Click versus Share: A Feature-driven Study of Micro-Video Popularity and Virality in Social Media

Jingtao Ding* Yanghao Li* Yong Li* Depeng Jin*

Abstract

Micro-video has recently become an important form of user generated contents in the social media of microblogging. It is propagated by sharing and reaches the other users through being clicked and watched. Besides the traditional popularity metric for a micro-video such as click (or view) count, share count can indicate its virality in social domain. Understanding the differences between clicking and sharing behaviors is fundamental when evaluating the actual influence of micro-videos in social media. However, since that click data is usually not public available, above question has not been investigated in most studies. Thanks to a massive set of anonymized data from a major operator covering the whole China, we jointly study both clicking and sharing behaviors of over 10,000 micro-videos in Sina Weibo, the largest microblogging service and micro-video platform in China. Having extracted a rich set of features covering micro-video publishers, description texts and those shared users, we are able to identify the most influential features for click and share. From our studies, we observe that publisher-related features (*post* and *followee* counts) as well as the video *duration* have more impact on click, while video-description-related features including topical features and *emoticon* count are more correlated to share. Impacted by different features, the received clicks and shares of a micro-video may differ a lot from each other. Based on above observations, we build a prediction model for existing deviations among these two metrics, which can aid the development of a more effective and attractive micro-video platform.

Keywords— Click; Micro-Video; Share; Social Media

1 Introduction

Micro-video, a new form of user generated contents (UGCs), is gaining increasing enthusiasm due to its short-length and viewing convenience on mobile platforms. Recently, it has become an important source of profit for both platforms and video publishers. For example, Facebook is starting to put ads in the middle of its videos[1]. Therefore, it is important to study the information diffusion of micro-videos in social media.

Unlike texts or photos which can be displayed to users directly through sharing without further actions, micro-videos require users' actual clicks. When quantifying the influence of micro-videos, click and share are intrinsically different. A micro-video with high clicks has more viewers, i.e., a high popularity, and thus has important commercial value to both social media and potential advertisers. As for its received shares, it depends on the will of propagating the content for each user who has seen it in social media, indicating the social virality. Intuitively, shares may bring some more clicks from other users. However, *do those trending videos with high shares always receive more clicks?* Or in other words, *what is the difference between these two different metrics of micro-video popularity and virality?* The above questions cannot be overlooked when studying the information diffusion of micro-videos in social media. To answer this, we need to jointly analyze both clicking and sharing behaviors of micro-videos.

However, unlike shares, capturing these click events in social media is notoriously difficult. To our knowledge, since the dataset containing both shares and clicks is not publicly accessible, no previous work has simultaneously investigated clicking behavior and sharing behavior in the same social media. Indeed, Vallet *et al.* [18] considered a classification task about predicating video popularity (click) in YouTube and video virality (share) in Twitter. Since YouTube and Twitter are two different social media, a large proportion of the video clicks are not contributed by the links shared in Twitter. Thus one cannot compare clicking behavior and sharing behavior using the above data. To tackle the problem of data unavailability, Gabielkov *et al.* [6] developed a crawling strategy that captures the click statistics of the hyper-links in Twitter via a third-party URL shortening service (*bit.ly*). However, this largely depends on whether a specific social media uses URL shortening service, making the statistics not always available.

In this work, by collaborating with one of the largest Chinese mobile operators, we obtained an anonymous dataset containing millions of subscribers' HTTP traffic collected by using a deep packet inspection (DPI) system. Our dataset was collected between April 21st and April 26th, 2016 in China. Using multiple data collecting and preprocessing techniques, we are able to jointly

*Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Electronic Engineering, Tsinghua University, Beijing, China. {dingjt15, liyanghao14}@mails.tsinghua.edu.cn, {liyong07, jindp}@tsinghua.edu.cn.

[1]https://www.recode.net/2017/2/23/14707484/facebook-video-ad-test-midroll

study both clicking and sharing behaviors of over 10,000 micro-videos in Sina Weibo (`weibo.com`), the largest microblogging service in China. Weibo itself provides the micro-video service, where each video link is embedded in a post. Therefore, both sharing and clicking behaviors of the micro-videos can be captured in Weibo. Moreover, there are rich information about both video publishers and videos on the Weibo website, which can be utilized in our study.

The present work: This paper offers a feature-driven study of comparing clicking and sharing behaviors of micro-videos in Weibo, which correspond to popularity and virality, respectively. Besides the basic temporal dynamics of click and share, a rich set of features from multiple categories are extracted to comprehensively characterize the micro-videos, covering micro-video publishers, description texts and those shared users. Using statistical hypothesis testing methods, we are able to show the different impacts of our extracted features on clicks and shares, respectively. After that, we define a metric called click to share ratio (CSR), which measures the conversion rate from share count to click count. A micro-video with high or low CSR usually has a deviation between its click count and share count. We further show how our previous observations about influential features for click and share can be used to develop a model for automatically predicting the CSR level (high, medium, low) of micro-videos.

