Viral Video Style: A Closer Look at Viral Videos on YouTube

Lu Jiang, Yajie Miao, Yi Yang, Zhenzhong Lan, Alexander G. Hauptmann
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{lujiang, ymiao, yiyang, lanzhzh, alex}@cs.cmu.edu

ABSTRACT

Viral videos that gain popularity through the process of Internet sharing are having a profound impact on society. Existing studies on viral videos have only been on small or confidential datasets. We collect by far the largest open benchmark for viral video study called CMU Viral Video Dataset, and share it with researchers from both academia and industry. Having verified existing observations on the dataset, we discover some interesting characteristics of viral videos. Based on our analysis, in the second half of the paper, we propose a model to forecast the future peak day of viral videos. The application of our work is not only important for advertising agencies to plan advertising campaigns and estimate costs, but also for companies to be able to quickly respond to rivals in viral marketing campaigns. The proposed method is unique in that it is the first attempt to incorporate video metadata into the peak day prediction. The empirical results demonstrate that the proposed method outperforms the state-of-the-art methods, with statistically significant differences.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences; H.4 [**Information Systems**]: Applications

General Terms

Human Factors, Algorithm, Measurement

Keywords

CMU Viral Video Dataset, YouTube, Popular Video, Social Media, Gangnam Style, View Count Prediction

1. INTRODUCTION

The recent total view count of the Gangnam Style video on YouTube is approaching 1.8 billion, accounting for approximately one fourth of the world's population. Videos of this type, which gain popularity through the process of Internet sharing, are known as *viral videos* [2]. Viral videos are usually user-generated amateur videos and shared typically through sharing web sites and social media [17]. A video is said to go/become viral if it spreads rapidly by being frequently shared by individuals.

Viral videos have been having a profound social impact on many aspects of society, such as politics and online marketing. For example, during the 2008 US presidential election, the pro-Obama video "Yes we can" went viral and received approximately 10 million views [2]. We found that during the 2012 US Presidential Election, Obama Style and Mitt Romney Style, the parodies of Gangnam Style, both peaked on Election Day and received approximately 30 million views within one month before Election Day. Viral videos also play a role in financial marketing. For example, Old Spice's recent YouTube campaign went viral and improved the brand's popularity among young customers [17]. Psy's commercial deals has amounted to 4.6 million dollars as a result from his single viral video Gangnam style. [1]

Due to the profound societal impact, viral videos have been attracting attention from researchers in both industry and academia. Researchers from GeniusRocket [14] analyzed 50 viral videos and presented 5 lessons to design a viral video from a commercial perspective. They found that viral videos tend to have short title, short duration and appear on hundreds of blogs. More recently, Broxton et al. analyzed viral videos on a large-scale but confidential dataset in Google [2]. Specifically, they computed the correlation between the degree of social sharing and the video view growth, the video category and the social sites linking to it. One important observation they found is that viral videos are the type of video that gains traction in social media quickly but also fades quickly. In [17], West manually inspected the top 20 from Times Magazine's popular video list and found that the length of title, time duration and the presence of irony are distinguishing characteristics of viral videos. Similarly, Burgess concluded that the key of textual hooks and key signifiers are important elements in popular videos [3].

Existing observations are valuable and greatly enlighten our work. However, there are two drawbacks in the previous analyses. First, the datasets used in the study are either confidential [2] or relatively small, containing only tens of videos [14, 17, 3]. When using small datasets, it remains unclear whether or not the insufficient number of samples would lead to biased observations. Regarding confidential datasets, it is barely possible to further inspect and develop the analysis outside Google. Second, the previous analysis

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

ICMR '14, April 01-04, 2014, Glasgow, United Kingdom
Copyright 2014 ACM 978-1-4503-2782-4/14/04 ...$15.00
http://dx.doi.org/10.1145/2578726.2578754.

[1] http://www.canada.com/entertainment/music/much+money+make+with+Gangnam+Style/7654598/story.html

mainly concentrates on qualitative rather than quantitative analysis. For example, different studies reach the consensus that short title and short duration are characteristics of viral videos, but their statistics and the derivation from other types of videos are unknown.

Our first objective is to further understand the characteristics of viral videos by experimenting on by far the largest open dataset of viral videos named CMU Viral Video Dataset. The videos are collected from YouTube which is the focal platform for many social media studies [19]. We share the dataset with researchers from both academia and industry. Our statistical analysis verifies already existing observations, and even leads to new discoveries about viral videos. The analysis results, which are consistent with the existing observations on Google's dataset [2], suggest that the dataset is less biased. Our analysis not only sheds light on how to design a video design with a better chance to become viral, but also benefits other applications that need to detect viral videos more accurately. In this paper, we follow the definition in [2, 17, 14], where viral videos are characterized by the degree of social sharing. The term as used in [5] refers to a different meaning where "viral" videos are a type of video with certain views on the peak day, relative to its total views. We choose to avoid the term used in [5], as a considerable number of viral videos, including Gangnam Style, would be judged non-viral under this criterion.

