higher activeness than those who joined in the earlier years. The (cumulative) activeness levels of existing uploaders decay over time, and this decaying behavior can be well *modeled and predicated* via an iterative nonlinear least square regression method. All in all, the contributions of new uploaders drive the overall growth of YouTube videos (see Sec 6).
- To further distinguish and evaluate the contributions of users with varying number of uploads, we consider two extreme groups: the so-called *pop-up* uploaders who upload only a few videos and whose "life span" lasts less than a week vs. YouTube *partners* who frequently upload videos with the option to display YouTube ads. We find that the portions of videos contributed by the pop-up uploaders over the years decrease steadily, while those by the partners increase significantly. Our analysis suggests that the growth of YouTube increasingly relies on – and will likely be sustained by – the continued contributions of existing "heavy" uploaders, as the growth in the number of new uploaders slows significantly (see Sec 7).

2 Related Work

The popularity success of YouTube has attracted a flurry of academic studies over the years. Much of the research attention has focused on studying the video statistics such as video popularity, life-cycle of videos as well as viewing behavior of users; there is a large literature related to these topics, here we cite a few as representative examples [10, 11, 12, 1, 2, 3]. Most of these studies are either based on statistics crawled from the YouTube site or statistics extracted from network traces collected from one or few vantage points, with exception of [6], which utilizes YouTube's internal data. The authors in [4] study the YouTube video popularity over time, whereas Ahmed et al. [5] develops a dynamic model to predict the temporal evolution of YouTube video popularity. Other studies have examined social relations and user behaviors and how they impact user viewing or uploading behavior. For example, Spathis et al. [8] investigate how the social relations (i.e., social community structure) impact the video popularity of YouTube videos, while Ding et al. [7] demonstrate the strong tie between online social behavior and video uploading behavior. The YouTube video delivery infrastructure, traffic dynamics and their impact on ISPs and the underlying networking substrate have also been investigated through passive and active network measurement, see, e.g., [13, 14, 15].

3 Sampling & Statistic Estimation

In this section, we first briefly present the preliminaries of YouTube, and random prefix sampling algorithm proposed by [9]. We then introduce unbiased estimators we have developed to estimate YouTube statistics such as the number of videos and daily new uploaders.

3.1 Preliminaries & Random Prefix Sampling

- **YouTube ID space.** Each YouTube video has a unique identifier, with 11 characters. The first 10 characters of a valid ID contains any of the characters in $S = \{0-9, \_, A-Z, a-z\}$, and the last (11-th) character v_{11} only comes from $T = \{0,4,8,A,E,I,M,Q,U,Y,c,g,k,o,s,w\}$. Hence, the video ID space can be expressed as $\mathcal{S} = S^{10} \times T$. The empirical study in [9] reveals that an unused video ID is randomly generated for each new uploaded video.
- **YouTube video search API.** YouTube provides a video search API. Given an input 11-character string (if it is a valid YouTube ID,) the API returns the video profile information, including the uploader, view counts, uploading date, etc. In particular, when searching using a string xy...z as a prefix with length $1 \leq L \leq 11$, that does not contain "-", YouTube will return a list of videos whose IDs begin with this prefix followed by "-", if they exist.
- **Random Prefix Sampling.** The authors in [9] developed a novel prefix sampling technique which exploits the features of the YouTube ID space and video search API. The technique works as follows. At each sampling step, we randomly generate a prefix with length of four and the fifth character as a dash "-" to perform a video search. The return of the searched query, which we refer to as a "query sample", is a list of video profiles whose video IDs all start with that prefix. It is shown that the prefix length of $L = 5$ is the best length for video search query [9]. The authors in [9] employs this technique to estimate the total number of YouTube videos and show it yields an unbiased estimator for the total number of videos.

3.2 YouTube Statistic Estimation

We generalize and extend the prefix sampling technique developed in [9] by estimating the number of daily video uploads, the number of uploaders and other statistics associated with the sampled videos in an *unbiased* manner. In the following we briefly describe the estimation method.

Estimating the number of daily uploaded videos and other video statistics. We first note that when a YouTube video is uploaded, it is assigned a random video ID. Given a prefix of length L, $L = 1,\ldots,11$, the probability, p_L, that a randomly generated ID from $\mathcal{S}$ matches this given L-length prefix is given as follows: $p_L = \frac{1}{|S|^L}$ for $L = 1,...,10$ and $p_L = \frac{1}{|S|^{10}|T|}$ if $L = 11$. When performing the random prefix sampling using a randomly generated prefix of length L, the returned video profiles in each "query sample" contain rich information of the videos, such as uploading time, video category, video length, uploader of the video, etc.

Those information components can be viewed as video "labels", which allow us to estimate video statistics in a broader range, e.g., the number of *daily* uploaded videos, the number of uploaded videos in the *Music* category, etc. To be precise, let N_ℓ denote the total number of videos with a specific label ℓ, e.g., a video category. We can estimate N_ℓ with an *unbiased* estimator $\hat{N}_\ell$ as follows:

$$\hat{N}_\ell = \frac{1}{mp_L}\sum_{i=1}^{m} X_{i,\ell}^{L}, \tag{1}$$

where $X_{i,\ell}^{L}$ is the random variable denoting the number of videos with label ℓ in i-th "query sample" ($1 \le i \le m$) from the total of m "query samples" and for a prefix with length of L. Note that the estimation for the *total* number of videos is the special case of eq. (1) where no label is specified and all the videos in the query samples are counted: $\hat{N} = \frac{1}{mp_L}\sum_{i=1}^{m} X_i^L$.

