An Analysis on a YouTube-like UGC site with Enhanced Social Features

Adele Lu Jia
Information and Electrical Engineering Department
China Agricultural University
ljia@cau.edu.cn

Siqi Shen
School of Computer
National University of Defense Technology, China
shensiqi@nudt.edu.cn

Shengling Chen
School of Computer
National University of Defense Technology, China
waitsl@126.com

Dongsheng Li
School of Computer
National University of Defense Technology, China
dsli@nudt.edu.cn

Alexandru Iosup
Distributed Systems Group
Delft University of Technology, the Netherlands
a.iosup@tudelft.nl

ABSTRACT

YouTube-like User Generated Content (UGC) sites are nowadays entertaining over a billion people. Resource provision is essential for these giant UGC sites as they allow users to request videos from a potentially unlimited selection in an asynchronous fashion. Still, the UGC sites are seeking to create new viewing patterns and social interactions that would engage and attract more users and complicate the already rigorous resource provision problem. In this paper, we seek to combine these two tasks by leveraging social features to provide the reference for resource provision.

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 capture the complete view of BiliBili—containing over 2 million videos and over 28 million users—we characterize its video repository and user activities, we demonstrate the positive reinforcement between on-line social behavior and upload behavior, we propose graph models that reveal user relationships and high-level social structures, and we successfully apply our findings to build machine-learnt classifiers to identify videos that will need priority in resource provision.

Keywords

User Generated Content sites; social features; graph model; prediction

1. INTRODUCTION

YouTube-like User Generated Content (UGC) sites are nowadays a major Internet phenomenon that entertain over a billion users–almost one-third of all people on the internet—and form a billion dollar global industry [3]. Given the scale, dynamics, and decentralization of the contents provided by individual users and the gradually shifting user interests, two fundamental questions for maintaining and growing such UGC sites are that, in the system level, how to perform resource provision so as to achieve smooth user viewing experiences, and that, in the application level, how to design enhanced features that will benefit user retention and attraction. In this paper, we seek to combine these two tasks by leveraging social features to provide the reference for the resource provision problem.

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 [1], 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 enhanced social features, including non-reciprocal user following, chat replay, and virtual money donation, which provide valuable information for the resource provision problem.

Secondly, unlike previous studies that are often carried on sampled datasets [8, 9, 10], we were able to capture the complete view of BiliBili with over 2 million videos and over 28 million users. Moreover, the information we obtained includes not only the static repository characteristics such as the video duration and the user gender, but also dynamic user activities, for example, how users view and donate to the videos, when and who left what comments, and how users follow each other. This complete view and fine-grained information provide a ground truth for the system in analysis and therefore avoid potential defects, e.g., under- and over- estimations of certain network properties, caused by sampling and sampling biases [5, 14, 16].

Our analysis of BiliBili mainly consists of three parts. First, we reveal, quantitively, the scale and characteristics of BiliBili by examining its video repository and user ac-

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3053901

tivities, which allows us to derive a number of qualitatively interesting findings including the positive reinforcement between on-line social behavior and upload behavior. Then, we propose two graph models based on the social features provided by BiliBili. These graph models capture both the direct user relationships and the higher-order social structures, by looking not only at who a user is connected to, but also how those connected users are linked amongst each other. Finally, applying our findings, we develop machine-learned classifiers that can successfully identify videos that will gain great popularity and therefore need priority in the resource allocation process. In this way, by leveraging the social features in BiliBili, we provide valuable reference for the resource provision problem.

We summarize our contributions as follows:

- We collect, use, and offer public access[1] to the dataset that contains the complete view of BiliBili, 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 3.2), (iii) the upload activity, (iv) the follow activity, and (v) the comment activity of the users (Section 3.3).
- We propose two graph models 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 4).
- We build machine-learned classifiers to predict with high accuracies the videos that will need priority in resource provision (Section 5).

2. BILIBILI DATASET

In this section, we first give a brief introduction on BiliBili. Then, we introduce the dataset used throughout this article.

