

Beyond Views: Measuring and Predicting Engagement in Online Videos

Siqi Wu, Marian-Andrei Rizoio, Lexing Xie

Australian National University and Data 61, CSIRO, Australia
 {siqi.wu, marian-andrei.rizoio, lexing.xie}@anu.edu.au

Abstract

The share of videos in the internet traffic has been growing, therefore understanding how videos capture attention on a global scale is also of growing importance. Most current research focus on modeling the number of views, but we argue that video engagement, or time spent watching is a more appropriate measure for resource allocation problems in attention, networking, and promotion activities. In this paper, we present a first large-scale measurement of video-level aggregate engagement from publicly available data streams, on a collection of 5.3 million YouTube videos published over two months in 2016. We study a set of metrics including time and the average percentage of a video watched. We define a new metric, *relative engagement*, that is calibrated against video properties and strongly correlate with recognized notions of quality. Moreover, we find that engagement measures of a video are stable over time, thus separating the concerns for modeling engagement and those for popularity – the latter is known to be unstable over time and driven by external promotions. We also find engagement metrics predictable from a *cold-start* setup, having most of its variance explained by video context, topics and channel information – $R^2=0.77$. Our observations imply several prospective uses of engagement metrics – choosing engaging topics for video production, or promoting engaging videos in recommender systems.

1 Introduction

Attention is a scarce resource in the modern world. There are many metrics for measuring attention received by online content, such as page views for webpages, listen counts for songs, view counts for videos, and the number of impressions for advertisements. Although these metrics describe the human behavior of *choosing* one particular item, they do not describe *how users engage with this item* (Van Hentenryck et al. 2016). For instance, an audience may become immersed in the interaction or quickly abandon it – the distinction of which will be clear if we know how much time the user spent interacting with this given item. Hence, we consider *popularity and engagement* as different measures of online behavior.

In this work, we study online videos using publicly available data from the largest video hosting site YouTube. On YouTube, popularity is characterized as the willingness to

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

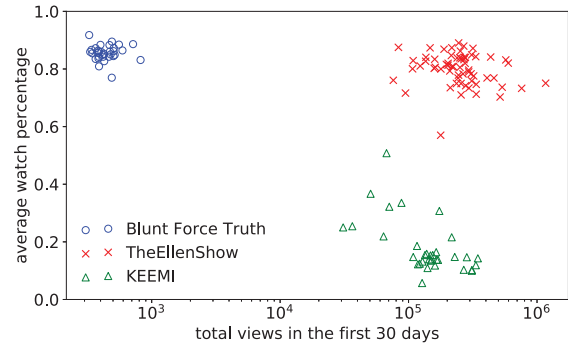


Figure 1: Scatter plot of videos from three YouTube channels: Blunt Force Truth (political entertainment, blue circle), KEEMI (cooking vlog, green triangle), and TheEllenShow (comedy, red cross). x-axis: total views in the first 30 days; y-axis: average watch percentage.

click a video, whereas engagement is the watch pattern after clicking. While most research have focused on measuring popularity (Pinto, Almeida, and Gonçalves 2013; Rizoio et al. 2017), engagement of online videos is not well understood, leading to key questions such as: How to measure video engagement? Does engagement relate to popularity? Can engagement be predicted? Once understood, engagement metrics will become relevant targets for recommender systems to rank the most valuable videos.

In Fig. 1, we plot the number of views against the average percentage watched for 128 videos in 3 channels. While the entertainment channel Blunt Force Truth has the least views on average, the audience tend to watch more than 80% of each video. On the contrary, videos from the cooking vlogger KEEMI have on average 159,508 views, but they are watched only 18%. This example illustrates that videos with a high number of views do not necessarily have high watch percentages, and prompts us to investigate other metrics for describing engagement.

Recent progress in understanding video popularity and the availability of new datasets allow us to address three open questions about video engagement. Firstly, *on an aggregate level, how to measure engagement?* Most engagement literatures focus on the perspective of an individual user, such

as recommending relevant products (Covington, Adams, and Sargin 2016), tracking mouse gestures (Arapakis, Lalmas, and Valkanas 2014) or optimizing search results (Drutsa, Gusev, and Serdyukov 2015). Since user-level data is often unavailable, defining and measuring average engagement is useful for content producers on YouTube. Secondly, **within the scope of online video, can engagement help measure content quality?** As shown in Fig. 1, video popularity metric is inadequate to estimate quality. One early attempt to measure online content quality was taken by Salganik, Dodds, and Watts (2006), who studied music listening behavior in an experimental environment. For a large number of online contents, measuring quality from empirical data still remains unexplored. Lastly, **in a cold-start setup, can engagement be predicted?** Online attention is known to be difficult to predict without early feedback (Martin et al. 2016). For engagement, Park, Naaman, and Berger (2016) showed the predictive power of user reactions such as views and comments. However, these features also require monitoring the system for a period of time. In contrast, **if engagement can be predicted before content is uploaded**, it will provide actionable insights to content producers.

We address the first question by constructing 4 new datasets that contain more than 5 million YouTube videos. We build two 2-dimensional maps that visualize the internal bias of existing engagement metrics – *average watch time* and *average watch percentage* – against video length. Building upon that, we derive a novel **relative engagement metric**, as the duration-calibrated rank of average watch percentage.

Answering the second question, we demonstrate that relative engagement is stable over time, and strongly correlates with established quality measures in Music and News categories, such as Billboard songs, Vevo artists, and top news channels. This newly proposed relative engagement metric can be a target for recommender systems to prioritize quality videos, and for content producers to create engaging videos.

Addressing the third question, we predict engagement metrics in a *cold-start* setting, **using only video content and channel features**. With off-the-shelf machine learning algorithms, we achieve $R^2=0.77$ for predicting average watch percentage. We consider this as a significant result that shows the predictability of engagement metrics. Furthermore, we explore the predictive power of video topics and find some topics are strong indicators for engagement.

The main contributions of this work include:

- Conduct a large-scale measurement study of engagement on 5.3 million videos over two-month period, and publicly release 4 new datasets and the engagement benchmarks¹.
- Measure a set of engagement metrics for online videos, including average watch time, average watch percentage, and a novel metric – **relative engagement, which is calibrated with respect to video length, stable over time, and correlated with video quality**.
- Predict relative engagement and watch percentage from video context, topics, and channel reputation in a *cold-*

¹The code and datasets are publicly available at <https://github.com/avalanchesiqi/youtube-engagement>

| Dataset | #Videos | #Channels |
|------------------|-----------|-----------|
| TWEETED VIDEOS | 5,331,204 | 1,257,412 |
| VEVO VIDEOS | 67,649 | 8,685 |
| BILLBOARD VIDEOS | 63 | 47 |
| TOP NEWS VIDEOS | 28,685 | 91 |

Table 1: Overview of 4 new video datasets.

| Category | #Videos | Category | #Videos |
|---------------|-----------|----------|---------|
| People | 1,265,805 | Comedy | 138,068 |
| Gaming | 1,079,434 | Science | 110,635 |
| Entertainment | 775,941 | Auto | 84,796 |
| News | 459,728 | Travel | 65,155 |
| Music | 449,314 | Activism | 58,787 |
| Sports | 243,650 | Pets | 27,505 |
| Film | 194,891 | Show | 1,457 |
| Howto | 192,931 | Movie | 158 |
| Education | 182,849 | Trailer | 100 |

Table 2: Breakdown of TWEETED VIDEOS by category.

start setting (i.e., before the video gathers any view or comment), achieving $R^2=0.45$ and 0.77 respectively.