It is not hard to see that $X_{i,\ell}^{L}$'s are samples drawn from a Binomial distribution, $Binomial(N_\ell, p_L)$.[1] As $X_{i,}$'s are *i.i.d*, $\sum_{i=1}^{m} X_{i,\ell}^{L}$ is also a Binomial random variable but with a different success probability, $Binomial(N_\ell, mp_L)$. The following equation shows that eq. (1) yields an unbiased estimator:

$$E[\hat{N}_\ell] = \frac{1}{mp_L}\sum_{i=1}^{m} E[X_{i,\ell}^{L}] = \frac{1}{mp_L}\sum_{i=1}^{m} N_\ell p_L = N_\ell, \tag{2}$$

The variance of this unbiased estimator is:

$$Var[\hat{N}_\ell] = \frac{1}{m^2 p_L^2} Var[\sum_{i=1}^{m} X_{i,\ell}^{L}] = N_\ell(\frac{1}{mp_L} - 1) \tag{3}$$

Estimating the number of uploaders. Each video profile collected contains information of its uploader, such as the number of videos uploaded. It enables us to estimate statistics of YouTube uploaders, e.g., the number of daily (new) YouTube uploaders. Let K denote the maximum number of videos an uploader has in YouTube, and $n = \sum_{k=1}^{K} n_k$ be the total number of YouTube uploaders, with each n_k as the number of uploaders who uploaded k videos. Consider n_k as a video "label", using eq. (1), we can estimate, N_{n_k}, the total number of videos whose uploader had k videos, as $\hat{N}_{n_k} = \frac{1}{mp_L}\sum_{i=1}^{m} X_{i,n_k}^{L}$. By definition, $N_{n_k} = kn_k$ holds, thus we obtain an unbiased estimator $\hat{n}$ as

$$\hat{n} = \sum_{k=1}^{K} \hat{n}_k = \sum_{k=1}^{K} \frac{\hat{N}_{n_k}}{k} = \frac{1}{mp_L}\sum_{k=1}^{K}\sum_{i=1}^{m} \frac{X_{i,n_k}^{L}}{k}, \tag{4}$$

and to show it is unbiased, we have:

$$E[\hat{n}] = \sum_{k=1}^{K} E[\hat{n}_k] = \sum_{k=1}^{K} \frac{E[\hat{N}_{n_k}]}{k} = \sum_{k=1}^{K} \frac{N_{n_k}}{k} = \sum_{k=1}^{K} n_k = n, \tag{5}$$

The variance of $\hat{n}$ is calculated as follows:

$$Var[\hat{n}] = Var[\sum_{k=1}^{K} \hat{n}_k] = \sum_{k=1}^{K} \frac{Var[\hat{N}_{n_k}]}{k^2} = (\frac{1}{mp_L} - 1)\sum_{k=1}^{K} \frac{1}{k} n_k, \tag{6}$$

where the second equality is the result of n_k's independency from each other (the number of uploaders who have uploaded k_1 videos is independent of number of uploaders with k_2 number of uploads) and the relation $N_{n_k} = kn_k$. The forth equality is a simple substitution of eq. (3).

The next Theorem helps us to figure out the least number of "query samples" to bound the estimation error:

Theorem 1 *(Confidence Interval of the Estimators $\hat{N}$ and $\hat{n}$). Given any $0 < \varepsilon \ll 1$ and $0 < \alpha \le 1$, if the number of "query samples" follows $m \ge \frac{1}{p_L(\alpha\varepsilon^2 n+1)}$, the following confidence intervals for the estimators $\hat{N}$ and $\hat{n}$, given in eq. (1) and eq. (4) respectively, can be guaranteed* ($0 < c < 1$)

$$Pr[N(1-c\varepsilon) \le \hat{N} \le N(1+c\varepsilon)] = 1-\alpha, \tag{7}$$

$$Pr[n(1-\varepsilon) \le \hat{n} \le n(1+\varepsilon)] \ge 1-\alpha, \tag{8}$$

Proof. The Binomial random variable $\sum_{i=1}^{m} X_i^L \sim Binomial(N, mp_L)$ can be well approximated by a Normal random variable, $Normal(Nmp_L, Nmp_L(1-mp_L))$, when both Nmp_L and $N(1-mp_L)$ are larger than 10 [17], which is the case in our problem. A Normal random variable X has the confidence interval $Pr[-z_{\alpha/2} \le \frac{X-\mu_X}{\sigma_X} \le z_{\alpha/2}] = 1-\alpha$, where μ_X is the mean, σ_X is the standard deviation, and $z_{\alpha/2}$ is the $100(1-z_{\alpha/2})$-th percentile of the standard normal distribution [17]. Therefore, the confidence interval $Pr[N(1-\varepsilon') \le \hat{N} \le N(1+\varepsilon')] = 1-\alpha$ conditioned on having $m \ge \frac{z_{\alpha/2}^2}{p_L(\varepsilon'^2 N + z_{\alpha/2}^2)}$ holds for $\hat{N}$. Substituting $\varepsilon' = c\varepsilon$ and $c = \sqrt{\alpha} z_{\alpha/2}\sqrt{\frac{n}{N}}$, the confidence interval in eq. (7) and the corresponding inequality for m are obtained. To see that $0 < c < 1$, consider the Chebyshev's inequality for Normal random variables X, which is $Pr[-\kappa \le \frac{X-\mu_X}{\sigma_X} \le \kappa] \ge 1 - \frac{1}{3\kappa^2}$, $\kappa > 0$. Comparing this inequality with the mentioned confidence interval for Normal random variable implies that in the case of $\kappa = z_{\alpha/2}$ the relations $\frac{1}{3\kappa^2} \le 1-\alpha$ and so $\sqrt{\alpha} z_{\alpha/2} \le \frac{1}{\sqrt{3}} < 1$ hold true. In addition, since each uploader has at least one uploaded video, i.e. $N \ge n$, the constant $c = \sqrt{\alpha} z_{\alpha/2}\sqrt{\frac{n}{N}}$ is a positive number with value smaller than 1.

To find a confidence interval for $\hat{n}$, we use the *general* form of Chebyshev's inequality $Pr[-\kappa \le \frac{X-\mu_X}{\sigma_X} \le \kappa] \ge 1 - \frac{1}{\kappa^2}$, since $\hat{n}$ does not follow a normal distribution necessarily (n_k's have different mean and variances). Thus, random variable $\hat{n}$ with mean value of n and variance's upper bound of $(\frac{1}{mp_L} - 1)n$ has the following confidence interval, where $\frac{1}{\kappa^2}$ has been substituted by α:

$$Pr[n - \sqrt{(\frac{\frac{1}{mp_L} - 1}{\alpha})n} \le \hat{n} \le n + \sqrt{(\frac{\frac{1}{mp_L} - 1}{\alpha})n}] \ge 1-\alpha$$

Note that the variance's upper bound is simply derived from $Var[\hat{n}] = (\frac{1}{mp_L} - 1)\sum_{k=1}^{K} \frac{1}{k} n_k \le (\frac{1}{mp_L} - 1)\sum_{k=1}^{K} n_k = (\frac{1}{mp_L} - 1)n$. The substitution of $m = \frac{1}{p_L(\alpha\varepsilon^2 n+1)}$ in the above equation results in eq. (8). ■

Datasets. Our datasets were crawled from 06/30/2013 to 07/25/2013 using random prefix sampling, with video prefixes starting from random combinations of [a-z], [A-Z], or [0-9], and the fifth letter as a dash "-". We recorded $m = 7,752,384$ "query samples" with prefix length of $L = 5$, which include $7,977,651$ videos in total. Rearranging the inequality for m in Theorem 1, we find the lower bound for α and ε: $\alpha\varepsilon^2 \ge \frac{1-mp_L}{mp_L n}$. Having about $n = 10^8$ number of uploaders, $\alpha\varepsilon^2$ can be in the order of 10^{-6} which can provide as small relative error as 0.01 (at most) for our estimation of number of uploaders and 0.001 for our estimation of number of videos, with high probability of 99%.

