The Lifecyle of a Youtube Video: Phases, Content and Popularity

Honglin Yu, Lexing Xie, Scott Sanner
Australian National University, NICTA
Canberra, Australia
{honglin.yu, lexing.xie}@anu.edu.au, scott.sanner@nicta.com.au

Abstract

This paper proposes a new representation to explain and predict popularity evolution in social media. Recent work on social networks has led to insights about the popularity of a digital item. For example, both the content and the network matters, and gaining early popularity is critical. However, these observations did not paint a full picture of popularity evolution; some open questions include: what kind of popularity trends exist among different types of videos, and will an unpopular video become popular? To this end, we propose a novel phase representation that extends the well-known endogenous growth and exogenous shock model (Crane and Sornette 2008). We further propose efficient algorithms to simultaneously estimate and segment power-law shaped phases from historical popularity data. With the extracted phases, we found that videos go through not one, but multiple stages of popularity increase or decrease over many months. On a dataset containing the 2-year history of over 172,000 YouTube videos, we observe that phases are directly related to content type and popularity change, e.g., nearly 3/4 of the top 5% popular videos have 3 or more phases, more than 60% *news* videos are dominated by one long power-law decay, and 75% of videos that made a significant jump to become the most popular videos have been in increasing phases. Finally, we leverage this phase representation to predict future viewcount gain and found that using phase information reduces the average prediction error over the state-of-the-art for videos of all phase shapes.

1 Introduction

How did a video become viral? – this is one of the well-known open research questions about social media and collective online behavior. An online information network is known to have bursts of activities responding to endogenous word-of-mouth effects or sudden exogenous perturbations (Crane and Sornette 2008). A number of studies revealed that a video's long-term popularity is often determined, and can be predicted from its early views (Cheng, Dale, and Liu 2008; Szabo and Huberman 2010; Pinto, Almeida, and Gonçalves 2013), and that early-mover's advantage exists in the competition for attention (Borghol et al. 2012). Recently different groups of researchers studied the relationship between content popularity and various factors, including network actor properties (Cheng et al. 2014), content features (Cheng et al. 2014; Bakshy et al. 2011), and effects of complex contagion (Romero, Meeder, and Kleinberg 2011), among many others. However, some questions remain: what does a video's lifecycle look like? Is there one, or multiple endogenous or exogenous shocks?

One well-known model of social media popularity was proposed by Crane and Sornette (2008), in which the observed popularity over time consists of power-law precursory growth or power-law relaxations. Such rising and falling power-law curves are indeed observed in large quantities – Figure 1(a) and (b) contains one example of each type, respectively. Note however, that real popularity cycles are rather complex – a video can go through multiple *phases* of rise and fall, as shown in Figure 1(c) and (d).

To address such limitations, we propose a novel representation, popularity *phases*, to describe the rich patterns in a videos lifecycle. We propose a method to jointly segment *phases* from the popularity history and find the optimal parameters to describe their shapes. We present statistical descriptions for 172K+ videos over 2 years, measuring their phases, content types, and popularity evolution. We found that the number of phases is strongly correlated to a video's popularity – nearly 3/4 of the top 5% popular videos have 3 or more phases, whereas only 1/5 of the least popular videos do; different content categories (e.g. *news*, *music*, *entertainment*) exhibit very different phase profiles – more than 60% *news* videos are dominated by one long power-law decay, whereas only 20% of music videos do. Overall, this work unveils a rich and multi-faceted view of popularity dynamics – consisting of successive rising and falling phases of collective attention and their close relationship to content types and popularity. Although our focus is on YouTube videos – one of the few sources where the popularity history is publicly available – the method for extracting phases and observations about viral content are potentially applicable to other, similar media content.

The main contributions of this work are as follows:

- We propose *phases* as the new description for the bursty popularity lifecycle of a video, and present a method to extract phases from popularity history – i.e., simultaneous segmentation and recovery of their power-law shapes

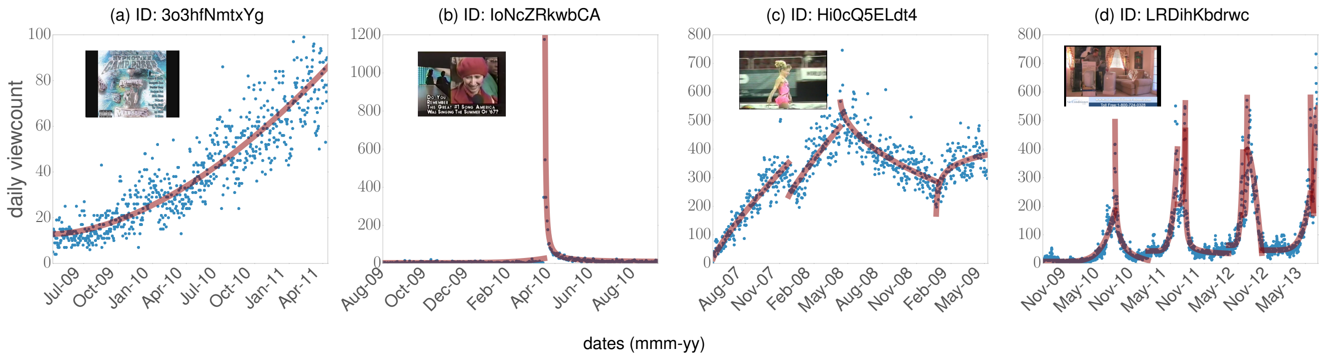


Figure 1: The complexity of viewcount dynamics: the lifecycles of four example videos. Blue dots: daily viewcounts; red curves: phase segments found by our algorithm in Section 3. (a) A video with one power-law growth trend. (b) A video with one power-law decay. (c) A video with many phases, including both convex and concave shapes – this video contains a Gymnastic performance. (d) A video with seemingly annual growth and decay – this video demonstrates how to vent a portable air-conditioner, and reaches viewcount peaks during each summer. Viewcount shapes such as (a) and (b) are explained by Crane and Sornette's model, but (c) and (d), and many more like them, are not.

without needing to pre-determine the number of phases (Section 3).

- We present a large-scale measurement study of *phases* and long-term popularity on hundreds of thousands of videos (Sections 2 and 4). We directly relate phases to popularity, content types, and the evolution of popularity over time.
- We learn predictive models of future popularity using the phase representation – this method out-performs prior methods across videos with all types of lifecycles (Section 5).
- We publicly release the YouTube popularity dataset and the software for phase segmentation online[1].

2 A dataset of long-term video popularity

We construct a dataset containing a large and diverse set of YouTube videos using Twitter feeds. We extract video links from a large Twitter feed (Yang and Leskovec 2011) of 184 million Tweets from June 1st to July 31st in 2009, roughly $20-30\%$ of total tweets in this period. We extracted URLs from all tweets and resolved shortened URLs, retaining those referring to YouTube videos. This yields 402,740 unique YouTube videos, among which 261,391 videos are still online and having their meta-data publicly available. We remove videos that have less than 500 views in its first two years (not enough views for meaningfully extracting phases), our final dataset includes 172,841 videos.