2 Datasets

We curate 4 new publicly available video datasets, as summarized in Table 1 and Table 2. We also describe three daily series available for all videos: shares, views and watch time.

2.1 Video datasets

TWEETED VIDEOS contains 5,331,204 videos published between July 1st and August 31st, 2016 from 1,257,412 channels. The notion of *channel* on YouTube is analogous to that of *user* on other social platforms, since every video is published by a channel and belonging to one user account. Using Twitter mentions to sample a collection of YouTube videos has been used in previous works (Abisheva et al. 2014; Yu, Xie, and Sanner 2014). We use the Twitter Streaming API to collect tweets, by tracking the expression "YOUTUBE" OR ("YOUTU" AND "BE"). This covers textual mentions of YouTube, YouTube links and YouTube's URL shortener (youtu.be). This yields 244 million tweets over the two-month period. In each tweet, we search the `extended_urls` field and extract the associated YouTube video ID. This results in 36 million unique video IDs and over 206 million tweets. For each video, we extract its meta-data and three attention-related dynamics, as described in Sec. 2.2. A non-trivial fraction (45.82%) of all videos have either been deleted or their statistics are not publicly available. This leaves a total of 19.5 million usable videos.

We further filter videos based on recency and their level of attention. We remove videos that are published prior to this two-month period to **avoid older videos**, since being tweeted a while after being uploaded may indicate higher engagement. We also filter out videos that receive less than 100 views within their first 30 days after upload, which is

the same filter used by Brodersen, Scellato, and Wattenhofer (2012). Videos that do not appear on Twitter, or have extremely low number of early views are unlikely to accumulate a large amount of attention (Rizoiu and Xie 2017; Pinto, Almeida, and Gonçalves 2013), therefore, they do not provide enough data to reflect collective watch patterns. Our proposed measures can still be computed on these removed videos, however the results might have limited relevance given the low level of user interaction with them. Table 2 shows a detailed category breakdown of TWEETED VIDEOS. **QUALITY VIDEOS.** We collect three datasets containing videos deemed of high quality by domain experts, two of which are on Music and one is on News. These datasets are used to **link engagement and video quality** (Sec 3.3).

- **VEVO VIDEOS.** Vevo is a multinational video hosting service which syndicates licensed music clips from three major record companies on YouTube (Wikipedia 2018b). VEVO artists usually come from professional music background, and their videos are professionally produced. We consider VEVO VIDEOS to be of higher quality than the average Music videos in the TWEETED VIDEOS dataset. We collect all the YouTube channels that contain the keyword “Vevo” in the title and a “verified” status badge on the profile webpage. In total, this dataset contains 8,685 Vevo channels with 67,649 music clips, as of August 31st, 2016.

- **BILLBOARD VIDEOS.** Billboard acts as a **canonical ranking source** in the music industry, aggregating music sales, radio airtime and other popularity metrics into a yearly Hot 100 music chart. The songs that appear in this chart are usually perceived as having vast success and being of high quality. We collect 63 videos from 47 artists based on the 2016 Billboard Hot 100 chart (Wikipedia 2018a).

- **TOP NEWS VIDEOS** features a list of top 100 most viewed News channels, as reported by an external ranking source (vidstatx 2017). This list includes traditional news broadcasting companies (e.g., CNN), as well as popular on-line talk shows (e.g., The Young Turks). For each channel, we retrieve its last 500 videos published before Aug 31st, 2016. This dataset contains 91 publicly available News channels and 28,685 videos.

2.2 YouTube metadata and attention dynamics

For each video, we use the YouTube Data API to retrieve video metadata information – video id, title, description, upload time, category, duration, definition, channel id, channel title and associated Freebase topic ids, which we resolve to entity names using the latest Freebase data dump² (Figueiredo et al. 2014).

We then develop a software package³ to extract three daily series of video attention dynamics: **daily volume of shares, view counts and watch time**. Throughout this paper, we denote the number of **shares and views** that a video receives on the t^{th} day after upload as $s[t]$ and $x_v[t]$, respectively. Similarly, $x_w[t]$ is the total amount of **time of video being watched** on the t^{th} day. Each attention series is observed for at least 30 days, i.e., $t = 1, 2, \dots, 30$. Most prior research

²<https://developers.google.com/freebase>

³<https://github.com/computationalmedia/youtube-insight>

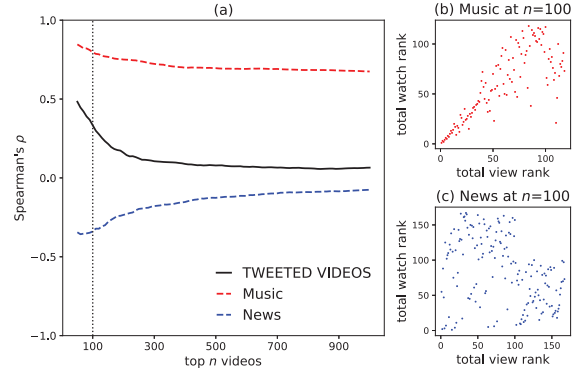


Figure 2: (a) Disagreement between the union set of top n most viewed and top n most watched videos in TWEETED VIDEOS at the age of 30 days, measured with **Spearman's ρ** . (b-c) Scatter plots of video ranking in view and in watch at $n=100$ in **Music** ($\rho=0.80$) and **News** ($\rho=-0.34$).

on modeling video popularity dynamics (Szabo and Huberman 2010; Figueiredo et al. 2016) study only view counts. To the best of our knowledge, our work is the first to perform large-scale measurements on video watch time.

3 Measures of video engagement

In this section, we measure the interplay between view count, watch time, watch percentage and video duration. We first examine their relation in a new visual presentation – **engagement map**, then we propose **relative engagement**, a novel metric to estimate video engagement (Sec. 3.2). We show that relative engagement calibrates watch patterns for videos of different lengths, demonstrates correlation to external notions of video quality (Sec. 3.3), and remains stable over time (Sec. 3.4).