[1] For further details please refer to [9]. To be more precise, $X_{i,\ell}^{L}$ has a Hypergeometric distribution which converges to Binomial distribution when the population size ($|\mathcal{S}|$ here) is very large [16].

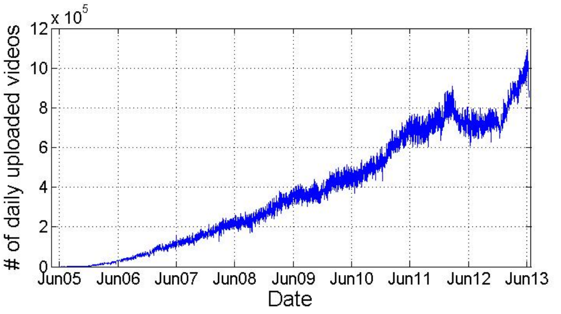

Figure 1: Daily uploads

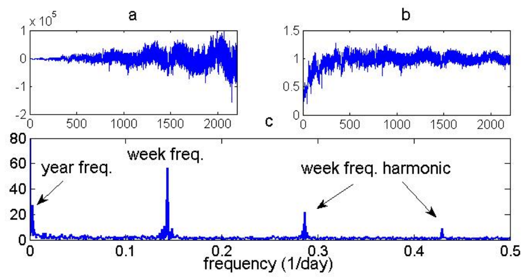

Figure 2: (a) detrended data by subtraction, (b) detrended data by division, (c) FFT of detrended (by division) data

4 Growth Dynamics of YouTube

By applying the estimation methods presented earlier to the datasets, in this section we examine and model the evolution dynamics of YouTube from the perspectives of daily uploaded videos, video categories and daily uploaders.

4.1 Daily Uploads and Growth Phases

In Fig. 1, we plot the (estimated) numbers of YouTube video uploaded each day (*daily uploads*) since the inception of YouTube until June 2013. Clearly, with the exception of the early part of 2012, the number of daily uploaded YouTube videos has grown significantly over time. Based on our estimates, YouTube received $\approx 2 \times 10^4$ uploads per day around June 2006, and this number has reached to 10^6 videos per day 7 years later. The *cumulative* number of uploaded YouTube videos is about 1.1×10^9 by July 2013.

To better understand the evolution dynamics of YouTube, we apply time series analysis to model the growth of YouTube videos over time. We find that the growth dynamics of YouTube videos can be best captured via a *multiplicative* time series decomposition model, $X_t = T_t S_t R_t$, where T_t represents the *trend*, S_t the *seasonality*, and R_t daily variations (the *residual*). The trend T_t can be obtained by applying a smoothing operation such as the moving average method; we apply a moving window $w = 365$, i.e., a yearly smoothing window, to smooth out the seasonality effects. The smoothing operation is in effect a low pass filter which rejects the high frequency components belonging to the seasonal and irregular components ($S_t W_t$). We then apply the Fast Fourier Transform (FFT) to the *detrended* data (obtained via division by the trend) to identify the significant seasonalities in the data. The results are shown in Fig. 2(c), where we see that two most dominant frequency components are the (stronger) weekly and (weaker) yearly seasonalities. Further analysis reveals that users tend to upload more videos during the weekends and Mondays, and the uploads generally taper off from Tuesday to Friday; this behavior leads to the (stronger) *weekly* frequency component (the two smaller frequency components are simply the harmonics of the weekly frequency component); In addition, holidays (e.g., the Christmas season) and summer vacations also tend to have some effect on user uploads, giving rise to the (weaker) *yearly* frequency components. The detrended data is shown in Fig. 2(b). For comparison, we also plot the residuals obtained via the *additive* model, $X_t = T_t + S_t + R_t$, where it can be seen that R_t is not stationary, as it grows larger over time (Fig. 2(a)). The seasonal component S_t can be constructed using sinusoidal functions with the mentioned important frequencies observed in FFT analysis, and be further removed from the detrended data to obtain the residual component R_t. Performing some statistical tests (e.g., Augmented Dickey-Fuller test for stationarity and Kolmogorov-Smirnov test for normality) on the residual component R_t, we find that the residual is stationary and has a normal distribution, but it is not white noise.

Given this multiplicative time series decomposition model for YouTube video growth dynamics, we perform curve fitting to model the growth rate of YouTube videos as represented the trend T_t. To remove the effect of the interruption that occurs in early 2012, we consider only the trend data from Nov 2005 to Nov 2011. We find that the YouTube growth trend can be best approximated by two (parsimonious) functions: a quadratic function of the form $0.086t^2 + 104.3t^1 - 2691$ and an exponential function of the form $42550e^{0.001354t}$, with the transition period occurring circa the beginning of 2010 (see the two top curves in Fig. 3). Interestingly, if we ignore the YouTube video data due to the disruptions in 2012 and apply the model to the data from Jan to June 2013, we find that the growth trend in early 2013 fits the same exponential trend as predicted by our model. In summary, we find that the growth of YouTube videos undergoes several phases: i) an initial growth phase best fitted by a quadratic curve, ii) an exponential growth phase that started circa late 2009/early 2010, interrupted by iii) a sudden drop that lasts a few months in the early part of 2012, and followed by iv) a (resumed) exponential growth phase again.