While recent other YouTube datasets were constructed with standard feeds ("most recent", "most popular", "deleted") (Figueiredo, Benevenuto, and Almeida 2011; Pinto, Almeida, and Gonçalves 2013), within-category search (Cha et al. 2007), text search (Xie et al. 2011), or trying random video IDs (Pinto, Almeida, and Gonçalves 2013), constructing a Twitter-driven YouTube dataset will not be biased to the most popular videos, nor will it be biased towards a small list of topics or keywords. Moreover, this approach will mostly return videos that received more than a minimum amount of attention, assuming people who tweet the video likely watched it. Studying videos that are at least a few years old provides a long enough history to observe the different popularity phases. Choosing videos that are mentioned in (a random sample of) Twitter will yield a set of videos covering diverse topics. Furthermore, discussions that happened on Twitter naturally engenders both endogenous and exogenous evolution of popularity.

For each video v, we obtain from YouTube API[2] its meta-data such as *category, duration, uploader* as well as its daily viewcount series, denoted as $\mathbf{x}_v = [x_v(1), \ldots, x_v(T)]$. We present the analysis for these videos up to two years of age, i.e., T=735 days. Compared to related recent work, this dataset is notable in two aspects. In terms of data resolution, most prior work use a 100-point interpolated cumulative viewcount series over the lifetime of a video (Figueiredo, Benevenuto, and Almeida 2011; Ahmed et al. 2013; Borghol et al. 2012; Yu, Xie, and Sanner 2014). this dataset is one of the first to contain fine-grained history of daily views. In terms of the time span, recent work examines popularity history during a video's first month (Szabo and Huberman 2010; Pinto, Almeida, and Gonçalves 2013; Abisheva et al. 2014) or up to 1 year (Crane, Sornette, and others 2008), this dataset is also the first to enable longitudinal analysis over multiple years.

Table 1 summarizes the number of unique videos per user-assigned category in this dataset. We can see that *music* videos are the most tweeted, 7 categories (until *sports*) have more than 7,500 (or 5%) unique videos, and 15 categories (until *animals*) have more than 1,700 (or 1%) unique videos. The categories *movies, shows* and *trailers* are at least an order of magnitude less frequent than other categories, likely resulting from a change in YouTube category taxonomy – these 435 videos are excluded from statistics across categories in Section 4 and later.

[1]http://yuhonglin.github.io

[2] https://developers.google.com/youtube/

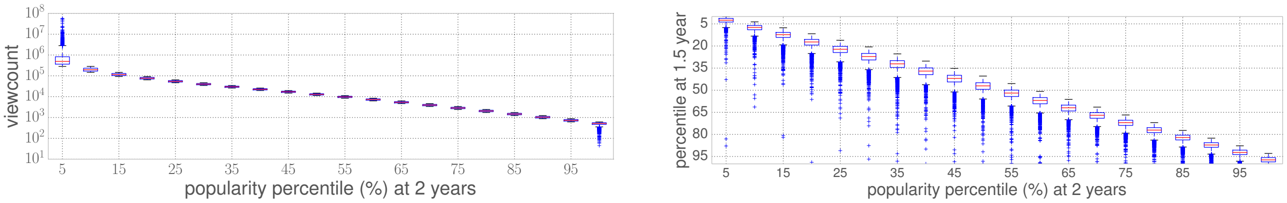


Figure 2: Left: Boxplots of video viewcounts at $T = 735$ days, for *popularity percentiles* quantized at 5%, or 8000+ videos each. Viewcounts of the 5% most- and least- popular videos span more than three orders of magnitude, while videos in the middle bins (from 10 to 95 percentile) are within 30% views of each other. Right: The change of popularity percentile from 1.5 years (y-axis, from 0.0% to 100.0%) to 2 years (x-axis, in 5% bins). While most videos retain a similar rank, videos from almost any popularity at 18 months of age could *jump* to the top 5% popularity bin before it is 24 months old (left most boxplot).

Table 1: The number of videos broken down by user-assigned categories. We can see that *Music* videos are the most-tweeted (64,152 unique videos), over twice as many as *Entertainment* (26,622) and over five times as many as *News* (10,429). 15 distinct categories (from *Music* to *Animals*) have more than 1,700, or 1% of all videos.

Category	#videos	Category	#videos
Music	64096	Howto	4357
Entertainment	26602	Travel	3379
Comedy	14616	Games	3299
People	12759	Nonprofit	2672
News	10422	Autos	2398
Film	8356	Animals	2375
Sports	7872	Shows	407
Tech	4626	Movies	15
Education	4577	Trailers	13
Total number: 172841			

We rank all videos by the total viewcounts they receive at age t-days, i.e. $sum([x_v(1), \ldots, x_v(t)])$, for $v = 1, \ldots, |V|$. The rank for each video is converted to a percentile scale, i.e. video v at 1% will be less popular than exactly 1%, or ∼1720 other videos in the collection. We quantize this percentile into *bins*, each of which contains 5%, or ∼8,642 videos. Figure 2(left) shows a boxplot of video viewcounts in each bin after $T = 735$ days. We can see that viewcounts of the 5% most popular (leftmost bin) and least popular (rightmost bin) videos span more than three orders of magnitude. The rest of the collection shows a linear trend of rank versus log-viewcount, and videos in the same bin are within 30% views of each other.

In Figure 2(right) we explore the change of popularity from 1.5 years (y-axis) to 2 years (x-axis). While most videos retain a similar rank, video from any bin can *jump* to the top popularity bucket in 6 months (as seen in the leftmost boxplot). One asks- how did these videos go viral? We will present some observations in Section 4.2.

3 Detecting popularity phases

We define a *phase* as one continuous time period in which a video's popularity has a salient rising or falling trend. In this section we present a model to describe such phases, and propose efficient algorithms to simultaneously find both phase segments and their shape parameters from a time series.

Given the daily viewcount for video v: $\mathbf{x}_v = x_v[1 : T]$, the goal is to segment this time series as a set of successive *phases* ρ_v, where each phase $\rho_{v,i}$ is uniquely determined by its starting time $t^s_{v,i}$, with $1 = t^s_{v,1} < t^s_{v,2} < \ldots < t^s_{v,n} < T$. In the rest of this section, we omit subscript v since the segmentation algorithm works on each video independently. For convenience, we include in ρ_i its ending time as t^e_i.

$$t^e_i = t^s_{i+1} - 1, \text{ if } i < n; \;\; t^e_i = T, \text{ if } i = n.$$

3.1 Generalized power-law phases

We use a generalized power-law curve to describe viewcount evolution for a phase of length $\bar{T}$:

$$x[t] = at^b + c, \; t = 1, 2, \ldots, \bar{T} \tag{1}$$

with a power-law exponent b, scale a and shift c. The power-law shapes are suitable for describing general popularity evolutions for the following reasons: (1) They can result from epidemic branching processes (Sornette and Helmstetter 2003) with power-law waiting times (Crane and Sornette 2008). (2) Such generalized power-law shape is sufficiently expressive for describing a wide range of monotonic curves that are either accelerating or decelerating in their rise (or fall). A change in rising/falling or acceleration/deceleration indicate either an external event or a changed information diffusion condition, hence they are conceptualized as different phases. (3) The optimal fit is efficiently computable, as described in Section A.1. Note that the proposed power-law shape generalizes popularity model by (Crane and Sornette 2008) – there $a = 1$, $c = 0$, and b is in the range of $[-1.4, -0.2]$. In particular, our generalized power-law model addresses two crucial aspects for capturing real-world popularity variations: the first is to account for multiple peaks in the same video's lifetime, potentially generated by a number of exogenous or endogenous events of different strengths – hence varying a; the second is to account for different background random processes that are super-imposed onto the power-law behavior – hence varying c. Our model will rely on the phase-finding algorithm to determine a, b and c from observations. We also allow two temporal directions in Equation (1), in order to capture all monotonically accelerating or decelerating power-law shapes (Table 2), i.e.,

$$x[\tau] = a\tau^b + c, \text{ with either} \quad (2)$$
$$\tau = t, \text{ denoted as } \rightarrow, \text{ or}$$
$$\tau = \bar{T} - t, \text{ denoted as } \leftarrow .$$

