Computer Networks 140 (2018) 112–125

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Predicting the implicit and the explicit video popularity in a User Generated Content site with enhanced social features

Adele Lu Jia[a], Siqi Shen[b,c,*], Dongsheng Li[b,c], Shengling Chen[b,c,d]

[a] College of Information and Electrical Engineering, China Agricultural University, China
[b] School of Computer, National University of Defense Technology (NUDT), China
[c] Parallel and Distributed Processing Laboratory, NUDT, China
[d] Baidu Inc., China

ARTICLE INFO

Article history:
Received 31 January 2018
Revised 13 April 2018
Accepted 7 May 2018

Keywords:
User Generated Content (UGC) sites
Social features
Popularity prediction

ABSTRACT

User Generated Content (UGC) sites like YouTube are nowadays entertaining over a billion people. Identifying popular contents is essential for these giant UGC sites as they allow users to request contents from a potentially unlimited selection in an asynchronous fashion. In this work, we conduct an analysis on the popularity prediction problem in UGC sites and complement previous work with two new aspects, namely differentiating contents that attract a lot of attention and that users really appreciate, and leveraging built-in social features to predict the content popularity immediately upon publication.

To this end, we conduct an extensive measurement and analysis of BiliBili, a YouTube-like UGC site with enhanced social features including user following, chat replay, and virtual money donation. Based on datasets that contain over 2 million videos and over 28 million users, we characterize the video repository and the user activities, we analyze the video popularities, we propose graph models that reveal user relationships and high-level social structures, and we successfully apply our findings to build machine-learned classifiers to identify popular videos.

1. Introduction

User Generated Content (UGC) sites exemplified by YouTube are nowadays a major Internet phenomenon that entertains over a billion users—almost one-third of all people on the internet—and form a billion-dollar global industry [1]. Given the scale, the dynamics, and the decentralization of the contents provided by individual users, one fundamental question for maintaining and growing such UGC sites is to understand and to identify contents that will gain great popularities. The content popularity, implicitly measured by the number of views, has been extensively studied before, ranging from revealing the popularity characteristics [2–7] to popularity predictions based on features [8–12] and generative models [13,14]. In this paper, we revisit this problem and complement previous studies with two new aspects, as follows:

- *Implicit and explicit popularity*: Implicit popularity in terms of the number of views reflect user's attention but not necessarily appreciations. Nowadays, UGC sites provide many opportunities for users to interact with the contents. Can we infer the content popularity explicitly from these interactions and predict the contents that users really like?
- *The benefits of the built-in social features*: For user retention and attraction, UGC sites are coupled with and have been introducing new social features like *donation* and *chat replay*. Can we leverage these new social features to refine the popularity prediction problem and particularly to make predictions immediately upon publication?

To this end, we need to choose a UGC site with enhanced social features as our research vehicle and obtain preferably the complete view or a representative sample of the whole system. And the dataset should be multi-dimensional and contains not only the content (video) information but also the user information including their relationships and interactions. For our analysis, we have chosen BiliBili [15], a Youtube-like UGC site with enhanced social features, for the following two reasons:

First, beyond traditional UGC functions like video sharing/viewing, voting, commenting and channel subscription, BiliBili implements a number of social features, including *non-reciprocal user following, chat replay*, and *virtual money donation*. Features extracted from social relationships like following have been shown to be informative for popularity prediction [9–14]. However, these

* Corresponding author at: School of Computer, National University of Defense Technology (NUDT), China.
E-mail addresses: ljia@cau.edu.cn (A.L. Jia), shensiqi@nudt.edu.cn (S. Shen), dsli@nudt.edu.cn (D. Li), chenshengling@baidu.com (S. Chen).

https://doi.org/10.1016/j.comnet.2018.05.004
1389-1286/

studies only focus on a small subset of videos and predictions can only be made after they get promoted in external social networks. In contrast, the built-in social features in BiliBili provide the possibility and the extra information for predicting the popularity, even immediately upon publication, of any video.

Secondly, unlike previous studies that are often carried out on sampled datasets [11,13,14,16–18], we were able to capture the entire repository of BiliBili (at the time of our crawling) with over 2 million videos and over 28 million users. Moreover, the information we obtained includes not only the repository characteristics such as the video duration and the user gender, but also user activities and interactions, for example, how users view and donate to the videos, when and who left what comments, and how users follow each other. This global view and fine-grained information avoid potential defects, e.g., under- and over- estimations of certain network properties, caused by sampling and sampling biases [19–21].

Our analysis of BiliBili mainly consists of three parts:

Implicit and explicit popularity. We first quantitively reveal the scale and the characteristics of BiliBili by examining its video repository that contains over 2 million videos. We analyze the implicit and the explicit popularity and we study the influence of the video type. Similar to previous studies, we find that both the implicit and the explicit video popularity is *highly skewed*, with a small number of videos collecting a large portion of the total popularity. Interestingly, we further find that popularity metrics that measure user's attention but not necessarily appreciations (i.e., the number of views and the number of replayed chat messages) are best fitted by Log-Normal distributions, whereas metrics that directly reflect user's appreciations (the amount of donated virtual money and the number of favorites) are best fitted by more skewed Power-Law distributions. Moreover, while users prefer to view videos shared from other sites, they are more generous in donating virtual money to the videos uploaded locally (highly likely to be user-generated).

Social features. Intuitively, social features provide complementary information for inferring both the explicit and the implicit popularity. Based on a dataset containing information on over 28 million users, we first examine the characteristics of three types of user activities and interactions, i.e., uploading, following, and commenting. We derive a number of interesting findings including that uploaders not only get more followers but are also more active in following others and that in general male users are more active while female users are more popular. Then, we propose two graphs that look not only at who a user is connected to, but also how those connected users are linked amongst each other. These graphs capture both the direct user relationships and the higher-order social structures and they provide valuable information for the popularity prediction problem.

Popularity prediction. Finally, applying our findings, we build feature-based predictors that can successfully predict popular videos, in terms of the number of views (implicit popularity) and the amount of virtual money users donated (explicit popularity). As it turns out, both the social features and the graphs we proposed are informative for the popularity prediction problem, e.g., on a balanced dataset where random guessing would yield an accuracy of 50%, our predictors achieve 86% even without knowing the early view patterns, and further improves the accuracy to 92% with only one-day observation.

We summarize our contributions as follows:

- We collect, use, and offer public access[1] to the dataset that contains the whole repository of BiliBili (at the time of the crawling), with detailed statistics for 2,858,844 videos and 28,962,041 users (Section 2).
- We provide a characterization on BiliBili. Our analysis includes (i) the repository scale, (ii) the statistical properties of the video popularity (Section 3), (iii) the uploading activity, (iv) the following activity, and (v) the commenting activity of the users (Section 4).
- We propose two graphs to analyze the user relationships, i.e., a *follow graph* that contains 10,749,726 users and a *comment graph* that contains 6,677,456 users (Section 5).
- We build machine-learned classifiers to identify with high accuracies the popular videos (Section 6).

[1] https://sites.google.com/view/bilibilidataset.

2. Methodology and the BiliBili dataset

In this section, we first give a brief introduction on the basic operations of BiliBili. Then, we identify important and informative characteristics and metrics for the popularity prediction problem. Finally, we introduce our measurement methodology and the dataset used throughout this article, and we describe the scale of BiliBili.

2.1. An overview of BiliBili

BiliBili is a YouTube-like UGC site with enhanced social features. As in traditional UGC sites, users in BiliBili can consume and share videos, vote and leave comments to videos, and subscribe to channels (of a series of videos). In addition, BiliBili provide three (unique) social features:

(i) *Non-reciprocal user following* that allows users to follow each other, for social purposes or merely getting updates on videos that they are interested in.
(ii) *Chat replays*, named *danmu* in BiliBili, are comments flying over the screen on exactly the video time when they are left before by various users. Chat replays allow the later users to understand and to communicate with their ancestors. They provide immersive viewing experiences and are adopted by a number of popular UGC sites including Twitch.tv [22]. An example of a BiliBili video page with chat replays is shown in Fig. 1.
(iii) *Virtual money donations* are made to uploaders by users who appreciate their contribution. In BiliBili, the virtual money (named coin) is used in various circumstances, including upgrading user membership and exchanging for new emojis.

Terminology. As users in BiliBili can take various roles, to simplify our arguments, we define the following user types:

(i) *uploaders*, users that have uploaded at least one video,
(ii) *viewers*, users that have not uploaded any videos,
(iii) *commenters*, users that have left at least one danmu, and
(iv) *social users*, users that have followed or have been followed by at least one other user. Specifically, if a user A follows a user B, then user A is named the *follower* of user B and user B is named the *followee* of user A.

2.2. Characteristics and metrics

To characterize BiliBili videos and users, we identify the following three important aspects that make up the basic operations of BiliBili, which provide important knowledge for the popularity prediction problem and will be later discussed in detail in Sections 3, 4, and 5, respectively.