Our contributions are two-fold:

- The feature analysis of click and share identifies those informational features. In terms of the publisher-related features, high social influence ($follower$) can provide an effective promotion in terms of both click and share. Besides, significant impacts exist in the features related to the video descriptions, with differences between click and share. More importantly, the most influential features for click and share are totally different. On the one hand, publisher-related features, like $followee$ count and *post* count, as well as the video *duration* length have more impacts on the click count of micro-videos. On the other hand, video-description-related features including the topical features as well as the number of emoticons have more impacts on the share count of micro-videos (Section 4).

- Our established CSR prediction model achieves the high performance when using publisher-related features and video-description-related features in the early period (73.67% macro F1 score, 90.20% macro AUC),as well as the robust performance gain of 10% when varying the length of this period. This helps promoting the micro-videos with a high potential of popularity, and thus aids the development of a more effective and attractive micro-video platform (Section 5).

2 Related Work

(Micro-)Video popularity: Considering the potential guideline and commercial value that it can provide, the analysis of (micro-)video popularity has long attracted great attention from both industry and academia. One line of work has been focused on the characterization of video popularity. Popularity growth pattern of a video is always related to the type of its referrer [5], and its content [20]. Another line of work is to predict the (micro-)video popularity. The two main types of recent approaches to this problem are feature-driven and generative models. By connecting video popularity to an extensive set of features, feature-driven models can provide insightful knowledge about the informative features, like early view patterns [13] and evolution patterns [1]. Considering that micro-videos can be described by heterogeneous channels, multi-modal features are utilized, including social modality, textual modality, visual modality and so on [2]. As for generative models, which focus on providing both explanation and prediction, mostly point-process based models are used [3, 16].

In this work, we consider click (view) and share as the metrics of video popularity and virality, respectively. In order to investigate the differences between them, we analyze informative features for each metric and apply these knowledge in identifying micro-videos with significant deviations between their click count and share count.

Information diffusion in social media: Turning to research that studies information diffusion in social media, one hot topic is to trace paths of diffusion and influence through the social network [7, 8]. Apart from observing the structure and dynamics, there emerges another hot topic about how to predict the phenomenon of information diffusion from user to user, like information cascade [17]. Most works have investigated the share-based information diffusion, while the click-based studies are rare. Recently, Twitter users' clicking behavior on news-related links has been studied [6]. However, it was limited by the click statistics provided by a third-party URL shortening service.

Our work adds to this line of research by firstly comparing clicking and sharing behaviors of micro-videos in the same social media, based on a rich set of features from multiple categories.

3 Data and Features

We start by introducing the raw data collection. We then describe a general method for extracting a rich set of features, targeted to both clicks and shares of the micro-video.

3.1 Raw Data Collection We first lay out the terminology that we use in the remainder of the paper.

Micro-Video and Post. In Weibo, each micro-video link is embedded into a post. The unique id of the post, denoted as *PostID*, is a 16-digit number.

Click. Number of times a video-link, embedded in a post *PostID*, has been clicked by a user in Weibo.

Share. Number of times a post *PostID* that contains the video-link has been shared (i.e., reposted) by other Weibo users.

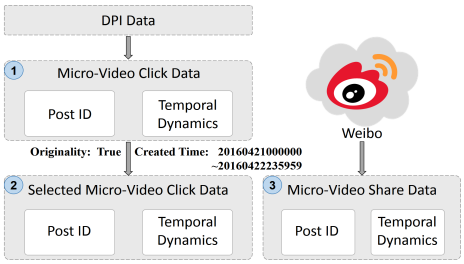


Figure 1: An illustration of the raw data collection.

Click data retrieved from operator network. The raw data was collected by one of the largest mobile operators in China using Deep Packet Inspection (DPI) appliances, between April 21*st* and 26*th* in 2016. This week includes no holiday or special events, thus can be considered as a typical week of Weibo usage. The coverage of data collection is the whole nation, with a uniform random sample of users. The detailed information about users' web visits are recorded in this dataset, including anonymized user IDs, timestamps for all HTTP requests and responses, and corresponding HTTP headers.

As shown in Figure 1, by referring to the destination URL field from the dataset, we extract the information related to micro-video play in Weibo (Step 1). Note that we only track users' actual clicks, filtering out those auto-play events. After that, we consider eliminating two main bias existed in our data (Step 2). One is non-original videos whose publishers are not the original uploaders. Another is micro-videos without a sufficient temporal range of click trace. Finally, we obtain the click data containing 10,473 original micro-videos, which are filtered from 0.32 million micro-videos with 30.9 million clicks. Each one has a full 4-day click trace and its final click count is the aggregation during these 4 days.