3.2 The phase-finding problems

Given the daily viewcount series $\mathbf{x}[1 : T]$, the PHASE-FINDING problem can be expressed as simultaneously determining the parameter set $\mathcal{S}$ for a phase segmentation $\{\rho_i, 1 \leq i \leq n\}$ with n being the number of phases, and the optimal phase parameters $\{\theta_i, 1 \leq i \leq n\}$:

$$\begin{aligned} \text{Find } & \mathcal{S} = \{n;\ t_i^s, \theta_i,\ i = 1, \ldots, n\} \\ \text{to minimize } & E\{\mathbf{x}_{1:T}, \rho_{1:n}, \theta_{1:n}\} \\ = & \sum_{i=1}^{n} E_i\{x[t_i^s : t_i^e], \theta_i\} . \end{aligned} \quad (3)$$

A sub-problem of the PHASE-FINDING problem is to find the optimal phase shape parameters θ_i^* for a given starting and ending time t_i^s, t_i^e of a segment, called the PHASE-FITTING problem. This is done by minimizing a loss function $E_i\{\cdot,\cdot\}$ between the observed and fitted volumes as shown in problem (4). Here the parameter set of the generalized power-law is: $\theta = [a, b, c, \tau]^T$.

$$\text{given } t_i^s, t_i^e, \quad \text{find } \theta_i^* = \arg\min_{\theta_i} E_i\{x[t_i^s : t_i^e], \theta_i\} . \quad (4)$$

3.3 Solution Summary

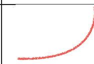
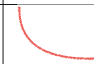
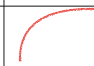
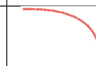
For the PHASE-FITTING problem (4), we minimize the sum-of-squares loss between the observations and the fitted sequence. Our solution includes a technique called *variable projection* to reduce the search space, and an initialization strategy especially suited for power-law fitting. We validate this component using synthetic data, and find the algorithm is able to recover the original parameters across a broad range of curve shapes. Details are in Section A.1.

We solve the PHASE-FINDING problem (3) by embedding the PHASE-FITTING algorithm in a dynamic programming setting to jointly find the best sequence segmentation and power-law parameters. We use validation data to tune the trade-off between fitting error and the number of phases, the fitting generally works well, see examples in Figures 1, 8 and on project website[1]. This solution overcomes the limitations of not accounting for more than one phase with arbitrary timing and shape (Crane and Sornette 2008), and enables us to examine the persistence of popularity trends, as well as the evolution history of viral videos. Details are in Section A.2.

3.4 Four types of phases

We intuitively categorize power-law phases into four types, according to whether the trend over time is increasing or decreasing, and whether the rate of change is accelerating or decelerating. These types are intuitively named as convex/concave curves that are either increasing or decreasing.

Table 2: Four types of phase shapes and their basic statistics. See Section 3.4 for notations and discussions.

Phase-type	Convex increasing	Convex decreasing	Concave increasing	Concave decreasing
Shorthand	**vex.inc**	**vex.dec**	**cav.inc**	**cav.dec**
Sketch				
Parameter $(a; b; \tau)$	+; >1; $\rightarrow$ +; <0; $\leftarrow$ −; [0,1]; $\leftarrow$	+; <0; $\rightarrow$ −; [0,1]; $\rightarrow$ +; >1; $\leftarrow$	+; [0,1]; $\rightarrow$ −; <0; $\rightarrow$ −; >1; $\leftarrow$	−; >1; $\rightarrow$ +; [0,1]; $\leftarrow$ −; <0; $\leftarrow$
Phase count	172,329 (30.6%)	286,070 (50.8%)	67,862 (12.0%)	37,363 (6.6%)
Length (days)	3.0×10^7 (23.7%)	8.2×10^7 (64.6%)	1.0×10^7 (8.0%)	4.6×10^6 (3.7%)
Views	3.5×10^9 (28.2%)	5.8×10^9 (46.8%)	2.2×10^9 (17.4%)	9.6×10^8 (7.6%)

See the shape sketches in Table 2. Furthermore, each type is uniquely identified by three parameters: the power-law scaling factor $a > 0$ or < 0, short-handed as $+/-$; exponent b being < 0, within $[0, 1]$, or > 1; and the temporal direction of τ as in Equation (2), short-handed as $\rightarrow$ or $\leftarrow$.

We segmented phases for all 172K+ videos using the PHASE-FINDING algorithm described in Section 3. There are 563,624 phases in total, with an average of 3.3 phases per video. Table 2 presents a profile of these shapes. We can see that roughly half of the segments are convex-decreasing – these phases span more than 60% of the duration and account for less than half of the view counts. Convex-increasing is the second-most common shape, accounting for another 30% segments, while concave-decreasing is the least common.

4 Phase statistics

In this section, we will first examine phase statistics with respect to content category and popularity, and then discuss observations about how phases (and popularity) evolve over time.

4.1 Phases, popularity, and content types

How many phases does a video have? Figure 3(a) breaks down videos in each popularity bin by the number of phases they contain, and Figure 3(b) does so for each content category. We can see that among the top 5% most popular videos, more than 95% have more than one phase, and about 45% of them have four or more phases. We observe a general trend of more popular videos having larger number of phases (hence more complex lifecycle). Across different content categories, over 70% of news videos have only one or two phases, whereas videos related to art and entertainment (*music, comedy, animal, film, entertainment*) have the most complex life-cycles. Intuitively, the need for consuming a *news* item decreases drastically after a few days, while arts and entertainment content not only retains interest over time, but is also suitable for re-consumption.