The second objective is to utilize the discovered characteristics to forecast the peak day of viral videos. The proposed method allows for estimating the number of days left before the viral video reaches its peak views. Forecasting the peak day for viral videos is of great importance to support and drive the design of various services [11]. For example, accurately estimating the peak day of the viral video is of great importance to advertising agencies to plan advertising campaigns or to estimate costs. Knowing the peak day in advance puts a company in an advantageous position when responding to their rivals in viral marketing campaigns. Existing popularity prediction methods only use the accumulated views [15, 11]. The proposed method, however, is the first attempt to incorporate video metadata in the peak day prediction. In summary, the contributions of this paper are as follows:

- We establish by far the largest open dataset of viral videos that provides a benchmark for the viral video study.
- We discover several interesting characteristics about viral videos.
- We propose a novel method to forecast the peak day for viral videos. The experimental results show significant improvements over the state-of-the-art methods.

2. CMU VIRAL VIDEO DATASET

2.1 Overview

The collected dataset consists of 20,000 videos divided into three categories: viral, quality or background. The categories are established based on the classification in [4], which may not be ideal. For two videos, including Gangnam Style, they can belong to both the viral and the quality category. The quality and background categories are included as the comparison groups to study viral videos.

Table 1: Overview of the dataset.

Category	Viral	Quality	Background
#Total Videos	446	294	19,260
#Videos with insight data	304	270	7,704

The viral videos come from three sources: Time Magazine's popular videos, YouTube Rewind 2010-2012 and Equals Three episodes. Time Magazine's list contains 50 viral videos selected by its editors[2]; YouTube Rewind is YouTube's annual review on viral videos crowdsourced by YouTube editors and users[3]; Equals Three is a weekly review of Internet's latest viral videos featured by William Johnson[4]. In total, we collect 653 candidates and exclude 207 videos with less than 500,000 views. The remaining 446 videos cover many of the viral videos 2010-2012. Table 2 lists some representative videos in the dataset. As mentioned in the introduction, the viral videos are selected by experts in terms of their degree of social sharing [2] rather than their peak views [5]. The second category includes quality videos. Music videos are selected to represent quality videos since they seem to be easily confused with viral videos [2]. The official videos from the Billboard Hot Song List 2010-2012[5] are used. In addition, background videos, which are randomly sampled from YouTube, are also included for comparison. For each video, we attempt to collect all types of information and the ones absent are the video contents and comments, which are not released due to copyright issues. All data are collected by our Deep Web crawler [7, 8]. In summary, the released information includes:

Thumbnail: For each video a thumbnail (360×480) along with its captured time-stamp is provided.

Metadata: Two types of metadata are provided, namely video metadata and user metadata. The video metadata includes video ID, title, text description, category, duration, uploaded time, average rate, #raters, #likes, #dislikes and the total view count. The user metadata includes user ID, name, subscriber count, profile view count, #uploaded videos, etc.

Insight data: YouTube insight data [20] provides the historical information about the view, comments, likes and dislikes information in the chart format. In our dataset, the chart is converted into plain text to make it easier to work with. It also includes the key events, locations and demographics about the audience [20]. Insight data is only available for the videos whose uploaders agree to publish the information, and the number of videos with the insight data can be found in Table 1.

Social data: The number of inlinks of the video returned by Google is collected as the social data. To obtain this information, we first issue a query using video ID to Google and then count the number of returned documents as the indicator of #inlinks outside of YouTube.

Near duplicate videos: Near duplicate video IDs are provided (see Section 2.2). For a given video, we issue a query using the bigram and trigram of the title to collect a set of candidate videos on which the near duplicate detection is automatically performed.

2.2 Multimodel Near Duplicate Detection

[2] http://tinyurl.com/ybseots

[3] http://en.wikipedia.org/wiki/Rewind_YouTube_Style_2012

[4] http://en.wikipedia.org/wiki/Equals_Three

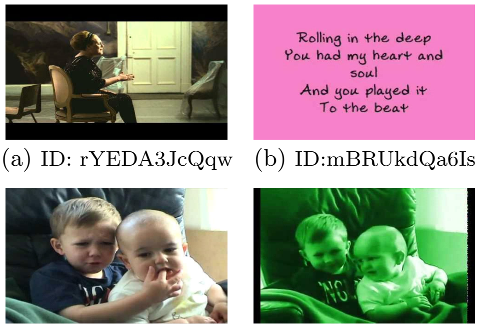
[5] http://en.wikipedia.org/wiki/Billboard_Year-End_Hot_100_singles_of_2011

Table 2: The representative viral videos in the dataset.

Description	Name	YouTube ID	Comments
Most Viewed	Gangnam Style	9bZkp7q19f0	1,206,879,985 views
Most Liked	Gangnam Style	9bZkp7q19f0	6,739,715 likes
Most Disliked	Friday - Rebecca Black	kfVsfOSbJY0	940,615 dislikes
Longest	Randy Pausch Last Lecture	ji5_MqicxSo	1 hour and 16 minutes
Shortest	A Cute Hamster with a Cute Shock	C3JPwxQkiug	3 seconds
Earliest	Best Fight Scene of All Time	uxkr4wS7XqY	Uploaded on 2006-02-10

An important type of information included is the automatically detected near duplicate videos [18] or video clones [1], which refer to the videos with essentially the same content. Borghol et al. proposed to manually annotate video clones [1]. Instead, we utilize the state-of-the-art visual and acoustic techniques to automate near duplicate detection. Then we manually inspect the detection results as a double check. This proposed detection paradigm significantly alleviates heavy manual annotation, and seems to be beneficial for the large-scale analysis.