4.2 Growth Dynamics per Video Categories

YouTube provides 15 categories that a user can specify when uploading videos (we note in case a user does not specify a category, YouTube applies *People & Blogs* as the default category). The categories of videos provide hints regarding the interest areas of uploaders. The pie chart in Fig. 4 shows the percentages of videos in each category by the end of June 2013. The top six categories are *People & Blogs*, *Music*, *Entertainment*, *Gaming*, *Comedy* and *Sports*; collectively they constitute about 75% of the total uploaded videos. In Fig. 5 we show the (*smoothed*) growth rate of YouTube videos in the top 6 categories over time, where we use a smoothing window of 6 months (instead of one year) to obtain the trend curves so as to better illustrate the inflections in the growth rates of individual video categories. We see that *People & Blogs*, *Music*, *Entertainment*, *Comedy* and *Sports* have stayed among the top six most popular categories since YouTube

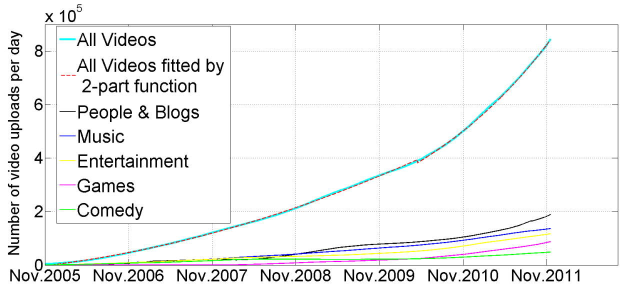

Figure 3: Daily upload's increasing trend and the fitted piecewise function

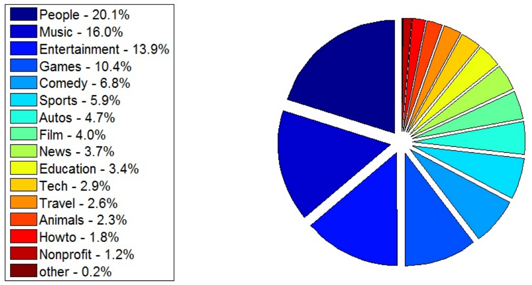

Figure 4: Video category portions

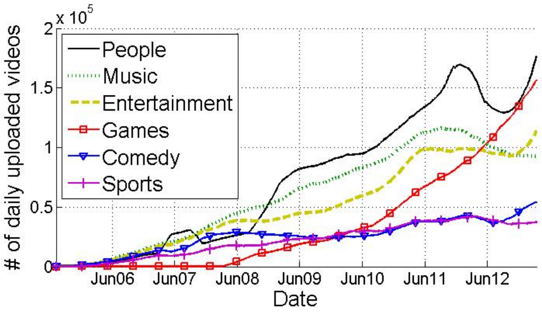

Figure 5: Daily uploads in categories

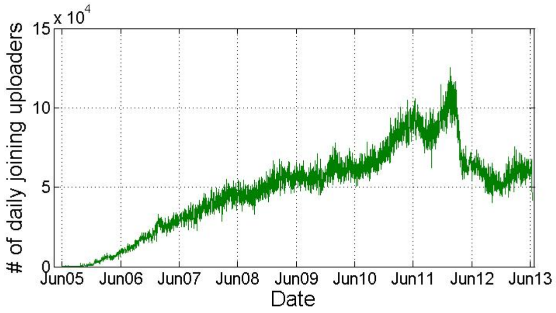

Figure 6: Daily new uploaders

was created in 2005. It is interesting to note that the first three categories, *People & Blogs*, *Music* and *Entertainment* grow much faster than the latter two, *Comedy* and *Sports*, the growth rates of which remain relatively flat. We note that since 2009, *People & Blogs* has been the top category (being the default category may have partly contributed to this status), but its growth rate has experienced a few large inflections over time. Among the current top 6 categories, *Game* grows the fastest, starting from nearly zero before 2008 to nearly taking over "People & Blogs" in terms of daily uploads. We have also applied the multiplicative time series decomposition models to study the growth dynamics of these categories and fit the growth trends. In Fig. 3, we show the smoothed (using a yearly moving window) growth trends for the top five categories together with the overall video growth trend, using the data from Nov 2005 to Nov 2011. Due to space limit, we do not elaborate on these models. Note that all of the top 6 categories have video samples in our dataset belonging to 2005 which indicates that these categories have been defined and existed from YouTube's inception.

Perhaps more interesting is how the turbulence period observed in the overall daily YouTube video uploads during the early part of 2012 manifests itself in the growth dynamics of the top 6 video categories. We see that *People & Blogs*, *Music* and *Comedy* experience similar disruptions, with those in *People & Blogs* more pronounced and those in *Comedy* far less so. In contrast, the disruptions in *Entertainments* and *Sports* are barely visible, while *Game* continues to grow exponentially without any disruptions in the same period. The disparate manifestations of these disruptions in the top 6 categories point to some plausible cause that affects the video uploaders of these categories differently, as we will discuss further in Section 5.

4.3 Growth Dynamics in Daily New Uploaders

We consider the date that a user uploads its first video as the date that he/she "joins" YouTube, and at this date the user is labeled as a *new* uploader. By examining the number of new uploaders in the sampled datasets, we apply the same estimation method presented in Section 3 to estimate the number of *daily new uploaders* for each day. Fig. 6 shows the growth dynamics of daily new uploaders over time. We see that the growth dynamics of daily new uploaders also undergo several growth phases, similar to the growth dynamics of daily uploaded videos, with a key exception. The growth rate exhibits a quickly increasing trend from June 2005 to late 2011, where more than 10^5 new uploaders joining YouTube every day. Based on our estimate, by the end of June 2013, YouTube has had about 1.4×10^8 uploaders (as measured by the unique user account id's). We have applied the multiplicative time series decomposition models to study the growth dynamics of daily new uploaders, and find that the growth trend between 2005 to 2011 can also be fitted parsimoniously with two curves, a quadratic curve and exponential curve with a phase transition circa the beginning of 2010. (We omit the details here due to space limitations.) The growth of daily new uploaders is highly correlated with the growth of daily uploaders (with a correlation coefficient of 0.85), suggesting that the growing new uploaders have played a significant role in driving the growth of YouTube. However, this correlation is broken since early 2012, where we see that the growth rate of daily new uploaders also experiences a significant sudden drop in early 2012. In contrast to the growth rate of the daily uploaded videos which resumes the same exponential growth after the turbulence period in the early part of 2012, we see that the growth rate of the daily new uploaders never fully recovers. We will further discuss the implication of this observation in Section 5.