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 [2]. 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 useful in various circumstances, including upgrading user membership and exchanging for new emojis.

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

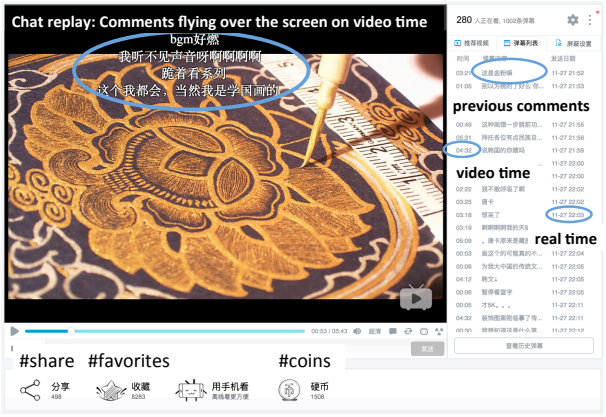

Figure 1: An example of a BiliBili video page with chat replays.

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) *commentors*, 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 Dataset

The primary challenge in crawling large online communities is covering, if not the entire repository, the giant connected component consisted by related contents or 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 [5, 16].

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 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 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 uers, 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), 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 follow-

Table 1: BiliBili scale

	BiliBili	**YouTube** [9]
#videos	2,858,844	448 million
aggregate video length	122 years	2,649 years
aggregate viewing time	2.9 million years	9.9 million years
aggregate #views	23 billion	1.5 trillion
mean #views	**8,220**	**3,348**
#registered users	28,962,041	NA
#uploaders	417,834	47.3 million
#commentors	6,677,458	NA
#social users	10,749,720	NA

Table 2: Video statistics

	min	**median**	**mean**	**max**
dur (min)	0.02	5.93	22.45	9,915
#views	1	849	7,643	10,290,411
#favorites	0	8	124.7	340,547
#coins	0	2	35.5	362,660
#danmus	0	16	250.7	1,109,635
#commentors	0	8	52.2	4,963

ers/followees, and the repository of its uploaded videos (can be cross-referenced with the video dataset). The basic statistics of our datasets are introduced in Table 1.

3. BILIBILI CHARACTERISTICS

In this section, we first introduce the scale of BiliBili. Then we provide a characterization on the video repository and the user activities.

3.1 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 [9]. Compared to YouTube, BiliBili has a smaller scale in terms of the number of videos and uploaders, possibly due to the fact that BiliBili is mainly targeted at Chinese users. Nevertheless, it achieves a higher level of user activity: on average, BiliBili videos are viewed twice more compared to YouTube videos. We conjecture that *the enhanced social features in BiliBili provide valuable opportunities for users to interact and hence boost the community engagement.*

3.2 Video characteristics

3.2.1 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 [10, 13].

3.2.2 Video duration

Video duration directly measures the scale of the repository in terms of the workload of its contents. As shown in Table 2, most videos in BiliBili are of short durations: over half of the videos are within 10 minutes, and over 95% videos are within 20 minutes. This result is consistent with YouTube [10] but not with Twitch [13, 15], possibly because

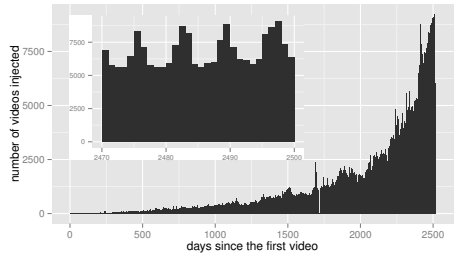

Figure 2: Video injection rate

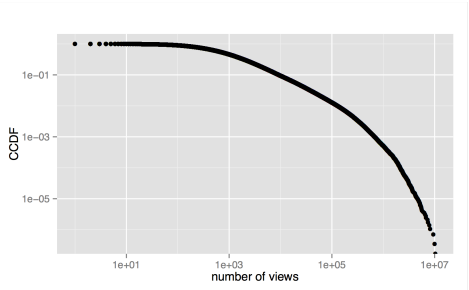

Figure 3: CCDF of the number of views.