Share data crawled from Weibo. With the *PostID*s, we are able to crawl the full share trace of each selected micro-video from the Weibo website. The share count is computed as the total number of shares after 4 days, corresponding to the click data. This share data contains 10,473 original micro-videos with 4.48 million shares in total.

3.2 Feature Extraction As shown in the Figure 2, when a new video-post is pushed into a user's Weibo feeds, there are three main categories of features that can affect her will of clicking or sharing. The first one is the publisher influence. The second one is the text description of this post, which provides key content information before the user watches this micro-video. Apart from these, the third category is the influence of those users who have shared this micro-video in the first several hours after publishing.

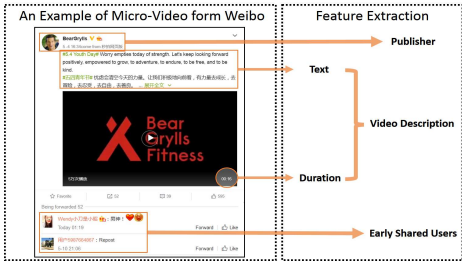


Figure 2: An illustration of the feature extraction.

Publisher category. We characterize the publisher influence via the social features crawled from the Weibo website. More specifically, these are follower count, followee count, post count and account verification[2].

Video description category. The basic characterizations of the video description include time duration, count of words (Chinese or English) and count of other special marks widely used in microblogs (hashtags and emoticons).

In addition to above basic characterizations, we also conduct a content analysis over the textual descriptions of micro-videos. Labeled-LDA [15] has been proved to be useful in the topic extraction of short-text corpora, like microblog text [14]. After building a training corpora from labelled data on the Weibo website, for each post d in our data, we are able to infer the probability of belonging to a given label z[3]. We denote

[2] A binary value indicating whether the publisher has been verified by Weibo.

[3] 8 categories: Fun, Music, Baby&Pet, star&show, Society, Life, Movie, Sport.

these features as $P_d\,(label{=}z)$.

Early shared user category. Similar to the publisher category, we characterize the influence of those early shared users from the following aspects: follower count and account verification. More specifically, for each micro-video post, we extract the follower counts and verification statuses of the shared users in the $1st$, $2nd$, $3rd$ and $4th$ hour, respectively. Then we compute the median and the quartiles (25%, 75%) of follower count, as well as the proportion of the verification.

To summarize, we extract the above three categories of features for 10,473 micro-videos. The extracted features are listed in Table 1 below.

Table 1: Features considered in our analysis.

Category	Name	Detail
Publisher	follower count	
	followee count	
	post count	
	verification status	
Video description	duration	
	word count	Chinese or English
	mark count	hashtag or emoticon
	$P_{label=z}$	$z \in \{fun, music, life$ $baby\&pet, star\&show,$ $society, movie, sports\}$
Early shared users	follower count	mean and quartiles (25%, 50%, 75%) for the 1st, 2nd, 3rd and 4th hour
	verification proportion	for the 1st, 2nd, 3rd and 4th hour

4 Feature Comparison between Click and Share

We now turn to analyze correlations of click and share with our extracted features. We study what different impacts these features have on click count and share count of a micro-video. More importantly, we investigate how these impacts differ between click and share.

4.1 Methodology Analysis of variance (ANOVA) is used to analyze the differences among group means by testing the null hypothesis that samples in all groups are drawn from populations with the same mean values [9]. Compared to the multiple two-sample t-tests, it is more conservative and thus results in less type I error. In our study, we first group 10,473 micro-videos into 5 groups, based on the scales of the final share count and click count, respectively. For example, in terms of the share, we denote five groups from $S1$ to $S5$, corresponding to the micro-videos with final share count falling in $[1, 10)$, $[10, 100)$, $[100, 1000)$, $[1000, 10000)$ and $[10000, +\infty)$. Similarly, five groups based on click counts are denoted from $C1$ to $C5$. To analyze the impact of a specific feature on the final count of click or share, we test if there is a strong evidence to reject the null hypothesis in one-way ANOVA, i.e., a correlation between the feature and the click count or the share count.

However, the normality assumption of ANOVA may be violated due to the fact that some of our extracted features follow a right-skewed distribution in the whole population, like the $follower$, $followee$ and $post$ that have been studied in other work [11]. Although the ANOVA is a relatively robust procedure with respect to violations of the normality assumption [10], we tackle this problem from the following two aspects. Firstly, we apply a log-transformation on those right-skewed features when performing the one-way ANOVA. Secondly, Kruskal-Wallis H test, which is a non-parametric variant of the one-way ANOVA, is also performed to act as a complement. Unlike the ANOVA, this test does not assume a normal distribution of the features.

Table 2: ANOVA results of the different features.