Before we leave this section, we briefly discuss the geograph-

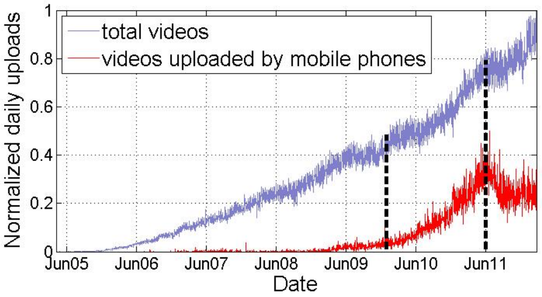

Figure 7: Daily uploaded videos by mobile devices

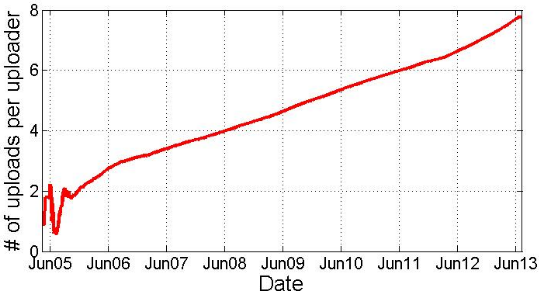

Figure 8: Cumulative average uploads per uploader

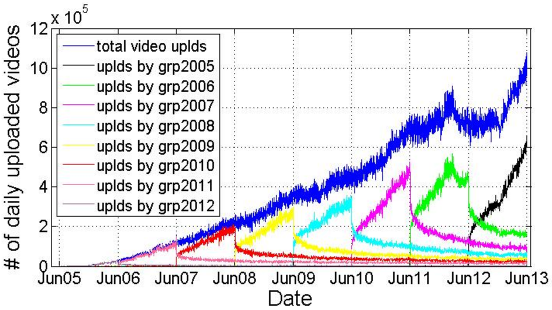

Figure 9: Uploads in uploader groups

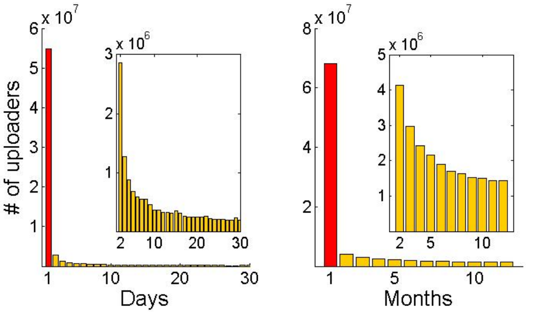

Figure 10: Daily and monthly uploading behaviors (uploaders' life span)

ical distribution of YouTube uploaders. By crawling the individual user profiles of YouTube uploaders seen in our sampled data, extracting and inferring their locations (in terms of countries), we estimate the number of uploaders from each country. We find that the top 20 countries encompass 80% of total uploaders; and uploaders from these countries contribute 82% of the total uploaded videos. The top 10 countries are the same in terms of the number of uploaders and the number of videos they have uploaded (with slightly different ordering): United States, Great Britain, Brazil, Germany, Canada, Japan, France, Spain, Italy, and Mexico. Not surprisingly, about one third of all uploaders come from United States, who collectively contribute more than one third of the total uploaded videos. Japan has the average activeness of 16 uploads per uploader, which is the highest among the 10 countries. The drop in the daily new uploaders during the turbulent period of early 2012 is most severe for United States, but can also be observed in the daily new uploaders from other countries.

5 Plausible Causes for Multiple Growth Phases in YouTube Evolution

To understand what may have contributed to the distinct growth phases in the YouTube video evolution dynamics, in particular, the exponential growth phase starting in late 2009 and the subsequent major interruption in the early part of 2012, in this section we explore a few plausible major factors and events that occur within or coincide with the observed growth phases in YouTube evolution. In particular, we zero in on two plausible major factors or events: i) the widespread popularity and growing adoption of camera-equipped (smart) phones and other mobile devices which make it easy to capture and upload videos to social media sites such as YouTube; and ii) the promulgation of Google's new privacy which forces users to merge their YouTube accounts with other Google accounts(e.g., Gmail or Google+), and explore their effects on the growth dynamics of YouTube.

5.1 Exponential Growth Phase since 2010

Examining the meta-data associated with the sampled videos, we observe more and more videos with a description such as "video uploaded via my mobile phone" starting in 2009. This is perhaps not too surprising, as the same period also sees the widespread popularity and increasing adoption of camera-equipped mobile smart devices such as iPhone and Android phones. In addition, many of these smart phones also come with pre-installed YouTube apps that allow users to upload videos with only a few finger touches. This observation suggests that the exponential growth of YouTube videos may be propelled by the emergence and rapid growth in mobile smart phones. To investigate and confirm this conjecture, we start by manually collecting a list of the most common default signatures generated by devices, applications, and services used to upload a video to YouTube. From this list, we identify a set of signatures indicating that the videos are either captured or uploaded via mobile devices. Examples of such descriptions include "Sent from my iPhone," "This video was uploaded from an Android phone," "Recorded using iVidCam on my iPhone," just to name a few. Note that not all the uploaded videos via these devices (or applications) come with such default signatures necessarily. Finding a sufficiently large number of videos coming with these signatures would enable us to derive a good estimate of (and a lower bound on) the number of videos uploaded by mobile devices.

We in total extracted 77,000 videos in our unbiased sampled video datasets that contain one of the default signatures in our list in the video descriptions. From these, we apply the estimation method presented in Section 3 to estimate the number of videos uploaded daily via mobile devices over time, which is shown in Fig. 7 (together with the total number of daily uploaded videos). Here the y-axes are normalized for ease of comparison. The figure clearly demonstrates that the exponential growth phase of YouTube coincides and is strongly correlated with the emergence and exponential growth of videos captured and/or uploaded via mobile devices. Our analysis provides a strong evidence that the exponential growth phase of YouTube which starts circa late 2009/early 2010 is likely driven by the increasing video uploads by (new) users with mobile devices who can now easily capture videos and upload them.

5.2 Sudden Drops in the Early Part of 2012

We speculate that the sudden drops in the growth of YouTube are caused by the promulgation of Google's new privacy policy, and provide some circumstantial evidences here to support this conjecture. We start by first briefly explaining the Google's new privacy policy [18]. In the new privacy policy announced by Google in Dec 2011, private data collected by one Google service can be shared with its other platforms including YouTube, Gmail and Blogger. Google announced the new setup in order to enable it to tailor search results more effectively, as well as offering better targeted advertising to users. Google also announced that if a user continues to use Google services or YouTube after March 1, 2012, he/she will be doing so under the new privacy policy, and the only option of not accepting the new privacy policy was to close your YouTube account. For this purpose, it also began to force users to merge their YouTube account and Google account and also began asking users to provide additional personal information such as mobile phone numbers. This new privacy policy raised many complains around the world.