How many increasing and decreasing phases? Figure 3(c)(d) report the fraction of each of the four types of phases (Section 3.4) found in each popularity bin and each

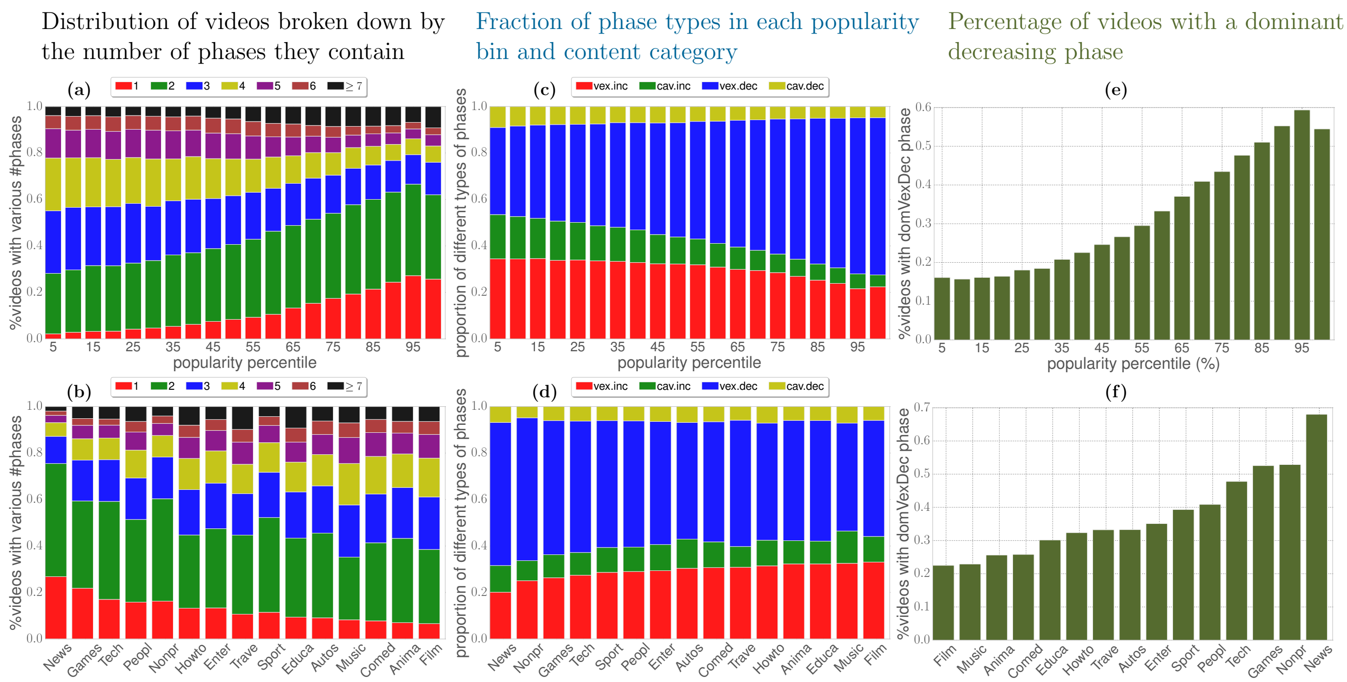


Figure 3: Left: Percentage of videos broken down by the number of phases they have, over (a) popularity percentile and (b) content categories. Middle: Percentage of the four phase types, broken down by (c) popularity percentile and (d) content categories . Right: Percentage of videos with a dominant convex-decreasing phase ($\geq 90\%T$), broken down by (e) popularity percentile and (f) content categories. A general trend is that popular videos and entertainment content (e.g. *music* videos) have more phases overtime, and more than half of *news* videos and the least popular videos have one dominant decreasing phase. See Section 4.1 for discussion.

category, respectively. Overall, popular videos have more increasing phases (both convex and concave, 53.5%, see Figure 3(c)), with this ratio decreasing to 27.5% for the least popular videos. Across different content categories, *news* has the least number of increasing phases, while entertainment and instructional videos such as *music*, *howto* and *autos* have the most increasing categories ($\geq$ 42%). This is also explained by the persistent consumption value of entertainment and how-to videos, e.g., recall the viewcount periodicity of the air-conditioner venting video in Figure 1(d).

Do videos revive from an initial exogenous shock? We further examine videos that have a *dominant* convex-decreasing phase, with $t^e - t^s \geq 0.90T$. These videos typically receive a burst of attention from an exogenous shock (e.g. News), and ceased to attract further attention, such as Figure 1(b). Figure 3(e)(f) plot the fraction of videos characterized by a dominant decreasing phase, for each popularity bin and content category, respectively. We can see that more than 60% of *news* videos have a dominant decreasing phase, in other words, more than half of news videos do not start a new phase after a main attention shock. On the other hand, only ∼20% of *film* and *music* contain the dominant decreasing phase, with the remaining 80% enjoying “revival” of attention over their life-cycles. Moreover, over 50% of the least popular videos have a dominant decreasing phase, while only 15% of the most popular ones do. This also shows the inherent uncertainty of popularity: despite having one long decreasing phase, 0.75% of all, or ∼1275 videos still made it to the top 5% in the popularity chart.

4.2 Phase evolution over time

How long do phases last? Figure 4 examines the distribution of phase durations, broken down into increasing and decreasing phases, with popularity and category as co-variates. In Figure 4(a), we can see that popular videos tend to have longer increasing phases, while the increasing phases for videos in the least popular bins tend to be short. In Figure 4(b), while there is a fair amount of long ($\geq$ 160 days) decreasing phases across the entire popularity scale, the least popular videos are still the most likely to have a long and dominant decreasing phase, this is consistent with Figure 3(e). In (c) and (d), on the other hand, we can see that the probability of having longer phases of either type spread over different categories. With *music* slightly more likely to have longer increasing phases than other categories, and *news* notably more likely to have a decreasing phase lasting more than 320 days, consistent with Figure 3(f).

Are older videos *forgotten*? Since new phases tend to be triggered by external events, one may ask whether there are less activities and attention on older videos – in other words, are they *forgotten*? Surprisingly, the data says no. Figure 5 plots the number of new phases that commence over the age of a video (red curves broken down by phase types, in 15-day intervals) and the amount of total attention (views) re-

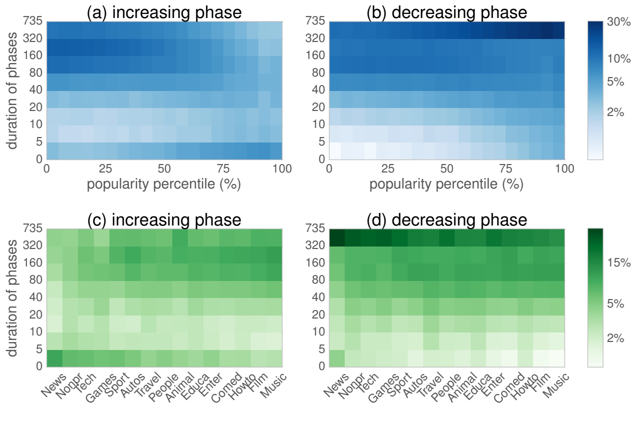


Figure 4: Distribution of phase durations. X-axis: covariates – popularity percentile (20 values) and 15 content categories. Y-axis: duration in days, with log-scaled bins. Intensity: the fraction of phases of a certain property (x) and duration (y). We can see from (a) that many popular videos have long and sustained (> 80 days) increasing phases, and from (b) that unpopular videos have longer decreasing phases (> 320 days). In (c), entertainment-related videos are more likely to have long increasing phases. In (d), while *news* videos have by far the most amount of decreasing phases over a year (also see Figure 3(f)), long decreasing phases exist across all categories.