Near duplicate videos are judged according to both visual and acoustic similarity. We observed that many music videos that greatly differ in visual content tend to use exactly the same or similar soundtrack. For example, Figure 1 (a) and (b) are near duplicates of the song Rolling in the Deep where (a) is the official video, and (b) is the same song with lyrics. Therefore acoustic near duplicate detection is performed for videos whose category is music (in its metadata), and visual near duplicate detection is applied on the rest of the videos.



(a) ID: rYEDA3JcQqw (b) ID:mBRUkdQa6Is

(c) ID: _OBlgSz8sSM (d) ID: Y13oB-IhE7I

Figure 1: Visual and acoustic near duplicate videos. (a) and (b) are the acoustic near duplicates of the video Rolling in the Deep. (c) and (d) are the visual near duplicates of Charlie Bit my Finger.

Table 3: Basic statistics about viral videos.

Statistics	Viral	Quality	Background
View Count Median	3,079,011	55,455,364	7,528
Title length	5.0±0.1	5.4±0.1	7.0±0.1
Duration(s)	138.6±16.0	248±3.9	252±24.6
Average Rate	4.69±0.03	4.75±0.03	4.04±0.08
Rater/View	0.54±.03%	0.38±.01%	0.87 ±.07%
Cor(inlinks, view)	0.54	0.25	0.28
Days-to-peak Median	24	63	30
Lifespan Median	7	166	10

The first task is to detect acoustically similar videos i.e. videos sharing similar soundtracks. Specifically, the Dynamic Time Warping (DTW) algorithm [6, 12] is adopted to measure acoustic distance between music videos. DTW was originally proposed for template-based speech recognition [12] and has been widely used in speech utterance matching and spoken term detection [6]. The audio track of each music video is represented as a set of 13-dimensional MFCC vectors, with a frame length of 25 milliseconds. To reduce potential mismatch due to recording conditions and environments, mean and variance normalization of feature vectors is performed on the whole audio file. The near duplicate detection consists of two steps. First, the DTW distance for each pair of the given video and its near duplicate video candidates is computed and normalized by the length (total number of MFCC vectors) of the base video. Second, a threshold (0.2 in the experiment) is applied to the DTW scores. The candidates whose distances fall below the threshold are taken as near duplicate videos.

The second task is to detect visually similar videos. Given a video, we first extract key frames by a shot boundary detection algorithm [16] based on the color histogram difference between consecutive frames. The frame in the middle of shot is used to represent the shot. SIFT [9] is extracted using Harris-Laplace key point detector from the detected key frames. Following [16], we cluster the SIFT descriptors into 4,096 clusters using k-means, and quantize them into a standard bag-of-feature representation. Then, the video is represented by the averaged bag-of-features of all key frames. To query a video's near duplication, we calculate the intersection similarity between the given video and the candidate videos, and return the top k (k=20) most similar videos as the detected near duplicate videos.

The above methods are applied to the videos in our dataset. To evaluate the performance we manually inspect the detection results of 50 viral videos and 38 music videos. 398 visual near duplicate videos and 1296 acoustical duplicate videos are detected, i.e. on average 7.96 and 34.1 per video. Generally, the precision@10 is above 0.9.

3. STATISTICAL CHARACTERISTICS

Table 3 summarizes the statistical characteristics of viral videos in our dataset. The results are consistent with existing observations, including the observations on Google's dataset, suggesting the dataset is less biased. For example, generally, viral videos are less popular than quality videos [2]. They have short titles, durations and lifespans [14, 17, 2, 3]. In the rest of the section, we highlight some interesting characteristics that were revealed in our study.

Observation 1: The correlation between the #inlinks returned by Google and #views is a proxy to the socialness. The socialness is a metric to measure the level of dissemination on social media. Broxton et al. define it as a fraction between the views coming from social sources such as Facebook and non-social sources such as search engines [2]. They found that viral videos tend to have the higher socialness compared with other types of videos. Although the observation is interesting, it is impossible to obtain the information on how the user came to watch a video outside Google. We

Table 4: Evolution of viral videos.

Statistics	2010	2011	2012
Days-to-peak Median	24	14	9
Lifespan Median	8	6	3

found a publicly available metric that measures the correlation between the #inlinks returned by Google (see social data in Section 2.1) and #views. The correlation reflects the dependency between the social views and the total views. As listed in Table 3, Pearson Correlation Coefficient (PCC) of viral videos is approximately the twice of the others indicating that the correlation can reasonably estimate the socialness.

Observation 2: The days-to-peak and the lifespan of viral videos decrease over time. The days-to-peak is the number of days before reaching the peak view since uploaded. The lifespan refers to the number of consecutive days maintaining certain views, which is set to 30% of the peak views in our analysis. The medians of both metrics are presented in Table 3. As found in [2], viral videos tend to have both a shorter lifespan and shorter days-to-peak. We found that these statistics are not static but evolve over time. Table 4 lists the evolution of the metrics from 2010 to 2012. In 2012, on average, it only took 9 days to reach the peak views in contrast to 24 days in 2010. The decreasing days-to-peak suggests social media has been becoming more efficient in disseminating viral content. The decrease in the lifespan, on the other hand, perhaps stems from the increasing number of interesting videos available on social media.

Observation 3: Significant correlation can be observed in the metadata of viral videos. Figure 3 plots the PCC matrix, with each entry indicating the correlation between the corresponding two variables. The magnitude of the correlation is represented by the slope of the ellipse, and the color indicates the correlation type: blue for positive and orange for negative correlations. Two coefficient clusters can be identified. The first one is about users, including the user profile views, the subscriber count and the user's total upload views. The second one is about videos, including the number of dislikes, raters and likes. The observation about these correlations is important as it will influence the choice of viral videos detection models. Due to the high correlation, a model explicitly considering the feature correlation, such as the Bayesian Network [10], may lead to a superior performance. The experimental result in Table 5 substantiates this argument.