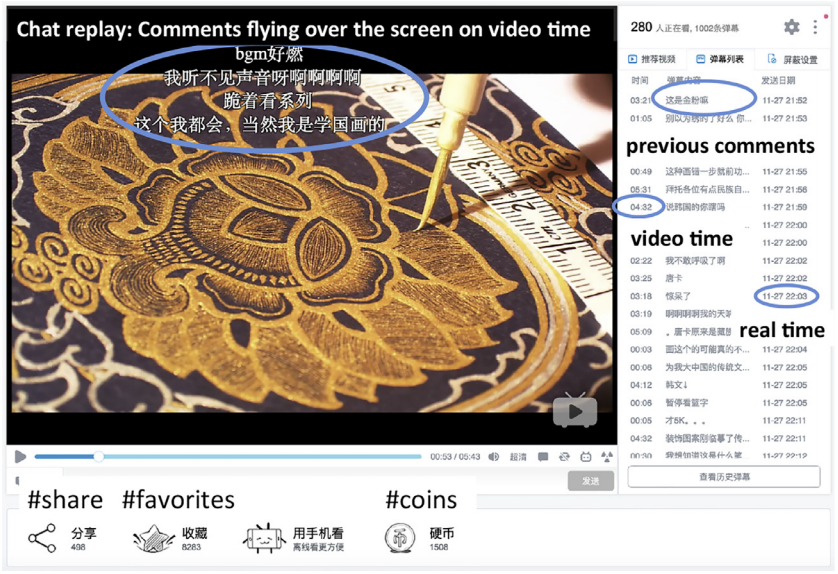


Fig. 1. An example of a BiliBili video page with chat replays.

2.2.1. Video repository characteristics

We consider in this article (i) video injection and duration that measure the scale of the repository and (ii) explicit and implicit popularity that measure the preference of the users, which are defined as the number of donations and the number of views, respectively. We report the injection rate and the duration for videos in the entire repository of BiliBili, we analyze the statistical properties of the video popularities, and we study their correlations with other features including the type of the videos.

2.2.2. User characteristics

We identify three aspects that cover most user activities in BiliBili, including (i) the upload activity, (ii) the follow activity, and (iii) the comment activity. In this article, we report the number of uploads and the number of views collected by videos uploaded by each user, the number of followers and the number of followees of each user, and the number of videos and the number of danmus commented by each user. Further, we study the mutual influences between the upload and the follow activity, and we examine their correlations with other features like the gender of the uploaders.

2.2.3. User relationships

We propose two graph models to capture user relationships based on their follow activity and comment activity. For both graphs we calculate three properties for each user that potentially influence the popularities of the videos he uploads, including (i) the degree, measuring his number of "friends", (ii) the clustering coefficient, measuring the closeness of his friends, and (iii) the PageRank score, measuring the structural importance of the user.

2.3. The BiliBili dataset

The primary challenge in crawling large online communities is covering, if not the entire repository, the giant connected component consisted of related contents or related users. For most online communities, such as YouTube, the user and the content identifiers do not follow a standard format. For such communities, the *snowball* method is commonly adopted for collecting the (ideally) complete list of identifiers, which later is used for fetching the user and the content information. However, based on an early-ended breadth-first search, the snowball method is known to produce a biased sample of nodes and to cause defects in representing the community structure [19,21].

Fortunately, BiliBili identifies each of its video and users with a unique numerical number in the increasing order. Each identifier corresponds to a webpage with detailed video or user information. By gradually increasing the identifier number from 1 to the maximal identifier we obtained at the time when we performed the crawling (May 19, 2016), we were able to obtain the complete view of BiliBili since it was first launched in September 2009. After removing the pages that are broken or are removed by the community or the users, in total we have collected information on 2,858,844 videos and 28,962,041 users, which we name the *video dataset* and the *user dataset*, respectively.

The video dataset contains, for each video, the uploader id (can be cross-referenced with the user dataset), the duration, the number of views, the number of favorites, and the amount of virtual money it collected, the repository of its chat replays (when and who left what comments at what video time[2]), and the video type (uploaded locally or shared from other sites, original or copied). The user dataset contains, for each user, the gender, the number of followers/followees, and the repository of its uploaded videos (can be cross-referenced with the video dataset).

2.4. BiliBill scale

Table 1 introduces the scale of BiliBili derived from our datasets. For the reference, we have also included the scale of YouTube as estimated in [16].[3] We note that this is a relatively early dataset and it was obtained when YouTube was roughly at the same age as BiliBili is now. We did not find any recent studies provide such statistics on the whole YouTube repository. Compared to YouTube, BiliBili has a smaller scale regarding the number of videos and uploaders, possibly due to the fact that BiliBili is mainly targeted at Chinese users. Nevertheless, it achieves a higher user

[2] BiliBili adopts two separate identifiers for users with general purposes (uploading and following) and users leaving danmus, the unique viewing experience and social feature provided by BiliBili. We do now know how these two identifiers link and therefore the commenter information cannot be cross-referenced with the user dataset.

[3] It should be noted that these numbers were estimated since crawling YouTube requires a search based on the snow-ball method and it is highly unlikely to capture the whole user base.

Table 1
BiliBili scale.

Features	BiliBili	YouTube [16]
no. videos	2,858,844	448 million
aggregate video length	122 years	2649 years
aggregate viewing time	2.9 million years	9.9 million years
aggregate no. views	23 billion	1.5 trillion
mean no. views	**8220**	**3348**
no. registered users	28,962,041	NA
no. uploaders	417,834	47.3 million
no. commenters	6,677,458	NA
no. social users	10,749,720	NA

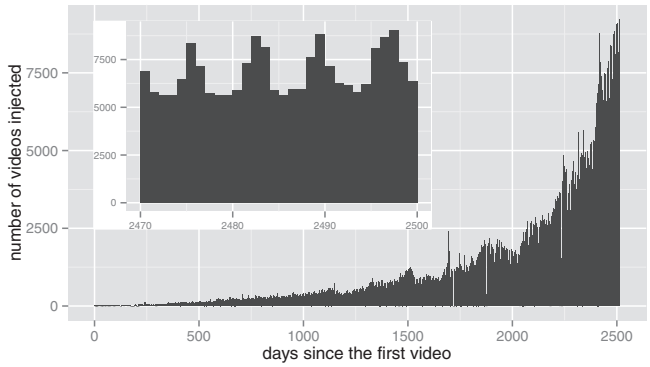


Fig. 2. Video injection rate.

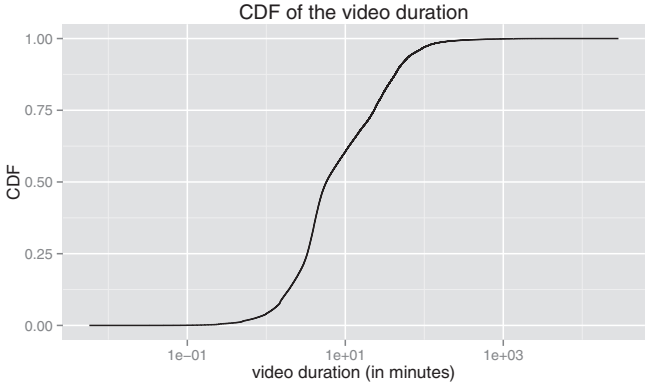


Fig. 3. CDF of the video duration.

activity level: on average, BiliBili videos are viewed twice more compared to YouTube videos.

2.4.1. Video injection

Fig. 2 shows the number of videos injected each day since the start of BiliBili. We find that BiliBili is growing dramatically over the years, with an exponentially increasing daily video injection rate. Further, the video injection exhibits a clear weekend effect, with a larger number of videos injected on the weekend than in the week-days. Similar patterns have also been observed in other UGC sites such as YouTube and Twitch [6,7,18].

2.4.2. Video duration

As shown in Fig. 3, most videos in BiliBili are of short durations: over half of the videos are within 10 min, and over 95% videos are within 20 min. This result is consistent with YouTube [18] but not with Twitch [4,6], possibly because YouTube and BiliBili cover a wide range of topics whereas Twitch is exclusive for gaming videos.

Table 2
Video statistics.

Features	Min	1st Qu.	Median	Mean	3rd Qu.	Max
no. views	1	282	849	7643	2747	10,290,411
no. favorites	0	1	8	124.7	35	340,547
no. coins	0	0	2	35.5	6	362,660
no. danmus	0	3	16	250.7	69	1,109,635
no. commenters	0	2	8	52.2	32	4,963

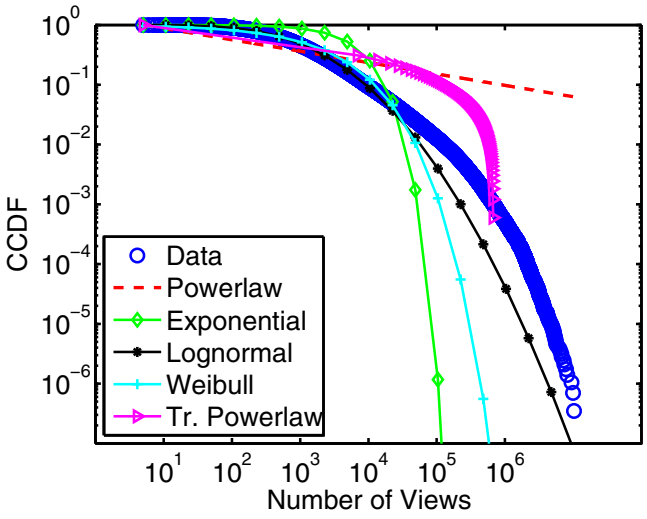


Fig. 4. CCDF of the number of views and the curve fittings.