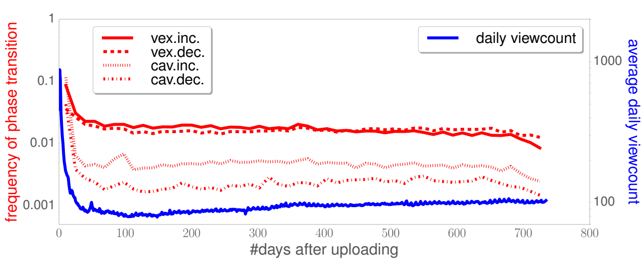


Figure 5: Red: The probability of a video having a new phases in 15-day intervals over time, broken down by phase types. Blue: Average daily viewcount for each video.

ceived by all videos over video age (blue curve). We can see that after an initial period (90 days) of higher occurrences of new phases and views, $\sim 2\%$ of the 172K videos has a new convex increasing and decreasing phases in any given 15-day period. It is notable that (1) this trend holds constant from 3 months to 2 years into videos' lifecyles and (2) for videos older than ($>$ 3 months of age), the amount of new convex-increasing phases is about the same as the amount of new convex-decreasing phases, despite the latter being much more popular overall (ref Table 2). The same temporal trend holds for concave-increasing or decreasing phases, except with lower occurrence.

How did videos become viral? Figure 6 explores the most popular videos and the phases they went through over time. We examine the top 5% (or 8,642) videos at 180, 360, 540, and 720 days of age (called the "after" dates t_a), and collect statistics about their popularity percentile on a "before" dates t_b about 6 months prior to t_a – one visualization of this is the left-most boxplot in Figure 2(right), with $t_b = 540$ days and $t_a = 720$ days. We examine the types of phases that are present between t_b and t_a. We graph the data according to four types in the change of popularity percentile: decreased (-), increased by (0-30%, 30%-60%, or >60%); and four types of phase history: having (one) continued increasing phase, continued decreasing, with at least one new increasing phase, and other (one or more decreasing phases). We can see that the most popular, "viral" videos are highly volatile, with about half (4000+) *jumped* more than 60% in percentile to join this group between 30 and 180 days. Furthermore, new and increasing *phases* plays an important role in the videos whose rank increase significantly. Among the 5,948 videos with improved popularity percentile between 180 and 360 days, for example, only 5% (312 videos) is in a continued decreasing phase, the majority either had a new phase (75%), or are in in a continued increasing phase (20%). In other words, the most popular videos tend to have (new and) increasing phases.

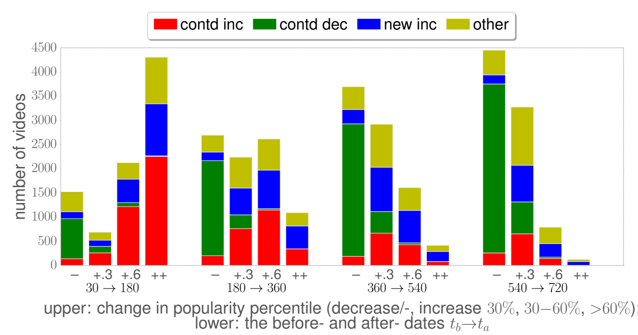


Figure 6: Phase type and popularity evolution for the top 5% videos over time. See Section 4.2 for explanation and discussions.

5 Predicting popularity

Recent work (Pinto, Almeida, and Gonçalves 2013; Szabo and Huberman 2010) showed that future popularity is fairly strongly correlated with past popularity, however popularity prediction remains a difficult problem due to many real-world uncertainties. We propose to use the novel popularity phase representation to predict popularity, and discuss the method and results here.

Problem setting The input to the prediction system is viewcount history for each video up to a *pivot date* t_p, i.e. $\mathbf{x}_{1:t_p}$. The prediction target is the total viewcount χ in a *prediction horizon* of the next Δt days, i.e. $\chi = \sum_{\tau=1}^{\Delta t} x[t_p + \tau]$.

Prediction method We adopt the linear regression predictor. The prediction output is $\chi^* = \mathbf{w}^T \mathbf{x}^* + w_0$, here $\mathbf{x}^*$ is a feature vector, $\mathbf{w}$ and w_0 are weights and bias term learned from training data (with L2 regularization). The *baseline* algorithm (Pinto, Almeida, and Gonçalves 2013) uses $\mathbf{x}_{1:t_p}$ as feature vector, and learns one set of parameters $\{\mathbf{w}, \mathbf{w_0}\}$ for all videos. Our *phase*-informed method, first sorts a video into one of eight subsets, according to (a) whether there

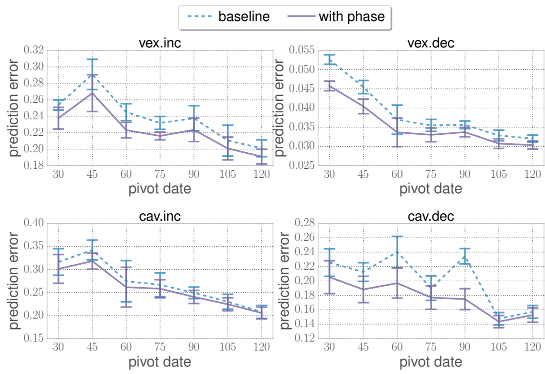


Figure 7: Mean normalized MSE for the baseline and phase-informed prediction over different pivot dates (x-axis) for videos with less than 5 phases, broken down by the shape of the last phase of $\mathbf{x}_{1:t_p}$, Δt=15 days.

are more than 4 phases in $\mathbf{x}_{1:t_p}$, intuitively this accounts for the complexity and uncertainty of popularity evolution for this video, and (b) the shape of the last phase of the observable viewcount (four types), this is useful information about how the phase and popularity will evolve. We then learn a separate predictor for each subset. When the last phase is *convex-decreasing*, we add *phase-extension*, i.e. $a(t_p + \tau)^b + c,\ \tau = 1, \ldots, \Delta t$, as additional features.

Performance metric and experimental setup To avoid being dominated by the most popular videos that have orders of magnitude more views than others, we minimize a normalized mean-square-error (MSE). For each video, let $x_{max} = max\{\mathbf{x}_{1:t_p}\}$, denote $\hat{x} = x/x_{max}$, $\hat{\chi} = \chi/x_{max}$. For a set of videos $\mathcal{V}$, denote the normalized MSE as

$$\epsilon = \frac{1}{\Delta t|\mathcal{V}|} \sum_{v \in \mathcal{V}} (\chi^* - \hat{\chi})^2.$$

We use a sample of 155K videos for evaluation, and report performance and its 95% confidence interval with 5-fold cross-validation. Regularization parameters are tuned on hold-out cross-validation sets.

Results Table 3 summarizes prediction performance across all phase-induced subsets. We can see that in all subsets, the *phase* method reduces the prediction error. The improvement is most significant for videos that end in convex-increasing or concave-decreasing phases – this shows that the additional phase information indeed helps predict future viewcount. Both methods yield higher error when the number of phases is > 4 – this is the small fraction of videos with highly complex dynamics, indicating that predicting popularity is still a challenging problem. Figure 7 shows prediction performance across pivot dates from 30 to 120 days – the *phase* method outperforms baseline in all cases. Figure 8 (a)(b)(c) contains representative examples where *phase* predictor works much better than the baseline, due to the phase change that happened just before the pivot date; (d) contains an example where *baseline* works better, because the sharp decline just before the pivot date seems to be noise on a long-term rising trend, rather than a new phase.