3.1 Discrepancy between views and watch time

Fig. 1 illustrates that watch patterns (e.g., average percentage of video watched) can be very different for videos with similar views. We examine the union set of top n videos in TWEETED VIDEOS dataset, respectively ranked by total views and total watch time at the age of 30 days. For n varying from 100 to 1000, **we measure their agreement using Spearman's ρ** . With value between -1 and +1, a positive ρ implies that as the rank in one variable increases, so does the rank in the other variable. A ρ of 0 indicates no correlation exists in these two ranked variables. Fig. 2a shows that in TWEETED VIDEOS, video ranks in total view count and total watch time correlate at the level of 0.48 when n is 50, but this correlation declines to 0.08 when n increases to 500 (solid black line). Furthermore, the level of agreement varies across different video categories: for Music, a video that ranks high in total view count often ranks high in total watch time ($\rho=0.80$ at $n=100$, Fig. 2b); for News, the two metrics have a weak negative correlation ($\rho = -0.34$ at $n = 100$, Fig. 2c).

This observation indicates that **total view count and total watch time provide different aspects** of how audience in-

teract with YouTube videos. One recommender system optimizing for view count may generate remarkably different results with one that drives watch time (Yi et al. 2014). In the next section, we analyze their interplay to construct more diverse set of measures for video engagement.

3.2 Engagement map and relative engagement

Recent studies show that the quality of a digital item is linked to the audience’s decision to continue watching or listening after first opening it (Salganik, Dodds, and Watts 2006; Krumme et al. 2012). Therefore, the average amount of time that the audience spend on watching a video should be indicative of video quality. For a given video, we compute two aggregate metrics:

- *average watch time* $\bar{\omega}_t$: the total watch time $x_w[1 : t]$ divided by the total view count $x_v[1 : t]$ up to day t

$$\bar{\omega}_t = \frac{\sum_{i=1}^t x_w[i]}{\sum_{i=1}^t x_v[i]} \quad (1)$$

- *average watch percentage* $\bar{\mu}_t$: the average watch time $\bar{\omega}_t$ normalized by video duration D

$$\bar{\mu}_t = \frac{\bar{\omega}_t}{D} \quad (2)$$

$\bar{\omega}_t$ is a positive number bounded by the video length, whereas $\bar{\mu}_t$ takes values between 0 and 1 and represents the average percentage of video watched.

Engagement map. We observe that video duration is an important covariate on watch percentage. In the TWEETED VIDEOS dataset, duration alone explains more than 58% of the variance of watch percentage. Intuitively, longer videos are less likely to be fully watched compared to shorter videos due to the limited human attention span.

We construct two 2-dimensional maps, where the x-axis shows video duration D , and the y-axis shows average watch time $\bar{\omega}_{30}$ (Fig. 3a) and average watch percentage $\bar{\mu}_{30}$ (Fig. 3b) over the first 30 days. We project all videos in the TWEETED VIDEOS dataset onto both maps. The x-axis is split into 1,000 equally wide bins in log scale. We choose 1,000 bins to trade-off enough data in each bin and having enough bins. We have also tried discretizing to smaller or larger number of bins, and the results are visually similar. We merge bins containing a very low number of videos (<50) to nearby bins. Overall, **each bin contains between 50 and 38,508 videos**. The color shades correspond to data percentiles inside each bin: the darkest color corresponds to the median value and the lightest correspond to the extremes (0% and 100%). Both maps calibrate watch time and watch percentage against video durations: highly-watched videos are positioned towards the top of allocated bin, while barely-watched videos are at the bottom compared to other videos with similar length.

Those two maps are logically identical because the position of each video in Fig. 3b can be obtained by normalizing with its duration in Fig. 3a. It is worth noticing that a linear trend exists between average watch time and video duration in the log-log space, with an increasing variance as

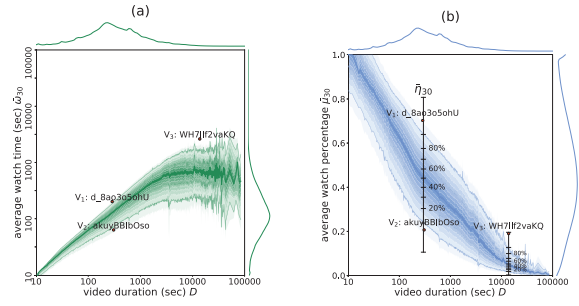


Figure 3: Video engagement in the TWEETED VIDEOS dataset at the age of 30 days. (a) video duration D vs average watch time $\bar{\omega}_{30}$; (b) the engagement map: video duration D vs average watch percentage $\bar{\eta}_{30}$.

duration grows. In this work, we predominantly use the map of watch percentage (Fig. 3b) given its y-axis is bounded between $[0,1]$, making it easier to interpret. We denote this map as the **engagement map**.

Note that our method of constructing the engagement map resembles the idea of non-parametric quantile regression, which essentially computes a quantile regression fit in an equally spaced span (Koenker 2005). For smaller datasets, using quantile regression may result in a smoother mapping. We tried quantile regression on TWEETED VIDEOS dataset, and we found that the values on both tails are inaccurate as the polynomial fits do not accurately reflect nonlinear trends. Our binning method works better in this case. Finally, we remarks that the engagement map can be constructed at different ages, which allows us to study the temporal evolution of engagement (Sec. 3.4).

Relative engagement $\bar{\eta}_t$. Based on the engagement map, we propose the relative engagement $\bar{\eta}_t \in [0,1]$, defined **as the rank percentile of video in its duration bin**. This is an average engagement measure in the first t days. Fig. 3b illustrates the relation between video duration D , watch percentage $\bar{\mu}_{30}$ and relative engagement $\bar{\eta}_{30}$ for three example videos. Video v_1 (d_8a03o5ohU) shows kids doing karate and v_2 (akuyBBibOso) is about teaching toddlers colors. They are both about 5 minutes, but have different watch percentages, $\bar{\mu}_{30}(v_1) = 0.70$ and $\bar{\mu}_{30}(v_2) = 0.21$. These amount to very different values of the relative engagement: $\bar{\eta}_{30}(v_1) = 0.96$, while $\bar{\eta}_{30}(v_2) = 0.07$. Video v_3 (WH7l1f2vaKQ) is a much longer video ($D=3$ hours 49 minutes) showing a live fighting show. It has a relatively low watch percentage ($\bar{\mu}_{30}(v_3) = 0.19$), similar to v_2 . However, its relative engagement $\bar{\eta}_{30}(v_3)$ amounts to 0.99, positioning it among the most engaging videos in its peer group.

