LARM: A Lifetime Aware Regression Model for Predicting YouTube Video Popularity

Changsha Ma
State University of New York at Buffalo
338 Davis Hall
Buffalo, NY 14260-2500
changsha@buffalo.edu

Zhisheng Yan
Georgia State University
P.O. Box 3994
Atlanta, GA 30302-3994
zyan@gsu.edu

Chang Wen Chen
State University of New York at Buffalo
316 Davis Hall
Buffalo, NY 14260-2500
chencw@buffalo.edu

ABSTRACT

Online content popularity prediction provides substantial value to a broad range of applications in the end-to-end social media systems, from network resource allocation to targeted advertising. While using historical popularity can predict the near-term popularity with a reasonable accuracy, the bursty nature of online content popularity evolution makes it difficult to capture the correlation between historical data and future data in the long term. Although various existing efforts have been made toward long-term prediction, they need to accumulate a long enough historical data before the prediction and their model assumptions cannot be applied to the complex YouTube networks with inherent unpredictability.

In this paper, we aim to achieve fast prediction of long-term video popularity in the complex YouTube networks. We propose LARM, a lifetime aware regression model, representing the first work that leverages content lifetime to compensate the insufficiency of historical data without assumptions of network structure. The proposed LARM is empowered by a lifetime metric that is both predictable via early-accessible features and adaptable to different observation intervals, as well as a set of specialized regression models to handle different classes of videos with different lifetime. We validate LARM on two YouTube data sets with hourly and daily observation intervals. Experimental results indicate that LARM outperforms several non-trivial baselines from the literature by up to 20% and 18% of prediction error reduction in the two data sets.

CCS CONCEPTS

•**Information systems** → **Social networks** ; **Data mining;**

KEYWORDS

social media; YouTube; popularity prediction; regression model; time series analysis

1 INTRODUCTION

With the prevalence of social media, the ever-increasing YouTube videos have played a dominant role in Internet traffic. For example, 300 hours of videos are uploaded to YouTube every minute, which results in YouTube's 18% share of Internet traffic in North America [1]. A key characteristic of these enormous amount of user-generated content (UGC) is that they enjoy extremely diverse popularity and popularity evolution patterns [1]. A small part of the videos attract over-billion views while the majority of the videos are rarely viewed. Similarly, some videos can stay popular for a long time while some others freeze in only several hours after being uploaded. Therefore, it is of substantial value to predict the future popularity of UGC in order to benefit the end-to-end social media system in network resource allocation, personalized content recommendation, targeted advertising, etc.

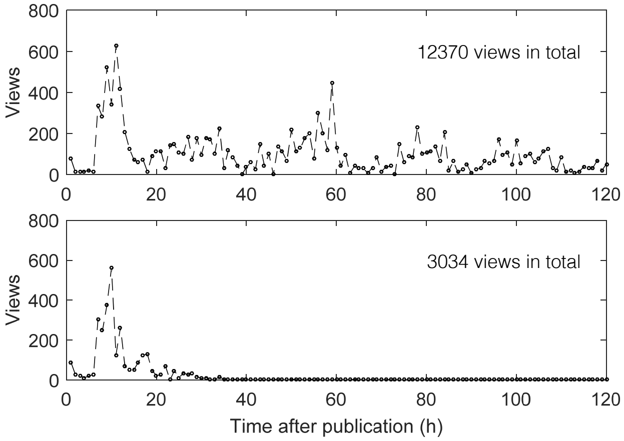


Figure 1: YouTube videos with similar early-stage popularity exhibit distinct long-term popularity

Historical popularity has been proven to be a strong predictor for future popularity [10]. It can be employed in either the classification problem that differentiates popular and unpopular contents [3, 25], or the regression problem that predicts the exact value of the popularity metric (e.g. views for YouTube videos) [4]. Unfortunately, we observe that simple historical popularity based prediction cannot accurately predict the UGC popularity *in the long term*. For example, considering that future popularity is decided by early popularity by a single factor, the simple model proposed in [18] can only produce reasonable results within a few steps beyond the prediction moment. The fundamental reason behind this observation is that

CIKM'17, November 6–10, 2017, Singapore.
© 2017 ACM. ISBN 978-1-4503-4918-5/17/11…$15.00
DOI: https://doi.org/10.1145/3132847.3132997

[1] https://www.sandvine.com/pr/2015/12/7/sandvine-over-70-of-north-american-traffic-is-now-streaming-video-and-audio.html

the bursty nature of online content popularity evolution makes it difficult to directly capture the correlation between historical data and future data for a long term. According to our study shown in Figure 1, two YouTube videos that have similar initial popularity can achieve totally distinct popularity in the end.

The goal of this paper is to predict the long-term popularity of YouTube videos. In addition to the prediction accuracy, we identify two key requirements for such a prediction to benefit YouTube. First, it is critical to support fast popularity prediction, i.e., predicting the popularity as soon as possible after the video is uploaded. Waiting a long time to accumulate historical data may negatively impact the system management decisions such as advertisement and recommendation. Second, the popularity prediction should be generic such that it can be applied into the complex YouTube networks with inherent unpredictability. Classic assumptions of content propagation may not work under such a complicated network structure [20, 26].

Although long-term popularity prediction of UGC has been recently studied, these efforts fail to satisfy one or more of the key requirements. Time series based approach usually leverages the dynamics of a extended period of historical popularity [4–8] to predict the popularity. However, they suffer the cold-start problem since they have to accumulate enough historical data before achieving a satisfactory prediction performance [13]. Furthermore, generative model based approach models the popularity evolution with additional assumptions such as the structure of networks over which the content propagates [14–17], e.g., propagating as cascades. They are not applicable for such a complex social network as YouTube and would cause barriers in explaining content propagation, popularity evolution, and parameter inference.

In this paper, we aim at achieving fast prediction of long-term popularity for online videos in complex YouTube networks. We propose LARM, a lifetime aware regression model, representing the first work that leverages content lifetime to compensate the insufficiency of historical data without underlying assumptions of the network structure. The intuition is that future popularity is not only impacted by past popularity, but also driven by the video's ability to keep its attractiveness for users over a long period, which is reflected by lifetime. As exemplified in Figure 1, if we have the knowledge of video lifetime (instead of the early-state dynamics in the first 20 hours) in advance, we would achieve a timely long-term prediction for the video popularity.

To realize LARM, we are facing two technical challenges. First, how to define and predict the lifetime of a video in advance is unclear, which is the key to the lifetime aware prediction. Second, a prediction method that utilizes the lifetime needs to be deliberately designed for videos with a diverse range of lifetime.

To tackle these challenges, we first define the lifetime metric as the α-lifespan of videos, i.e., the time point when videos have accumulated an α ratio of their views. The reason for this choice is that α-lifespan is highly predictable when utilizing early-accessible features. It is also flexible and can be adapted to different observation intervals. Furthermore, we develop a robust classifier to divide the studied online contents into classes with various lifetime, where a specialized regression model is further trained for each class of contents to precisely predict their future popularity.