Observation 4: The popularity of the uploader is a factor that affects the popularity of the viral video which seems to be more important than the upload time. It is an interesting phenomenon that some particular videos in a near duplicate group receive significantly higher views than others, considering that these videos share essentially the same content. For example, the most popular Gangnam style video receives 95% of the total views in its near duplicate group of 28 videos. As the near duplicate videos are uploaded at different time, the one uploaded earlier has better chance to become viral. Borghol et al. found background videos have this property and coined it as the first mover advantage [1]. We found that for viral videos, though the advantage still holds, the popularity of the uploader plays a more substantial role. In order to measure the importance, we calculate PCC between the numeric metadata and the total views in each near duplicate group which consists of videos with essentially the same content. Specifically, two lists can be obtained within a group; one is the list of metadata e.g. the upload time for each video, the other contains the total views for each video; PCC is calculated between the two lists.

Figure 2 illustrates the macro-PCC across all near duplicate groups. As we see, the correlation of the upload time is not as high as one expects. The factor with the strongest correlation is the uploader per-video-views, which is the average views of all videos uploaded by the uploader, prior to the viral video. The result shows that the popularity of the uploader is a more important factor in disseminating viral content. The most viewed video in a near duplicate group is not necessarily the first uploaded video. In fact, the most viewed video is, on average, the 2nd to 3rd uploaded video in the group. This phenomenon is more evident for music videos where the official music video usually receives the most views irrespective of its upload time.

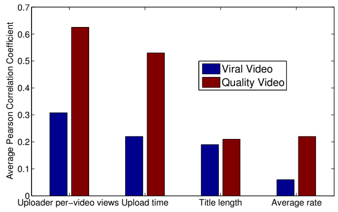


Figure 2: The macro Pearson Correlation Coefficient between the numeric metadata and the view count in near duplicate groups.

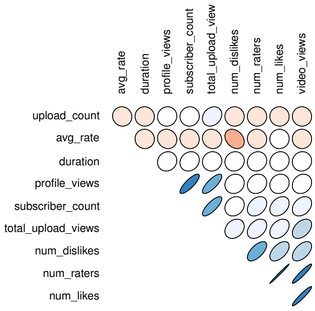


Figure 3: The correlation coefficient matrix in viral videos' metadata. Each entry indicates the correlation between the corresponding two variables. The magnitude of the correlation is represented by the slope of the ellipse. The color indicates the correlation type: blue for positive and orange for negative correlations.

4. FORECASTING THE PEAK DAY

Based on the observations discussed in Section 3, we introduce a novel method to forecast the peak day of viral videos. The peak day refers to the date on which a video reaches its highest views. The ground truth peak day is known because the daily view count of a video can be obtained in its insight data. Our method takes a video as input, and outputs the estimated number of days left before the viral video reaches its peak views. The proposed method is novel in that it first incorporates the metadata in the peak day prediction.

According to [2], viral videos usually experience sudden burstiness in their views. We model the burstiness by a modified HMM which has a single hibernating state h and a series of active states $a_1, ..., a_n$. h depicts the state before entering or after leaving the burstiness in the view pattern, and the active states model the situation within the burstiness. For example, the state sequence h, h, h, a_1, a_2, h describes a video that hibernates for the first three days and peaks on the fifth day. Since the task is to predict the peak day, we are more interested in the states prior to or at the peak day than the states after it.

A state transits either to the hibernating state or the next active state, as illustrated in Figure 4. The hibernating state and the last active state (a_n) have a self-loop. The symbol emitted by the state is the daily view count which is uniformly quantized into a set of discrete levels. In other words, the emission probability of each state is a distribution of the view count levels. The hibernating state models the insignificant number of views outside the burstiness, and thus, the probability of emitting higher views at the hibernating state should be less than that at an active state. Following the standard notation, let q_t denote the state at time t and $P(q_{t+1}|q_t)$ denote the transition probability from time t to $t + 1$. The emission probability is represented by $P(o_t|q_t)$ where o_t is a variable of the emitted symbol, i.e. the view count at time t. We assume that the starting state is the hibernating state.

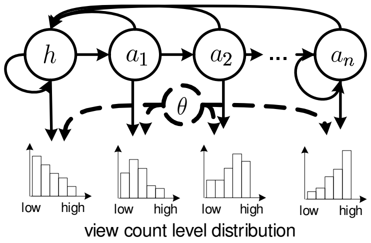


Figure 4: The proposed model. θ is a variable on the video type and the emitted symbols are the quantized view counts.