We denote the mapping from watch percentage $\bar{\mu}_t$ to relative engagement $\bar{\eta}_t$ as f , and its inverse mapping as f^{-1} . Here f is implemented as a length-1,000 look up table with a maximum resolution of 0.1% (or 1,000 ranking bins). For a given video with duration D , we first map it to corresponding bin on the engagement map, then return the engagement percentile by watch percentage. Eq. 3 describes the mapping between relative engagement and average watch percentage using engagement map.

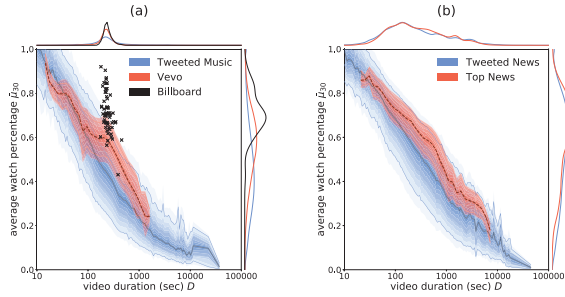


Figure 4: Relative engagement and video quality for Music (a) and News (b). Videos in QUALITY VIDEOS dataset are shifted towards **higher relative engagement** compared to that in TWEETED VIDEOS. Best viewed in colors.

$$\bar{\eta}_t = f(\bar{\mu}_t, D) \Leftrightarrow \bar{\mu}_t = f^{-1}(\bar{\eta}_t, D) \quad (3)$$

While researchers have observed that watch percentage is affected by video duration (Guo, Kim, and Rubin 2014; Park, Naaman, and Berger 2016), to the best of our knowledge, this work is the first to quantitatively map its non-linear relation with video duration and present measurements in a large-scale dataset.

3.3 Relative engagement and video quality

We examine the relation between relative engagement and video quality. We place the QUALITY VIDEOS datasets (Sec. 2.1) on the engagement map. Fig. 4a plots the engagement map of all Music videos in the TWEETED VIDEOS (blue), that of the VEVO VIDEOS (red), and the videos in the BILLBOARD VIDEOS as a scatter plot (black dots). Similarly, Fig. 4b plots the engagement map of all News videos in the TWEETED VIDEOS in blue and that of the TOP NEWS VIDEOS in red. All the maps are built from observations in the first 30 days.

Visibly, the QUALITY VIDEOS are skewed towards higher relative engagement values in both figures. Most notably, 44 videos in the BILLBOARD VIDEOS dataset (70% of the dataset) possess a high relative engagement of over 0.9. The other 30% of videos have an average $\bar{\eta}_{30}$ of 0.83 with a minimum of 0.54. For QUALITY VIDEOS, the 1-dimensional density distribution of average watch percentage $\bar{\mu}_{30}$ also shifts to the upper end as shown on the right margin of Fig. 4. Overall, relative engagement values are high for content judged to be high quality by experts and the community. **Thus, relative engagement is one plausible surrogate metric for content quality.**

Relative engagement within channel. Fig. 5 shows the engagement mapping results of 25 videos within one channel (PBABowling). This channel uploads sports videos about *Professional Bowlers Association* with widely varying lengths – from 2-minute player highlights to 1-hour event broadcasts. Video length has a significant impact: the short video cluster has mean average watch percentage $\bar{\mu}_{30}$ of 0.82, whereas the long video cluster has mean $\bar{\mu}_{30}$ of 0.21. However, after mapping to relative engagement, those two clusters have mean $\bar{\eta}_{30}$ of 0.92 and 0.78 – much more

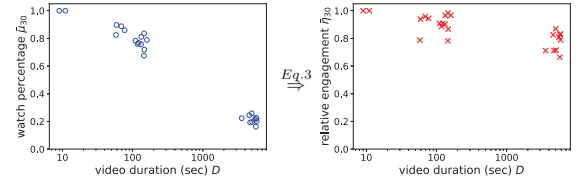


Figure 5: Watch percentage $\bar{\mu}_{30}$ (left) and relative engagement $\bar{\eta}_{30}$ (right) for videos in channel PBABowling. While it appears that $\bar{\mu}_{30}$ has a linear relation with the logarithmic duration $\log_{10} D$, $\bar{\eta}_{30}$ can be reasonably explained by only using the mean value of $\bar{\eta}_{30}$.

consistent for this channel than measured by watch percentage. Overall, the mean relative engagement of channel PBABowling is 0.86, which suggests this channel is likely to produce more engaging videos than an average YouTube channel, regardless of the video length. This example illustrates **video relative engagement tends to be stable within the same channel**, and sheds some light on using past videos to predict future relative engagement.

3.4 Temporal dynamics of relative engagement

How does engagement change over time? This question is important because popularity dynamics tend to be bursty and hard to predict (Cheng et al. 2014). If engagement dynamics can be shown to be stable, it is useful for content producers to understand watch patterns from early observation. Note that the method for constructing the engagement map is the same, but one can use data at different ages t to build different mapping function $f(\bar{\mu}_t, D)$.

Relative engagement is stable over time. We examine the temporal change of relative engagement at two given days t_1 and t_2 ($t_1 < t_2$) in TWEETED VIDEOS. We denote the cumulative distribution function (CDF) as $F_x(\Delta\bar{\eta})$, where $x = \bar{\eta}_{t_2} - \bar{\eta}_{t_1}$. This computes the fraction of videos with relative engagement changing *less* than $\Delta\bar{\eta}$ during t_1 to t_2 . Fig. 6a shows $\Delta\bar{\eta}$ distribution of day 7 vs day 14 and day 7 vs day 30. There are 4.6% of videos that increase more than 0.1 and 2.7% that decrease more than 0.1, **yielding 92.7% of the videos with an absolute relative engagement change of less than 0.1** between day 7 and day 30. Such a small change results from the fact that relative engagement $\bar{\eta}_t$ is defined as average measure over the past t days. It suggests that future relative engagement can be predicted from early watch patterns within a small margin of error. Similarly, this observation extends to both average watch percentage $\bar{\mu}_t$ and average watch time $\bar{\omega}_t$.

Next we examine relative engagement on a daily basis. To avoid days with zero views, we use a 7-day sliding window, i.e., changing the summations in Eq. 1 to between $t-6$ and t , yielding a *smoothed* daily watch percentage $\bar{\mu}_{t-6:t} = \frac{\sum_{i=t-6}^t x_w[i]}{D \sum_{i=t-6}^t x_v[i]}$. We then convert $\bar{\mu}_{t-6:t}$ to smoothed daily relative engagement $\bar{\eta}_{t-6:t}$ via the corresponding engagement map. For $t < 7$, we calculate relative engagement from all prior days before t .