We have crawled two YouTube data sets with different observation intervals, i.e. hourly and daily, and validated the designs of LARM via extensive experiments in various practical conditions. Experimental results show that LARM significantly outperforms several non-trivial baselines from the literature, with up to 20% and 18% of prediction error reduction in hourly and daily data set, respectively.

To summarize, we make the following three main contributions:

- We leverage lifetime estimation to compensate the insufficiency of historical data in long-term online content popularity prediction, which gains us the advantage of fast prediction and easy interpretation.
- We formally define the lifetime metric that is both predictable via early-accessible features and applicable for model specialization. We believe it throws light on how to exploit related features in popularity prediction.
- We present a practical demonstration of LARM, which validates its prediction performance as well as its flexibility to adapt to data sets with different observation intervals.

This paper is organized as follows. We present the related work in Section 2. In Section 3, we describe our data sets that are used to validate the proposed schemes. We introduce how we predict the video lifetime in Section 4. Then we proceed to show how the lifetime knowledge is utilized in YouTube video popularity prediction in Section 5. In Section 6, we present the empirical results and we conclude this paper in Section 7.

2 RELATED WORK

The technical challenges of long-term popularity prediction for online contents are mainly resulted from two aspects, i.e. the uncertainty of the driving factors for the popularity evolution, and the insufficiency of available data used for prediction [9]. There are a plenty of recent efforts towards tackling the above challenges. We discuss these works by categorizing them into generative model based approaches that attempt to explain the popularity evolution, time series based approaches that utilize the historical popularity dynamics, as well as feature driven approaches that explore the hidden factors impacting popularity.

Generative Model: This type of methods model each content propagation as a stochastic event under some specific assumptions about the popularity evolution. In [27], the spread of tweets is treated as a reinforced Poisson process. In [14], cascading is modeled as a Hawkes process with two components, i.e. the human reaction time of sharing and the time-varying post infectiousness of tweets. The parameters are inferred utilizing the network information including the the time point of each sharing and the out-degree of each sharer. In [16], a factor indicating the influence of tweet publishers is additionally modeled. A recent work [21] assumes that UGC popularity is a combination of endogenous factors such as user interactions within the system and exogenous factors such as external events. Although these methods have been successfully applied to study the cascades on social networks such as Twitter and Facebook [3], they are not applicable for complex YouTube systems where the content propagation process is less explainable and more difficult to model. For example, the main sources that attract video views in YouTube are not the sharing. Instead, the

Table 1: Data set statistics

Data set	No. videos	Study period	Mean views	Mean video length	Top-3 categories
Daily	28,190	First 100 days	25730	702 seconds	People&blog, gaming, entertainment
Hourly	14,933	First 120 hours	5194	750 seconds	Gaming, people&blog, entertainment

searching and related video recommendation [2], which are very difficult if not impossible to infer [20], play an important role in video popularity evolution.

Time Series Approach: Time series approach explores the effective ways of using available historical popularity series for the prediction targets. H. Pinto et. al proposes to predict the future popularity of YouTube videos by assigning different weights to video popularity in past different days through a multiple-linear regression (MR) model [4]. They further improve the MR model by considering the differences of videos in popularity evolution. Specifically, they first select a number of videos as radial basis function (RBF) centers. They then measure the early-stage popularity similarity between videos by computing the RBF values to the centers, and feed the RBF features in the regression model (MRBF). The problem of MRBF is that both the optimal number of RBF features and the selection of them require exhaustive experiments, resulting in tremendous training cost. In [7], the popularity series is modeled as a pure birth process (PBP), where the growth of popularity is related to their historic dynamics. However, it does not differentiate the popularity evolution of different videos and infers the model parameters using the whole training data set. It is proposed in [22] that videos can be first classified into groups based on their popularity phases. The information can then improve the precision of the regression based popularity prediction method. Unfortunately, it does not elaborate on how to identify the future popularity phases of a video at the early stage of the video. In fact, this process needs to accumulate enough historical data, making it infeasible for fast prediction. M. Ahmed et. al. model the past popularity as a chain of states, which are determined by time and the defined popularity rate of change. The future popularity is then predicted by inferring the transition probability among states [6]. This method is able to coarsely predict the popularity level of online contents but cannot predict the exact value. One common issue of time series based approaches is that they only depend on the historical popularity, which easily result in a cold-start problem.

Feature Driven Approach: This type of method exploits the potential features that impact popularity evolution. Except for the historical popularity series, content features such as the language, length, and sentiment of tiltles [10], original poster features like the number of followers, the number of past posts, and the average popularity of past posts [12], and the user interactions including the user comments, likes, and dislikes [28] are all shown to be correlated to content popularity. Additionally, cross-platform information such as related tweets on twitter is considered to impact the video views [19, 23, 24]. A recent research shows that how a YouTube video is discovered also affects its views [2]. Directly learning based on these features can achieve comparable results by only utilizing historical popularity series in popularity classification task. In [16], these features are applied to add a predictive layer on the generative model to improve the prediction accuracy in popularity regression task. However, how to combine these features with historical popularity series for popularity prediction in complex social system is still an open problem. To bridge this gap, the proposed work reasonably combines the historical popularity series with the lifetime, which is estimated by all other popularity related features.

3 DATA SET

To motivate and validate the proposed LARM, including both the video lifetime prediction and video popularity prediction, we have crawled two data sets from YouTube Data API v3[2]. In this section, we start with describing the statistics of our data sets. We then examine the differences of the two data sets in terms of video popularity evolution patterns, which provides important insights for the design of LARM.

3.1 Data Set Statistics

The first data set is a daily data set that contains 631,459 videos. We track each video's popularity (views) every day for 100 consecutive days since it is uploaded. The other data set is an hourly data set containing 172,602 videos, where each video's popularity is tracked every hour for 120 consecutive hours. Therefore, we have a popularity series for each video with the length of 100 and 120 in the daily data set and the hourly set, respectively. We observed that some videos are revoked from YouTube after being uploaded for some time, resulting in a non-increasing popularity series. Additionally, a large portion of videos have never been watched since their upload, leading to an all-zero popularity series. We filter out both types of videos, and focus our studies on the remaining 28,190 videos in the daily data set and 14,933 videos in the hourly data set.

Table 1 summarizes the statistics of the two data sets. Note that mean views are the average of the cumulative views received by all videos at the end of the study. We can see that the hourly data set exhibits smaller mean views than the daily data set since it has a relatively short study period. On the other hand, all other video statistics of the two data sets, such as video length and categories, are very similar.