6 Related Work

This work relates to a few areas of active research; we will structure our discussion along three topics: empirical profile of YouTube statistics, models for describing popularity dynamics, and predicting social media popularity.

The first topic is large scale empirical analysis on YouTube. In two pioneering papers, (Cha et al. 2007) measured video metadata statistics in nearly 2 million videos, and (Gill et al. 2007) examined video usage and file properties from network traces. Subsequent metadata analysis have concentrated on the relationship of video popularity with other observable metrics. (Cheng, Dale, and Liu 2008) found that most videos' popularity are determined in its early stage, (Chatzopoulou, Sheng, and Faloutsos 2010) examined 37 million videos and observed that while popularity and user activity metrics are strongly correlated, a video's average rating is not correlated with them. (Figueiredo, Benevenuto, and Almeida 2011) compared viewcount dynamics of "top", "deleted" and "random" videos, and found that bursts of popularity tend to be caused by external search traffic and referrals. (Borghol et al. 2012) took a unique angle to examine duplicate videos, and found that there is a distinct "early-mover advantage". Inspired by such rich literature in measurement studies, we set off to derive a fine-grained representation of popularity evolution overtime.

Among models that describe social media popularity, our work is inspired by Crane and Sornette's model on endogenous growth and exogenous shocks (Crane and Sornette 2008), measured on thousands of videos. The same authors (Crane, Sornette, and others 2008) also found that the shapes of popularity dynamics are related to inherent interestingness – quality videos often relax slower than junk videos. Our proposal of popularity phases operationalize the notion of a shock from only one to multiple throughout a video's lifetime, and our generalized power-law model captures a richer class of shapes. More recently, the SpikeM model (Matsubara et al. 2012) adopts a heuristic search method for power-law (and other) shapes with fixed exponent, and (Tsytsarau, Palpanas, and Castellanos 2014) propose a deconvolution approach on fixed shapes for modeling sentiment events. Our power-law phase model captures a physical event process - self-excited Hawkes processes - and reaches a globally optimal segmentation for these shapes. In terms of general time series modeling and segmentation, this work is most related to parametric fitting with dynamic programming segmentation (Bellman 1961), rather than non-parametric, bottom-up approaches (Keogh et al. 2004). Our proposed PHASE-FINDING algorithm is an extended case of the former, with optimization strategy specially-designed for power-law shapes.

Predicting social media popularity is another area of active investigation. (Szabo and Huberman 2010) found strong linear correlations between (the log of) viewcount in the first week and those after the first month. (Pinto, Almeida, and Gonçalves 2013) built on this insight and used multi-linear regression on the shape of popularity in the first few days to further improve medium-term prediction. Recently, (Cheng et al. 2014) showed that information cascades are predictable (about whether or not they cross the median)

Table 3: Mean normalized MSE on different video subsets, with $\Delta t = 15, 30$ days, $t_p = 60$ days. * denotes a significant improvement (t-test, $p < 0.05$); † denotes relative error reduction $> 5\%$.

Δt	Method	Performances on different subsets							
		#phase ≤ 4 (79.5% videos)				#phase > 4 (20.5% videos)			
		vex.inc	**vex.dec**	**cav.inc**	**cav.dec**	**vex.inc**	**vex.dec**	**cav.inc**	**cav.dec**
15	baseline	0.2450 ± 0.0103	0.0370 ± 0.0038	0.2745 ± 0.0447	0.2402 ± 0.0216	0.2555 ± 0.0105	0.1754 ± 0.0075	0.2722 ± 0.0095	0.2676 ± 0.0138
	phase	$0.2232 \pm 0.0093^{*\dagger}$	$0.0337 \pm 0.0037^{\dagger}$	0.2614 ± 0.0432	$0.1969 \pm 0.0208^{*\dagger}$	0.2456 ± 0.0134	0.1745 ± 0.0072	0.2670 ± 0.0090	0.2654 ± 0.0124
30	baseline	0.5013 ± 0.0386	0.0852 ± 0.0027	0.5953 ± 0.0562	0.5085 ± 0.0552	0.5146 ± 0.0288	0.3880 ± 0.0095	0.5719 ± 0.0388	0.5633 ± 0.0108
	phase	$0.4642 \pm 0.0373^{*\dagger}$	$0.0771 \pm 0.0011^{*\dagger}$	0.5734 ± 0.0598	$0.4241 \pm 0.0428^{*\dagger}$	0.4948 ± 0.0286	0.3865 ± 0.0106	0.5559 ± 0.0321	0.5594 ± 0.0118

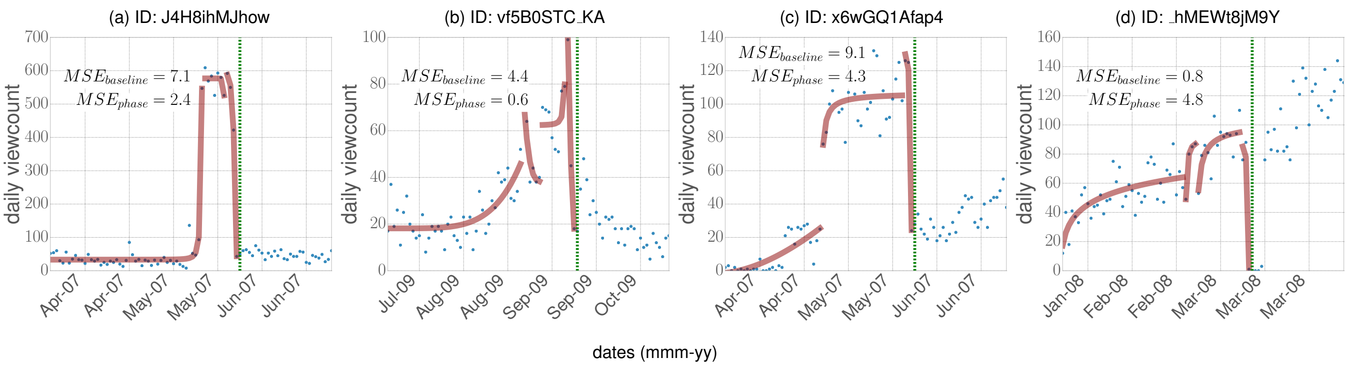


Figure 8: (a)(b)(c): Three examples that phase-informed prediction performs much better than the baseline; (d): An example our method performs worse than the baseline ($t_p = 60, \Delta t = 30$). Blue dots: daily viewcounts; Red curves: phase segments detected; Green lines: indicating the pivot date.

with a set of rich descriptors about the network, timing, and users. This work proposes a detailed representation of the temporal dynamics, and showed that phases correlate with video properties across the whole *popularity scale*, in addition to the median (see Section 4). In addition, recent work (Yu, Xie, and Sanner 2014) showed that a sudden rise in popularity can be predicted with high accuracy from external Twitter feeds, and popularity phases here is a new and natural way to capture such sudden changes.

