# Statistical Study of View Preferences for Online Videos With Cross-Platform Information

Yipeng Zhou, Xuhong Gu, Di Wu, *Senior Member, IEEE*, Min Chen, *Senior Member, IEEE*, Terence H. Chan, and Siu Wai Ho

***Abstract*—The knowledge of view preferences of users is crucial for online video providers to improve their system operations and video recommendations. However, it is challenging to accurately acquire this knowledge by merely relying on a single online video system. In this paper, we conduct a joint statistical study using the cross-platform information obtained from Douban, the largest online video database with video rating functionality in China, and Youku, one of the largest online video streaming systems in China. The Douban dataset includes feedbacks (e.g., movie ratings, comments, and reviews) from all users of different online video systems, and movie metadata (e.g., release date, actors, and directors), based on which we can statistically explore effective and significant factors attributing to video view counts. Meanwhile, our study unveils user behaviors that are latent when only observing a single video system. Finally, a multiple correlation analysis reveals that factors extracted from Douban can significantly increase our ability to predict video view counts. Our study can benefit video caching, video procurement, and advertisement campaign for online video providers.**

***Index Terms*—User view preference, view count, movie rating, comments.**

Manuscript received May 25, 2017; revised August 26, 2017 and October 13, 2017; accepted October 14, 2017. Date of publication November 3, 2017; date of current version May 15, 2018. This work was supported in part by the National Key R&D Program of China under Grant 2016YFB0201900, in part by the National Natural Science Foundation of China under Grant 61572538, in part by the Fundamental Research Funds for the Central Universities under Grant 17LGJC23, and in part by the Australian Research Council under Discovery Projects DP150103658. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yonggang Wen. This paper was presented in part at the 3rd ACM Conference on Information-Centric Networking, Kyoto, Japan, September 26–28, 2016. *(Corresponding author: Di Wu.)*

Y. Zhou is with the School of Data and Computer Science, Sun Yat-Sen University, Guangzhou 510006 China, and also with the Institute for Telecommunications Research, University of South Australia, Adelaide, SA 5095, Australia (e-mail: yipeng.job@gmail.com).

X. Gu and D. Wu are with the School of Data and Computer Science, Sun Yat-Sen University, Guangzhou 510006 China, and also with the Guangdong Province Key Laboratory of Big Data Analysis and Processing, Guangzhou, 510006, China (e-mail: 1965963961@qq.com; wudi27@mail.sysu.edu.cn).

M. Chen is with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China, and also with the Wuhan National Laboratory for Optoelectronics, Wuhan 430074, China (e-mail: minchen@ieee.org).

T. H. Chan and S. W. Ho are with the Institute for Telecommunications Research, University of South Australia, Adelaide SA 5095, Australia (e-mail: hlchan6@gmail.com; siuwai.ho@unisa.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2017.2769807

## I. Introduction

ONLINE video streaming is fast becoming the dominated way how people consume videos.[1] In comparison to the traditional video distribution channels (e.g., cinema, TV ), online video streaming systems are much more flexible and convenient, offering interactive functionalities (e.g., pause, browsing, fast forward) to users. Furthermore, fuelled by the astronomical growth in the population of mobile devices and smart TVs, there is a quantum leap in demands for online video streaming. For online video providers (e.g., Netflix, Youku), one of their biggest challenges is how to understand and interpret view preferences of users, which can be measured by video view counts in a certain degree, so as to provide intelligent online services (e.g., video recommendation). This challenge is rooted from the ambiguity and limitation of user feedback information collected by a single video provider. Realising the limitation, we tackle this challenge from a novel perspective by jointly analyzing the data collected from Douban (a well-known video database in China which offers more comprehensive user feedbacks than any single online video provider) and Youku (one of the largest online video systems in China).

Analogous to IMDB [1], Douban provides two essential functionalities: (1) supporting registered raters[2] to post comments and ratings for any movie they are interested in, and (2) publishing movie metadata (e.g., release dates, directors, casts). All the above data can be conveniently accessed by the public. Thus, Douban can provide more comprehensive user feedbacks. In contrast, online video systems focusing on recommending and delivering high quality videos can tell us the current view counts for movies provided by them.[3]

The idea that information from Douban can boost the accuracy of estimating user view preferences is inspired by the following fact. The information from either Douban or Youku is biased, since neither of them can acquire complete information from all users. However, Douban is specially designed for video ratings , which is quite neutral to different online video streaming systems, e.g., YouTube, Tencent Video. This fact means that Douban provides multimodal user feedback information, with

[1] Videos are used interchangeably with movies in this work.