Category	Feature	Click (Sig.)	Share (Sig.)
Publisher	# of followers	$***$	$***$
	# of followees	$***$	$***$
	# of posts	$***$	$***$
Video Description	duration	$***$	$***$
	# of characters (zh.)	$***$	$***$
	# of words (en.)	$***$	$***$
	# of hashtags	$***$	
	# of emoticons	$***$	$***$
	$P_{label=fun}$	$***$	$***$
	$P_{label=music}$	$***$	$***$
	$P_{label=baby\&pet}$		$***$
	$P_{label=star\&show}$	$***$	
	$P_{label=society}$		$**$
	$P_{label=life}$	$**$	$***$
	$P_{label=sports}$	$**$	$***$
Early Shared Users	% of verification (1h)	$***$	$***$
	% of verification (2h)	$***$	$***$
	follower (3h)	$**$	
	follower (4h)	$***$	$**$

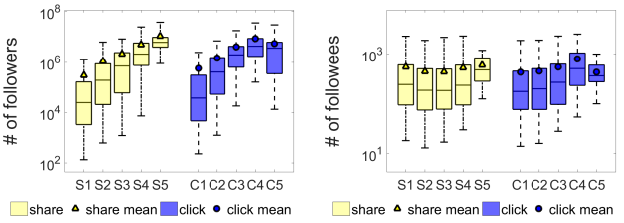
Significance (p) $*$:$p < 0.05$, $**$: $p < 0.01$, $***$: $p < 0.001$

For every feature extracted in Section 3.2, we perform both the ANOVA and the Kruskal-Wallis H test. Having obtained the similar significance p for rejecting

the null hypothesis under two tests of each feature, we believe the ANOVA results are fairly credible, i.e., with a low type I error. Thus we only present the ANOVA results in Table 2. Note that only the features with significant differences ($p < 0.05$) among at least one set of groups (click $\{C_i\}$ or share $\{S_i\}$, $i = 1, 2, 3, 4, 5$) are presented.

4.2 Results

4.2.1 Publisher Category As shown in Table 2, there are three features (*follower*, *post* and *followee*) that have impact on both clicks and shares of a micro-video ($p < 0.001$)). More specifically, Figure 3 plots the distribution quantiles (5%, 25%, 50%, 75%, 95%) and means of both *follower*s and *followee*s, where two sets of five-groups ($\{C_i\}$ and $\{S_i\}$) are presented together. It can be easily observed that these features of the micro-video publisher have positive effect on both clicks and shares of the micro-video. With the higher social influence, i.e., more *follower*s, the publisher is more possible to obtain large numbers of shares and clicks for her micro-videos. However, there are a few high-click micro-videos that are published by less influential users. The corresponding 5-percentile values of *follower* are about 10^4 in C_5, while those in S_5 are an order of magnitude higher, about 10^6 .

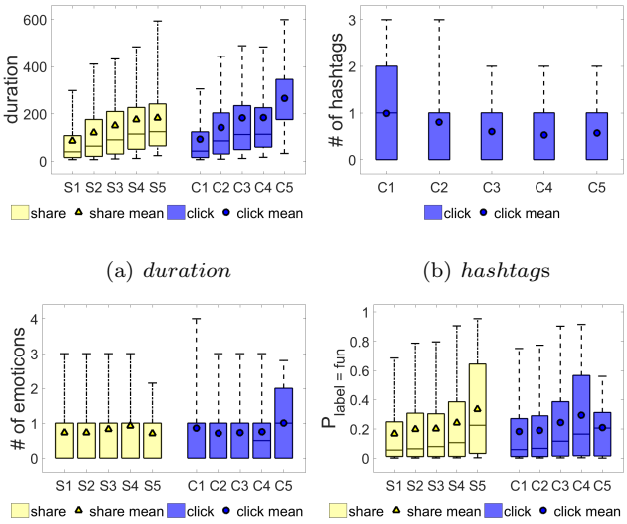


(a) *followers* (b) *followees*

Figure 3: Comparing differences among groups in terms of quantiles (5%, 25%, 50%, 75%, 95%) and means of the Publisher category features.

4.2.2 Description Category According to the ANOVA results in Table 2, differences are observed between two sets of 5-group. Several features can only affect either clicks or shares of a micro-video. Due to the space limit, we only present the 4 most significant features in Figure 4.

First of all, in terms of the micro-video *duration*, both high-click and high-share micro-videos are longer, with the means $240s$ versus $100s$ (C_5 vs. C_1, $p < 0.001$) and $180s$ versus $100s$ (S_5 vs. S_1, $p < 0.001$). This indicates that users prefer those micro-videos lasting



(a) *duration* (b) *hashtags*

(c) *emoticons* (d) $P_{label=fun}$

Figure 4: The distribution of the Video Description category features within different groups.

several minutes in Weibo.