3.2 Video Popularity Evolution Pattern

In this study, we investigate the difference of popularity evolution shapes among different videos. Specifically, we represent each video v in the data set as a time series $(\frac{N_v(t_1)}{N_v(t_e)}, \frac{N_v(t_2)}{N_v(t_e)}, ..., 1)$, where $N_v(t_i)$ is the total views that v receives at time t_i, and t_e is the end of the study time. We then use K-means to cluster the videos in a data set into six clusters, and repeat this clustering for the other data set. We select six representative evolution patterns from each cluster for each data set and show them in Figure 2. The fraction

[2] https://developers.google.com/youtube/v3/

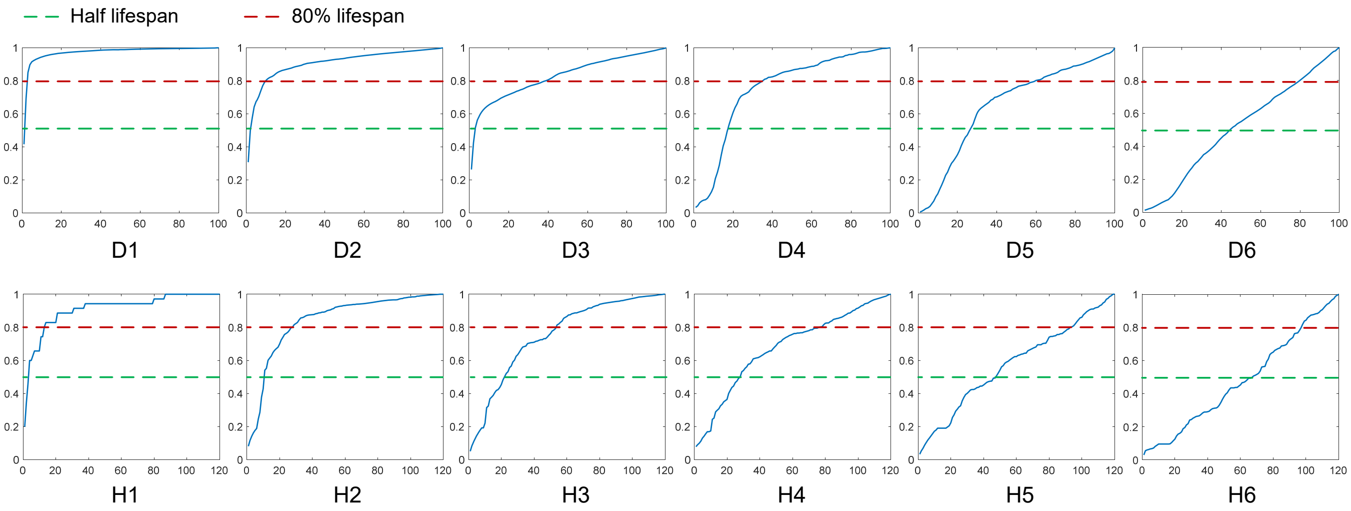


Figure 2: Six representative patterns in daily data set (first row) and hourly data set (second row)

Table 2: Percentage of different video popularity patterns (pattern # 1-6) in data sets (D: daily; H: hourly)

	1	2	3	4	5	6
D	25.7%	25%	19.8%	15%	10%	4.5%
H	11.4%	25.5%	28.3%	21%	10%	3.8%

of the six patterns in the two data sets are summarized in Table 2. We can observe that only a small portion of videos (4.5% in daily data set, and 3.8% in hourly data set) enjoy a linear increase in their popularity, while most of the videos have a dynamic popularity increasing rate. For example, for videos in D1, the increasing rate of video popularity drops dramatically at the very early stage, and videos in D4 exhibit a low-high-low pattern of the increasing rate. It is hence unlikely to accurately capture the long-term future popularity by training a single prediction model to adapt to such kind of pattern dynamics and diversity. This motivates us to build specialized models for different subsets to improve the prediction performance in Section 5.

Additionally, we find that the two data sets have very different sets of popularity evolution patterns. For example, for as many as 25.7% of videos in the daily data set, the increasing rate of video popularity drops dramatically in the very early stage. However, such a pattern is not common in hourly data set. This observation requires us to design a prediction scheme that is able to adapt to different observation intervals.

4 VIDEO LIFETIME PREDICTION

Video lifetime captures the persistence of a video in keeping its attractiveness to users. It is different from popularity, which instead captures the extent of the video attractiveness. In this section, we first formally define the metric for video lifetime that can address the insufficient historical data for fast prediction. Then we proceed to investigate its predictability utilizing a set of early-accessible features.

Table 3: The α-lifespan statistics of video subsets with different popularity patterns

$\alpha = 0.5$

variance	1	2	3	4	5	6	all
D	16.0	3.0	30.4	0.5	48.7	87.8	245.9
H	137.5	40.3	30.1	16.0	2.4	93.1	307.0

mean	1	2	3	4	5	6	all
D	1.5	3.3	8.7	20.8	37.2	60.2	12.5
H	2.5	7.6	15.9	26.6	47.5	77.2	19.7

$\alpha = 0.8$

variance	1	2	3	4	5	6	all
D	60.0	37.9	72.3	6.2	115.4	118.0	666.3
H	136.5	106.7	103.3	75.0	55.0	134.2	668.4

mean	1	2	3	4	5	6	all
D	4.5	18.0	37.8	54.8	70.8	83.6	32.5
H	13.0	32.0	50.9	69.9	87.3	101.8	51.0

4.1 The Lifetime Metric

A naive lifetime metric would be the time interval from a video's upload to the moment when no view is received by the video for a certain period of time. However, it is not practical to predict video lifetime under such a metric. This is because the observation time is always finite and sometimes short due to the cost of crawling data sets, some videos may have not experienced any freeze during the whole study period [11]. The incomplete knowledge of videos will unavoidably add noises on the lifetime estimation. To avoid this issue, we define the lifetime metric in LARM as the α-lifespan of videos. It refers to the time point when videos have accumulated an α ratio of their views during the study period, which can be as short as several hours or as long as multiple days. In general, a smaller value of α leads to a heavier head in the lifetime distribution, while a larger one leads to a more even distribution. Figure 3 shows the

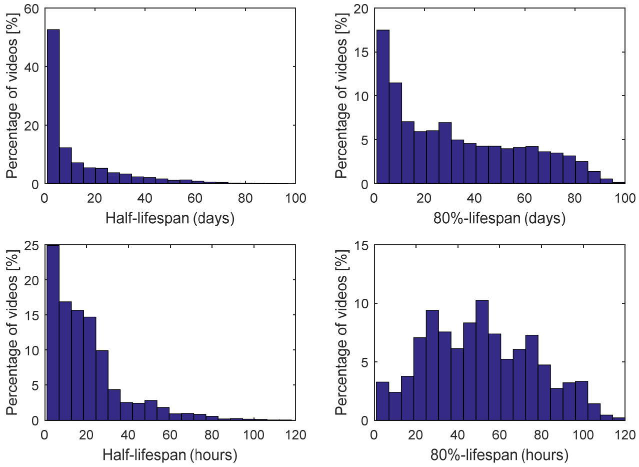


Figure 3: The distribution of video lifetime in two data sets

lifetime distribution of the two data sets when α is set as 50% and 80%, respectively. We can see that more than 50% videos in the daily data set and 25% videos in the hourly data set have a small lifetime (the first bin) under the half-lifespan metric. More videos tend to have a longer lifetime under the 80%-lifespan metric. For simplification, we will use the same settings of α (50% and 80%) to demonstrate LARM.