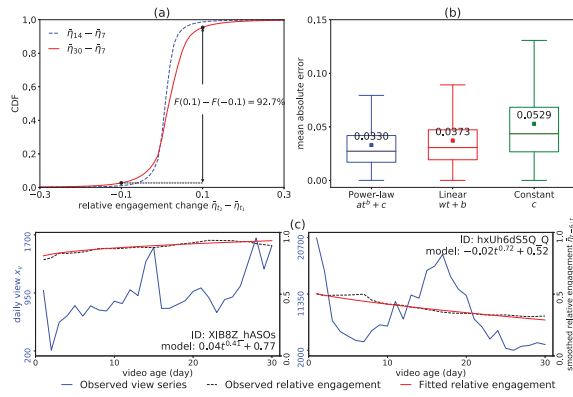


Figure 6: **Relative engagement is stable** over time. (a) CDF of temporal change in relative engagement of day 7 vs day 14 (blue dashed), day 7 vs day 30 (red solid). (b) Fitting error of power-law model (blue), linear regressor (red) and constant function (green) in TWEETED VIDEOS. (c) Temporal view series (blue) and smoothed daily relative engagement (black dashed) fitted by generalized power-law model $at^b + c$ (red).

Fig. 6c shows the daily views and smoothed relative engagement over the first 30 days of two example videos. While the view series has multiple spikes (blue), **relative engagement is stable** with only a slightly positive trend for video XIB8Z_hASOs and a slightly negative trend for hxUh6dS5Q_Q (black dashed). View dynamics have been shown to be affected by external sharing behavior (Rizoiu and Xie 2017), the stability of relative engagement can be explained by the fact that it measures the average watch pattern but not how many people view the video.

Fitting relative engagement dynamics. We examine the stability of engagement metrics across the entire TWEETED VIDEOS dataset. If the engagement dynamics can be modeled by a parametric function, one can forecast future engagement from initial observations. To explore which function best describes the gradual change of relative engagement $\bar{\eta}_t$, we examine generalized power-law model ($at^b + c$) (Yu, Xie, and Sanner 2015), linear regressor ($wt + b$), and constant (c) function. For videos in TWEETED VIDEOS, we fit each of the three functions to smoothed daily relative engagement series $\bar{\eta}_{t-6:t}$ over the first 30 days. Fig. 6b shows that **power-law function fits best** on the dynamics of relative engagement, with an average mean absolute error of 0.033.

To sum up, we observe that relative engagement $\bar{\eta}_t$ is stable throughout lifetime, which implies that early watch pattern is a **strong predictor** for future engagement. Therefore, in the next section, we set up a prediction task to examine whether engagement can be predicted *before* upload.

4 Predicting engagement

In this section, we predict relative engagement and watch percentage of a video *before* it is uploaded. We further analyze the relation between **video features and engagement metrics**.

4.1 Prediction tasks setup

We observe that relative engagement and watch percentage are stable over time (Sec. 3.4), which makes them attractive prediction targets. Furthermore, it is desirable to predict them before videos get uploaded, and viewing or commenting behavior is observed.

Prediction targets. We setup two regression tasks to predict average **watch percentage** $\bar{\mu}_{30}$ and **relative engagement** $\bar{\eta}_{30}$. Watch percentage is intuitively useful for content producers, while relative engagement is designed to calibrate watch percentage against duration as detailed in Sec. 3.2. It is interesting to see whether such calibration changes prediction performance. We report three evaluation results: predicting relative engagement and watch percentage directly, and predicting relative engagement then mapping to watch percentage via engagement map by using Eq. 3. We do not predict average watch time because it can be deterministically computed by multiplying watch percentage and duration.

Training and test data. We split TWEETED VIDEOS at 5:1 ratio over publish time. We use the first 51 days (2016-07-01 to 2016-08-20) for training, containing 4,455,339 videos from 1,132,933 channels; and the last 11 days for testing (2016-08-21 to 2016-08-31), containing 875,865 videos from 366,311 channels. 242,017, or 66% of channels in the test set have appeared in training set, however, none of the videos in the test set is in the training set. The engagement map between watch percentage and relative engagement is built on the training set over the first 30 days. We split the dataset in time to ensure that learning is on past videos and prediction is on future videos.

Evaluation metrics. Performance is measured with two metrics:

- Mean Absolute Error $MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$
- Coefficient of Determination $R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$

Here y is the true value, \hat{y} the predicted value, \bar{y} the average; i indexes samples in the test set. MAE is a standard metric for average error. R^2 quantifies the proportion of the variance in the dependent variable that is predictable from the independent variable (Allen 1997), and **is often used to compare different prediction problems** (Martin et al. 2016). A lower MAE is better whereas a higher R^2 is better.

4.2 Features

We describe each YouTube video with 4 types of features as summarized in Table 3.

Control variable. Because video duration is the primary source of variation for engagement (Fig. 3), we use *duration* as a control variable and include it in all predictors. In TWEETED VIDEOS dataset, durations vary from 1 second to 24 hours, with a mean value of 12 minutes and median of 5 minutes. We take the logarithm (base 10) of duration to account for the skew.

Context features are provided by video uploader. They describe basic video properties and production quality (Hessel, Lee, and Mimno 2017).

| Control variable (D) | |
|---|---|
| <i>Duration</i> | Logarithm of duration in seconds |
| Context features (C) | |
| <i>Definition</i> | Binary, high definition or not |
| <i>Category</i> | One hot encoding of 18 categories |
| <i>Language</i> | One hot encoding of 55 languages |
| Freebase topic features (T) | |
| <i>Freebase topics</i> | One hot sparse representation of 405K topics |
| Channel reputation features (R) | |
| <i>Activity level</i> | Mean number of daily upload |
| <i>Past engagement</i> | Mean, std and five points summary of previously uploaded videos |
| Channel specific predictor (CSP) | |
| One predictor for each channel using available features | |

Table 3: Overview of features for predicting engagement.

- *Definition*: “1” represents high definition (720p or 1080p) and “0” represents low definition (480p, 360p, 240p or 144p). High definition yields better perceptual quality and encourages engagement (Dobrian et al. 2011).
- *Category*: broad content identifications assigned by video producers, the full list is shown in Table 1 (bottom). Here we encode it as an 18-dimensional one-hot vector.
- *Language*: we run `langdetect` package on the video description and choose the most likely language. `langdetect` implements a Naive Bayes classifier to detect 55 languages with high precision (Shuyo 2010). The language is indicative of audience demographics.