7 Conclusion

We proposed a novel representation, popularity *phases*, for describing the lifecycle of an online media item (i.e., a YouTube video in this investigation). Along with this representation, we devised a method for segmenting and estimating power-law phases from viewcount traces and presented observations that uses phase history to explain content type and long-term popularity. Furthermore, phase information was used in predicting future popularity, and our approach consistently outperforms prediction approaches using viewcount representations alone. Future work includes explicit modeling of the underlying process with out-of-network input, and better prediction strategies for popularity. Overall, such multi-phase representation has potential to become a tool for understanding the dynamics of other online media, such as hashtags and online memes, and we believe this is a promising direction for further uncovering the laws governing online collective behavior.

Acknowledgements NICTA is funded by the Australian Government as represented by the Department of Communications and the Australian Research Council (ARC) through the ICT Centre of Excellence program. This research was supported in part by the National Computational Infrastructure (NCI) and ARC DP140102185. We would like to thank Xinhua Zhang for his help on optimization.

References

Abisheva, A.; Garimella, V. R. K.; Garcia, D.; and Weber, I. 2014. Who watches (and shares) what on youtube? and when?: Using twitter to understand youtube viewership. WSDM '14, 593–602. New York, NY, USA: ACM.

Ahmed, M.; Spagna, S.; Huici, F.; and Niccolini, S. 2013. A peek into the future: Predicting the evolution of popularity in user generated content. WSDM '13, 607–616.

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In *WSDM '11*, 65–74. ACM.

Bellman, R. 1961. On the approximation of curves by line segments using dynamic programming. *Commun. ACM* 4(6):284–.

Borghol, Y.; Ardon, S.; Carlsson, N.; Eager, D.; and Mahanti, A. 2012. The untold story of the clones: content-agnostic factors that impact youtube video popularity. KDD '12.

Cha, M.; Kwak, H.; Rodriguez, P.; Ahn, Y.-Y.; and Moon, S. 2007. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *IMC '07*.

Chatzopoulou, G.; Sheng, C.; and Faloutsos, M. 2010. A first step towards understanding popularity in youtube. In *INFOCOM Workshops*.

Cheng, J.; Adamic, L.; Dow, P. A.; Kleinberg, J. M.; and Leskovec, J. 2014. Can cascades be predicted? In *WWW '14*, 925–936.

Cheng, X.; Dale, C.; and Liu, J. 2008. Statistics and social network of youtube videos. In *IWQoS '08*.

Crane, R., and Sornette, D. 2008. Robust dynamic classes revealed by measuring the response function of a social system. *PNAS* 105(41):15649–15653.

Crane, R.; Sornette, D.; et al. 2008. Viral, quality, and junk videos on youtube: Separating content from noise in an information-rich environment. In *AAAI Spring Symposium: Social Information Processing*, 18–20.

Figueiredo, F.; Benevenuto, F.; and Almeida, J. M. 2011. The tube over time: Characterizing popularity growth of youtube videos. In *WSDM '11*.

Gill, P.; Arlitt, M.; Li, Z.; and Mahanti, A. 2007. Youtube traffic characterization: a view from the edge. In *IMC'07*.

Golub, G., and Pereyra, V. 2003. Separable nonlinear least squares: the variable projection method and its applications. *Inverse problems* 19(2):R1.

Keogh, E. J.; Chu, S.; Hart, D.; and Pazzani, M. 2004. Segmenting time series: A survey and novel approach. In *Data Mining In Time Series Databases*, volume 57. 1–22.

Matsubara, Y.; Sakurai, Y.; Prakash, B. A.; Li, L.; and Faloutsos, C. 2012. Rise and fall patterns of information diffusion: Model and implications. KDD '12.

Pinto, H.; Almeida, J. M.; and Gonçalves, M. A. 2013. Using early view patterns to predict the popularity of youtube videos. In *WSDM '13*, 365–374.

Rabiner, L. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257–286.

Romero, D. M.; Meeder, B.; and Kleinberg, J. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *WWW '11*, 695–704.

Sornette, D., and Helmstetter, A. 2003. Endogenous versus exogenous shocks in systems with memory. *Physica A* 318(3-4):577 – 591.

Szabo, G., and Huberman, B. A. 2010. Predicting the popularity of online content. *Commun. ACM* 53(8):80–88.

Tsytsarau, M.; Palpanas, T.; and Castellanos, M. 2014. Dynamics of news events and social media reaction. KDD '14.

Xie, L.; Natsev, A.; Kender, J. R.; Hill, M.; and Smith, J. R. 2011. Visual memes in social media: Tracking real-world news in youtube videos. In *ACM Multimedia*, 53–62.

Yang, J., and Leskovec, J. 2011. Patterns of temporal variation in online media. In *WSDM '11*.

Yu, H.; Xie, L.; and Sanner, S. 2014. Twitter-driven youtube views: Beyond individual influencers. MM '14, 869–872. New York, NY, USA: ACM.

Zhu, C.; Byrd, R. H.; Lu, P.; and Nocedal, J. 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.* 23(4):550–560.

A The PHASE-FINDING algorithm

A.1 Estimating a generalized power-law phase

In this work, we use sum-of-squares loss in problem (4). Denote the relative time and duration in a phase as time elapsed since the end of the previous phase $\bar{t} = t - t^s + 1$ and $\bar{T} = t^e - t^s + 1$.

$$\min_{\{a,b,c\}} \; E\{x[t^s : t^e]; a, b, c\} = \frac{1}{2}\sum_{\bar{t}=1}^{\bar{T}} (a\bar{t}^{\,b} + c - x[\bar{t}])^2 \tag{5}$$

Notice that this loss function is differentiable everywhere, but non-convex in $\{a, b, c\}$ – it can be optimized with a general unconstrained optimization technique such as Newton's method, but it will be prone to local minima, and slow to converge. We adopt a technique called *variable projection* (Golub and Pereyra 2003) to address this problem. The basic idea is to separate the nonlinear parameter b and the linear parameter a, c, by re-writing the loss function as follows.

$$E = \frac{1}{2}\sum_{\bar{t}=1}^{\bar{T}} (a\bar{t}_i^{\,b} + c - x[\bar{t}])^2 = \frac{1}{2}||\Phi(\bar{t}) \cdot \beta - \mathbf{x}||^2 \tag{6}$$

where

$$\Phi(\bar{t}) = \begin{bmatrix} 1^b, \; 1 \\ 2^b, \; 1 \\ \dots \\ \bar{T}^b, \; 1 \end{bmatrix}, \beta = \begin{bmatrix} a \\ c \end{bmatrix} \tag{7}$$

$\Phi(\bar{t})$ includes the nonlinear parameter b, and β includes the linear parameters a and c. Given b, there is a unique minimum for the quadratic equation (6), with a, c given by the following closed-form via the Moore–Penrose pseudoinverse:

$$\begin{bmatrix} a \\ c \end{bmatrix} = (\Phi(\bar{t})^T \Phi(\bar{t}))^{-1} \Phi(\bar{t})^T \mathbf{x} \tag{8}$$

Equation (8) is a *necessary* condition of the optimal solution of Equation (6). Substituting a and c by it, the loss function becomes:

$$E = \frac{1}{2}||\Phi(\bar{t})(\Phi(\bar{t})^T \Phi(\bar{t}))^{-1} \Phi(\bar{t})^T \mathbf{x} - \mathbf{x}||^2 \tag{9}$$

Now, we have reduced the parameter space from $\{a, b, c\} \in \mathcal{R}^3$ to $b \in \mathcal{R}$. And the optimal solutions of Equation 9 is the same with that of Equation (6).