A noticeable feature of α-lifespan metric is that it provides efficiency and flexibility in separating data, which is essential for us to build specialized models for a data set. We provide the explanations from the following two aspects.

First, the α-lifespan metric characterizes different popularity evolution patterns. From Figure 2, we can observe that different patterns exhibit different values of α-lifespan. For example, the half-lifespans of D1, D2, D3, D4, D5 and D6 are 2, 3, 5, 16, 25 and 43 days, respectively. Their 80%-lifespans are also different, which are 3, 10, 39, 37, 60 and 77 days, respectively. Similarly, H1, H2, H3, H4, H5 and H6 have 4, 10, 20, 30, 47 and 68 hours for half-lifespan, and 15, 29, 55, 76, 95, 96 hours for 80%-lifespan, respectively. We further examine the variance and mean of α-lifespan values for the K (K = 6) clusters, and summarize the results in Table 3. For both of the two data sets, each cluster has a distinct mean of α-lifespan values. In addition, the variance of α-lifespan values for a cluster is much lower than that for the whole data set. These results indicate the correlation between the videos' alpha-lifespan and popularity evolution pattern.

Second, the α-lifespan metric is able to adapt to the distribution of popularity patterns of a data set by adjusting the value of α. For example, in our daily data set, D1 and D2 are the top two patterns, and we want to address more about their difference. In this case, the 80%-lifespan metric is preferred since the means of lifespan values of D1 and D2 are more distinct than that under the half-lifespan metric. We will discuss more details about the impact of the selection of α in popularity prediction results in Section 6.

Table 4: Early-accessible features for lifetime prediction

Predictor	Description	Used in Model
c_v	Video category	1-10
h_v	Time length of video	2-10
u_p	Number of channel views	3-10
u_m	Number of channel comments	4-10
u_s	Number of channel subscribers	5-10
u_v	Number of channel videos	6-10
$N_v(t_1)$	Number of initial views	7-10
$N_m(t_1)$	Number of initial comments	8-10
$N_l(t_1)$	Number of initial likes	9-10
$N_d(t_1)$	Number of initial dislikes	10

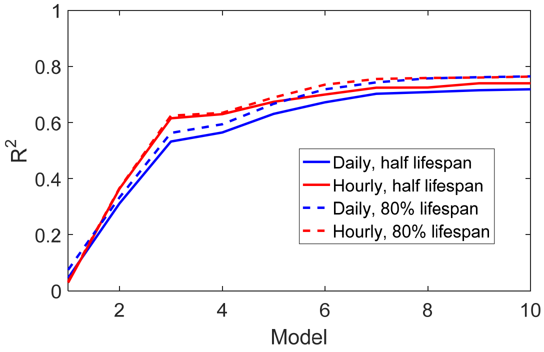


Figure 4: Predictability of lifetime using different sets of early-accessible features

4.2 The Predictability of Lifetime

In this section, we present how to predict the α-lifespan using early-accessible features in order to support the fast popularity prediction in LARM. We study three types of features that may potentially impact the video lifetime, i.e., channel features, popularity features, and content features. Specifically, channel features capture the social impact and the past success of the channel that uploads the target video. We collect the number of channel subscribers u_s, the total videos uploaded by the channel u_v, and the total views u_p and total comments u_m received by these videos. Popularity features capture the activity of user interactions with the target video, including number of views $N_v(t_1)$, comments $N_m(t_1)$, likes $N_l(t_1)$, and dislikes $N_d(t_1)$. Since we focus on fast prediction, we only collect the popularity features at the initial observation, i.e. the first hour or the first day upon the upload of videos. Finally, we collect the content features including the video length and the video category.

We proceed to evaluate how well the α-lifespan of a video can be predicted using these features. In particular, we choose regression tree as the prediction model, since it is appropriate to capture the non-linearity between features and the prediction target. Additionally, we use R^2, the coefficient of determination, which indicates the fraction of variance explained by the regression model, to evaluate the predictability of α-lifespan. In order to isolate effects of different features, we train multiple models using different subsets of the features. Table 4 summarizes the notations of the features and their usages in models.

We train the models using half of our daily data set and hourly data set, and evaluate the α-lifespan ($\alpha = 50\%, 80\%$) using the other half of the data. From the prediction results shown in Figure 4, we can see that model 2 achieves a notable performance improvement over model 1. This indicates that adding content features like time length of video h_v have significant impact on video lifetime. Similarly, some channel features, such as the number of channel views u_p, have even more significant effect on α-lifespan. While there exist popularity features that can further improve the prediction, e.g., the number of initial views $N_v(t_1)$, most other popularity features do not bring significant improvements. This may be interpreted as that the popularity features are highly correlated with each other. Moreover, it is interesting to see that 80% lifespan is generally more predictable than half lifespan. We will show how this effect may impact popularity prediction in Section 5.

In summary, α-lifespan is predictable using the early-accessible features. This is highly desirable since we can predict video lifetime and utilize it for fast popularity prediction when the video is uploaded.

5 VIDEO POPULARITY PREDICTION

In this section, we introduce the proposed video popularity prediction scheme in LARM, which utilizes video lifetime to build a set of specialized prediction models for videos. Specifically, we first present an overview of LARM. Then we discuss the details of LARM by answering two fundamental questions: 1) How to specialize models in order to improve prediction performance; 2) What affects the performance of the long-term popularity prediction when using specialized modes.

5.1 Overview of LARM

LARM predicts the future popularity of a video as a linear function of its observed popularity series as in [4]. Unlike a large body of existing linear prediction works that uses a single set of parameters for all videos as in related work, LARM separates the videos into multiple subsets and trains a specialized model for each of these subsets.