Compared with a plain HMM, the proposed model incorporates two modifications to improve the accuracy. First, it introduces a variable named θ to describe the video type so that the emission probability can be jointly determined by the current state and the video type. Formally, the emitted view count is written as:

$$o_t = P(\theta|q_t) \times \hat{o}_t \tag{1}$$

where $P(\theta|q_t)$ describes the probability of being a viral video at the state q_t. $\hat{o}_t$ is the observed view count at time t, and o_t is the estimated view count given the video is a viral video. Since our task is to predict the peak day for viral videos, $P(o_t|q_t)$ is designed to emphasize the views of viral videos. Given a known o_t, $P(o_t|q_t)$ can be easily estimated using standard Maximum Likelihood Estimation (MLE) [13]. Note that the variable θ is incorporated to discount the view count, rather than conventionally incorporated as a latent variable in the graphic model, because we found that this strategy works reasonably well in practice. Besides, estimating latent variable in the graphic model, usually by Expectation Maximization (EM), only utilizes the view count information. According to Observation 2 and 4, however, ignoring the metadata leads to a suboptimal solution. For example, according to Observation 2, the lifespan cannot be accurately estimated without knowing the upload time. Therefore we regard $P(\theta|q_t)$ as a prior, and estimate it using a classifier trained on the features available at the current state q_t. The features are divided into the following groups:

Bag-of-Words (BoW): Bag-of-words of the video's title and text description.

Time-invariant Metadata: The category, duration, upload time, title length, description length and uploader.

Time-variant Metadata: The accumulated view count, likes, dislikes and comments up to the current time.

The BoW and the time-invariant metadata are time invariant features because their classification values are independent of the state q_t. The accuracy of the above features will be discussed in Section 5.3. Note that we cannot use the features that contain any information after the peak day, e.g. #raters, #inlinks and the average rate.

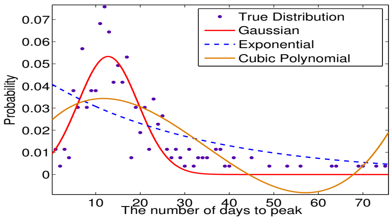


Figure 5: The distribution of the number of days to peak (after entering in active states) in our dataset. The best fitting of three curves are plotted, namely Gaussian, Exponential and Cubic Polynomial.

The second modification regards the transition probability. The MLE approach estimates the transition probability from state a_i to a_{i+1} as the fraction between the number of transitions from a_i to a_{i+1}, and the number of transitions from state a_i. However, due to the short lifespan of viral videos in Observation 2, the number of observations decreases sharply as the active state ID becomes larger. For example, our dataset has hundreds of observations for a_1 whereas only less than 10 observations for a_{12}. A potential problem is that the insufficient number of observations may fail to estimate accurate transition probabilities for the active state with a larger ID. To solve this problem, we smooth the transition probability by Gaussian distribution. Empirically, we found the days-to-peak of viral videos (after entering the active state) follows the discrete Gaussian distribution $\mathcal{N}(\mu, \delta^2)$. This assumption may not be optimal but reasonably captures the distribution (see Figure 5).

The transition probability is then calculated from:

$$P(q_{t+1} = a_{i+1}|q_t = a_i) = 1 - \sum_{k=1}^{i} \alpha\mathcal{N}(k; \mu, \delta^2) \tag{2}$$

where α is a parameter and the second term of the right-hand side is the accumulated probability. Eq. (2) indicates that the probability transiting to a_{i+1} equals the probability not peaking before the state a_i. Only three parameters i.e. α, μ and δ need to be estimated and can be derived by regression on the observations in the training set.

Since the ground truth daily views of a video can be accessed in its insight data, the hidden state can be automatically assigned, and the proposed model can be trained on viral videos by the standard Baum-Welch algorithm [13]. Once the model is trained, the next task switches to the peak day prediction. Given an observation sequence $o_1, ..., o_t$, Viterbi algorithm can be applied to estimate the most likely sequence of corresponding hidden states up to the current time. A video is estimated to peak at state q_{t+k}, if $q_{t+k} \neq h \wedge q_{t+k+1} = h$, which is the last state before transiting back to the hibernating state. Physically, the peak state corresponds to the day with the highest views. As mentioned, we are more interested in modeling the states before the peak day. We do not explicitly model the states after the peak day because they are less important for the peak day prediction. More formally, suppose at the current time t, the state estimated by Viterbi algorithm is q_t, the prediction task is to find the peak state q_{peak} in the most likely sequence starting with q_t, i.e.

$$q_{\text{peak}} = \arg\max_k P(q_t, q_{t+1}, ..., q_{t+k}, h), \tag{3}$$

where k is the number of days left before the peak. To calculate the joint probability we rewrite it as

$$q_{\text{peak}} = \begin{cases} \arg\max_k P(h|a_{i+k}) \prod_{j=i}^{i+k-1} P(a_{j+1}|a_j) & q_t = a_i (1 \leq i < n) \\ q_t & \text{otherwise} \end{cases} \tag{4}$$

where $i + k \leq n$, and n is the total number of active states[6]. Eq. (4) is more computationally efficient than Eq. (3), and can be solved by enumerating all possible $k < n$. It indicates that when the current state is one of the active states except a_n, the peak state is in the most likely sequence. Otherwise, when the current state is either the hibernating state or the last active state a_n, the peak state is the current state. To prove when $q_t = a_n$, $q_{\text{peak}} = q_t = a_n$, we have:

$$\begin{aligned} & P(q_t = a_n, q_{t+1} = a_n, q_{t+2} = h) \\ & = P(q_t = a_n) P(q_{t+1} = a_n | q_t = a_n) P(q_{t+2} = h | q_{t+1} = a_n) \\ & = P(q_t = a_n) P(q_{t+1} = a_n | q_t = a_n) P(q_{t+1} = h | q_t = a_n) \\ & = P(q_t = a_n, q_{t+1} = h) P(q_{t+1} = a_n | q_t = a_n) \\ & \leq P(q_t = a_n, q_{t+1} = h). \end{aligned} \tag{5}$$

Similarly we can also verify that $q_t = h$ then $q_{\text{peak}} = h$.