As for the basic textual features of the video description text, interestingly, the number of *hashtag*s only has effect on the clicks. More specifically, more *hashtag*s are used in texts of the low-click micro-videos, while *emoticon*s are more common in both high-click and high-share ones (except for S_5). Since each hashtag is generally related to a trending topic in Weibo, multiple ones in one post can be confusing and thus obtain less clicks. In contrast, rich expressions in micro-video descriptions, like emoticons, attract more clicks and shares.

Apart from the above observations, the high-level topical features (P_{label}) also differs significantly among different groups in terms of both click and share. Both high-click and high-share micro-videos have a higher possibility of belonging in *label-fun* (means of 0.35 in S_5 and 0.30 in C_4, $p < 0.001$). As for the differences, *label-baby&pet* and *label-star&show* only have impacts on share and click, respectively (Table 2). High-share micro-videos have the lower means of $P_{label=baby\&pet}$, while the high-click micro-videos have the higher means of $P_{label=star\&show}$.

4.2.3 Early Shared User Category The social influence of those users who have shared the micro-videos in the first several hours has always been considered as an important feature to the video popularity. However, according to the Table 2, *follower*-related features of the early shared users in the first two hours have no impact on either click count or share count, while the

verification proportion of these early shared users are observed to have differences among groups. In the further analysis, we observe that micro-videos which have been shared by more verified users in the first two hours tend to obtain more clicks and shares.

4.3 Further Comparison Our previous analysis has demonstrated that different features have different impacts on click count and share count of a micro-video, respectively. Moreover, the feature impacts differ between click and share. In order to quantitatively compare the different impacts of our extracted features on final click count and share count, we perform a multiple linear regression analysis using the features above as predictor variables. Note that the response variables are scales of these two counts, i.e., 10-based logarithm values of click count and share count, and right-skewed features are applied with a log-transformation, as is in the ANOVA analysis. Additionally, all features are normalized to have zero means, making the intercept to represent the average scale of click/share count.

Table 3 shows the coefficients and the ranks of predictor features that rank top 5 in terms of click or share in the linear regression, where the ranks are obtained using the absolute values of coefficients. From the table we can tell that, the most correlated features for click are $follower$, *post*, *Chinese character*, *duration* and $followee$, while those for share are $follower$, *Chinese character*, $P_{label=society}$, $P_{label=music}$ and *emoticon*. Corresponding to our aforementioned observation, numbers of $follower$s and *Chinese character*s have positive and significant impacts on both click count and share count, as their coefficients all rank top 5 in Table 3. Specifically, *post*, *emoticon* and $P_{label=music}$ rank differently between click and share, with a difference at 10 or more, while they all have $p < 0.001$ in terms of both click and share in ANOVA analysis. Comparatively, except for $follower$ and *Chinese character*, the most influential features for clicks and shares are quite different. Publisher-related features, like $followee$ and *post*, as well as the *duration* have more impacts on the click count of micro-videos. Large number of $followee$s or long *duration* can increase the click count of a micro-video, while number of *post*s is negatively correlated to it. As for the share count of a micro-video, video-description-related features including $P_{label=society}$, $P_{label=music}$ and *emoticon* are highly correlated. Except for the $P_{label=society}$, the rest two both have positive impacts on the share count. Overall, this indicates that publisher-related influence plays an important role in the click count of a micro-video. When it comes to the share, the video description about its content has more impact on users' sharing behaviors.

Table 3: LR regression of the different features.

	Response					
	Click Sacle			Share Scale		
Predictor	**Coef.**	**Sig.**	**Rank**	**Coef.**	**Sig.**	**Rank**
*follower*s	4.76E-01	$***$	1	3.94E-01	$***$	1
*followee*s	6.01E-02	$***$	5	-4.79E-02	$***$	9
*post*s	-1.42E-01	$***$	2	-1.59E-03		18
duration	6.29E-02	$***$	4	5.30E-02	$***$	8
Ch. char.	1.11E-01	$***$	3	1.24E-01	$***$	2
*emoticon*s	2.31E-04		20	6.08E-02	$***$	5
$P_{label=music}$	-3.86E-03		16	6.20E-02	$***$	4
$P_{label=society}$	-2.45E-02	$*$	8	-6.59E-02	$***$	3
(Intercept)	1.27	$***$		1.84	$***$	
	R^2=0.32			R^2=0.23		

Summary. In terms of the Publisher category features, high social influence ($follower$) can provide an effective promotion in terms of both click and share. Significant impacts exist in the Video Description category features, with differences between click and share. A long duration (>180s) and text (>40 characters) have positive effect on click count and share count. A use of emoticons can slightly increase both click count and share count, while a use of hashtags can only affect click count. Besides, *label-*fun micro-videos are more preferred in both two metrics, while low shares and high clicks are observed in *label-baby&pet* micro-videos and *label-star&show* micro-videos, respectively. Surprisingly, $follower$-related features under the Early Shared User category have no impact on either click count or share count, while micro-videos shared by more verified users in the first several hours are always connected to high clicks and shares.