Let t_0 be the upload time of video v, L_v be the estimated lifetime of the video, $N_v(t_i)$ be the total views received by v at t_i ($N_v(t_0) = 0$), and $x_v(t_i)$ be the number of views of the i-th time interval, i.e., $x_v(t_i) = N_v(t_i) - N_v(t_{i-1})$. LARM attempts to predict the actual popularity $N_v(t_t)$ at the future time t_t ($t_t > t_r$) utilizing L_v and $X_v(t_r) = (x_v(t_1), x_v(t_2), ..., x_v(t_r))$, and outputs the predicted value $\hat{N}_v(t_r, t_t)$. Specifically, LARM first identifies the specialized model k for v based on L_v, and then makes prediction as in (1), where $\Theta_k(t_r, t_t) = (\theta_1^k, \theta_2^k, ..., \theta_r^k)$ is the parameter of model k, and K is the total number of specialized models used by LARM.

$$\hat{N}_v(t_r, t_t) = \Theta_k(t_r, t_t) \cdot X_v(t_1, t_r) \quad (k = 1, 2, ..., K) \tag{1}$$

We use mean absolute percentage error (MAPE) as the prediction performance criterion, as in related work [14]. MAPE is defined as follows.

$$MAPE = \frac{|\hat{N}_v(t_r, t_t) - N_v(t_t)|}{N_v(t_t)} \tag{2}$$

To train the model parameters, we formulate the optimization problem as in (3), where C_k is the training set for model k. We minimize the prediction error defined by MAPE in order to obtain the parameters. By treating $\frac{1}{N_v(t_t)^2}$ as the weight, this can be easily solved as a weighted least square problem.

$$\arg\min_{\Theta_k(t_r, t_t)} \frac{1}{C_k} \sum_{v \in C_k} \frac{|\hat{N}_v(t_r, t_t) - N_v(t_t)|}{N_v(t_t)} \tag{3}$$

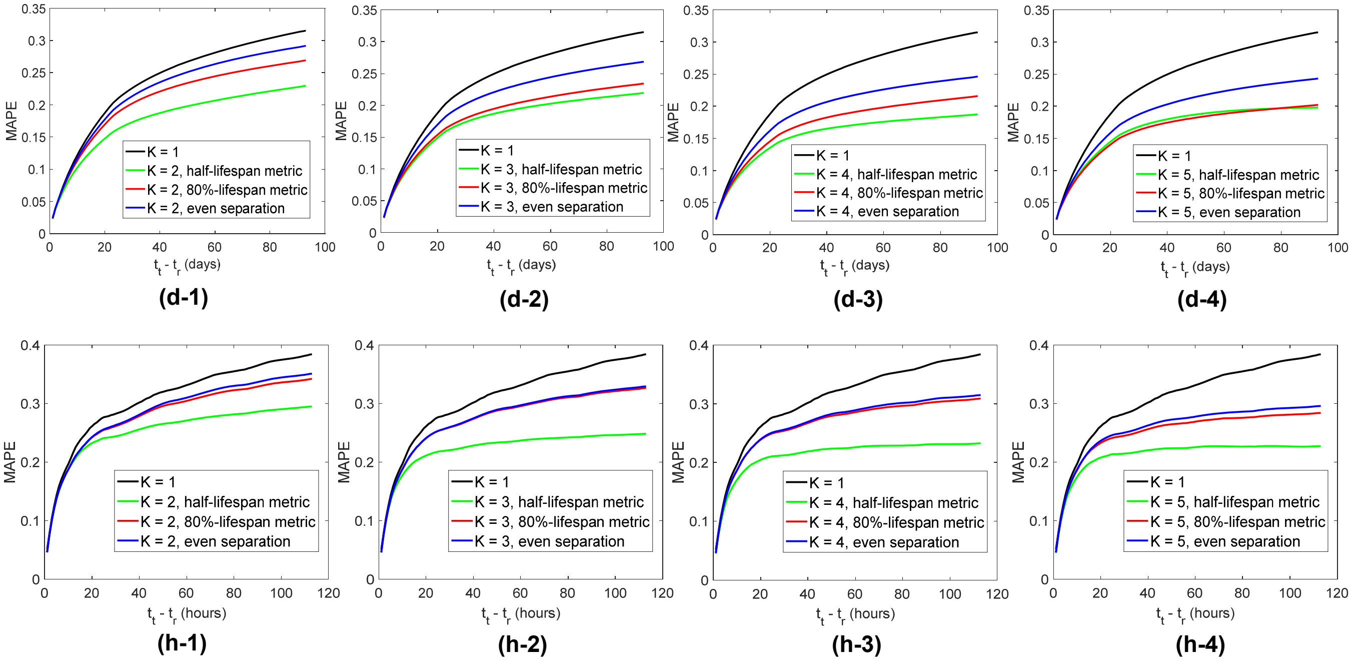
5.2 Model Specialization

To build K specialized models, we need to first divide the training data set into K subsets. Formally, it requires us to identify an increasing sequence of lifetime boundaries $L = (l_1, l_2, ..., l_{K-1})$. The optimal L^* should lead to the best overall popularity prediction performance. If we have the longest possible lifetime as N, then the searching space for L^* is as large as N^{K-1}. Training a regression model to evaluate an estimated $\hat{L^*}$ is an process with $O(M)$ complexity, where M is the size of the data set. Repeating such a process for N^{K-1} times would introduce tremendous cost and would not be feasible in practice.

In this paper, we switch to search for a near-optimal L with efficient computation. In particular, we use K-means to first cluster all videos from the data set into K clusters, using the same similarity measurement as in Section 3, i.e. the video popularity evolution pattern. We then sort the K clusters according to the average α-lifespan of the videos in the cluster in an increasing order. We also denote l_i^* as the largest α-lifespan in cluster i, i.e., the cluster boundary. The reason of this choice is that α-lifespan can be treated as a representation of popularity evolution pattern, which inherently impacts the popularity predictions. We can then utilize the $K - 1$ lifetime boundaries to divide the training data sets into K subsets, and train $\Theta_k(t_r, t_t)$ for a given pair of (t_r, t_t) under the optimization problem in (3).

To valid the effectiveness of the proposed specialization scheme, we show the prediction performance using $K = 2, 3, 4, 5$ specialized models, respectively. Specifically, we set $t_r = 7$ days and t_t from 8 to 100 days for daily data set, and $t_r = 7$ hours and t_t from 8 to 120 hours in the hourly data set. Note that $t_r = 7$ generally indicates a small amount of historical data. For comparison, we choose two baselines. The first baseline uses a single model for the whole training data set (K = 1). The second one employs the same amount of specialized models but configures an L that evenly divides the study time into K segments. For example, if K is set as 4, the baseline set L as $(30, 60, 90)$ for hourly data set, and $(25, 50, 75)$ for daily data set, respectively. Additionally, in order to isolate the prediction error caused by lifetime prediction, we use the ground truth value of video lifetime to perform the comparison.

The comparison results are shown in Figure 5. We can observe that model specialization are able to decrease MAPE compared to using a single model. For example, the MAPE is reduced by up to 0.13 in the long term when four specialized models based on half-lifespan metric are used in daily data set, and up to 0.19 when five specialized models based on half-lifespan metric are used in hourly data set. Using 80%-lifespan metric based specialization also reduces the MAPE in both data sets. We can observe that the proposed model specialization scheme outperforms the naive



(d-1) (d-2) (d-3) (d-4)

(h-1) (h-2) (h-3) (h-4)

Figure 5: The expected MAPE reduction of using specialized models (note that this is not the ultimate performance of LARM, since the lifetime prediction error is isolated)

specialization scheme, which achieves a MAPE reduction up to 0.12 in the daily data set, and up to 0.09 in the hourly data set.