As a result of this new privacy policy, when a new user uploads a video to YouTube, she cannot simply create an (anonymous) account and upload it. Instead she has to use her existing Google account (or create one if she does not have one). An immediate and direct impact of this new privacy policy can be seen in the number of new uploaders (as measured in terms of the unique user account id's) experience similar sudden drops during the same turbulence period in the early part of 2012. But unlike the growth rate of the daily uploaded videos which resumes the same exponential growth afterwards, we see that the growth rate of the daily new uploaders never fully recovers, as discussed earlier in Section 4. Furthermore, when examining the possible impact of Google's new privacy policy on the categories of videos being uploaded during the turbulence period, we see that the top categories that experience most visible drops are those that are most likely associated with new, casual users who upload a few (user-generated) personal videos, such as *People & Blogs*, *Music* and *Comedy* (see Fig. 5). That the top category *People & Blogs* experiences the most pronounced drops is particularly telling, as it also serves as the default category when uploaders do not bother to select a category (we believe that these uploaders are more often "casual" new uploaders). More interestingly, the new privacy policy also has a major impact on the number of videos uploaded by mobile devices, as shown in Fig. 7. All of these point to Google's new privacy policy as the plausible culprit for the sudden drops in the number of total uploaded videos during the turbulent first part of 2012. That the total number of daily YouTube video uploads resumes its exponential growth afterwards (while the number of daily new uploaders does not) suggests that the sustained growth of YouTube videos in the past year or so is more likely due to the increasing number of uploaders by existing "heavy" uploaders (e.g., YouTube partners) or more active "new" uploaders who have joined the system in more recent years (see Sections 6 and 7 for further discussions).

6 Uploaders and Uploading Behaviors

We saw in Section 4 that the overall growth in number of uploaders has had a significant impact on YouTube video growth. To further analyze what contributes to the growth of YouTube videos, we conduct an in-depth study of the uploading behavior of users. In particular, we group uploaders based on the year that they upload their first YouTube video, and study how the activeness of these groups has changed over time and how it has affected the growth of YouTube videos. Furthermore, we propose a model which describes (and predicts) the YouTube growth in terms of number of uploaders and their uploading behavior.

6.1 New vs. Existing Uploaders

We first investigate how *cumulatively* the average number of uploaded videos *per uploader* evolves over time. Let $t \geq 1$ denote the number of days since 06/15/2005 (the inception day of YouTube). Define $N_c(t)$ as the cumulative number of videos that have been uploaded to YouTube up to t (inclusive), and $n_c(t)$ is the total number of uploaders up to t. Define $A_c(t) = N_c(t)/n_c(t)$, which we refer to as the *cumulative activeness* in terms of upload per uploader; $A_c(t)$ reflects in a sense how "active" an "average" uploader who has joined the system by time t is. We plot $A_c(t)$ in Fig. 8 over time, with the x-axis labeled by the actual dates, which shows a clear increasing trend over time. This indicates that an "average" uploader uploads more videos over time, for instance, from about 2 videos by the end of 2005 to about 8 videos by June 2013. However, such increased "activeness" levels may be due to the fact that new uploaders who recently joined YouTube are more active in video uploading, or that existing uploaders who joined the system earlier upload an increasing number of videos over time, or even both. To test these hypotheses, we conduct further analyses which are discussed next.

In order to distinguish the activeness of "new" vs. "existing" uploaders and study the impact of their activeness levels on the growth of YouTube videos, we divide the YouTube uploaders into groups based on the years when they joined YouTube. Since we have complete 8 years of YouTube data from 2005-06-15 to 2013-06-15, we group uploaders into eight groups as shown in Table 1. For $i = 1, \ldots, 8$ (with $i = 1$ representing "grp2005" and $i = 8$ representing "grp2012"), we use $N_i(t)$ to denote the number of videos uploaded by uploaders in group i at time t. Fig. 9 plots $N_i(t)$ for each of the eight groups, together with the overall daily uploaded videos over time. It shows how different groups of uploaders have contributed in the total uploads and how this contribution has evolved over time. Furthermore, by defining $A_{ci}(t) = N_{ci}(t)/n_{ci}(t)$ as the cumulative activeness of group i at time t, we analyze and evaluate the activeness of the groups over time (the marks in Fig. 14 show $A_{ci}(t)$ for coarser granularity

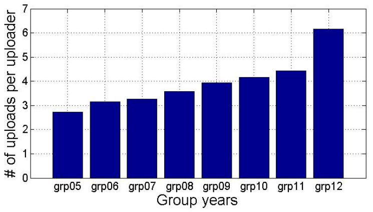

Figure 11: First year uploading behavior

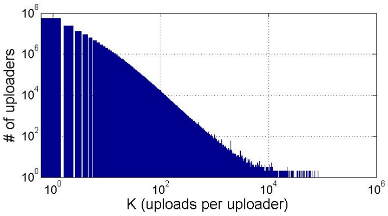

Figure 12: distribution of uploaders over number of uploads

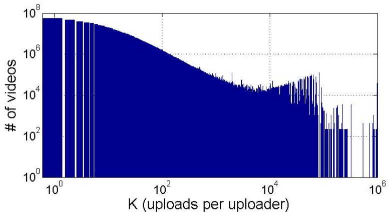

Figure 13: distribution of videos over number of uploads

Table 1: Uploader Groups

Group	Joining Date
grp2005	2005-06-15 to 2006-06-14
grp2006	2006-06-15 to 2007-06-14
grp2007	2007-06-15 to 2008-06-14
grp2008	2008-06-15 to 2009-06-14
grp2009	2009-06-15 to 2010-06-14
grp2010	2010-06-15 to 2011-06-14
grp2011	2011-06-15 to 2012-06-14
grp2012	2012-06-15 to 2013-06-14

of t to be years). From these two analyses we make the following two key observations: (i) *increasing (average) activeness of new uploaders* who joined YouTube in more recent years than earlier years; and (ii) *decreasing (average) activeness of existing uploaders* in the sense that the average activeness of uploaders in all groups decreases over time. Based on these observations, the first aforementioned hypothesis is confirmed and the second one is rejected.