Comparing click and share, $follower$ and *Chinese character* have positive impacts on both two metrics. The biggest difference is that the publisher-related features and video-description-related features have more impacts on the click count and share count, respectively.

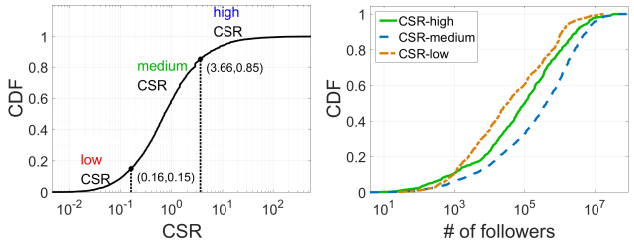
5 Deviation between Clicks and Shares

Generally, click count of a micro-videos tends to increase with share count. In fact, we have verified that early shares of the micro-videos are fairly correlated to the final count of clicks, which is close to the correlation between early clicks and final clicks (Spearman's ρ, 0.58 versus 0.69). However, our previous analysis found that the click count and the share count of a micro-video are affected by different features. Thus there must exist some micro-videos with deviations between their own click count and share count. In this section, we move forward into building a prediction model which is able to

identify these micro-videos. More specifically, based on the obtained knowledge, we consider a classification task that is related to the prediction of a micro-video with a significant deviation between its click count and share count. After that, we also demonstrate the possible applications.

5.1 Click to Share Ratio

5.1.1 Definition For a micro-video with final click count and share count after 4 days, denoted as N_{cf} and N_{sf}, it is necessary to define a deviation metric where the impact of the difference between click scale and share scale has been eliminated. Thus we first normalize N_{cf} and N_{sf} by the sums of all micro-videos, respectively. The metric, called click to share ratio (CSR), are defined as the normalized N_{cf} divided by the normalized N_{sf}. It indicates the conversion rate from share count to click count, i.e., average clicks received per share. The empirical CDF of CSR is plotted in Figure 5 (a), and the curve can be roughly divided into three parts. Based on this observation, we group these micro-videos into three classes, i.e., *CSR-high* (top 15% ranked by CSR, $CSR > 3.66$), *CSR-medium* (15%-85%, $0.16 < CSR < 3.66$) and *CSR-low* (last 15%, $CSR < 0.16$). A micro-video in *CSR-high* has relatively more clicks than shares, while that in *CSR-low* is opposite. Note that the CSR value can be less than 1, as we normalize both click count and share count before computing it.



(a) click to share ratio (b) *followers*

Figure 5: Distribution of the click to share ratio (CSR) and the most significant features.

5.1.2 Feature Impacts To investigate whether the aforementioned features have impacts on the CSR, we perform the similar ANOVA on every feature listed in Table 2. As we have demonstrated that some of these features have different impacts on clicks and shares, 14 (out of 19) features[4] are considered to have a strong evidence rejecting the null hypothesis, i.e., a significance $p < 0.001$. Due to the space limit, we only present the most significant feature, *followers*, in Figure 5 (b). Interestingly, both high-CSR and low-CSR micro-videos are more possible to be generated by less influential publishers, as $CDF(followers < 10^5)$ are 0.50, 0.35 and 0.60 for *CSR-high*, *CSR-medium* and *CSR-low*, respectively.

5.2 Micro-Video CSR Prediction Model

Given a micro-video with the 3 main categories of features, is it possible to predict its CSR level? In this model, we consider the 3-class classification task of the mirco-video CSRs in our data. According to Figure 5 (a), these 3 classes are unbalanced, with the proportion of *CSR-medium* being 70% and those of the rest two classes being 15% respectively. Thus the random guessing based on these proportions results in an unweighted average accuracy of 33%.

Since the ANOVA has indicated the significance of feature impacts among three classes, we select the same 14 features, which cover 3 categories of the Video Publisher, the Video Description and the Early Shared User. Considering the dominance of early periods in temporal evolutions of both clicks and shares, click count and share count in the first t hours, i.e., $NC(t)$ and $NS(t)$, are also used in the training process. We define this early period after the upload of micro-videos as *training window*. However, our prediction of CSR not only requires the accuracy, but also considers the timeliness as a key performance indicator. Thus when evaluating our model, we analyze the effect of the *training window* size in detail.

We perform 10-fold cross validation using a random forest classifier [12] which can inherently deal with multi-class problems. To tackle the problem of unbalanced classes, we employ a threshold-moving technique and train the classifier on the original data [4].