In summary, the proposed model specialization scheme is able to significantly improve the long-term popularity prediction performance.

5.3 Model Optimization

Now we examine what factors affect the performance of specialized models based prediction, so that we can optimize the scheme. First, α plays an important role as it defines the lifetime of videos. Figure 5 clearly indicates that model specialization using the half-lifespan metric outperforms that using the 80%-lifespan metric. This is especially true for the hourly data set. Besides, we even find that the proposed specialization scheme does not significantly reduces MAPE compared to even specialization scheme for the hourly data set if 80%-lifespan metric is used. As we have discussed in Section 4, 80%-lifespan metric does not differentiate the video popularity evolution patterns and therefore it is critical to select the proper alpha ratio for the lifetime metric.

Apart from the selection of α ratio, we can observe from Figure 5 that a larger K tends to bring better results than smaller ones. However, it is still not wise to set an extremely large K, since it can easily result in the over-fitting issue. By comparing the prediction performance in the cases (d-4) and (d-3), we see that using five specialization models leads to worse performance than using four models when half-lifespan metric is used.

To sum up, we should deliberately choose α and K to achieve optimal performance of LARM according to the characteristics of the data set.

6 PERFORMANCE EVALUATION

In this section, we compare the popularity predication performance of LARM to state-of-the-art schemes.

6.1 Baselines for Comparison

We consider two different time series based prediction methods for comparison. Furthermore, we also compare LARM with a feature driven method to validate the effectiveness of LARM on combining historical popularity and other early-accessible features.

MRBF [4]: The MRBF method is an extension of MR model that considers the differences of videos in popularity evolution, which are measured by RBF defined as in (4). The prediction model of MRBF is formulated as in (5), where C is the set of videos selected as RBF centers.

$$RBF(v, v_c) = e^{(-\frac{||X_v(t_1,t_r)-X_{v_c}(t_1,t_r)||^2}{2\cdot\sigma^2})} \tag{4}$$

$$\hat{N_v}(t_r, t_t) = \Theta_{t_r,t_t} \cdot X_v(t_1, t_r) + \sum_{v_c \in C} \omega_{v_c} \cdot RBF(v, v_c) \tag{5}$$

VCDM [7]: The view counts dynamic model (VCDM) models the popularity evolution as a pure birth process (PBP) and infers the parameter using the historical popularity time series.

Regression Tree: It feeds the historical popularity as well as a subset of features listed in Table 4 into the prediction model. Specifically, we set the feature set as $(X_v(t_r, t_t), u_v, u_s)$ since it consistently provides the most accurate predictions.

Table 5: MAPE performance comparison in daily data set

Algorithms	LARM-0.5	LARM-0.8	MRBF	VCDM	Tree
$t_r = 1$	**0.4462**	0.4570	0.5035	0.5051	0.5106
$t_r = 2$	0.3983	**0.3904**	0.4532	0.4452	0.4623
$t_r = 3$	0.3573	**0.3505**	0.4104	0.4042	0.4161
$t_r = 4$	0.3242	**0.3185**	0.3773	0.3681	0.3810
$t_r = 5$	0.2976	**0.2922**	0.3512	0.3400	0.3512
$t_r = 6$	0.2744	**0.2686**	0.3281	0.3166	0.3283
$t_r = 7$	0.2590	**0.2540**	0.3011	0.2974	0.3152

Table 6: MAPE performance comparison in hourly data set

Algorithms	LARM-0.5	LARM-0.8	MRBF	VCDM	Tree
$t_r = 1$	**0.4228**	0.4450	0.5105	0.5170	0.5330
$t_r = 2$	**0.3786**	0.4050	0.4833	0.4746	0.5018
$t_r = 3$	**0.3466**	0.3691	0.4441	0.4336	0.4664
$t_r = 4$	**0.3248**	0.3438	0.4173	0.4090	0.4424
$t_r = 5$	**0.3123**	0.3290	0.4012	0.3943	0.4229
$t_r = 6$	**0.2985**	0.3136	0.3889	0.3794	0.4003
$t_r = 7$	**0.2773**	0.2910	0.3671	0.3564	0.3842

Table 7: EPA performance (t_r) comparison in daily data set

Algorithms	LARM-0.5	LARM-0.8	MRBF	VCDM	Tree
$t_t = 50$	**4**	**4**	5	5	5
$t_t = 60$	**4**	**4**	6	6	6
$t_t = 70$	5	**4**	6	6	7
$t_t = 80$	**5**	**5**	7	6	7
$t_t = 90$	**5**	**5**	7	7	8
$t_t = 100$	6	**5**	7	7	8

Table 8: EPA performance (t_r) comparison in hourly data set

Algorithms	LARM-0.5	LARM-0.8	MRBF	VCDM	Tree
$t_t = 60$	**5**	6	8	8	8
$t_t = 70$	**5**	6	8	8	9
$t_t = 80$	**6**	7	9	9	10
$t_t = 90$	**6**	7	9	9	10
$t_t = 100$	**6**	7	9	9	12
$t_t = 110$	**6**	7	11	11	13
$t_t = 120$	**7**	**7**	11	11	14

6.2 Evaluation Setup

We evaluate LARM and the baseline approaches using a 3-fold cross validation. We set up two experiments in order to highlight the fast prediction advantage of LARM. Specifically, we first evaluate the accuracy of long-term popularity prediction for LARM and baselines under various amount of historical data. In this experiment, we use MAPE as the evaluation metric. On the other hand, the second experiment measures the amount of early data that is needed to achieve a given prediction target and a tolerable error.

6.3 Prediction Improvement of LARM

We set this historical data amount t_r as 1 to 7, and the target prediction moment t_t as 100 and 120 for the daily data set and the hourly data set, respectively. In this way, t_r is small enough, and $t_t - t_r$ is large enough to guarantee that this is a long-term popularity prediction problem. According to the empirical study, we use the feature set $(N_v(t_1), N_c(t_1), u_s, c_v, h_v)$ in Table 4 to predict the half-lifespan and 80%-lifespan of videos in the data set. Furthermore, we set up six specialization models for LARM so that it can achieve its best performance.

We measure the prediction performance of LARM using both half-lifespan metric and 80%-lifespan metric, which is denoted as LARM-0.5, and LARM-0.8, respectively. Table 5 and Table 6 show the performance comparison of all schemes. We can see that both LARM-0.5 and LARM-0.8 significantly outperform other schemes for all t_r in the two data sets. It brings up to 20% of MAPE reduction in the hourly data set, and 18% of MAPE reduction in the daily data set. LARM outperforms MRBF and VCDM essentially because it introduces lifetime in prediction while the other two only depend on historical popularity. Additionally, although regression tree also introduces other features, directly feeding these features with the historical popularity by a single non-specialized model does not help at all. In contrary, its prediction performance even worse than the pure historical data based approaches.