Freebase topics features. YouTube labels videos with Freebase entities (Bollacker et al. 2008). These labels incorporate user engagement signals, video metadata and content analysis (Vijayanarasimhan and Natsev 2018), and are built upon a large amount of data and computational resources. With the recent advances in computer vision and natural language processing, there may exist more accurate methods for annotating videos. However, **one can not easily build such an annotator at scale, and finding the best video annotation technique is beyond the scope of this work.** On average, each video in the TWEETED VIDEOS dataset has 6.16 topics. Overall, there are **405K topics** and 98K of them appear more than 6 times. These topics vary from broad categories (Song), to specific object (Game of Thrones), celebrities (Adele), real-world events (2012 Seattle International Film Festival) and many more. Such fine-grained topics are descriptive of video content. While learning embedding vectors can help predict engagement (Covington, Adams, and Sargin 2016), using raw Freebase topics enables us to interpret the effect of individual topic (Sec. 4.4).

Channel reputation features. Prior research shows that user features are predictive for product popularity (Martin et al. 2016; Mishra, Rizoiiu, and Xie 2016). Here we compute feature from a channel’s history to represent its reputation.

We could not use social status indicators such as the number of subscribers, because it is a time-varying quantity and the value when a video is uploaded can not be retrospectively obtained. Thus, we compute two proxies for describing channel features.

- *Activity level*: mean number of daily published videos by channels in the training data. Intuitively, channels with higher upload rates reflect better productivity.
- *Past engagement*: relative engagement of previously uploaded videos from the same channel in the training set. Here we compute mean, standard deviation and five points summary: median, 25th and 75th percentile, min and max.

Several features used in prior works are interesting, but they do not apply in our setting. Network traffic measurement (Dobrian et al. 2011) requires access to the hosting backend. Audience reactions such as likes and comments (Park, Naaman, and Berger 2016) can not be obtained before a video’s upload.

4.3 Prediction methods and results

Prediction methods. We use linear regression with L2-regularization to predict engagement metrics, $\bar{\eta}_{30}$ and $\bar{\mu}_{30}$, both lie between 0 and 1. Since the dimensionality of Freebase topics features is high (4M x 405K), we convert the **feature matrix to a sparse representation**, allowing the predictor to be trained on one workstation. We adopt a fall-back strategy to deal with missing features. For instance, we use the context predictor for videos for which the channel reputation features are unavailable. The fall-back setting usually results in a lower prediction performance, however it allows to predict engagement for *any* video. We also tried KNN regression and support vector regression, but they did not yield better performances.

Channel specific predictor (CSP). In addition to the shared predictor, we train a separate predictor for each channel that has at least 5 videos in the training set. This fine-grained predictor covers 61.4% videos in the test data and may capture the “on-topic” effect within channel (Martin et al. 2016). Intuitively, a channel might have specialty on certain topics and videos about those attract the audience to watch longer. For the remaining 38.6% videos, we use the shared linear regressor with all available features.

Prediction results. Fig. 7a summarizes the results of predicting the relative engagement $\bar{\eta}_{30}$. Context (C) and Freebase topics (T) alone are weak predictors, explaining 0.04 and 0.19 variance of $\bar{\eta}_{30}$ in the test set. Combining the two (C+T) yields a slight gain over Freebase topics. Channel reputation (R) is the strongest feature, achieving $R^2=0.42$, and is slightly improved by adding context and Freebase topics. Channel-specific predictor (CSP) performs similarly to the All-feature predictor (All), suggesting that one can use a shared predictor to achieve similar performance with finer-grained per-channel model for this task.

Average watch percentage $\bar{\mu}_{30}$ is easier to predict, achieving R^2 up to 0.69 (Fig. 7b) by using all features. Interestingly, predicting $\bar{\eta}_{30}$ then mapping to $\bar{\mu}_{30}$ consistently outperforms direct prediction of $\bar{\mu}_{30}$, achieving R^2 of 0.77. This

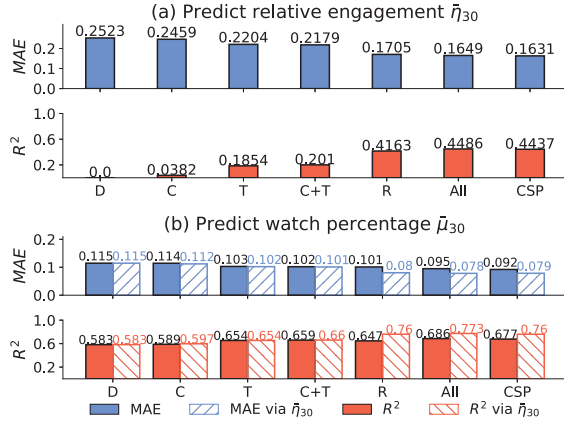


Figure 7: Summary of engagement prediction with two metrics, MAE : lower is better; R^2 : higher is better. (a): Performance for predicting $\bar{\eta}_{30}$ in different feature combinations. (b): Performance for predicting $\bar{\mu}_{30}$ in different feature combinations, directly (solid bars, left) or via relative engagement $\bar{\eta}_{30}$ (shaded bars, right). Predicting **watch percentage** via converting relative engagement performs better than predicting watch percentage directly in all predictors.

shows that removing the influence of video duration via engagement map is beneficial for predicting engagement.

To understand why predicting via $\bar{\eta}_{30}$ performs better, we examine the shared linear regressors in both tasks. For simplicity, we include video duration and channel reputation features as covariates, and exclude the (generally much weaker) context and Freebase topics features for this example. In Fig. 8, we visualize the two shared channel reputation predictors (R) at different video lengths for channel PBABowling (also shown in Fig. 5): one predicts $\bar{\mu}_{30}$ directly (blue dashed), and the other predicts $\bar{\eta}_{30}$, then maps to $\bar{\mu}_{30}$ via the engagement map (red solid). **The engagement map captures the non-linear effect for both short and long videos.** In contrast, predicting $\bar{\mu}_{30}$ directly does not capture the bimodal duration distribution here: it overestimates for longer videos and underestimates for shorter videos.

Analysis of failed cases. We investigate the causes of failed prediction for each predictor. The availability of channel information appears important – for most poorly predicted videos, their channels have only one or two videos in the training set. Moreover, some topics appear more difficult to predict than others. For example, videos that are labeled with *music* obtain a MAE score of 0.175 ($\bar{\eta}_{30}$ using the All-feature predictor). This amounts to an error increase of 28% compared to videos labeled with *obama* ($MAE = 0.136$). Lastly, the prediction performance varies considerably even for videos from the same channel and identically labeled. For example, the channel Smyth Radio (UC79quCUqSgHyAY9Kwt1V6mg) released a series of videos about “United States presidential election”, 8 of which are in our dataset: 6 are in the training set and 2 are in the test set. These videos have similar lengths (3 hours) and they are produced in a similar style. The 6 videos in training set are watched on average between 3 and 10 min-

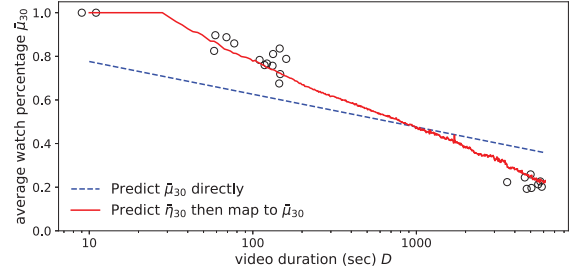


Figure 8: Shared linear regressors with channel reputation features on channel PBABowling, for predicting $\bar{\mu}_{30}$ (blue dashed) and predicting $\bar{\eta}_{30}$ then mapping to $\bar{\mu}_{30}$ (red solid).