3. Video popularity

In any UGC sharing site, content popularity provides important knowledge for the activity level of the users and the potential workload for maintaining the site. Content popularity in UGC sites has been extensively studied before [2–7,11,13,14,23,24]. In this section, we complement previous studies with an in-depth analysis on not only the implicit but also the explicit popularity, measured by the number of views, the number of favorites, the number of coins, the number of danmus, and the number of commenters a video attracted, respectively. The basic statistics are shown in Table 2.

We see that, for any definition, the video popularity is highly skewed, with a small number of videos collect a large portion of the total popularity. We provide a more detailed analysis in the following sections.

3.1. Implicit popularity: the number of views

Most previous studies use the number of views as the measure of the content popularity. It is indeed an important metric as it reflects the interests of the users and implies the potential workload for maintaining the site (videos with many viewers will need more servers to support them). However, viewing a video (especially for only a small part of it) does not necessarily mean that the users like the video: they may simply be exploring, or they may dislike it and will quickly turn it off. Therefore, the number of views only implicitly reflects the video popularity and we name it the *implicit popularity*. We show in Fig. 4 the Complementary Cumulative Distribution Function (CCDF) of the number of views collected by each video.

Method for distribution fitting. We attempt to fit the empirical data with a set of well-known probability distributions that are available in most simulation and experimental toolboxes, namely the Exponential, the Weibull, the Log-Normal, the Power-Law, the Gamma, and the Normal distributions. The fitting is performed us-

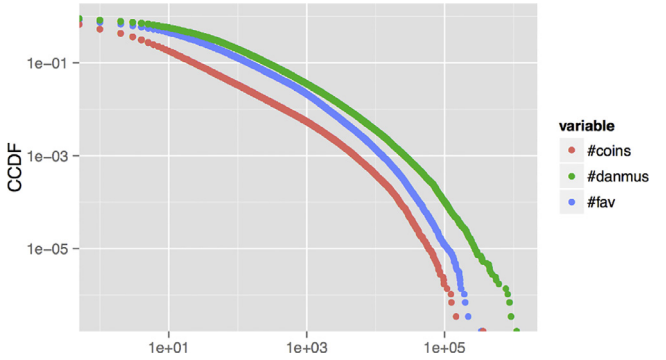


Fig. 5. CCDF of the number of coins, the number of favourites, and the number of danmus.

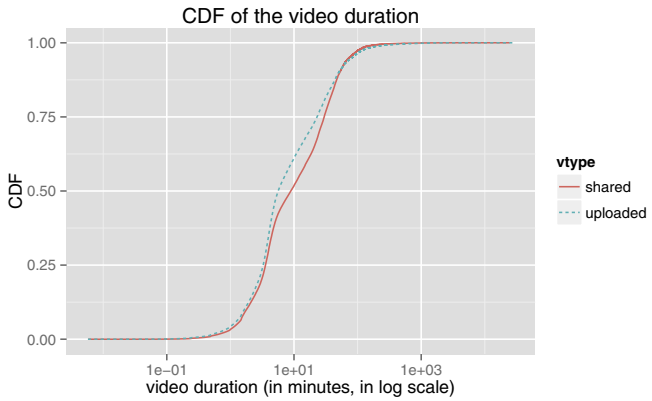


Fig. 6. CDF of the video duration.

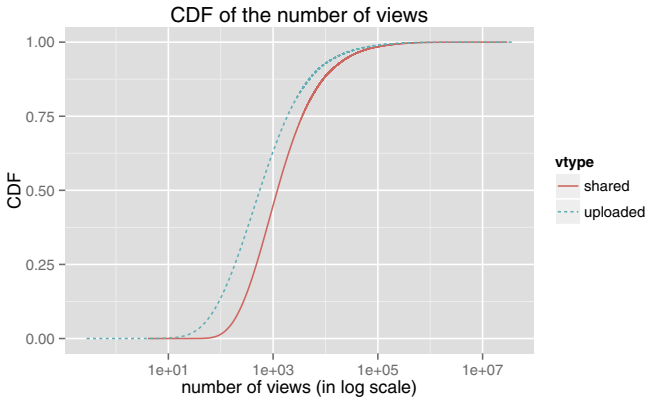


Fig. 7. CDF of the number of views.

ing maximum likelihood estimation (MLE), which determines for a distribution the parameters that lead to the best fit with given empirical data. Then, we use a method for assessing the goodness-of-fit (GoF) by using a combination of Kolmogorov-Smirnov (KS) [25] and the Anderson-Darling (AD) [26] GoF tests. We determine a distribution is the best fit for the empirical data if it passes the two GoF tests and has the smallest D value (the maximum vertical deviation between the two curves). We call the probability distribution of data is "long- tail" if the tail of the probability distribution is longer than the fitted exponential distribution of data.

We find that the best fit for the number of views is the Log-Normal distribution (mean $\mu = 6.85$, variation $\sigma = 1.76$, $x_{min} = 5$), rather than a power-law distribution that is observed in many other UGC sites such as YouTube and Twitch [4,18], indicating that the skewness of the implicit popularity in BiliBili is not as severe as in YouTube and in Twitch.

3.2. Explicit popularity: user appreciation

Users express their appreciations explicitly through favoring and donating to the video. While favoring is a traditional function provided by many video sharing sites, donating to other users is a relatively rare service. Yet it is a key social feature provided by BiliBili, aiming to encourage users to show their support explicitly for the effort of other users.

Fig. 5 shows the CCDF of the number of coins and the number of favorites collected by each video. Again, we observe highly skewed popularities: while 80% videos have received fewer than 10 coins and less than 20 favorites, 3% videos have earned more than 100 coins and 500 favorites, with top videos accumulating thousands of coins and favorites. Based on the same fitting method as introduced above, we find that the number of coins and the number of favorites are both best fitted by Power-law distributions, with $\alpha = 1.76$, $C = 5.00$, $x_{min} = 5$ and $\alpha = 1.52$, $C = 5.00$, $x_{min} = 5$, respectively.

3.3. Explicit popularity: the "buzz" makers.

Another way for users to explicitly express their interests is through leaving comments, or danmus in the case of BiliBili. As shown in Fig. 5, the number of danmus attracted by each video is highly skewed: while half of the videos have attracted fewer than 16 danmus and fewer than 10 commenters, 3% videos have received more than 1000 danmus and more than 200 commenters. Similar to the above analyses, we find that the number of danmus is best fitted by the Log-Normal distribution (mean $\mu = 3.83$, variation $\sigma = 1.60$, $x_{min} = 5$).

It is worthwhile to mention that leaving danmus does not necessarily mean that the users like the video—it is possible that the video is a "buzz" maker and attracts a lot of attention and conflicts from factional users. And the number of danmus only imply the amount of attention received by the videos, but not the nature of the attention. Using techniques form natural language processing, we will be able to identify the positive and the negative emotions associated with each danmu. We leave this as our future work.

Discussion: For the four popularity metrics we have considered, we find that the number of views and the number of danmus, i.e., the two metrics that reveal user's attention but not necessarily appreciations, are best fitted with Log-Normal distributions whereas the number of coins and the number of favorites, which directly shows user's admirations, are best fitted with Power-Law distributions. While both Log-Normal and Power-Law are heavy-tailed distributions, the latter one is more skewed than the former one. These results indicate that the "*rich-get-richer*" phenomenon is more profound for collecting user's appreciations than for collecting user's attention. Further reasoning on this distinction requires qualitative analyses which we leave as our future work.

3.4. Influence of the video type

In BiliBili, videos are either directly uploaded by the users or shared from other sites. In this section, we analyze the influence of the video type on the basic video characteristics. After removing videos that contain no explicit information on the video type, we find in total 557,414 videos (19.50%) are shared from other site and 1,830,382 (64.03%) videos are uploaded locally. Figs. 6–9 show the CDFs of the video duration, and of the number of views, the number of danmus, and the number of coins accumulated by each video from either type.

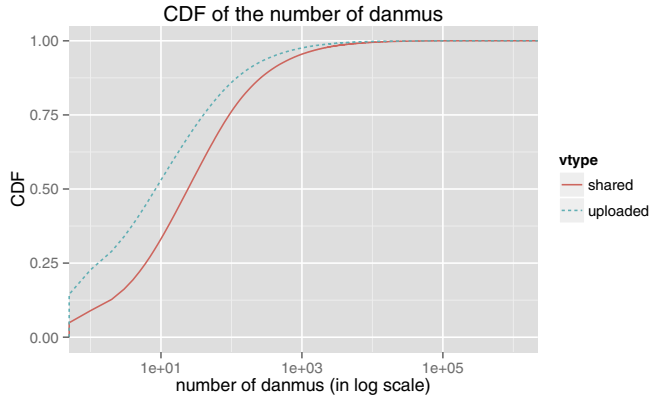


Fig. 8. CDF of the number of danmus.