It is worthwhile to notice that LARM-0.8 performs better in the daily data set. Although this is different from the observation shown in Figure 5, it is not out of expectation since lifetime prediction error is introduced and considered in this performance evaluation experiment. Besides, we have shown in Section 4 that 80%-lifespan metric is more predictable than the half-lifespan metric and thereby the results further confirm the effectiveness of LARM.

Furthermore, we examine how popularity prediction performance varies with different popularity evolution patterns. Specifically, we use the six popularity patterns introduced in Section 3 for each data set, and show their 50% and 80% percentile of APE when LARM-0.5, LARM-0.8, and VCDM are used as prediction models. We only show the results of VCDM since it outperforms the other two baselines. As shown in Figure 6, we can see that LARM significantly outperforms VCDM for some specific patterns, e.g. D2 and D4 in the daily data set, and H4 and H5 in the hourly data set. This shows the ability of LARM to capture the future popularity evolution of videos. Furthermore, LARM can steadily decrease APE for all patterns as t_r increases, while VCDM shows much more fluctuations. Sometimes more historical information does not help reduce prediction error but instead introduces more errors in VCDM. Therefore, we conclude that LARM is more adaptable to different types of data sets.

6.4 Early Prediction Advantage

Now we compare the early prediction advantage (EPA) of the prediction models. It is defined as the first time point t^* at which the MAPE for $N_v(t_r, t_t)$ $(t_r > t^*)$ is less than a threshold τ. The EPA performance is essential if we apply popularity prediction to make decisions such as targeted advertising. Specifically, we set the target date t_t as 50 to 100 for the daily data set, and 60 to 120

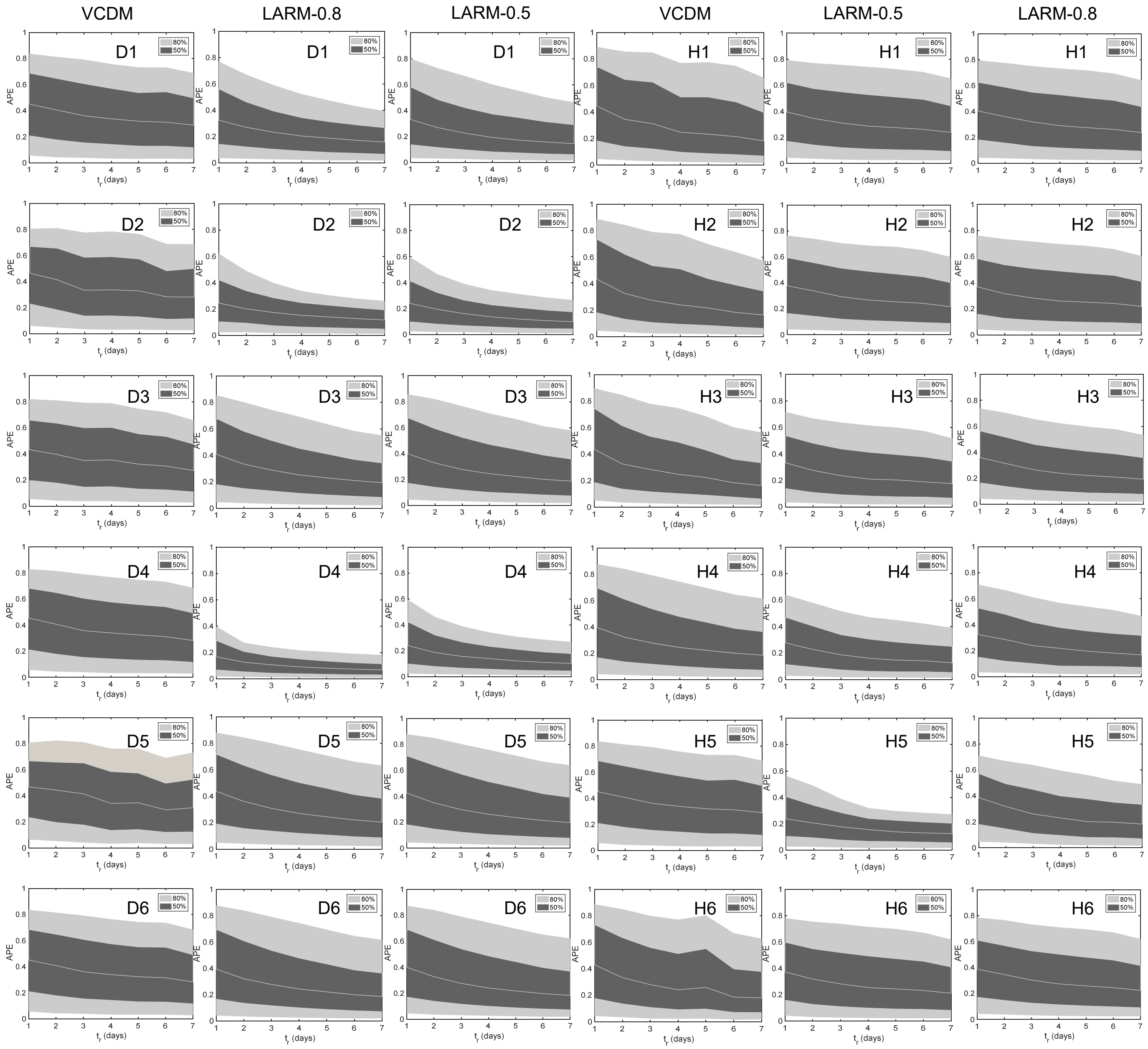


Figure 6: The 50th and 80th percentiles of the distribution of APE for different popularity evolution patterns.

for the hourly data set, with a step of 10. τ is set as 0.3 for both of the data sets.

Table 7 and Table 8 summarize the comparison results. We can see that LARM has better EPA performance in all different settings. It can achieve up to 3 steps and 7 steps ahead of baseline approaches in the daily data set and hourly data set, respectively. This advantage is attributed to LARM's awareness of video lifetime in a very early stage, which facilitates the popularity prediction.

6.5 Discussion

The performance evaluation have validated the effectiveness of lifetime based model specialization in long-term popularity prediction. However, the scheme still has two inherent performance barriers, which are resulted from the accuracy of lifetime prediction and the utilization of lifetime information. As shown in Figure 7, there is a significant gap between LARM and the scheme using the same model specialization mechanism but with the ground truth of lifetime, as well as between that mechanism and the ideal case with

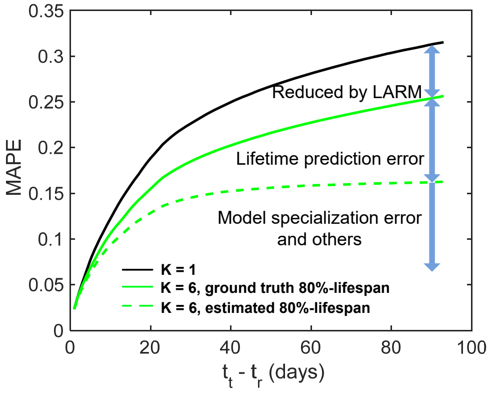


Figure 7: Performance bottleneck of LARM

zero prediction error. The result throws two interesting questions to us. First, how accurately can we predict the lifetime? Second, how effectively can we utilize the lifetime information? We believe it is imperative to investigate more advanced features such as the video topics and user interactions to improve the estimation of video lifetime. Additionally, lifetime estimation can be used as an additional layer on top of existing prediction models. In this paper, we have applied it on top of a simple linear regression model. Full-scale study on overlaying lifetime estimation as additional layer to other existing prediction models is needed to narrow the performance gap.