5.2.1 Overall Performance The prediction results are shown in Table 4, where we report the precision, recall, F1 score and area under ROC curve (AUC). For the first 3 metrics, we compute the unweighted average values of 3 classes (*Macro*), while the AUC is presented for each class. We observe that when combing all features and the early clicks and shares of the first 6 hours, the highest classification accuracy (73.67% macro F1 score, 90.20% macro AUC) can be obtained. This indicates a 11% gain in terms of the F1 score, comparing to the result of barely using early clicks and shares. Individually, the best performance is obtained by the Publisher category features (49.09% macro F1 score, 68.61% macro AUC), after which is the Video

[4]The rest 5 features are *duration*, $P_{label=fun}$, $P_{label=sports}$ and *follower*-related features of the shared users in first two hours.

Description category (40.93% macro F1 score, 60.02% macro AUC). On the one hand, the publisher and video description features can individually yield high accuracy without knowing the counts of early clicks and shares, i.e., before uploading the micro-videos. On the other hand, a great performance gain can be achieved when combining these features with the early statistics.

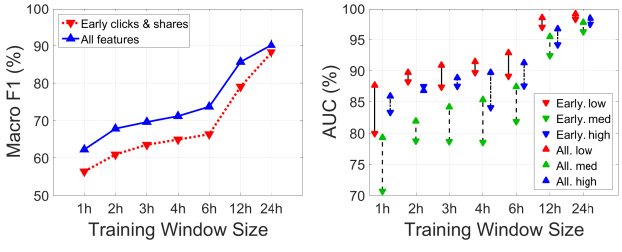
Table 4: Feature contribution analysis on the prediction of CSR level (%).

Features Used	M. Prec.	M. Rec.	M. F1	AUC-l/m/h
Random Guess	33.33	33.33	33.33	50.00/50.00/50.00
+Publisher	**51.37**	**47.87**	**49.09**	**70.30/66.86/68.67**
+Video Description	48.10	40.37	40.93	65.18/58.04/56.83
+Early Shared User	37.29	37.03	34.81	53.45/55.44/55.07
Clicks & Shares ($6h$)	67.67	65.17	66.31	89.39/82.15/87.82
+All Features	**80.55**	**69.63**	**73.67**	**92.54/87.12/90.94**

M.: Macro, l: CSR-low, m: CSR-medium, h: CSR-high

5.2.2 Effect of the Training Window Size It can be easily inferred that the predication performance increases with the *training window* size. Our previous investigation has highlighted the performance gain yield by adding the publisher and video description features. Consider the timeliness requirement of this prediction task, we analyze the relationship between performance gain of adding features and *training window* size and present it in Figure 6. In terms of both two metrics of F1 score and AUC, we all observe a higher performance gain with the shorter *training window* (1-6h), about 10% in F1 score. Though the prediction performance can be up to about 90% (F1 score) and 95% (AUC) when *training window* size is in $[12, 24]$ (hours), the gain of using features is limited due to fact that only using click and share counts is sufficient.

Overall, our detection model achieves the balance between the accuracy and timeliness. It not only can be used before the upload of micro-videos when barely using the publisher-related features and video-description-related features (49.09% macro F1 score, 68.61% macro AUC), but also achieves better performance when combining click and share statistics in the early period (73.67% macro F1 score, 90.20% macro AUC). Moreover, we demonstrate the robust performance gain of 10% by varying the length of early period, i.e., training window.



(a) Macro F1 score (b) AUC, in 3 classes

Figure 6: The performance gain of adding all features compared to barely using early statistics, with different training window size.

5.3 Application In social media, the click count and share count indicate popularity and virality of a micro-video, respectively. Generally, one would assume that the trending micro-videos, i.e., with high virality, are popular in terms of clicks. However, our previous investigation has shown that there exist a significant deviation between clicks and shares among some micro-videos. With help of the CSR prediction model, those *CSR-high* or *CSR-low* micro-videos can be identified, which may be overlooked when barely ranking them by the clicks. According to our statistics, when looking at the *CSR-high* micro-videos, i.e. top 15% ranked by the CSR, only 20.8% of them also rank top 15% by the click count, and 34.6% of them fall into the bottom 50%. As for the *CSR-low* micro-videos, 23.0% micro-videos rank top 50% by the click count. As we have demonstrated, most *CSR-high* micro-videos only have a few clicks and many popular ones do not have a high CSR.

After employing our detection model, for both publisher and micro-video service providers, they are able to detect those *CSR-high* yet low-click micro-videos in the early period. Then they can consider promoting these micro-videos by giving higher priority in recommendation (service providers) and incentivizing more users to share these in Weibo (publisher). Given the increasing shares and a high CSR, these micro-videos have a massive potential to become popular in terms of the clicks. In terms of those *CSR-low* micro-videos, publishers or service providers can consider some modifications before uploading the micro-videos, like replacing the uploader with a more influential user.

To summarize, by predicting the CSR of micro-videos in the early period or even before uploading, a more effective and attractive micro-video platform can be developed, where video-publishers and service providers can benefit from its commercial value, and ordinary users can enjoy its service.