We further illustrate the observation (i) above by comparing the activeness of the eight groups in their respective first year. Fig. 11 indicates that uploaders from more recent groups are more active than those from groups of early years. To better understand the uploading behaviors of uploaders in various groups, we compute the *life span* of each uploader, which is the time difference between the last and first video uploads crawled from the uploader's profile. (For this, we only consider uploaders who joined YouTube before June 2012, namely, those in groups 1-7; for uploaders in group 8, we may not be able to estimate their life spans accurately, as many of them may have joined the system less than a year.) Fig. 10 shows the distributions of "life spans" of uploaders in groups 1-7 in two time scales (days and months). From the left sub-figure, we see that a large number of uploaders have a life span less than a month, with a large majority of them having a life span of only 1 day; the inset shows the "magnified" version of the distribution with the uploaders of 1 day life span removed. (In Section 7 we will refer to the uploaders with a life span less than a week and uploading only a few videos as the "pop-up" uploaders.) The right sub-figure shows the life span distribution of uploaders in the monthly time scale, where the inset shows the "magnified" version of the distribution with the uploaders of 1 month life span removed. The life span distribution of uploaders in the first 7 groups explains why the average activeness of each group decreases significantly over time, as noted in the observation (ii) above. All in all, our analysis shows that new uploaders are driving force behind the growth dynamics of YouTube video uploads; in particular, new uploaders who join YouTube in more recent years tend to upload more videos, which contribute to the exponential growth in the number of daily uploaded videos we have observed.

6.2 Predicting YouTube Growth

In this section we develop a machine learning model to predict the YouTube video growth dynamics based on the observed uploading behaviors of the uploaders who joined the system in the previous years. In this model, we first find a pattern describing the uploading behavior (cumulative activeness) of different groups (generations) of the uploaders over time; then by estimating the size of each group, we can predict the number of uploads for the future years. Such model can be very useful in many applications such as helping the user generated content service site administrators with capacity planning, and the advertisers with marketing decisions.

Our earlier analysis shows that across all groups of uploaders, the uploading behavior in terms of (cumulative) average uploads per uploader (activeness) decays significantly over time. Further analysis reveals that the rate of decay is exponential and can be approximated and modeled by a general function of the form $(ai+b)(e^{-ct}+d)$ for group i. We have found that the rate of decay c is almost the same for all the groups, and the groups differ primarily in the multiplicative scaling factor $ai+b$. We perform a non-linear regression on the time series of the cumulative activeness for each group to learn the parameters of this model, namely, a, b, c, d. Furthermore, we take the data from the first 6 groups (grp2005-grp2010) in the first 6 years (2005-2010) as the train datasets, and the data of these 6 groups in the 7-th year (2011) as well as the data of the 7-th group (grp2011) as the test datasets. We apply the *iterative non linear least squares* method [19] on the train datasets to learn the parameters for the cumulative activeness of the first 6 groups in the first 6 years, which yields the following *learned model:*

$$(0.15i+2.15)(e^{-0.85t}+0.30) \tag{9}$$

The training error is $RMSE = 0.14$, where RMSE stands for Root Mean Square Error. Testing the learned model in eq. (9) on the test datasets, we find that the predicted results are highly precise, with an error as small as $RMSE = 0.15$. Fig. 14 shows the cumulative activeness (the markers) and our proposed model (the lines) for different i's (group indicator) over t (year indicator). It can be seen that the proposed model fits to the data points (cumulative activeness) accurately. The dashed line represents the envelop, which is the $(ai+b)$ part of the model; it indicates how the scaling factor of the exponential functions grows for the more recent years.

Let Z be a lower triangular matrix representing the cumulative activeness of groups over years, whose Z_{ti} entry indicates the cumulative activeness of group i in year t. Let x be a vector representing the group size (i.e., the number of uploaders in each group) and y be the vector of total uploads in each year. Vector y can be obtained from Z and x using the following relation: $y = Zx$. Given the entries Z_{ti}'s for $i = 1,\ldots,6$ and $t = 0,\ldots,5$, we can apply the proposed model discussed above (which resulted in eq. (9)) to predict the cumulative activeness of these 6 groups in the 7-th year (2011), namely Z_{7i}'s for $i = 1,\ldots,6$, as well as the cumulative activeness of the new group (grp2011) in 2011, namely Z_{77}. These entries together allow us to predict the total number of uploaded videos by all uploaders in 2011, i.e. y_7, given we know the number of new uploaders who joined in 2011. Applying this method, we can precisely predict the total number of videos uploaded in 2011 with a relative error of only 0.0025.

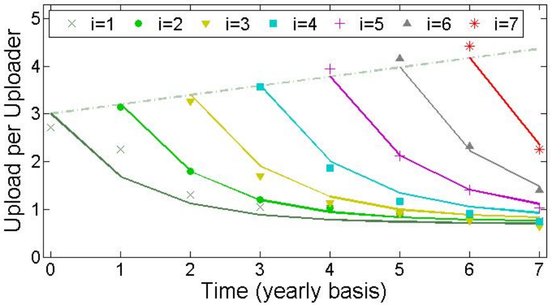

Figure 14: The proposed model in eq. (9) for cumulative activeness of different groups over years

7 "Heavy" vs. "Light" Uploaders

In this section we investigate the contribution of uploaders with varying number of uploads. Fig. 12 shows the distribution of uploaders based on their number of uploads. Notice that the group of uploaders with only one uploaded video has the maximum size among all the other groups. Moreover, the majority of YouTube uploaders are with relatively small number of uploads: about 82% of the uploaders uploaded less than the average uploads, 7.7 videos.

Fig. 13 shows the distribution of number of videos over number of uploads, representing the contribution of uploaders with varying upload size in the YouTube total uploads. From this figure we can infer that the contribution of uploaders with less than 7.7 uploads is only 24% of total uploaded videos. On the other hand, those uploaders with larger than 7.7 uploads have population of only 18% of total uploaders and contribute about 76% of YouTube uploads. The turning point in Fig. 13 around 5000 uploads shows that although the number of uplaoders with larger uploads (than 5000) is smaller than those with less uploads, their upload excess compensates for their smaller population and causes the contribution of these uploaders to be larger.