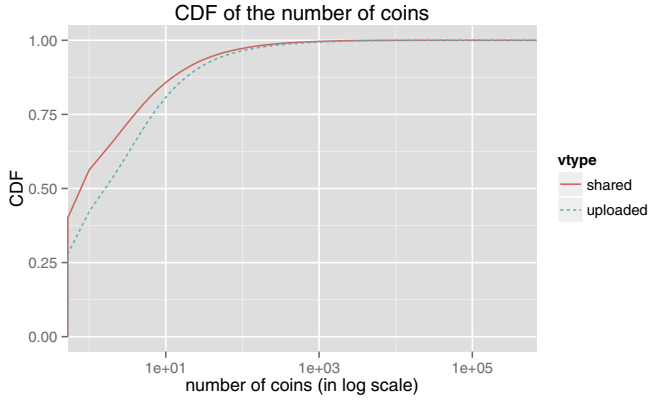


Fig. 9. CDF of the number of coins.

While the difference in the video duration between shared videos and uploaded videos is negligible, we find that shared videos usually attract a larger number of views and danmus: 50% uploaded videos attract fewer than 300 views and fewer than 10 danmus, whereas 50% shared videos attract more than 1000 views and more than 50 danmus.

Interestingly, though it seems that *users prefer shared videos in their viewing activities, they are more generous in donating virtual money to the locally uploaded videos*. As shown in Fig. 9, while 35% shared videos do not receive any coins, 20% uploaded videos receive more than 10 coins. One possible explanation is that users tend to support the local community. This hypothesis can only be tested by qualitative analyses such as surveys, which we leave as our future work.

4. User activities and interactions

As mentioned above, in this work we complement previous analyses on the content popularity with two new aspects, namely differentiating the implicit and the explicit popularity, and leveraging the new social features for the polarity prediction problem. To address the second issue, we first in this section provide a characterization on user activities and interactions in BiliBili, including the upload, the follow, and the comment activities. The basic statistics are shown in Table 3. The features we analyzed in this section will later be leveraged to predict video popularities.

4.1. Upload activity

In BiliBili, only 417,834 (1.44%) of all the 28 million users have ever uploaded a video (named *uploaders*), indicating that most users are merely viewers. The uploaders on average have uploaded 7.64 videos and have collected 59,390 views, resulting in an average of 4426 views per uploaded video. For comparison, [16] estimated that YouTube in 2011, at a similar age as BiliBili now, contains 47.3 million uploaders who on average each uploads 9.5 videos and collects 31,143 views. Again, we find that *BiliBili is a smaller but more active community compared to YouTube then: it has much fewer uploaders who nevertheless attract more views*.

Taking a closer look, we find a highly skewed upload activity level. While 87.24% uploaders have shared fewer than 10 videos, 0.81% (3,399) uploaders have shared more than 100 videos. On the extreme, the most active uploader have shared 22,324 videos.[4] Similarly, the total and the average number of views collected by all the videos shared by an uploader are also highly skewed.

Influence of the uploader gender. For this analysis, we consider only the uploaders who choose to reveal their gender on their homepage. In total, we find 71,994 female uploaders and 136,693 male uploaders. The basic statistics for their upload activities are shown in Table 4. We see that male and female uploaders do not exhibit prominent differences in the number of uploads, except that a few male uploaders are extremely active and the most active male uploader has uploaded 13,145 videos. Interestingly, we do find that *most female uploaders are more popular than male uploaders, while a few male uploaders are extremely popular*.

4.2. Follow activity

In BiliBili, 10,749,720 (37.12%) users have followed or have been followed by at least one user (named *social users*). For each social user, we have obtained the list of users that he follows and the list of users that follow him. The basic statistics can be found in Table 3. Particularly, for the ratio between the number of followers and followees, we have only considered users who have at least one follower and at least one followee, which in total counts for 1,079,027 users.

We find that on average each user follows and is followed by 5 users. To the extreme, a user follows 11,512 users and another user is followed by 1,420,203 users. The latter one turns out (after manual check) to be a popular internet streaming host and standup comedian in China. We did not find any identity information for the former user. Although the maximum number of followers is much larger than the maximum number of followees, for users with at least one follower and one followee, we find that more than 75% of these users follow more users than they have followers.

On the other hand, we observe that in total 18,238,014 (62.97%) users have not followed nor have been followed by any user, showing that follow activities in BiliBili are not substantial. This is probably because that, unlike Twitter, the running of BiliBili does not pivot on the follow function: even without following anyone, a user can still explore, efficiently or inefficiently, most of the contents in the network.

4.2.1. Influence of the upload activity

Unlike Twitter, the follow feature in BiliBili is mainly used for updating recently published videos instead of news/tweets. As a consequence, most of the follows are targeted at uploaders instead of viewers: as shown in Fig. 10(a), around 75% uploaders have obtained at least one follower while 98% viewers have no followers at all. Among the uploaders, the follower distribution is highly skewed: 62.5% uploaders have fewer than 10 followers while 2.7% (0.44%) uploaders have more than 1000 (10,000) followers.

As BiliBili is mainly used for video sharing, it is natural that uploaders would attract more followers. Nevertheless, we also find

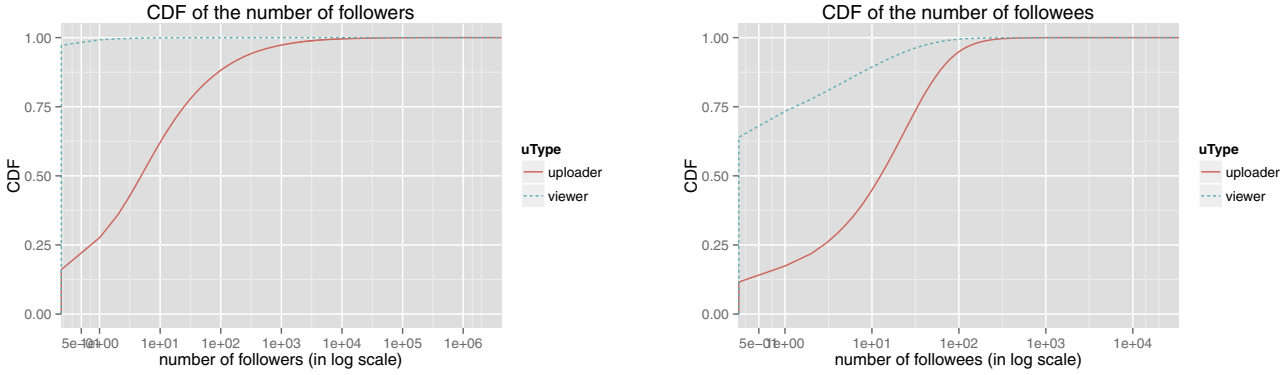
[4] After manual check, we find that this uploader is registered in 2012 and now have 341 thousand followers. We conjecture that its account is enterprise-maintained.

Table 3
Statistics on user activities.

Activity	Features	1st Qu.	Median	Mean	3rd Qu.	Max
upload	no. uploads	1	2	7.64	5	22,324
	no. views	515	2067	59,390	9375	1.570e+09
	avg. no. views	278	842	4426	2528	3.387e+06
follow	no. followers	0	0	14	0	1,420,203
	no. followees	1	4	13.86	14	11,513
	follower/followee ratio	0.0417	0.1429	19.9570	0.7273	619,524.00
comment	no. commented videos	2	6	22.35	20	65,450
	no. danmus	3	10	52.7	37	640,863
	avg. no. danmus	1.00	1.43	1.96	2	9153.00

Table 4
Influence of the user gender the upload activity.

Features	Gender	Min	1st Qu.	Median	Mean	3rd Qu.	Max
no. uploads	M	1	1	2	9.27	6	13,145
	F	1	1	2	7.73	6	3817
no.views	M	0	474	1983	85,343	9478	1.57+e10
	F	0	684	2775	54,057	12,672	1.23+e08
avg. no. views	M	0	226	706	4573	2239	3,142,000
	F	0	362	1011	4322	2890	3,387,000



(a) CDF of the number of followers

(b) CDF of the number of followees

Fig. 10. Influence of the user type on their following activities.

that uploaders are more active in following others. As shown in Fig. 10(b), while over 60% viewers have not followed anyone, over 50% (5%) uploaders have followed more than 10 (100) users. Further, we also observe that *uploaders with followers are more active in uploading*: their average number of uploads are 5 time as much as uploaders with no followers (7.84 versus 1.45 videos on average)

4.2.2. Influence of the user gender

For this analysis, we consider only the social users that choose to reveal their gender on their homepage. In total, we find 1,393,228 female social users and 1,893,312 male social users. We find that in general male users have more followers and more followees compared to female users. As shown in Fig. 11, 87% male users and 90% female users have no followers, and 75% male users have followed at least one user while only 67% female users have done so.