7 CONCLUSION

In this paper, we have presented a novel approach for fast prediction of the long-term popularity on YouTube videos by leveraging the knowledge of video lifetime in early stage. We explore the predictability of video life via early-accessible features and how it impacts the correlation between lifetime and long-term popularity. Inspired by our large-scale study, we propose a α-lifespan as the lifetime metric and a set of specialized regression models for the lifetime-aware popularity prediction. Through extensive comparison with non-trivial existing schemes in various practical conditions, we demonstrate that LARM can achieve up to 20% and 18% reduction in prediction error for our hourly and daily data sets, respectively. LARM represents a promising direction that combines minimal historical data and feature engineering to accomplish fast and satisfactory popularity prediction without any underlying model assumption. The success of LARM shall call for more effective features on top of existing prediction models to further improve the UGC popularity prediction.

ACKNOWLEDGMENTS

This work is supported by NSF Grant ECCS-1405594.

REFERENCES

[1] J. Yang and J. Leskovec, "Patterns of Temporal Varaiation in Online Media," *ACM WSDM*, pp.177-186, 2011.

[2] R. Zhou, S. Khemmarat, L. Gao, J. Wan, and J. Zhang, "How YouTube videos are discovered and its impact on video views," *Multimedia Tools and Applications*, vol. 75, no. 10, pp.6035-6058, 2016.

[3] K. Subbian, B. A. Prakash, and L. Adamic, "Detecting Large Reshare Cascades in Social Networks," *ACM WWW*, pp.597-606, 2017.

[4] H. Pinto, J. M. Almeida, and M. A. Goncalves, "Using early view patterns to predict the popularity of youtube videos," *ACM WSDM*, pp.365-374, 2013.

[5] G. Gursun, M. Crovella and I. Matta, "Describing and Forecasting Video Access Patterns," *IEEE INFOCOM*, pp.16-20, 2011.

[6] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini, "A peek into the future: predicting the evolution of popularity in user generated content," *ACM WSDM*, pp.607-616, 2013.

[7] Z. Tan, Y. Wang, Y. Zhang, and J. Zhou, "A Novel Time Series Approach for Predicting the Long-Term Popularity of Online Videos," *IEEE Trans. Broadcasting*, vol. 62, no. 2, pp.436-445, 2016.

[8] A. F. Costa, A. J. M. Traina, C. Traina and C. Faloutsos,, "Vote-and-Comment: Modeling the Coevolution of User Interactions in Social Voting Web Sites," *IEEE ICDM*, pp.91-100, 2016.

[9] T. Martin, J. M. Hofman, A. Sharma, A. Anderson, and D. J. Watts, "Exploring Limits to Prediction in Complex Social Systems," *ACM WWW*, pp.683-694, 2016.

[10] J. Cheng, L. A. Adamic, P. A. Dow, J. Kleinberg, and J. Leskovec, "Can Cascades be Predicted? " *ACM WWW*, pp.925-935, 2014.

[11] M. Cha, H. Kwak, P. Rodriguez, Y. Y. Ahn, S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," *ACM IMC*, pp.1-14, 2007.

[12] 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," *ACM KDD*, pp.1186-1194, 2012.

[13] C. Ma, Z. Yan, and C. W. Chen, "Forecasting Initial Popularity of Just-Uploaded User-Generated Videos," *IEEE ICIP*, pp.474-478, 2016.

[14] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, "SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity," *ACM KDD*, pp. 1513-1522, 2015.

[15] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li and C. Faloutsos, "Rise and fall patterns of information diffusion: model and implications," *ACM KDD*, pp.6-14, 2012.

[16] S. Mishra, M.A. Rizoiu, and L. Xie, "Feature Driven and Point Process Approaches for Popularity Prediction,"*ACM CIKM*, pp.1069-1078, 2016.

[17] Y. Rong, Q. Zhu, and H. Cheng, "A Model-Free Approach to Infer the Diffusion Network from Event Cascade," *ACM CIKM*, pp. 1653-1662, 2016.

[18] G. Szabo, and B. A. Huberman, "Predicting the popularity of online content," *Commun. ACM*, vol. 53, no. 8, pp.80-88, 2010.

[19] D. Vallet, S. Berkovsky, S. Ardon, A. Mahanti, and M. A. Kafaar, "Characterizing and Predicting Viral-and-Popular Video Content," *ACM CIKM*, pp.1591-1600, 2015.

[20] J. Wu, Y. Zhou, D. M. Chiu, Y. Hua, and Z. Zhu, "Modeling Dynamics of Online Video Popularity," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1882-1895, Sept. 2016.

[21] M. Rizoiu, L. Xie, S. Sanner, M. Cebrian, H. Yu, and P. Van Hentenryck, "Can this video be promoted? Endogenous and exogenous popularity processes in social media," *arXiv 1602.06033*, 2016.

[22] H. Yu, L. Xie, and S. Sanner, "The lifecyle of a youtube video: Phases, content and popularity," *AAAI*, 2015.

[23] Z. Wang, L. Sun, X. Chen, W. Zhu, J. Liu, M. Chen, and S. Yang, "Propagation-Based Social-Aware Replication for Social Video Contents," *ACM MM*, 2012.

[24] S. D. Roy, T. Mei, W. Zeng and S. Li, "Towards Cross-Domain Learning for Social Video Popularity Prediction," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1255-1267, 2013.

[25] S. Wang, Z. Yan, X. Hu, P. S. Yu, and Z. Li, "Burst Time Prediction in Cascades," *AAAI*, pp. 325-331, 2015.

[26] F. Figueiredo, "On the Prediction of Popularity of Trends and Hits for User Generated Videos," *ACM WSDM*, pp. 741-746, 2013.

[27] H. W. Shen, D. Wang, C. Song, A. L. Barabási, "Modeling and Predicting Popularity Dynamics via Reinforced Poisson Processes," *arXiv 1401.0778*, 2014.

[28] X. He, M. Gao, M. Y. Kan, Y. Liu, and K. Sugiyama, "Predicting the popularity of Web 2.0 Items Based on User Comments?," *SIGIR*, 2014.

[29] F. Figueiredo, J. M. Almeida, M. A. Goncalves, and F. Benevenuto, "On the dynamics of social media popularity: A YouTube case study," ACM Trans. Internet Technol., vol. 14, no. 4, 2014, Art. no. 24.