To further distinguish and evaluate the contributions of users with varying size of uploads, we consider two extreme groups: pop-up uploaders and YouTube partners. We define the pop-up uploaders to be those group of uploaders with life span shorter than a week and total uploads smaller than 7 uploads (less than the average of upload). YouTube partners are the uploaders who satisfy a certain conditions of YouTube, like having large view counts, and as a payoff, they are given the option to display ads on their videos and monetize the content [20]. Fig. 15 and Fig. 16 compare the contribution of these two groups in total number of uploaders and total uploads over the years. We observe that although the portion of pop-up uploaders increases over time, their contribution portions to uploaded videos decrease. In contrast, YouTube partners' contribution to uploaded videos has been growing significantly over time, although they form a very small portion of uploaders (approx 0.008). We observe that at early stage, pop-up uploaders dominate the contribution to the video uploads. But over time the contribution of YouTube partners is approaching the contribution of pop-up uploaders who form the majority of YouTube uploaders. This analysis suggests that the growth of YouTube increasingly relies on - and will likely be sustained by - the continued contributions of existing "heavy" uploaders, as the growth in the number of new uploaders slows significantly. In other words, YouTube is migrating from being driven by many people but small contribution each, to being driven by more professional uploaders and higher contributions.

8 Conclusion

In this paper, we measured and analyzed the growing dynamics of YouTube from its inception in 2005 up until now, which is the first of its kind in studying the evolution dynamics of YouTube to the best of our knowledge. We showed that the growth of YouTube videos undergoes several phases, starting with a quadratic growth followed by an exponential growth late 2009, and is interrupted by a sudden drop early 2012, and resumes the rapid growth phase after a few months again. We suggested two plausible factors, the ubiquity of mobile devices and Google's changing privacy policy, as the major contributors to the distinct growth phases of YouTube. By investigating the cumulative uploading behavior of users, we demonstrated that the higher activeness of more recent uploaders along with the increase of new joining uploaders over time have significantly contributed the growth of YouTube; we developed a model to predict the growth of YouTube videos in terms of these two factors. Our further analysis on two groups of uploaders, pop-up uploaders and YouTube partners, evidenced that YouTube is migrating from being driven by many people but small contribution each, to being driven by more professional uploaders and higher contributions.

Acknowledgment

This research was supported in part by DTRA grant HDTRA1-09-1-0050, DoD ARO MURI Award W911NF-12-1-0385, and NSF grants CNS-10171647 and CNS-1117536.

References

[1] Y. Borghol, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "The untold story of the clones: Content-agnostic factors that impact youtube video popularity," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1186–1194.

[2] T. Elvers and P. Srinivasan, "What's trending?: mining topical trends in ugc systems with youtube as a case study," in *Proceed-*

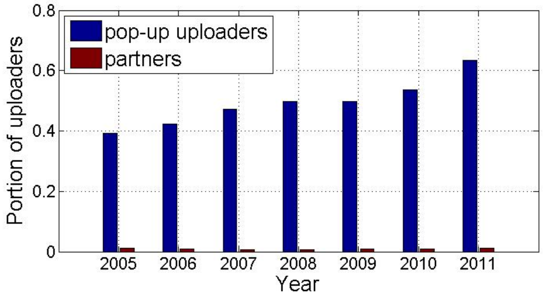

Figure 15: Contribution of pop-up uploaders and partners in total uploaders over years

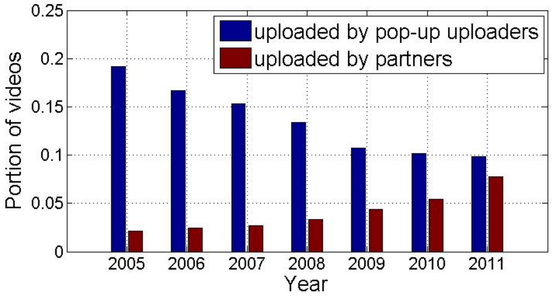

Figure 16: Contribution of pop-up uploaders and partners in total videos over years

ings of the Eleventh International Workshop on Multimedia Data Mining. ACM, 2011, p. 4.

[3] L. Xie, A. Natsev, J. R. Kender, M. Hill, and J. R. Smith, "Visual memes in social media: tracking real-world news in youtube videos," in *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 2011, pp. 53–62.

[4] F. Figueiredo, F. Benevenuto, and J. M. Almeida, "The tube over time: characterizing popularity growth of youtube videos," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 745–754.

[5] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini, "A peek into the future: Predicting the evolution of popularity in user generated content," in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 607–616.

[6] A. Brodersen, S. Scellato, and M. Wattenhofer, "Youtube around the world: geographic popularity of videos," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 241–250.

[7] Y. Ding, Y. Du, Y. Hu, Z. Liu, L. Wang, K. Ross, and A. Ghose, "Broadcast yourself: understanding youtube uploaders," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 2011, pp. 361–370.

[8] P. Spathis and R. A. Gorcitz, "A data-driven analysis of youtube community features," in *Proceedings of the 7th Asian Internet Engineering Conference*. ACM, 2011, pp. 12–18.

[9] J. Zhou, Y. Li, V. K. Adhikari, and Z.-L. Zhang, "Counting youtube videos via random prefix sampling," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 2011, pp. 371–380.

[10] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 1–14.

[11] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: a view from the edge," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 15–28.

[12] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of youtube network traffic at a campus network–measurements, models, and implications," *Computer Networks*, vol. 53, no. 4, pp. 501–514, 2009.

[13] V. K. Adhikari, S. Jain, Y. Chen, and Z.-L. Zhang, "Vivisecting youtube: An active measurement study," in *INFOCOM, 2012 Proceedings IEEE*. IEEE, 2012, pp. 2521–2525.

[14] V. K. Adhikari, S. Jain, and Z.-L. Zhang, "Youtube traffic dynamics and its interplay with a tier-1 isp: an isp perspective," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 2010, pp. 431–443.

[15] A. Finamore, M. Mellia, M. M. Munafò, R. Torres, and S. G. Rao, "Youtube everywhere: Impact of device and infrastructure synergies on user experience," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 2011, pp. 345–360.

[16] J. Rice, *Mathematical statistics and data analysis*. Cengage Learning, 2006.

[17] I. Hacking, *An introduction to probability and inductive logic*. Cambridge University Press, 2001.

[18] googleblog.blogspot.com, "Google new privacy policy in jan 2012," *http://googleblog.blogspot.com/2012/01/updating-our-privacy-policies-and-terms.html*.

[19] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2004.

[20] support.google.com, "What is the youtube partner program?" *https://support.google.com/youtube/answer/72851?hl=en*.