YouTube and BiliBili cover a wide range of topics whereas Twitch is exclusive for gaming videos.

3.2.3 Video popularity

In any UGC sharing site, content popularity provides important knowledge for the activity level of users and the potential workload for maintaining the site. Here, we measure the video popularity in terms of user's explicit and implicit interests in the videos. Users express their interests explicitly through favoring, commenting, and donating to the videos, and implicitly through viewing the videos. As shown in Table 2, for any definition, the video popularity is highly skewed, with a small number of videos attracting a large number of views, favorites, coins, danmus, and commentors.

We further show in Fig. 3 the Complementary Cumulative Distribution Function (CCDF) of the number of views collected by each video. When it is plotted on a log-log scale, we observe a curve rather than a straight line, indicating that the video popularity in BiliBili is better fitted with an exponential distribution, rather than a power-law distribution that is observed in many other UGC sites such as YouTube and Twitch [10, 15]. These results show that the skewness of the video popularity in BiliBili is not as severe as that in YouTube and Twitch. We conjecture that *the social features in BiliBili boost community engagement and hence reduce the disparity.*

Influence of the video type. In BiliBili, videos are either directly uploaded by users or shared from another site. In Fig. 4, we show the influence of the video type on the number of views they accumulated. We see that videos shared from other sites in general attract a larger number of views: 50% uploaded videos attract fewer than 300 views whereas 50% shared videos attract more than 1,000 views.

Table 3: Statistics on user's activity

activity		min	1^{st} quartile	median	mean	3^{rd} quartile	max
upload	no. uploaded videos	1	1	2	7.64	5	22,324
	no. views	0	515	2,067	59,390	9,375	1.570e+09
	avg. no. views per video	0	278	842	4,426	2,528	3.387e+06
follow	no. followers	0	0	0	14	0	1,420,203
	no. followees	0	1	4	13.86	14	11,513
	follower/followee ratio	0.0010	0.0417	0.1429	19.9570	0.7273	619,524.00
comment	no. commented videos	1	2	6	22.35	20	65,450
	no. danmus	1	3	10	52.7	37	640,863
	avg. no. danmus per videos	1.00	1.00	1.43	1.96	2	9,153.00

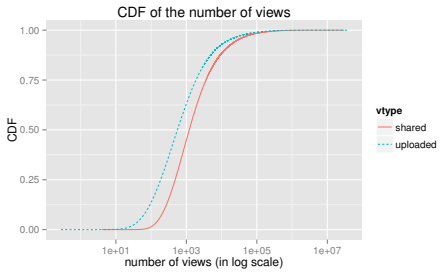

Figure 4: CDF of the number of views

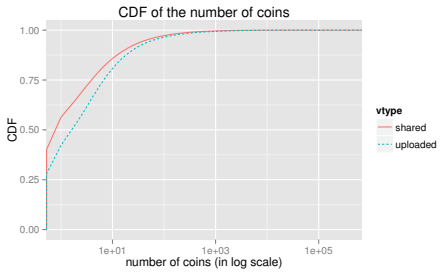

Figure 5: CDF of the number of coins

Interestingly, though it seems that *users prefer shared videos in their viewing activities, they are more generous in donating virtual money to the uploaded videos.* As shown in Fig. 5, while 35% shared videos do not receive any coins, 20% uploaded videos receive more than 10 coins. We conjecture that most uploaded videos are genuine and originally produced by the uploaders. Interesting or not (content-wise), users are more willing to donate to them, probably in a gesture of showing their support for the user generated contents.

3.3 User activities

3.3.1 *Upload activity*

In BiliBili, only 1.44% of all the 28 millions users have ever uploaded a video (named uploaders), indicating that most users are silent viewers. The uploaders on average have uploaded 7.64 videos and have collected 59,390 views, resulting an average of 4,426 views per uploaded video.

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. Similarly, the total and average number of views collected by all the videos shared by an uploader are also highly skewed.