If the current state is one of the active states other than a_n, according to Eq. (4) at most $n - 1$ sequences need to be calculated. It is because, as proved in Eq. (5), that when a state is at a_n, the most likely next state is h. It can also be shown that at any time t ($t \geq 1$), the probability of re-entering the active state is less than the probability of transiting back to the hibernating state. Formally we have:

$$\begin{aligned} P(q_t, q_{t+1} = h) = & \sum_{q_t, \cdots, q_{t+m}} P(q_t, q_{t+1} = h, ..., q_{t+m}) \\ > & P(q_t, q_{t+1} = h, q_{t+2} = a_1), \end{aligned} \tag{6}$$

where m ($m \geq 2$) is a parameter indicating the length of the state sequence. The left-hand side of Eq. (6) is the probability of peaking at the state q_t. The right-hand side indicates the probability of re-entering the active states so as to peak at a future state, and according to Eq. (6), it is less likely to happen. Therefore when q_t is an active state in $\{a_i | 1 \leq i < n\}$, Eq. (4) selects the most likely state sequence from all possible sequences. The following toy example shows how to calculate the peak state.

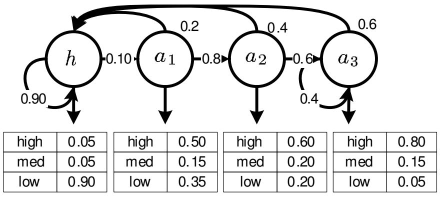


Figure 6: A toy example of the trained model.

EXAMPLE 1. *Suppose the symbols: "low", "med" and "high" represent the quantized view count levels using Eq.*(1) *and the trained HMM is in Figure 6. We have the following observation sequences:*

ID	observation sequence	state sequence
1	*low, low, med*	h, h, h
2	*low, low, high*	h, h, a_1
3	*low, low, high, med, high,*	h, h, a_1, a_2, a_3

Feeding the observation sequences to Viterbi algorithm, we get the corresponding most likely state sequences up to the current time. For the first case (ID=1) and last case (ID=3), the current state is h and a_3, according to Eq. (4)*, the peak state is the current state i.e. the video will not peak in future.*

For the second case, we calculate the probability of all possible sequence starting from a_1:

$P(h|a_1) = 0.2,$
$P(h|a_2)P(a_2|a_1) = 0.4 \times 0.8 = 0.32,$
$P(h|a_3)P(a_3|a_2)P(a_2|a_1) = 0.6 \times 0.6 \times 0.8 = 0.28.$

Since $a_1 a_2 h$ is the most likely sequence, the peak state $q_{peak} = a_2$ and the number of days left to peak is thus 1.

5. EXPERIMENTS

5.1 Experimental Setup

To validate the efficacy of the proposed method on the peak day prediction for viral videos, we conducted experiments on the videos with the insight data in our dataset. Since the daily view count is available in the insight data, the ground truth peak day of a video was set to the date with the largest view count. Following [11], the date at which a method makes predictions is called the *reference date*. Each method can access any information before and after the reference date in the training set. However, it only allow to use information before the reference date in the test set in order to forecast the peak day in future. Since the task is to forecast the peak day for viral videos, an ideal system should only report the peak day for viral videos and remain silence for non-viral videos. The viral and non-viral videos were equally divided into a training and a test set. A commonly used metric AP (Average Precision) was used to evaluate the precision of estimated peak days, which is immune to the threshold in prediction methods. We compared the proposed method with the following baseline methods:

Plain HMM [13]: Plain HMM without the variable θ, but with smoothed transaction probability discussed in Section 4.

S-H Model [15]: S-H (Szabo-Huberman) Model is a video popularity prediction model. The model assumes that

[6] The short lifespan observation in Section 3 indicates that $P(h|h) > P(h|a)$ that is a video stays in the hibernating state for the most of time. Relaxing this condition is trivial. Enumerating all the state sequences starting from h using the same method for active states in Eq. (4).

there exists a strong linear correlation between a video's accumulated views and views in future. Following the notation in [11], suppose $N(t_r)$ represents the accumulated views up to the reference date t_r. t_t ($t_t > t_r$) denotes the views in future. We have $\ln N(t_t) = \ln r(t_r, t_t) N(t_r) + \xi$, where $r(t_r, t_t)$ is a parameter and ξ is the Gaussian noise term. Given a reference date, we trained a regression model to estimate the weight $r(t_r, t_t)$ in the training data.

ML Model [11]: S-H model can be regarded as a regression model with a single explanatory variable, i.e. the accumulated view count. Multivariate Linear (ML) Model considers several explanatory variables, each of which is a delta view sampled before the reference date. Formally, the prediction is based on the regression function $N(t_t) = \Theta_{(t_r,t_t)} X_{t_r}$, where X_{t_r} is the feature vector of delta views, and Θ is a vector of parameters to estimate.

SVM Regression: SVM regression further extends the S-H model to incorporate the metadata (both time invariant and variant) into the explanatory variables. It also introduces a l_2 regularization term to avoid overfitting.