[2] We call users in Douban as raters and users in Youku as viewers to differentiate their roles.

[3] Although some online video systems also offer users the options to post their feedbacks (e.g., ratings and comments), it is severely restricted by users' view processes. This point will be clear in our analysis.

which we can more accurately estimate user view preferences and unveil user behaviors.

Existing studies mainly analyze view traces recorded by online video providers [2], [3]. This approach has some critical deficiencies. Firstly, the view record data may not be extensive enough. For instance, no data would be available for a movie if a video provider does not provide. This is the well-known "*cold start problem*"; Secondly, aggressively collecting user historical view data may lead to privacy issues as well, whereas, our study is merely based on public data for analysis which preserves user privacy; Finally, information contained in view records can be vague and noisy (e.g., view records cannot tell to what extent a user might like a viewed video). As a result, the resulting analysis may not be accurate enough.

The contributions of this paper are as follows:

1) We use Pearson correlation analysis to quantify factors/variables (e.g., movie ratings, release date) that are significantly correlated with view counts of movies;
2) We unveil numerous interesting user behaviors, which are latent and cannot be obtained by observing only a single system. For instances, we find out that users are more attracted to newly produced movies, even though there is a trend that the Douban scores for newer movies are descending. We also observe that high quality old movies can attain high Douban scores but not necessarily high view counts. Movies attaining higher view counts tend to receive lower Douban scores, and users' ratings are more divergent for movies attaining lower Douban scores.
3) We also conduct correlation analysis and discover that the Douban dataset can substantially improve our ability to predict whether a movie will become popular and attractive or not. This method helps online video providers to solve the cold start problem.

Our preliminary 6-page paper has been published in [4]. The current 12-page work has substantially extended the work in [4] by extracting and analyzing much more factors from the Douban dataset, and by conducting multiple correlation analysis to show the improvement of the ability to predict movie attractiveness.

The organization of the paper is as below. Related works are discussed in Section II. We introduce the background and the design principles of data crawlers in Section III. A comparative analysis is conducted in Section IV, before detailed statistical analysis is presented in Section V. A group-based analysis to explore how movie quality affects user view preferences is discussed in Section VI. We conduct the multiple correlation analysis in Section VII. Finally, conclusion and discussions for future work are presented in Section VIII.

## II. Related Work

Extensive studies have been conducted to understand and characterize user behaviors in online video systems. Zhang *et al.* [5] analyzed the access data collected from Twitch which is one of the leading video platforms in the world and found that view patterns are determined by both events and streaming sources. Cheng *et al.* [6] revealed that video duration, access pattern and active life span of YouTube short videos are very different from the traditional streaming videos. Krishnan *et al.* [7] studied the impact of video streaming quality on user engagement and abandonment rates. Ali-Eldin *et al.* [8] investigated user and session arrival rates and found that both rates do not follow a Poisson process. Krishnan *et al.* [9] measured the completion and abandonment rates of video advertisement by analyzing a large set of anonymous traces from Akamai's video delivery network. Sikdar *et al.* [10] studied the distribution of views on each video, and found that more than 10% videos on YouTube become very popular several years after being uploaded. Wu *et al.* [11] extended the epidemic model to explain the video popularity evolution patterns. Guo *et al.* [12] conducted a measurement study on video tweeting on Sina Weibo to investigate the characteristics of tweeted videos and user watching behaviors. Chen *et al.* [13] proposed a transfer learning model to infer viewer emotions, which is further extended as an approach to assess user QoE [14].

Deeper understanding of user behaviors can improve strategies for video and advertisement recommendations [15]. A mixed recommendation algorithm was proposed by taking social network and video content [16] into account. Zhou *et al.* [17] and Hossain *et al.* [18] separately proposed a distributed video recommendation system which achieved a proper trade-off between performance and privacy protection. In [19], [20] recommendation systems were refined by considering that user interests will change over time. Li *et al.* [21] designed a context-aware advertising framework by incorporating social news feeding to increase ad click-through rate. Han *et al.* [22] focused on dance video recommendation. The key challenge of their recommendation system is how to characterize the video content. Sun *et al.* [23] improved the traditional group recommendation algorithms by jointly utilizing the group-level interest and individual personality. Experimental results showed that it was very helpful for inactive group members. Zhang *et al.* [24] showed that information of access points can be used to improve the traditional factor-based CF algorithm.