3.3.2 *Follow activity*

In total, 10,749,720 (37.12%) users have followed or have been followed by at least one user (named social users). On average, each social user follows and is followed by 5 users. To the extreme, a user has followed 11,513 users and another user has been 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. For users with at least one follower and one followee, we find that over 75% of these users follow more users than they have followers, indicating that *most users in BiliBili are prone to consume rather than to provide information.*

Reinforcement between follow and upload activity. Unlike Twitter, the follow feature in BiliBili is mainly used for updating videos instead of news/tweets. As a consequence, most of the follows are targeted at uploaders rather than viewers (users who have not uploaded any videos). As shown in Fig. 6, 75% uploaders have obtained at least one follower while 98% viewers have no followers at all. Interestingly, we also find that uploaders are more active in following others. As shown in Fig. 7, 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), achieving a Spearman Ranking Correlation Coefficient (SRCC)[2] of 0.6418 between the number of followers and the number of uploads. These results indicate *a positive reinforcement between the social relationships and the upload activity.*

3.3.3 *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 commentor). On average, each commentor has commented on 22 videos and has left 52 danmus, resulting an average of 1.96 danmus per commented video. Similar to the previous observations, we find that user activity level in video commenting is also

[2]In brief, SRCC assesses how well the relationship between two variables can be described using a monotonic function [19].

Table 4: Graph properties. 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 Weakly Connected Component (WCC) and the Strongly Connecgted Copmponent (SCC) for directed graphs), the effective diameter, D, and the average clustering coefficient, c.

graph	n	e	l	d out/in	LCC, WCC/SCC (random)	D (random)	c (random)
follow	10,749,720	149,037,569	1.29e-16	13.86/13.90	9.50%/6.09% (100%)	3.86 (5.81)	0.0867 (0.0000)
comment	1,033,669	9,105,183	1.70e-05	17.00	96.90% (100%)	4.92 (5.58)	0.4016 (0.0000)

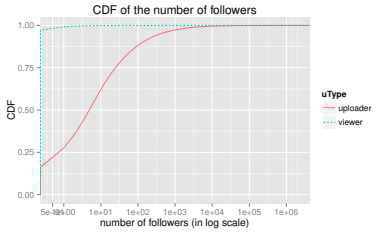

Figure 6: CDF of the number of followers

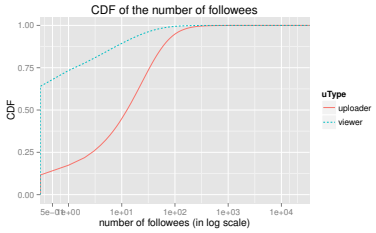

Figure 7: CDF of the number of followees

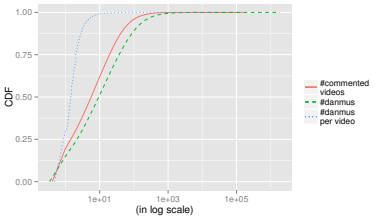

Figure 8: CDF of the number of commented videos

highly skewed: as shown in Fig. 8, while 19.59 % users have commented only on one videos, 4.08% users have commented on more than 100 videos.

4. GRAPH MODELS

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, which we will introduce in turn in the following sections.

4.1 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 and obviously to Facebook where all edges are reciprocated [17]—indicating that BiliBili is mainly an information network coupled with minor social implications. The graph properties are summarized in Table 4. 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, and reflected in the graph structure, we observe a very small Strongly Connected Components (SCC) consisted by only 6.09% of the nodes (for Twitter it is 68.7% [17]). 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 model 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 model, we conclude that the follow graph exhibit the *small-world* properties.

4.2 Comment graph

While the follow graph captures user's relationships explicitly through following, we propose another graph model that captures more subtle and implicit user relationships based on user's comment behavior, 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 co-commented on $\geq t$ videos that have $\leq m$ commentors. Here, the threshold t is used to control the user interaction level considered in the model, limiting the occasion of coincidence, that is, two users by accident co-comment on a small number of videos. The threshold m, on the other hand, is used to exclude videos with a large number of commentors—we conjecture that users co-commenting on such videos is due to the video popularity rather than individual common interests, and thus provides little social implications. In our model, we choose $t = 2$ and $m = 50$. A further analysis on the sensitivity of t and m is left for the future work.