The baseline methods were selected based on two considerations: (1) the methods cover both the well known [13] and the state-of-the-art methods [15, 11]; (2) the comparison between them helps to isolate the contribution of different components. For example, the contribution of the variable θ can be demonstrated by comparing the proposed method with the plain HMM. Likewise, the contribution of additional features can be shown by comparing SVM Regression with S-H Model. The parameters in the proposed method were selected in terms of the cross-validation performance on the training set. For example, the number of active states was set to 11; the estimated parameters in Eq. (2) were $\alpha = 0.0529$ $\mu = 12.79$ and $\delta = 9.665$; the views were uniformly quantized into five levels according to Eq. (1). By default, the prior in Eq. (1) was estimated using the late fusion of all three types of features discussed in Section 4. The words in the BoW features were stemmed using Porter Stemmer after removing the stop words, and only symbols and English words were included in the vocabulary.

5.2 Comparison with Baseline Methods

Figure 7 compares the performance of the baseline methods, where the x-axis denotes how many days the reference date is prior to the true peak day. Generally, all methods become more accurate while the reference date approaching the true peak day. The proposed method outperforms all baseline methods, and according to the paired t-test, the improvement is significant at the P-value 0.001 level. This result demonstrates that the proposed method's efficacy in forecasting viral videos. Figure 8 shows examples of the peak day prediction result.

Compared with a plain HMM, the superior performance of the proposed method stems from the consideration of the video type. The introduced variable θ leverages metadata to identify viral videos before the peak day. It can adjust the emission probability to emphasize the view counts of viral videos. In contrast, the plain HMM ignores this information and often confuses viral and quality videos. This argument can also be verified by comparing S-H model and SVM regression, where their difference lies in the usage of metadata in prior estimation. Therefore, the results substantiate our argument that considering metadata is beneficial in this task. S-H model and ML model have the poorest performance suggesting the linear correlation assumption [15] is violated in viral videos. According to Observation 3 in Section 4, viral videos usually experience sudden burstiness in their views, and obviously the view count before entering the burstiness does not correlate to the views in the burstiness. SVM Regression also adopts the same assumption. But because it uses the metadata, it is slightly better than S-H and ML model.

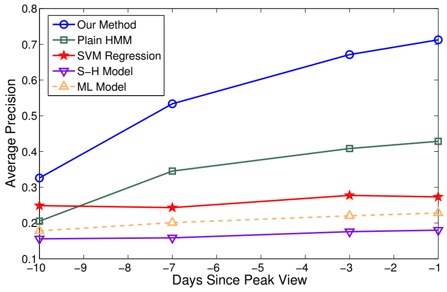
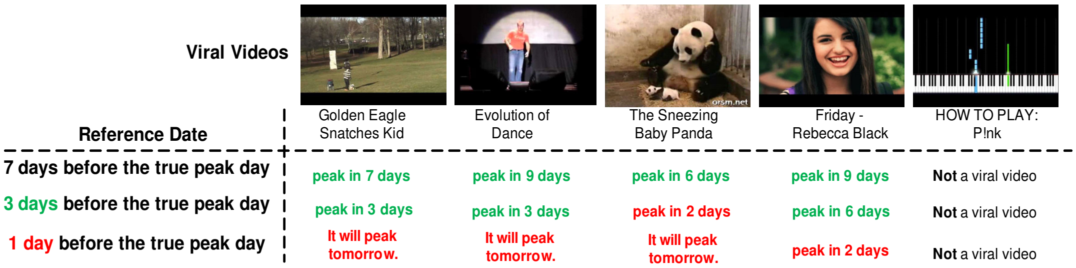


Figure 7: Performance comparison with baseline methods in terms of AP. The x-axis represents the date, at which the method makes predictions, relative to the true peak day.

5.3 Accuracy of Prior Estimation

To study the importance of each feature discussed in Section 4, we compared the accuracy of prior estimation with different types of features and classifiers. Since prior estimation can be regarded as a classification problem, F_1 was adopted to evaluate the accuracy. Table 5 lists the comparison results. Generally, time-invariant metadata turns out to be the single best feature. The result suggests that the characteristics in Table 3, including the duration, title length, are effective in distinguishing viral videos. While the reference date is approaching the true peak day, more information is available, resulting in the increased accuracy of the time-variant metadata. On the other hand, the BoW features and the time-invariant metadata, which are independent of the reference date, remain unchanged. Regarding the classifier, the linear SVM classifier outperforms others with BoW features. Bayesian Network achieves the best performance on both time-variant and time-invariant metadata. The result substantiates the analysis in Section 4 that a classifier modeling correlations in metadata leads to a better performance. Random Forest is the most robust classifier across features. The SVM classifier gets worse performance on the metadata features suggesting that the problem is nonlinear separable in the low dimensional feature space.

To study the impact of the accuracy of prior on the peak day prediction, we conducted experiments where the prior estimated by each feature was added, one at a time, to the proposed method. The experiment was designed to isolate the contribution brought by each feature. Figure 9 illustrates the comparison result. Generally, as we see, the prior plays an important role in the peak day prediction. A bad prior, e.g. Description BoW, may result in a worse model than the plain HMM. The best single feature is the time-invariant metadata, which seems to be consistent to its high F_1 discussed above. The best prior is the late fusion of all features suggesting the proposed features are of complementary information.



Reference Date	Golden Eagle Snatches Kid	Evolution of Dance	The Sneezing Baby Panda	Friday - Rebecca Black	HOW TO PLAY: P!nk
7 days before the true peak day	peak in 7 days	peak in 9 days	peak in 6 days	peak in 9 days	Not a viral video
3 days before the true peak day	peak in 3 days	peak in 3 days	peak in 2 days	peak in 6 days	Not a viral video
1 day before the true peak day	It will peak tomorrow.	It will peak tomorrow.	It will peak tomorrow.	peak in 2 days	Not a viral video

Figure 8: Examples of viral videos' peak day prediction, on different days before the ground truth peak day.