Implementation and solution quality. We use the *L-BFGS-B* algorithm (Zhu et al. 1997) to find a solution of this nonlinear objective. We search over the two temporal directions in Eq (2) and take the direction with a small mean-square-error. We observed significant improvement in speed and solution quality with the variable projection technique, consistent with the original proposal (Golub and Pereyra 2003). We also normalize $\mathbf{x}_{1:T}$ into $[0, 100]$ before running the phase-finding algorithm to be on the same order of magnitude with time stamp $\bar{t}$ – thus avoiding numerical issues in fitting a handful to a few million daily views. In addition, we employ the following initialization technique to start from a good "guess" of b — we use $\bar{t}=1$ with each observation $\bar{t}=2,\ldots,\bar{T}$ to solve for a value b by assuming each pair exactly follows power-law "$x=at^b$" (without c), and then average these estimates as the initial value[3]. As an initial validation for the solution quality of this curve-fitting problem, we generate 500 synthetic power-law curves (of length 200) with parameters randomly chosen from uniform distributions $a \sim U[-100,100]$, $b \sim U[-2,2]$, and $c \sim U[-500,500]$, and a and b bounded away from zero to avoid degenerate cases ($|a|>3$, $|b|>0.1$). We optimize Equation (9), and observed the relative fitting error in each coefficient ($E_a=|a^*-a|/|a|$, with the corresponding confidence intervals) as: $E_a=(1.8\pm0.3)\times10^{-3}$, $E_b=(1.1\pm0.6)\times10^{-5}$ $E_c=(1.8\pm0.3)\times10^{-3}$.

A.2 Simultaneous fitting and segmentation

A bruce-force enumeration approach to the joint segmentation and curve-fitting problem (3) will have a complexity exponential in T, the sequence length. Fortunately problem (3) is in a form suitable for induction with dynamic programming. We describe the algorithm in three stages, similar to (but extending) the description of the well-known Viterbi decoding algorithm (Rabiner 1989) with embedded curve-fitting.

As in problem (3), denote $1\le t'\le T$ as the current position in the recursion, n' as the number of optimal segments up to position t', and a shorthand $E^*(t')$ for the lowest segmentation and fitting error (under any segmentation) for the subsequence $\mathbf{x}_{1:t'}$

$$E^*(t')=\min E\{\mathbf{x}_{1:t'},\rho_{1:n'},\theta_{1:n'}\} \tag{10}$$

here the minimization is done over $\{n',t^s[1:n'],\theta_{1:n'}\}$.

In order to retrieve an optimal segmentation, we need to keep track of arguments that minimize Equation (10) for each t'. This is done via a pointer for each t' containing the starting position of the last phase and its parameters.

$$\delta(t')=\{t^{s*}_{n'},\theta^*_{n'}\} \tag{11}$$

The complete procedure for finding the best segmentation and their power-law fits is as follows:

Stage 1 Initialization:

$$\text{for } t=1,2,\ E^*(t)=0 \qquad \delta(t)=\emptyset \tag{12}$$

The reason we initialize a cost of zero and an empty parameter set for the first two positions (instead of only for $t=1$ as the Viterbi algorithm) is that the generalized power-law curve has three free parameters, and hence takes at least three observations to fit.

Stage 2 Recursion:

$$E^*(t')=\min_{t^s_{n'},\theta_{n'}}\{E^*(t^e_{n'-1})+E(x[t^s_{n'}:t'],\theta_{n'})\} \tag{13}$$

[3]This initialization heuristic is documented in the `power2start()` function of Matlab curve-fitting toolbox.

The step above computes the cumulative minimum error, for $t'=3,4,\ldots,T$. This is done by searching for an optimal starting point $t^s_{n'}=1,2,\ldots,t'$, for the *current* phase that ends at t', and obtaining optimal parameter $\theta^*_{n'}$ that minimizes fitting error on subsequence $x[t^s_{n'}:t']$ for each $t^s_{n'}$, using the PHASE-FITTING algorithm in Section A.1. We also populate the backtracking pointers:

$$\delta(t')=\arg\min_{t^s_{n'},\theta_{n'}}\{E^*(t^e_{n'-1})+E(x[t^s_{n'}:t'],\theta_{n'})\} \tag{14}$$

Stage 3 Backtracking: The set of segmentation parameters $\mathcal{S}^*=\{n^*,t^{s*}_{1:n},\theta^*_{1:n}\}$ is obtained using recursion:

- Initialize $\mathcal{S}^*\leftarrow\delta(T),\ t'\leftarrow t^s_{n'},\ n^*\leftarrow 1;$
- Recurse $\mathcal{S}^*\leftarrow\mathcal{S}^*\cup\delta(t'),\ t'\leftarrow t^s_{n'},\ n^*\leftarrow n^*+1;$
- Terminate $\mathcal{S}^*\leftarrow\mathcal{S}^*\cup n^*$.

How to avoid over-fitting Every three observations will provide a unique solution for the curve-fitting problem (4) with a set of a,b,c – this can easily lead to over-fitting by over segmentation. We introduce a segment regularizer, by adding a penalty constant η to every new segment introduced by the algorithm. That is, the objective for problem (3) is modified as:

$$\tilde{E}\{\mathbf{x}_{1:T},\rho_{1:n},\theta_{1:n}\}=\sum_{i=1}^{n}E_i\{x[t^s_i:t^e_i],\theta_i\}+(n-1)\eta \tag{15}$$

Minimizing the objective is still done with dynamic programming, by simply adding η to the iteration step (13),

$$\tilde{E}^*(t')=\min_{t^s_{n'},\theta_{n'}}\tilde{E}^*(t^e_{n'-1})+E\{x[t^s_{n'}:t'],\theta_{n'}\}+\eta \tag{16}$$

and also modify step (14) accordingly.

Implementation and solution quality. Hyper-parameter η controls the trade-off between fitting each phase well, and having a reasonable number of phases. To calibrate the value of η, we invite 6 people to choose their *preferred segmentation* for a random sample of 210 videos. This is done on a web interface, screenshot of which is at project website[1]. The labelers have about 72% agreement on the phase boundaries (that are ± 2 days of each other). Then we do a line search on η with the set of agreed boundaries and found when $\eta=2.3$, the algorithm achieves the highest F1-score 0.707 with $recall=0.779$ and $precition=0.648$.

The running time for step (13) above is $O(T\Gamma(T))$, where $O(T)$ is the time for searching over $t^s_{n'}$ and $\Gamma(T)$ is the T-dependent time complexity of power-law curve fitting and finding $\theta^*_{n'}$. The complexity of this entire dynamic programming algorithm is hence $O(T^2\Gamma(T))$. We implemented this algorithm in C++, the throughput for finding phases for one-year long viewcount sequences is about 400 per CPU per hour.