The basic graph properties are shown in Table 4. In total, the comment graph contains 1,033,669 nodes (capturing 15.48% of the commentors and 3.57% of all users) and 9,105,183 edges, resulting an average degree of 17, meaning that on average each users have repetitively co-commented on the same videos with 17 other users. Similar to the follow graph, the comment graph also exhibits a small-world phenomenon with a lage average clustering coefficient and a small diameter compared to the random model with the same number of nodes and edges.

5. PREDICTING PRIORITY VIDEOS IN RESOURCE PROVISION

Having gained several valuable insights on the characteristics of BiliBili and BiliBili users, we are now in a position to apply these findings by building machine-learned classifiers to predict videos that will gain great popularity and that therefore need priority in resource provision. More specif-

Table 5: Classification features

feature group	description
video characteristics (v)	uploaded/shared, original/copy, duration, age
uploader attributes (u)	gender, number of uploaded videos, total number of views collected by all uploaded videos of the uploader in analysis
commentor attributes (c)	number of commentors, number of commented videos and total number of comments made by the commentors prior to the video in analysis
follow graph properties (G)	degree (in and out), clustering coefficient, and PageRank score of the uploader in analysis
commentor graph properties (g)	degree, clustering coefficient, and PageRank score of all commentors of the video in analysis
viewing activities (a)	number of views (inside and outside BiliBili), number of coins, number of favorites, and number of comments collected within the first day

ically, our classification task is to predict whether a video will be one of the top 30% videos in terms of the number of views it collects.

To this end, we have followed all 2,854 videos that are uploaded on 24 May, 2016, for a period of 82 days. We find that the top 30% (868) videos have attracted more than 1,500 views at the end of this observation. We label these videos as positive examples and the rest as negative examples.

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

Features. Based on previous analysis, we extract six groups of features including the video characteristics (v), the uploader attributes (u), the commentor attributes (c), the follow graph properties (G), the comment graph properties (g), and the viewing activities (a). All these features have been extensively studied in Sections 3 and 4, and are summarized in Table 5.

Results. The classification results are shown in Table 6. 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 have a number of interesting findings as follows.

First, 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 model and the commentor graph model we proposed provide valuable information for the prediction. Adding the follow graph properties or the commentor 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 graph models can further increase the AUC to 87.03%.

Table 6: Classification results

features	AUC
v	59.29%
$v+u$	83.29%
$v+G$	75.94%
$v+u+G$	84.46%
$v+c$	79.22%
$v+g$	77.77%
$v+c+g$	79.12%
$v+u+c$	86.83%
$v+u+c+G+g$	87.03%
$v+u+c+a$	91.31%
$v+u+c+G+g+a$	91.47%

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 we can hope to achieve.

These results indicate that the insights we gained from previous analysis indeed provide valuable reference for identifying videos that need priority in resource provision. Once the priority videos are successfully identified, we can apply an arbitrary resource provision method to maintain and to improve the system performance such as the viewing quality.

6. RELATED WORK

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

Characterizing UGC sites. Traditional UGC sites like YouTube have been extensively studied before. Particularly, Cha *et al.* provide a complementary global view by crawling data of complete sets of video categories [7]. Ding *et al.* analyze in-depth the YouTube uploader behaviors [9]. Recent research trends have shiftted to new generation UGC sites such as Twitch. Kaytoue *et al.* and Pires *et al.* provide preliminary characterizations on Twitch.tv [15, 18]. Our previous work compares Twitch with a game replay downloading site [13]. Our analysis on BiliBili also contains video repository and user activity characterization, but emphasizes on the enhanced social features and the reinforcement between social behavior and upload/viewing activities.