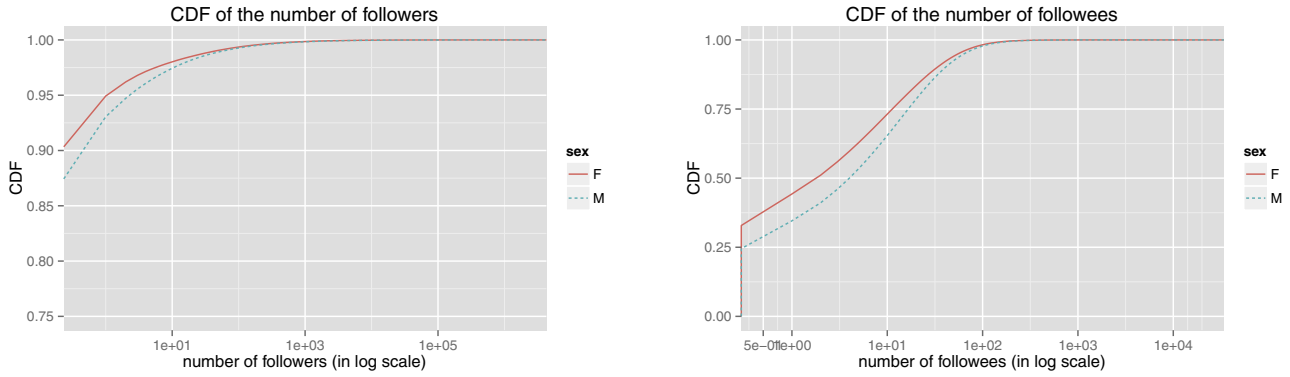
Together with the previous analysis, we conclude that *male users are more active in following each other and female users are more popular* (getting more attentions for the videos they share). Further reasoning on this phenomenon relies on a more subtle analysis (probably through surveys) which we leave as our future work.

4.3. Comment activity

In this section, we analyze user's comment activities in terms of the number of videos, the number of danmus, and the average number of danmus across all the videos commented by each user. In total, we find that 6,677,458 (23.06%) users have left at least one danmu (named *commenter*). On average, each commenter has commented on 22 videos and has left 52 danmus, resulting in an average of 1.96 danmus per commented video per user. Similar to the previous observations, we find that user activity level in video commenting is also highly skewed: while 19.59% users have commented only on one video, 4.08% users have commented on more than 100 videos. It is worthwhile to mention that BiliBili adopts separate identifiers for the commenters (different from their user id). So far, we did not find a way to link these two identifiers and hence a finer-grained analysis on user's commenting activity is left for our future work.

5. User relationships

While the content popularity is directly measured by the user-content interactions, we believe that indirect user relationships also provide valuable reference. For example, it is highly likely that friends (and friends of friends), or users co-commenting on the same videos repeatedly, will have similar viewing lists.



(a) CDF of the number of followers

(b) CDF of the number of followees

Fig. 11. Influence of the user gender on their following activities.

Table 5
Properties of the follow graph. The metrics we present include the number of nodes, n, the number of edges, e, the average degree, d, the percentage of nodes in the Weakly Connected Component (WCC) and the Strongly Connected Component (SCC), the effective diameter, D, and the average clustering coefficient, c.

n	e	d out/in	WCC/SCC (random)	D (random)	c (random)
10,749,720	149,037,569	13.86/13.90	99.50%/6.09% (100%)	3.86 (5.81)	0.0867 (0.0000)

To analyze user relationships in BiliBili, in this section we propose two graph models, i.e., a *follow graph* based on user's follow relationships, and a *comment graph* based on user's comment activities. These graph models capture both the direct user relationships and the higher-order social structures, and the graph properties will later be used as complementary features for the popularity prediction.

5.1. The follow graph

In the follow graph, a node represents a user and a directed edge from node A to node B represents that user A has followed user B in BiliBili. In total, the follow graph contains 10,749,720 nodes (counting for 37.12% registered users) and 149,037,569 directed edges, among which 583,888 (95%) nodes do not have a single mutual edge—these are instances where the user is exclusively using BiliBili for either information dissemination or consumption. In fact, only 0.76% of edges in the follow graph are reciprocated—this number is much lower compared to the case of Twitter where 42% edges are reciprocated [27] and obviously to Facebook where all edges are reciprocated [28]—indicating that BiliBili is mainly an information network coupled with minor social implications.

5.1.1. Overall graph properties

The graph properties we considered include the number of nodes, the number of edges, the average node degree, the fraction of nodes in the largest connected component (LCC, for follow graph which is directed, we have calculated both the WCC and SCC, abbreviated for strongly and weakly connected component, respectively), the diameter, the effective diameter (the 90th percentile of the distribution of shortest path lengths between all node pairs), and the clustering coefficient. The graph properties are summarized in Table 5. For the reference, we have also generated a random graph with the same number of nodes and edges as the follow graph. We have a number of interesting observations, as follows:

First, as discussed earlier, the level of reciprocation in user's follow activity is extremely low. Reflected in the graph structure, we observe a very small SCC consisted by only 6.09% of the nodes (for Twitter it is 68.7% [27]). This result again indicates that most BiliBili users are engaged in chain-like information consumption and dissemination. Secondly, though with minor reciprocations, the average clustering coefficients for the follow graph is much larger than the random graph in reference, showing that, social-driven or information-driven, user relationships established in BiliBili are spontaneous rather than purely random. Together with the fact that the diameter of the follow graph is smaller than the random graph, we conclude that the follow graph exhibit the *small-world* properties.

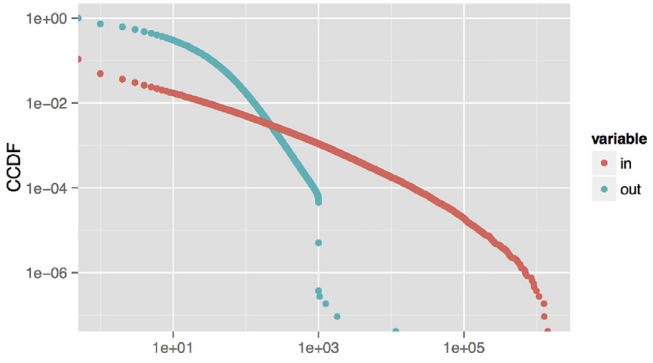


Fig. 12. CCDF of the node degree in the follow graph.

5.1.2. Node degree distribution

Fig. 12 shows the CCDF of the in- and the out-degree of all users. Not surprisingly, we find that some users are extremely active in following others, and some users are extremely popular and they get a large number of followers (they may not be the same users). These large in- and out-degrees are indications of non-social behavior, as it has been well-established that individuals are only able to maintain around 150 stable social relationships at a time [29]. The out-degree distribution shows a clear cut off at 1000, a limit imposed by BiliBili on the number of users one can follow at the time of our measurement.

Table 6
Properties of comment graphs with different values of the threshold t. The metrics we present include the number of nodes, n, the number of edges, e, the link density, l, the average degree, d, the percentage of nodes in the Largest Connected Component (LCC), the diameter, D, and the average clustering coefficient, c.

t	n	e	d	LCC (random)	D (random)	c (random)
1	3,694,739	259,821,515	140	99.9% (100%)	7 (4)	0.5581 (0.0000)
2	1,033,669	9,105,183	17	96.88% (100%)	12 (7)	0.4061 (0.0000)
3	411,923	2,594,483	12	95.72% (100%)	15 (8)	0.4111 (0.0000)
4	229,350	1,202,751	10	94.97% (100%)	14 (8)	0.4177 (0.0000)
5	148,352	680,420	9	94.24% (100%)	18 (8)	0.4182 (0.0000)
10	39,455	121,659	6	90.03% (99.82%)	14 (10)	0.4015 (0.0001)

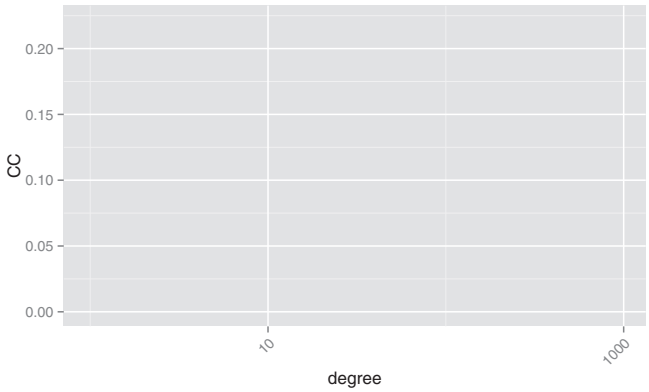


Fig. 13. The average clustering coefficient of users as a function of their degree in the mutual follow graph.

For most users, they still have moderate in- and out-degrees: 97.36% users have an in-degree smaller than 5 and 81.27% users have an out-degree smaller than 20. These numbers are much smaller compared to the Twitter follow graph [27] and the Facebook friendship graph [28]. As discussed earlier, it is potentially due to the fact that following is not a pivot function in BiliBili.

5.1.3. Connected components

For a directed graph like the follow graph we proposed, a Weakly Connected Component (WCC) contains nodes that are connected when ignoring the edge direction, whereas in a Strongly Connected Component (SCC), a pair of nodes must be reachable through a directed path. For both the WCCs and the SCCs in the BiliBili follow graph, there is a single dominant component that is much larger than the other components in size. Nevertheless, as stated earlier, the largest SCC contains only 6.09% nodes (for the largest WCC, it is 99.50%). This number is much smaller compared to that of the Twitter follow graph [27] and that of the Facebook friendship graph [28]. This is possibly due to the abundance of unreciprocated edges in the BiliBili follow graph.