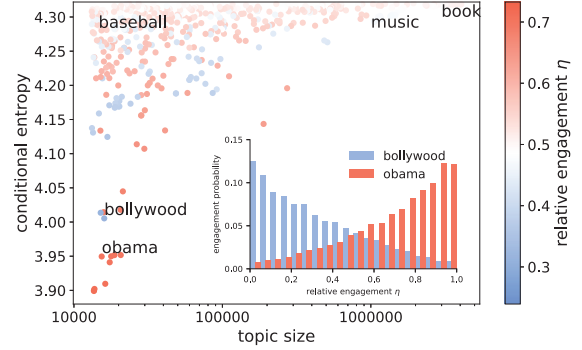


Figure 9: Informativeness for the most frequent 500 Freebase topics, measured by conditional entropy. (Inset) relative engagement distribution for two example topics: Obama - high engagement; bollywood - low engagement.

utes, yielding a $\bar{\eta}_{30}$ of 0.08. However, the 2 videos in the test set achieve considerable attention – 1.5 hours watch time on average, projecting $\bar{\eta}_{30}$ at 1.0. One possible explanation is that the videos in the test set discuss conspiracy theories and explicitly lists them in the title.

Overall, engagement metrics are **predictable from context, topics and channel information in a cold-start experiment setting.** Although channel reputation information is the strongest predictor, Freebase topics features are also somewhat predictive.

4.4 Are Freebase topics informative?

In this section, we analyze the Freebase topics features in detail and provide actionable insights for producing videos. Firstly, we group videos by Freebase topic and **extract the most frequent 500 topics.** Next we measure the amount of information gain with respect to relative engagement conditional entropy, defined in following equation:

$$H(Y|X_i = 1) = - \sum_{y \in Y} P(y|x_i = 1) \log_2 P(y|x_i = 1) \quad (4)$$

Each topic is represented as a binary variable $x_i \in \{0, 1\}$, for $i = 1, \dots, 500$. We divide relative engagement into 20 bins, and y is the discretized bin. A lower conditional entropy indicates the presence of current topic is informative

for engagement prediction (either higher or lower). Here we calculate $H(Y|X = 1)$ rather than $H(Y|X)$, because $X=0$ represents the majority of videos for most topics and the corresponding term will dominate. Using $H(Y|X = 1)$ quantifies its effect only when the topic is in presence (Sedhain et al. 2013). Fig. 9 is a scatter plot of topic size and conditional entropy. Here large topics such as `book` (3.2M videos) or `music` (842K videos) have high conditional entropy and mean relative engagement close to 0.5, which suggests they are not informative in predicting engagement. All informative topics (e.g., with conditional entropy 4.0 and lower) are relative small (e.g., appearing around 10K times in the training set). Fig. 9 (inset) plots two example topics that are very informative on engagement, from which we observe that videos about `bollywood` are more likely to have a low relative engagement while topic `obama` tends to keep audience watching longer. However, not all small topics are informative. A counter-example is `baseball`, which has a small topic size but a high condition entropy value.

In summary, watch percentage and relative engagement are predictable in a *cold-start* setting, before any behavioral data is collected. A few content-based semantic topics are predictive of low- or high- engagement. Such observation can help content producers make more engaging videos.

5 Related work

Measuring engagement in online content. Many researchers have analyzed engagement behavior towards web content. For example, the line of work that measures web page reading pattern often exploits auxiliary toolkit such as mouse-tracking (Arapakis, Lalmas, and Valkanas 2014) instrumented browsers. In search engine and recommender systems, dwell time, which is conceptually close to video watch time, has been widely used (Covington, Adams, and Sargin 2016). Interestingly, (Yi et al. 2014) compared two systems that optimize for clicks and dwell time, and found the one towards dwell time achieved better performance on ranking relevant products. All the above works focus on engagement with an individual user. However, user-level data is unavailable to content producers on YouTube platform. Our work measures engagement at an aggregate level, as complementary to individual engagement study.

The work most relevant to ours on measuring video aggregate engagement is from (Park, Naaman, and Berger 2016), in which the authors show the predictive power of collective reactions (e.g., view, like, and comment sentiment) for predicting average watch percentage. However, these features require observing videos for some period of time. Most importantly, a large fraction of videos do not have comments (Cheng, Dale, and Liu 2008), making this prediction setup inapplicable to a random YouTube video. In contrast, our work is the first to quantitatively measure the effect of video duration over a large-scale dataset and predict watch percentage in a *cold-start* setup. We further discuss related works in the following three directions.

Estimating quality of online content. MusicLab experiment is the first to measure online content quality in an experimental environment (Salganik, Dodds, and Watts 2006), in which they measure as the fraction of

download number over listening number. This experiment is further studied by (Krumme et al. 2012), who propose a two-step process to characterize user behavior in social systems. The key influencing factor in the first step is popularity such as product appeal and market position, while the second step is merely affected by content quality. (Stoddard 2015) has measured this process in Reddit and Hacker News. In this work, our notions of popularity and engagement are inspired by this two-step process, intuitively describing *the decision to click* and *the decision to interact* on YouTube. Moreover, (Van Hentenryck et al. 2016) show that popularity is a poor proxy to represent quality in online market. Thus, we propose a new metric *relative engagement* based on the engagement step, and formalize it to correlate with video quality.

Explaining popularity towards online videos. One of the most studied attributes is video popularity dynamics, defined as the number of times they are viewed. A number of models have been proposed to describe the popularity dynamics, such as a series of endogenous relaxations (Crane and Sornette 2008) or multiple power-law phases (Yu, Xie, and Sanner 2015). Other studies link popularity dynamics to epidemic contagion (Bauckhage, Hadiji, and Kersting 2015), external stimulation (Yu, Xie, and Sanner 2014) or geographic locality (Brodersen, Scellato, and Wattenhofer 2012). However, the amount of time that videos are watched has mainly been overlooked, despite becoming the centric metric for recommendation in YouTube (Meyerson 2012) and Facebook (Bapna and Park 2017). In this work, we provide an in-depth study on video engagement dynamics, and investigate key influencing factors.