Resource provision in UGC sites. Aparicio-Pardo *et al.* study the workload characteristics of two live streaming UGC services and propose a dynamic resource scheduler adaptively streaming videos using clouds with the goal to maximize users' quality of service (QoE) [4]. He *et al.* propose a live-streaming scheduler which jointly considers

the users' QoE and the availability and pricing of cloud resources for transcoding [12]. Similarly, Gao *et al.* propose a dynamic resource scheduler for transcoding with heterogeneous QoS criteria, which on average save about 50% of the resource consumption [11]. Different from these work, we do not study particular resource provision methods. Rather, we focus on the step before it, i.e., identifying videos that need priority in resource provision.

7. CONCLUSION AND FUTURE WORK

In this article, we conducted an analysis on a User Generated Content (UGC) site with enhanced social features named BiliBili. We presented the first publicly accessible dataset containing the complete view 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, and we applied our findings to build machine-learnt classifiers to identify videos that will need priority in resource provision.

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, though in a smaller scale compared to YouTube, 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 user generated contents. Thirdly, the uploaders not only get more followers, but they are also more active in following others. 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.

8. ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation for Young Scholars of China (NSFYSC) No. 61502500 and No. 61602500, and the Chinese Universities Scientific Fund No. 2017QC053 and No. 2017QC143.

9. REFERENCES

[1] Bilibili. https://www.bilibili.com.

[2] Twitch. http://www.twitch.tv/.

[3] Youtube statistics. https://www.youtube.com/yt/press/statistics.html.

[4] R. Aparicio-Pardo, K. Pires, A. Blanc, and G. Simon. Transcoding live adaptive video streams at a massive scale in the cloud. In *Multimedia Systems Conference (MMSys'15)*, 2015.

[5] L. Becchetti, C. Castillo, D. Donato, and A. Fazzone. A Comparison of Sampling Techniques for Web Graph Characterization. In *Proceeding of the Workshop on Link Analysis (LinkKDD'06)*, 2006.

[6] L. Breiman. Random forests. *Machine Learning*, 1:5–32, 2001.

[7] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Networking*, 17(5):1357–1370, 2009.

[8] X. Cheng, B. Burnaby, C. Dale, and J. Liu. Statistics and social network of youtube videos. In *Workshop on Quality of Service (WQoS'08)*, 2008.

[9] Y. Ding, Y. Du, Y. Hu, Z. Liu, L. Wang, K. Ross, and A. Ghose. Broadcast yourself: Understanding youtube uploaders. In *Proceedings of the 5th Internet Measurement Conference (IMC'11)*, 2011.

[10] F. Figueiredo, F. Benevenuto, and J. Almeida. The tube over time: characterizing popularity growth of youtube videos. In *Web search and data mining (WSDM'11)*, 2011.

[11] G. Gao, Y. Wen, and C. Westphal. Dynamic resource provisioning with qos guarantee for video transcoding in online video sharing service. In *Proceedings of the 2016 ACM on Multimedia Conference (MM'16)*, 2016.

[12] Q. He, J. Liu, C. Wang, and B. Li. Coping with heterogeneous video contributors and viewers in crowdsourced live streaming: A cloud-based approach. *IEEE Transactions on Multimedia*, 18:916–928, 2016.

[13] A. L. Jia, S. Shen, D. Epema, and A. Iosup. When game becomes life: The creators and spectators of online game replays and live streaming. *ACM Trans. on Multimedia Computing, Communications, and Applications*, 12(4), August 2016.

[14] L. Katzir, E. Liberty, and O. Somekh. Estimating Sizes of Social Networks via Biased Sampling. In *Proceeding of the 18th International World Wide Web Conference (WWW'11)*, 2011.

[15] M. Kaytoue, A. Silva, and C. Raissi. Watch me Playing, I am a Professional: a First Study on Video Game Live Streaming. In *World Wide Web Conference (WWW'12 Companion)*, 2012.

[16] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review E*, 73, 2006.

[17] S. A. Myers, A. Sharma, P. Gupta, and 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.

[18] K. Pires and G. Simon. Youtube live and twitch: A tour of user-generated live streaming systems. In *Multimedia Systems Conference (MMSys'15)*, 2015.

[19] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72-101:72–101, 1904.