5.1.4. Clustering coefficient versus degree

The clustering coefficient measures the fraction of users whose neighbors are also connected. As unreciprocated following is not a good indication of social relationships, here we consider only users who are connected with reciprocated edges. Fig. 13 shows the average clustering coefficient for these users, grouped by their degrees. For comparison, we have also considered the Facebook graph [28], the Twitter graph [27], and the MSN messenger graph [30]. It has been shown that in these graphs, for users with a degree of 5, the average clustering coefficient is around 0.4, 0.23, and 0.15, respectively; for users with a degree of 20, the average clustering coefficient is around 0.3, 0.19, and 0.1, respectively, and for users with a degree of 100, the average clustering coefficient is around 0.14, 0.14, and 0.05, respectively. We see that the clustering coefficient in the Facebook graph is larger than that of the Twitter graph, which is in turn larger than that of the MSN messenger graph, intuitively suggesting diminishing social indications in these communities. The BiliBili follow graph exhibits a clustering coefficient pattern similar to that of the MSN messenger graph.

5.2. Comment graph

The follow graph models the explicit relationships of the users. Here, we propose another graph that captures more subtle and implicit user relationships based on their co-commenting activities, which we name the *comment graph*. In this model, a node represents a user and an edge between two nodes represents that the two users have previously co-commented on the same videos. Here, we only consider videos with fewer than m commenters. The threshold m is used to exclude videos with a large number of commenters. We conjecture that users co-commenting on these videos is due to the video popularity rather than individual common interests, and thus it provides little social implications. In our model, we choose $m = 50$ and the resulting model covers 81.77% of the entire video repository.

Unlike the follow graph, the comment graph are weighted, with the edge weight equal to the number of co-commented videos between the two users. The comment graph contains in total 3,694,739 nodes and 259,821,515 edges, with 96.5% edges of a weight equal to 1, indicating that most users have only co-commented on the same video once. To remove the occasion of coincidence and to control the user interaction level that considered in our model, we further propose an unweighted version of the comment graph, that is, we introduce a threshold t and only edges with weight $\geq$ t are considered. We have tested different values of t and the properties of the resulting graphs are summarized in Table 6.

Consistent with our intuition, for $t = 1$, the graph is very dense since edges are added between any two users that have commented on the same video. In other words, commenters of the same video immediately form a complete graph, i.e., for a video with n commenters, $n(n-1)/2$ edges are established. Increasing t will dramatically reduce the scale of the graph, in terms of a smaller number of nodes and edges, a smaller average node degree, and a larger diameter. The clustering coefficient, however, only decreases when t increases from 1 to 2, and stays stable at around 0.4 (with minor fluctuations) afterwards. Clearly, $t = 2$ is a pivot point for our model—the resulting graph covers relatively a large fraction of the user population while still manages to keep a local structure as tightly clustered as those with higher constraints on the user interaction level (i.e., the cases of $t > 2$). In the following analysis, we only focus on the comment graph with $t = 2$.

Graph properties. Similar to the follow graph, we observe highly-skewed node degrees in the comment graph, indicating that a small number of users are extremely active in commenting and they have repeatedly co-commenting on the same videos with

hundreds of other users. On the other hand, most users have a moderate degree, with a median value of 4 and the third quartile of 13.

The comment graph is well connected compared to the follow graph, with the LCC containing in total 96.88% of all the users (for the follow graph the SCC is 6.09%), and its clustering coefficients are larger than those in the follow graph, as well as the Facebook graph, the Twitter graph, and the MSN messenger graph in reference (as used in Section 5.1.4). The possible reason is that the comment graph captures user interactions in a group, while the rest either consider only bilateral interaction (the MSN messenger graph), or focus merely on user relationships (the follow graph, the Facebook graph, and the Twitter graph).

6. Predicting video popularity

Having gained several valuable insights on the characteristics of BiliBili and BiliBili users, we are now in a position to apply these findings to the popularity prediction problem. As stated earlier, our work complements previous analyses with two new aspects, namely differentiating the implicit and the explicit popularity, and leveraging the new social features proposed in BiliBili for the popularity prediction problem.

6.1. Prediction tasks

Different from previous studies, we are not interested in predicting the exact value of the popularity of each individual video, rather we intend to build machine-learned classifiers to predict videos of top popularities. Here, we consider both the implicit and the explicit popularity, measured by the number of views and the number of coins, respectively.

Several previous work has focused on developing feature-based methods for predicting the number of views of a video [8,9,11,12]. We complement previous studies with a further prediction on user donations and we include new social features adopted in BiliBili. Recently, generative models such as the Hawkes Point Processes are also used to predict the popularity of online contents. However, they rely on the cascade information such as when and who re-shared the video [13,14] or re-tweeted the tweet [31–33]. Such information is rarely available for the whole video repository and therefore the generative models are more suitable for modelling the popularity of a small subset of videos that get promoted in external social media.

Our analysis is useful in various circumstances. For the implicit popularity, once the top videos (the ones with many viewers) are correctly identified, resource provision methods can be applied to guarantee smooth viewing experiences, for example, through allocating more servers to them. For the explicit popularity, the top videos are those that attract many virtual money donations. They are the ones that users really appreciate, and with high probability will be the targets of the social media marketing. Their success can be used to instruct other users who intend to attract fans and donations through video sharing—it certainly helps to maintain the community prosperity and sometimes will assist the users to gain a sizeable income [34,35].

More specifically, our prediction tasks are that, with the information obtained immediately or shortly (one day) after a video is uploaded, can we predict:

Task 1: will a video be one of the top videos in the number of views it attracts?

Task 2: will a video be one of the top videos in the number of coins it collects?

Table 7
Classification features.

Feature group	Description
video characteristics (v)	uploaded/shared original/copy duration
uploader attributes (u)	age gender number of uploads total number of views of all the uploads
commenter attributes (c)	number of commenters number of commented videos and number of danmus previously made by the commenters
follow graph properties (G)	degree (in and out) clustering coefficient and PageRank score of the uploader in analysis degree
comment graph properties (g)	clustering coefficient and PageRank score of all commenters
viewing activities (a) (within the first day)	number of views (inside and outside BiliBili) number of coins number of favorites number of comments

6.1.1. Features

Based on previous analysis, we extract six groups of features including the video characteristics (v), the uploader attributes (u), the commenter attributes (c), the follow graph properties (G), the comment graph properties (g), and the one-day viewing activities (a). All these features have been extensively studied in Sections 3–5, and are summarized in Table 7.

6.1.2. Methodology

For our analysis, we keep a record of all the 2854 videos that are uploaded on 24 May, 2016. We follow them for a period of 82 days. At the end of our observation, 868 (30.41%) videos have attracted more than 1500 views. We label these videos as the positive examples and the rest as the negative examples in Task 1. Meanwhile, 845 (29.61%) videos have collected more than 10 coins. We label these videos as the positive examples and the rest as the negative examples in Task 2.

For the feature groups as proposed above, the uploader attributes (u) and the follow graph properties (G) are calculated based on the past activities of the uploaders, i.e., before 24 May, 2016. Together with the video characteristics (v), models based on these feature groups are trained and predictions can be made immediately after the videos are uploaded. On the other hand, the commenter attributes (c), the comment graph properties (g), and the one-day viewing activities (a) are extracted after a one-day observation. Predictions can therefore be made one day after the videos are uploaded.

6.1.3. Classification algorithm

We experimented with a variety of classification algorithms—logistic regression, support vector machines, and random forests—and found the latter to work best. Hence all results reported here were obtained using random forests [36]. On average, the time complexity and the space complexity for training a random forest model is $O(mdn\log n)$, where m is the number of trees, d is the number of features, and n is the number of samples, and the time complexity and the space complexity for computing predictions of an object is $O(m\log n)$. For very large datasets, sampling and parallelization techniques can be leveraged to reduce the training time, which is beyond the scope of this article and we refer the interested readers to [37–39] for the details.

For each experiment, we run 5-fold cross-validation and report the area under the receiver operating characteristic (ROC) curve

Table 8
Classification results (AUC).

Features	Views	Coins
v	59.29%	61.59%
$v+u$	83.29%	78.40%
$v+G$	75.94%	77.98%
$v+u+G$	84.46%	83.77%
$v+c$	79.22%	80.78%
$v+g$	77.77%	78.93%
$v+c+g$	79.12%	80.57%
$v+u+c$	86.83%	85.27%
$v+u+c+G+g$	87.03%	86.78%
$v+u+c+a$	91.31%	92.79%
$v+u+c+G+g+a$	91.47%	92.73%

(AUC). We use balanced training and test sets containing equal numbers of positive and negative examples, so random guessing results in an AUC of 50%.

6.2. Results

The classification results are shown in Table 8. In order to understand which features are important for the prediction, we have progressively increased the group of features used in the classifier, which can be retrieved gradually during a video's lifetime. We find that the two tasks achieve similar performances. In the following discussions, we will take Task 1 as the example. We have a number of interesting findings as follows.