6 Conclusion

In this paper, we measure a set of aggregate engagement metrics for online videos, including average watch time, average watch percentage, and a new metric, *relative engagement*. We study the proposed metrics on a publicly available dataset of 5.3 million videos. We show that relative engagement is stable over the video lifetime, and strongly correlates with established notions of video quality. In addition, we show average watch percentage can be predicted (with $R^2=0.77$) from public information, such as video context, topics, and channel, without observing any user reaction. This is a significant result that separates the tasks of estimating engagement with predicting popularity over time.

Limitations. Our observations are only on publicly available videos. It is possible that untweeted, private and unlisted videos behave differently. The attention data used are aggregated over all viewers of a video. Therefore our observations are more limited than those from content hosting site that has individual user attributes and reactions. Hence our results do not directly translate to user-specific engagement.

Future work and broader implications. For future work, one open problem is to quantify the gap between aggregate and individual measurements. Another is to extract more sophisticated features and to apply more advance techniques to improve the prediction performance. The observations in this work provide content producers with a new set of tools to create engaging videos and forecast user behavior. For

video hosting sites, engagement metrics can be used to optimize recommender systems and advertising strategies, as well as to detect potential clickbaits.

Acknowledgments. This research is sponsored by the Air Force Research Laboratory, under agreement number FA2386-15-1-4018. We thank National eResearch Collaboration Tools and Resources (Nectar) for providing computational resources, supported by the Australian Government.

References

- Abisheva, A.; Garimella, V. R. K.; Garcia, D.; and Weber, I. 2014. Who watches (and shares) what on youtube? and when?: using twitter to understand youtube viewership. In *WSDM*.
- Allen, M. P. 1997. The coefficient of determination in multiple regression. *Understanding Regression Analysis*.
- Arapakis, I.; Lalmas, M.; and Valkanas, G. 2014. Understanding within-content engagement through pattern analysis of mouse gestures. In *CIKM*.
- Bapna, A., and Park, S. 2017. [News Feed FYI: Updating How We Account For Video Completion Rates](#).
- Baukhage, C.; Hadiji, F.; and Kersting, K. 2015. How viral are viral videos? In *ICWSM*.
- Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*.
- Brodersen, A.; Scellato, S.; and Wattenhofer, M. 2012. Youtube around the world. In *WWW*.
- Cheng, J.; Adamic, L.; Dow, P. A.; Kleinberg, J. M.; and Leskovec, J. 2014. Can cascades be predicted? In *WWW*.
- Cheng, X.; Dale, C.; and Liu, J. 2008. Statistics and social network of youtube videos. In *IWQoS*.
- Covington, P.; Adams, J.; and Sargin, E. 2016. Deep neural networks for youtube recommendations. In *RecSys*.
- Crane, R., and Sornette, D. 2008. Robust dynamic classes revealed by measuring the response function of a social system. *PNAS*.
- Dobrian, F.; Sekar, V.; Awan, A.; Stoica, I.; Joseph, D.; Ganjam, A.; Zhan, J.; and Zhang, H. 2011. Understanding the impact of video quality on user engagement.
- Drutsa, A.; Gusev, G.; and Serdyukov, P. 2015. Future user engagement prediction and its application to improve the sensitivity of online experiments. In *WWW*.
- Figueiredo, F.; Almeida, J. M.; Benevenuto, F.; and Gummadi, K. P. 2014. Does content determine information popularity in social media?: A case study of youtube videos' content and their popularity. In *CHI*.
- Figueiredo, F.; Almeida, J. M.; Gonçalves, M. A.; and Benevenuto, F. 2016. Trendlearner: Early prediction of popularity trends of user generated content. *Information Sciences*.
- Guo, P. J.; Kim, J.; and Rubin, R. 2014. How video production affects student engagement: An empirical study of mooc videos. In *L@S*.
- Hessel, J.; Lee, L.; and Mimno, D. 2017. Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity. In *WWW*.
- Koenker, R. 2005. *Quantile regression*.
- Krumme, C.; Cebrian, M.; Pickard, G.; and Pentland, S. 2012. Quantifying social influence in an online cultural market. *PloS one*.
- Martin, T.; Hofman, J. M.; Sharma, A.; Anderson, A.; and Watts, D. J. 2016. Exploring limits to prediction in complex social systems. In *WWW*.
- Meyerson, E. 2012. YouTube Now: Why We Focus on Watch Time.
- Mishra, S.; Rizoio, M.-A.; and Xie, L. 2016. Feature driven and point process approaches for popularity prediction. In *CIKM*.
- Park, M.; Naaman, M.; and Berger, J. 2016. A data-driven study of view duration on youtube. In *ICWSM*.
- Pinto, H.; Almeida, J. M.; and Gonçalves, M. A. 2013. Using early view patterns to predict the popularity of youtube videos. In *WSDM*.
- Rizoio, M.-A., and Xie, L. 2017. Online popularity under promotion: Viral potential, forecasting, and the economics of time. *ICWSM*.
- Rizoio, M.-A.; Xie, L.; Sanner, S.; Cebrian, M.; Yu, H.; and Van Hentenryck, P. 2017. Expecting to be hip: Hawkes intensity processes for social media popularity. In *WWW*.
- Salganik, M. J.; Dodds, P. S.; and Watts, D. J. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*.
- Sedhain, S.; Sanner, S.; Xie, L.; Kidd, R.; Tran, K.-N.; and Christen, P. 2013. Social affinity filtering: Recommendation through fine-grained analysis of user interactions and activities. In *COSN*.
- Shuyo, N. 2010. Language detection library for java.
- Stoddard, G. 2015. Popularity dynamics and intrinsic quality in reddit and hacker news. In *ICWSM*.
- Szabo, G., and Huberman, B. A. 2010. Predicting the popularity of online content. *Communications of the ACM*.
- Van Hentenryck, P.; Abeliuk, A.; Berbeglia, F.; Maldonado, F.; and Berbeglia, G. 2016. Aligning popularity and quality in online cultural markets. In *ICWSM*.
- vidstatsx. 2017. Youtube top 100 most viewed news and politics video producers.
- Vijayanarasimhan, S., and Natsev, P. 2018. Research Blog: Announcing YouTube-8M: A Large and Diverse Labeled Video Dataset for Video Understanding Research.
- Wikipedia. 2018a. Billboard Year-End Hot 100 singles of 2016.
- Wikipedia. 2018b. Vevo in Wikipedia.
- Yi, X.; Hong, L.; Zhong, E.; Liu, N. N.; and Rajan, S. 2014. Beyond clicks: dwell time for personalization. In *RecSys*.
- Yu, H.; Xie, L.; and Sanner, S. 2014. Twitter-driven youtube views: Beyond individual influencers. In *MM*.
- Yu, H.; Xie, L.; and Sanner, S. 2015. The lifecycle of a youtube video: Phases, content and popularity. In *ICWSM*.