6 Discussion and Conclusion

Based on the click data from operator network and the share data crawled from website, we jointly investigate both clicking and sharing behaviors among over 10,000 micro-videos in Weibo, via a rich set of features. Publisher *follower* count and video description *text* have positive impacts on both click count and

share count of a micro-video. As for the differences between click and share, we observe that publisher-related features (*post* and $followee$ counts) as well as the video *duration* are the most influential features for click, while those for share are video-description-related features including topical features and *emoticon* count. Based on above observations, we demonstrate how these publisher-related and video-description-related features can be utilized to build a model identifying micro-videos with large deviation between their clicks and shares, which can simultaneously benefit micro-video publishers, platforms and ordinary users.

This paper has several limitations. One is the 4-day length of data. Both final counts of click and share largely depend on the early characteristics, the growth of clicks or shares after several days should be slow and thus the 4-day length data is enough to obtain credible results. However, it is still interesting to see whether there will be different observations in terms of long-term statistics, which we aim to conduct in the future work. Another limitation is that, although we tried extensive features across three different categories, there can exist missing features which might be critically linked to our question, such as the latent features related to visual and acoustic content of micro-videos [2]. The right mentions ("@") in the tweets may also help to boost the diffusion of the video [19]. Due to limited data, we could not employ these features for our analysis, which we hope to investigate in the future.

7 Acknowledgments

This work was supported in part by the National Nature Science Foundation of China under 61621091 and 61673237, and research fund of Tsinghua University - Tencent Joint Laboratory for Internet Innovation Technology.

References

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

[2] J. Chen, X. Song, L. Nie, X. Wang, H. Zhang, and T.-S. Chua, *Micro tells macro: predicting the popularity of micro-videos via a transductive model*, in Proc. ACM MM, 2016, pp. 898–907.

[3] W. Ding, Y. Shang, L. Guo, X. Hu, R. Yan, and T. He, *Video popularity prediction by sentiment propagation via implicit network*, in Proc. ACM CIKM, 2015, pp. 1621–1630.

[4] P. Domingos, *Metacost: A general method for making classifiers cost-sensitive*, in Proc. ACM KDD, 1999, pp. 155–164.

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

[6] M. Gabielkov, A. Ramachandran, A. Chaintreau, and A. Legout, *Social clicks: What and who gets read on twitter?*, in Proc. ACM SIGMETRICS/IFIP Performance, 2016, pp. 179–192.

[7] M. Gomez Rodriguez, J. Leskovec, and A. Krause, *Inferring networks of diffusion and influence*, in Proc. ACM KDD, 2010, pp. 1019–1028.

[8] M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf, *Structure and dynamics of information pathways in online media*, in Proc. ACM WSDM, 2013, pp. 23–32.

[9] D. C. Howell, *Statistical methods for psychology*, Cengage Learning, 2012.

[10] R. E. Kirk, *Experimental design*, Wiley Online Library, 1982.

[11] H. Kwak, C. Lee, H. Park, and S. Moon, *What is twitter, a social network or a news media?*, in Proc. ACM WWW, 2010, pp. 591–600.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., *Scikit-learn: Machine learning in python*, Journal of Machine Learning Research, 12 (2011), pp. 2825–2830.

[13] H. Pinto, J. M. Almeida, and M. A. Gonçalves, *Using early view patterns to predict the popularity of youtube videos*, in Proc. ACM WSDM, 2013, pp. 365–374.

[14] D. Ramage, S. T. Dumais, and D. J. Liebling, *Characterizing microblogs with topic models.*, in Proc. ICWSM, 2010.

[15] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, *Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora*, in Proc. ACL EMNLP, 2009, pp. 248–256.

[16] M.-A. Rizoiu, L. Xie, S. Sanner, M. Cebrian, H. Yu, and P. Van Henteryck, *Expecting to be hip: Hawkes intensity processes for social media popularity*, in Proc. ACM WWW, 2017, pp. 735–744.

[17] K. Subbian, B. A. Prakash, and L. Adamic, *Detecting large reshare cascades in social networks*, in Proc. ACM WWW, 2017, pp. 597–605.

[18] D. Vallet, S. Berkovsky, S. Ardon, A. Mahanti, and M. A. Kafaar, *Characterizing and predicting viral-and-popular video content*, in Proc. ACM CIKM, 2015, pp. 1591–1600.

[19] B. Wang, C. Wang, J. Bu, C. Chen, W. V. Zhang, D. Cai, and X. He, *Whom to mention: expand the diffusion of tweets by@ recommendation on micro-blogging systems*, in Proc. WWW, 2013, pp. 1331–1340.

[20] H. Yu, L. Xie, S. Sanner, et al., *The lifecyle of a youtube video: Phases, content and popularity.*, in Proc. ICWSM, 2015, pp. 533–542.