Firstly, although the classifier that uses only the video characteristics achieves an AUC of 59.29% (only slightly better than random guessing), adding any feature group dramatically increases the prediction accuracy: the minimum improvement in AUC is about 15% (from 59.29% to 75.94%, achieved by adding the follow graph properties).

Secondly, the follow graph and the comment graph we proposed provide valuable information for the prediction. Adding the follow graph properties or the comment graph properties both significantly improves the prediction accuracy.

Thirdly, by using only the *start-up* features v, u and G, i.e., features that can be obtained immediately after the video is uploaded, the classifier can already achieve a relatively high AUC of 84.46%. Including more features from user's follow and comment activities and the graphs can further increase the AUC to 87.03%.

Finally, the AUC reaches 91.47% after including all feature groups, even the most demanding one (i.e., viewing activities a) that requires a one-day observation of the video and summarizes user's actions towards it (e.g., the number of coins and the number of comments)—in some sense, the AUC of 91.47% serves as a "gold-standard" that represents the best possible performance that we can achieve.

These results indicate that the insights we gained from previous analysis indeed provide valuable reference for the popularity prediction problem.

7. Related work

We summarize related work within each research topic our work covers as follows.

7.1. Characterizing UGC sites

Traditional UGC sites like YouTube have been extensively studied before. Cheng et al. investigated the video properties of YouTube. They found that YouTube videos have noticeably different statistics compared to traditional streaming videos and the related video network forms a small world network [17]. Cha et al. provided a complementary global view by crawling data of complete sets of video categories, and they presented a comprehensive analysis of the popularity distribution and the time evolution of UGC video requests and their implications [2,3]. Figueiredo et al. analyzed how video popularity in YouTube evolves since upload and how the referrers lead users to videos [7]. Ding et al. analyzed in-depth the behaviors of YouTube uploaders [16]. Gill et al. and Zink *et al* et al.investigated YouTube from the perspective of YouTube traffic [40,41]. They examined YouTube usage patterns, file properties, and transfer characteristics. Vasilakos et al. proposed a video caching mechanism based on the mobility prediction to optimize the mobile traffic usage [42]. We complement previous work with a novel dataset that contains the whole repository (at the time of the crawling) rather than a sample of the network.

7.2. New generations of UGC sites

Recently, a number of studies are dedicated to new generations of UGC sites with enhanced social features and sometimes exclusive for a particular topic, such as Twitch.tv for gaming videos. Kaytoue et al. [4] and Pires and Simon [5] provided preliminary characterizations on Twitch. They analyzed the dynamics of game spectators and proposed models for predicting video popularity. Deng et al. studied the workload of the Twitch streams and found that it is affected significantly by game tournaments [43]. Our previous work compared Twitch with a game replay downloading site, and investigated their repositories and user activities [6]. Twitch adopts a chat replay feature that is similar to the one used in BiliBili. However, we did not find any previous work on this topic.

7.3. Social networks in UGC sites

Recent surge of Online Social Network popularity has attracted the attention of researchers from a variety of fields, including UGC sites. Mislove et al. investigated several on-line social networks, including YouTube [44]. Their study mainly focuses on the properties of social graphs, e.g., the power-law, the small-world, and the scale-free properties. Benevenuto et al. investigated the user behavior in a social network created by interactions based on video responses in YouTube [45]. Chen et al. examined mobile social networks and proposed efficient multicast methods based on community and social features [46]. Closest to our work, Wattenhofer et al. analyzed the correlations between the popularity of YouTube videos and the properties of various social graphs created among the users. In particular, they found that characteristics of the graph built from links between YouTube users who comment each other's videos are more correlated to the popularity of a users video than to the characteristics of the subscription graph (though such correlation is strong) [47]. In our work, we further apply these findings to the classification tasks of identifying popular videos.

7.4. Predicting content popularity in UGC sites

Many efforts have been made to uncover the popularity temporal patterns. Most of these studies are based on early views records in predicting near-future popularity. Crane and Sornette proposed epidemic models to explain a burst in video popularity in terms of endogenous user interactions and external events [23]. Yang and Leskovec proposed a time series clustering algorithm to identify popularity trends [24]. A unifying analytical framework of the trends extracted by those studies was proposed in [48].

Recently, a large body of literature has utilized feature-based methods and generative models to predict the popularity of online contents, ranging from retweets in Twitter [31–33,49,50], shares of photos in Facebook [51], high impact academic papers [52], to videos with or without social media promotions [8–14,53]. Similar to our feature-based prediction on the viewcount, Pinto et al.

showed that Youtube video views were predictable from early view patterns [8]. Li et al. further included propagation network features. Yu et al., Vallet et al. and Roy et al. analyzed Youtube videos promoted in Twitter and extracted features relating the two social networks [10–12]. With the fine-grained cascade information provided by social media like Twitter, for example when and who re-tweeted the video, generative models such as the Hawkes Point Processes were also introduced and were shown to be effective for predicting the popularity of videos [13,14] and of tweets [31–33].

Though features from external social media are informative for both feature-based methods and generative models, such information is only available for a small subset of promoted videos and only after they are shared. For our analysis, we leverage features extracted from the built-in social network and the predictions can be carried out on the whole repository. And we can successfully identify popular videos immediately after they are uploaded.

Particularly, this article significantly extends and complements a previous conference publication [54] with (1) a finer-grained characterization on the video repository and the user activities, through differentiating the implicit and the explicit popularity, and adding statistics on the user gender and the video type; (2) a more comprehensive study on the graph properties and a new analysis of the impact of the thresholds on the characteristics of the resulting graphs; and (3) a new prediction task for identifying videos that users really appreciate (reflected by virtual money donations).

8. Conclusion and future work

In this article, we conducted an analysis on a UGC site with enhanced social features named BiliBili. We presented the first publicly accessible dataset containing the whole repository of a UGC site. Based on statistics for more than 2 million videos and more than 28 million users, we investigated the video repository and the user activities of BiliBili, we analyzed the implicit and the explicit popularity, and we applied our findings to build machine-learned classifiers that accurately identify popular videos.

Among our results, we find that BiliBili exhibits certain characteristics that are often observed in UGC sites, for example, the short video durations and the highly skewed video popularity. In addition, we find a number of fascinating distinctions in BiliBili. First, compared to YouTube, BiliBili is a much smaller but more active community: on average BiliBili videos are viewed twice more than YouTube Videos. Secondly, while users prefer to view videos shared from other sites, they are more generous to donate to videos uploaded locally from the users, probably in a gesture of showing their support for the efforts of other users. Thirdly, the uploaders not only get more followers, they are also more active in following others, and uploaders with followers are more active in uploading. We conjecture that the enhanced social features in BiliBili provide valuable opportunities for users to interact and hence boost the community engagement. We leave a further analysis on this topic as our future work.

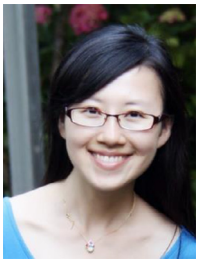
Acknowledgment

This work was partially supported by the National Science Foundation for Young Scholars of China (NSFYSC) nos. 61502500 and 61602500.