Table 5: Comparison of the prior estimation with different features and classifiers in terms of F_1. NB for Naive Bayes; BN for Bayesian Network; RF for Random Forest.

Time Unrelated	NB	SVM	BN	RF
Title BoW	0.36	**0.40**	0.06	0.19
Description BoW	0.28	**0.29**	0.02	0.04
Time-invariant metadata	0.37	0.16	**0.54**	0.45
Time-variant Metadata	**NB**	**SVM**	**BN**	**RF**
-10 days since the true peak day	0.05	0.03	**0.26**	0.18
-7 days since the true peak day	0.05	0.03	**0.41**	0.34
-3 days since the true peak day	0.05	0.05	**0.52**	0.36
-1 days since the true peak day	0.05	0.05	**0.53**	0.45
On the true peak day	0.05	0.05	**0.54**	0.50

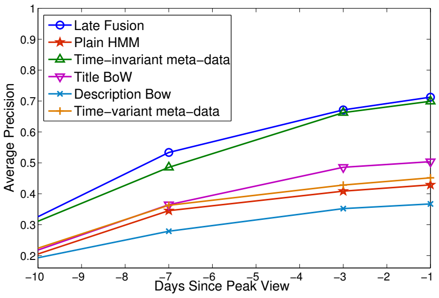


Figure 9: The influence of the different priors on the prediction precision.

6. CONCLUSIONS AND FUTURE WORK

We established by far the largest open dataset of viral videos to date. Having verified existing observations, we discovered some interesting characteristics of viral videos, including the metric of measuring socialness, the evolution of the lifespan, and the correlation residing in the metadata. By studying near duplicate videos, we found that the popularity of the uploader is a more important factor for a video go viral than the upload time. Inspired by our analysis, we introduced a novel approach to forecast the viral video's peak day in future. The proposed method is the first attempt to incorporate metadata in the peak day prediction. The empirical results demonstrate that the proposed method outperforms the state-of-the-art methods with statistically significant difference. We have only used the endogenous features available on YouTube. In future, we plan to exploit both endogenous and exogenous features such as the number of tweet mentions to forecast the peak day of viral videos.

7. ACKNOWLEDGMENTS

The authors are grateful to William Johnson and his colleagues, editors from Time Magazine, and YouTube for annotating viral videos. The authors would like to express gratitude to William Cohen, Bhavana Dalvi and Aasish Pappu for the useful comments and remarks.

8. 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 *SIGKDD*, pages 1186–1194, 2012.

[2] T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer. Catching a viral video. *Journal of Intelligent Information Systems*, 40(2):241–259, 2013.

[3] J. Burgess. All your chocolate rain are belong to us. *Video Vortex Reader: Responses to YouTube*, pages 101–109, 2008.

[4] R. Crane, D. Sornette, et al. Viral, quality, and junk videos on youtube: Separating content from noise in an information-rich environment. In *AAAI symposium on Social Information Processing*, 2008.

[5] F. Figueiredo, F. Benevenuto, and J. M. Almeida. The tube over time: characterizing popularity growth of youtube videos. In *WSDM*, pages 745–754, 2011.

[6] T. Hazen, W. Shen, and C. White. Query-by-example spoken term detection using phonetic posteriorgram templates. In *ASRU*, pages 421–426, 2009.

[7] L. Jiang, Z. Wu, Q. Feng, J. Liu, and Q. Zheng. Efficient deep web crawling using reinforcement learning. In *PAKDD*, pages 428–439, 2010.

[8] L. Jiang, Z. Wu, Q. Zheng, and J. Liu. Learning deep web crawling with diverse features. In *Web Intelligence*, pages 572–575, 2009.

[9] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[10] F. Pernkopf. Bayesian network classifiers versus selective k-nn classifier. *Pattern Recognition*, 38(1):1–10, 2005.

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

[12] L. Rabiner, A. Rosenberg, and S. Levinson. Considerations in dynamic time warping algorithms for discrete word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(6):575–582, 1978.

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

[14] G. Rocket. Emerging trends in viral video and the implications for advertising. http://www.slideshare.net/GeniusRocket/viral-video-research-presentation.

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

[16] W. Tong, Y. Yang, L. Jiang, S.-I. Yu, Z. Lan, Z. Ma, W. Sze, E. Younessian, and A. G. Hauptmann. E-lamp: integration of innovative ideas for multimedia event detection. *Machine Vision and Applications*, pages 1–11, 2013.

[17] T. West. Going viral: Factors that lead videos to become internet phenomena. *Elon Journal of Undergraduate Research*, pages 76–84, 2011.

[18] X. Wu, A. Hauptmann, and C. Ngo. Practical elimination of near-duplicates from web video search. In *ACM Multimedia*, pages 218–227, 2007.

[19] L. Xie, A. Natsev, J. R. Kender, M. Hill, and J. R. Smith. Visual memes in social media. In *ACM Multimedia*, pages 53–62, 2011.

[20] R. Zhou, S. Khemmarat, and L. Gao. The impact of youtube recommendation system on video views. In *SIGCOMM Conference on Internet Measurement*, pages 404–410, 2010.