References

[1] YouTube, YouTube statistics, 2017,. www.youtube.com/yt/press/statistics.html.
[2] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, S. Moon, Analyzing the video popularity characteristics of large-scale user generated content systems, IEEE/ACM Trans. Netw. 17 (5) (2009) 1357–1370.
[3] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, S. Moon, I Tube, YouTube, everybody tubes: analyzing the world's largest user generated content video system, Internet Measurement Conference (IMC'07), 2007.
[4] M. Kaytoue, A. Silva, C. Raissi, Watch me playing, I am a professional: a first study on video game live streaming, in: Proceedings of the 12th International World Wide Web Conference (WWW'12 Companion), 2012.
[5] K. Pires, G. Simon, YouTube live and Twitch: a tour of user-generated live streaming systems, in: Multimedia Systems Conference (MMSys'15), 2015.
[6] A.L. Jia, S. Shen, D. Epema, A. Iosup, When game becomes life: the creators and spectators of online game replays and live streaming, ACM Trans. Multim. Comput. Commun. Appl. 12 (4) (2016).
[7] F. Figueiredo, J. Almeida, M. Goncalves, F. Benevenuto, On the dynamics of social media popularity: a YouTube case study, ACM Trans. Internet Technol. 14 (24) (2014).
[8] H. Pinto, J. Almeida, M. Goncalves, Using early view patterns to predict the popularity of YouTube videos, in: Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM'13), 2013.
[9] H. Li, X. Ma, F. Wang, J. Liu, K. Xu, On popularity prediction of videos shared in online social networks, Proceeding of the 22th International Conference on Information and knowledge management (CIKM'13), 2013.
[10] S. Roy, T. Mei, W. Zeng, S. Li, Towards cross-domain learning for social video popularity prediction, IEEE Trans. Multim. 15 (2013) 1255–1267.
[11] H. Yu, L. Xie, S. Sanner, Twitter-driven YouTube views: beyond individual influencer, in: Proceedings of the 2014 ACM on Multimedia Conference (MM'14), 2014.
[12] D. Vallet, S. Berkovsky, S. Ardon, A. Mahanti, M.A. Kafaar, Characterizing and predicting viral-and-popular video content, Proceeding of the 24th International Conference on Information and Knowledge Management (CIKM'15), 2015.
[13] W. Ding, Y. Shang, L. Guo, R.Y. X. Hu, T. He, Video popularity prediction by sentiment propagation via implicit network, Proceeding of the 24th International Conference on Information and Knowledge Management (CIKM'15), 2015.
[14] M. Rizoiu, L. Xie, S.Sanner, Expecting to hip: Hawkes intensity processes for social media popularity, in: Proceeding of the 26th International World Wide Web Conference (WWW'17), 2017.
[15] BiliBili, 2017, www.bilibili.com.
[16] Y. Ding, Y. Du, Y. Hu, Z. Liu, L. Wang, K. Ross, A. Ghose, Broadcast yourself: understanding YouTube uploaders, in: Proceedings of the 5th Internet Measurement Conference (IMC'11), 2011.
[17] X. Cheng, B. Burnaby, C. Dale, J. Liu, Statistics and social network of YouTube videos, Workshop on Quality of Service (WQoS'08), 2008.
[18] F. Figueiredo, F. Benevenuto, J. Almeida, The tube over time: characterizing popularity growth of YouTube videos, in: Proceedings of the 4th ACM international conference on Web search and data mining (WSDM'11), 2011.
[19] L. Becchetti, C. Castillo, D. Donato, A. Fazzone, A comparison of sampling techniques for web graph characterization, Proceeding of the Workshop on Link Analysis (LinkKDD'06), 2006.
[20] L. Katzir, E. Liberty, O. Somekh, Estimating sizes of social networks via biased sampling, in: Proceeding of the 18th International World Wide Web Conference (WWW'11), 2011.
[21] S.H. Lee, P.-J. Kim, H. Jeong, Statistical properties of sampled networks, Phys. Rev. E 73 (2006).
[22] Twitch, Twitch, 2017, www.twitch.tv/.
[23] R. Crane, D. Sornette, Robust dynamic classes revealed by measuring the response function of a social system, Proc. Natl. Acad. Sci. 105 (2008) 15649–15653.
[24] J. Yang, J. Leskovec, Patterns of temporal variation in online media, in: Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11), 2011.
[25] J. Frank, J. Massey, The Kolmogorov-Smirnov test for goodness of fit, J. Am. Stat. Assoc. 46 (253) (1951) 68–78.
[26] T. Anderson, D. Darling, A test of goodness-of-fit, J. Am. Stat. Assoc. 49 (268) (1954) 765–769.
[27] S.A. Myers, A. Sharma, P. Gupta, J. Lin, Information network or social network? The structure of the Twitter follow graph, in: Proceeding of the 21th International World Wide Web Conference (WWW'14 Companion), 2014.
[28] J. Ugander, B. Karrer, L. Backstrom, C. Marlow, The anatomy of the Facebook social graph, arXiv:1111.4503 (2011).
[29] R. Dunbar, Neocortex size as a constraint on group size in primates, J. Hum. Evol. 22 (6) (1992) 469–493.
[30] J. Leskovec, E. Horvitz, Planetary-scale view on a large instant-messaging network, in: Proceeding of the 18th International World Wide Web Conference (WWW'05), 2008.
[31] P. Bao, H.-W. Shen, X. Jin, X.-Q. Cheng, Modeling and predicting popularity dynamics of microblogs using self-excited Hawkes processes, in: Proceeding of the 24th International World Wide Web Conference (WWW'15), 2015.
[32] Q. Zhao, M.A. Erdogdu, H.Y. He, A. Rajaraman, J. Leskovec, SEISMIC: a f-exciting point process model for predicting tweet popularity, Proceeding of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15), 2015.
[33] S. Mishra, M.-A. Rizoiu, L. Xie, Feature driven and point process approaches for popularity prediction, Proceeding of the 25th International Conference on Information and Knowledge Management (CIKM'16), 2016.
[34] W. Hamilton, O. Garretson, A. Kerne, Streaming on Twitch: fostering participatory communities of play within live mixed media, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'14), 2014.
[35] C. Kang, 2014, He wants to make it playing video games on Twitch. But will people pay to watch? http://www.washingtonpost.com/.
[36] L. Breiman, Random forests, Mach. Learn. 1 (2001) 5–32.

[37] N. Chawla, Data mining for imbalanced datasets: an overview, Data Mining and Knowledge Discovery Handbook, 2010.
[38] L. Mu, D. Andersen, J. Park, A. Smola, A. Ahmed, V. Josifovski, J. Long, E. Shekita, B. Su, Scaling distributed machine learning with the parameter server., in: Proceedings of the 11th USENIX conference on Operating Systems Design and Implementation (OSDI'14), 2014.
[39] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), 2016.
[40] P. Gill, M. Arlitt, Z. Li, A. Manhanti, YouTube traffic characterization: a view from the edge, in: Proceedings of the first Internet Measurement Conference (IMC'07), 2007.
[41] M. Zink, K. Suh, Y. Gu, J. Kurose, Characteristics of YouTube network traffic at a campus network: measurements, models, and implications, Comput. Netw. 53 (2009) 501–514.
[42] X. Vasilakos, V. Siris, G. Polyzos, Addressing niche demand based on joint mobility prediction and content popularity caching, Comput. Netw. 110 (2016) 306–323.
[43] J. Deng, F. Cuadrado, G. Tyson, S. Uhlig, Behind the game: exploring the twitch streaming platform, Network and System Support for Games (NetGames'15), 2015.
[44] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, B. Bhattacharjee, Measurement and analysis of online social networks, in: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement Conference (IMC'07), 2007.
[45] F. Benevenuto, F. Duarte, T. Rodrigues, V. Almeida, J. Almeida, K. Ross, Understanding video interactions in YouTube, in: Proceedings of the 16th ACM International Conference on Multimedia (MM'08), 2008.
[46] X. Chen, C. Shang, B. Wong, W. Li, S. Oh, Efficient multicast algorithms in opportunistic mobile social networks using community and social features, Comput. Netw. 111 (2016) 71–81.
[47] M. Wattenhofer, R. Wattenhofer, Z. Zhu, The YouTube social network, in: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM'12), 2012.
[48] Y. Sakurai, B. Prakash, L. Li, C. Faloutsos, Rise and fall patterns of information diffusion: model and implications, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12), 2012.
[49] T. Martin, J.M. Hofman, A. Sharma, A. Anderson, D.J. Watts, Exploring limits to prediction in complex social systems, in: Proceeding of the 25th International World Wide Web Conference (WWW'16), 2016.
[50] P. Bao, H. Shen, J. Huang, X. Cheng, Popularity prediction in microblogging network: a case study on sina weibo, in: Proceeding of the 22th International World Wide Web Conference (WWW'13 Companion), 2013.
[51] J. Cheng, L. Adamic, P.A. Dow, J.M. Kleinberg, J. Leskovec, Can cascades be predicted? in: Proceeding of the 23th International World Wide Web Conference (WWW'14), 2014.
[52] F. Davletov, A.S. Aydin, A. Cakmak, High impact academic paper prediction using temporal and topological features, Proceeding of the 23th International Conference on Information and Knowledge Management (CIKM'14), 2014.
[53] F. Fraile, J. Guerri, Simple models of the content duration and the popularity of television content, Comput. Netw. 40 (2014) 12–20.
[54] A.L. Jia, S. Shen, S. Chen, D. Li, A. Iosup, An analysis on a YouTube-like UGC site with enhanced social features, in: Proceedings of the 26th International World Wide Web Conference (WWW'17 Companion), 2017.



Adele Lu Jia received her B.S. degree from Harbin Institute of Technology, China, in 2007, her M.Phil. degree from The Chinese University of Hong Kong in 2009, and her Ph.D. degree from Delft University of Technology, the Netherlands, in 2013. She is currently an associate professor in the Computer Science Department at China Agricultural University. Her research interests include complex network analysis and data mining.

Siqi Shen received his B.S. and M.S. degree from National University of Defense Technology, China, in 2007 and 2009, respectively, and his Ph.D. degree from Delft University of Technology, the Netherlands, in 2015. He is currently an assistant professor at National Lab for Parallel and Distributed Processing, National University of Defense Technology, China. His research interests include complex network analysis, data mining and machine learning.

Dongsheng Li received the B.Sc. degree (with honors) and Ph.D. degree (with honors) in computer science from College of Computer Science, National University of Defense Technology, Changsha, China, in 1999 and 2005, respectively. He was awarded the prize of National Excellent Doctoral Dissertation of PR China by Ministry of Education of China in 2008. He is now a full Professor at National Lab for Parallel and Distributed Processing, National University of Defense Technology, China. His research interests include distributed computing, Cloud computing, computer network and large-scale data management.

Shengling Chen received his B.S. degree in Computer Science and Technology from North China University of Technology, China, in 2014. He is currently pursuing his Master degree at the School of Computer, National University of Defense Technology, China. His research interests include big data analysis and machine learning.