The video view patterns of users will also be helpful to improve the distribution efficiency and decrease the operational cost with multiple video versions generated by transcoding in online video systems [25]. Hu *et al.* [26] designed a dynamic algorithm based on video viewing behaviors to determine the video content placement in cloud centric CDN networks so as to reduce the operational cost. Hu *et al.* [27] proposed a joint content replication and request routing for social video distribution over cloud CDNs by classifying users into different communities based on user behaviors, interests, geolocations, etc. The transcoding cost is tremendous for adaptive video streaming [28], [29]. Thus, Gao *et al.* [30], [31] proposed a partial transcoding scheme to minimize the cost by taking user view behavior patterns into account. Gao *et al.* [32] proposed a transcoding video management system for adaptive streaming that is both cost-efficient and QoS-aware. Hu *et. al* studied the distribution and sharing of social videos among mobile users by incorporating the knowledge of video popularity [33], [34].

The study of user behaviors is also helpful for video popularity prediction, which can benefit system operation. Li *et al.*

[35] built a propagation-based model by taking into account user viewing and sharing behaviors to predict the popularity of videos. Yu *et al.* [36] revealed that user anomalous behaviors on Twitter were useful to predict sudden and early view cout increases on YouTube. Deng *et al.* [37] improved personalized video recommendation by proposing cross-platform user modeling. Yoshida *et al.* [38] detected similar videos by combining semantic and affective information for video recommendation [39], [40].

Lastly, the knowledge of user view preferences provides essential complementary information for personalized video recommendations. On one hand, video view count as an important reference has been incorporated by the video recommendation framework [41], which can work together with collaborative filtering, SVD (Singular-value Decomposition) [42] and other learning algorithms to more accurately infer personal video view interests. On the other hand, Douban collectively records the information from users using different online video platforms, and hence its information is multimodal fusion to a particular online video system. It has been extensively explored to solve the cold start problem in video recommendations with the multimodal fusion information by related works [37], [43], [44]. Our contribution to this field is to richen the sources providing such multimodality information.

## III. Preliminary

In this section, we give details about the background of our work, including the introduction of Douban and Youku platforms, the applicability of our cross-platform analysis, the design of data crawlers and the methodologies we use to process the dataset.

### A. Dataset Description

We acquire two datasets from two most representative video streaming/rating platforms in China, namely Youku [45] and Douban [46] for this study.

Youku, similar to Netflix [47], is a large-scale commercial video streaming system in China. During peak hours, it serves millions of concurrent users, and the number of monthly active users is more than tens of millions. For its large scale, the Youku dataset can provide us a very integral movie view count data from a single video streaming platform [48]. For each movie delivered by Youku, a user can not only watch the video, but also post his (or her) view comments, and vote *like* or *dislike*. All information about a movie (such as view counts, the number of likes/dislikes and the number of comments posted by users) is open to the public.

Douban is operated in a similar way as IMDB [1], except that it has a much simpler movie score computing method: each user can report 2 to 10 points for a particular movie as the movie rating to indicate how much the user likes the movie. Each star represents 2 points. The maximum rating is 5 stars equivalent to the maximum of 10 points. Then, a movie's Douban score is simply the mean star value by averaging all stars it receives from users. We regard the average of Douban ratings as the *Douban score*, which is used as the metric to evaluate *movie quality* in this paper. In addition, each Douban user can post his (or her) comments towards the movie he (or she) is interested in. Specifically, a user can post *want-to-see (WTS) comments, have-seen (HS) comments and review comments* for individual movies. From the Douban system, we collect ratings, comments and metadata information. Rating information includes individual the Douban score of each movie, rating distributions and rating times. Commenting information mainly refers to the number of users who post each kind of movie comments, while the metadata includes the release date, actors and director of each movie.

Note that to distinguish the users of Youku and Douban, they are sometimes named as viewers and raters respectively to make the context clear.[4] Due to the existence of multiple video streaming platforms in China, users have different options to view a movie. Thus, it is difficult to precisely link users between Douban and Youku. For simplicity, we just treat them as different users in this paper.

### B. Applicability

The applicability of the cross-platform analysis can be illustrated from the following three aspects.

Firstly, Douban is operationally independent from any online video system. In fact, if we limit ourselves to any consider a single online video system, our ability to interpret user view preferences will be greatly reduced due to the limited size of the collection of all movies served by any single system. In contrast, Douban covers almost all movies hence the results deduced from the Douban dataset are more general and applicable for systems mainly providing commercial videos.

Secondly, Douban provides more comprehensive information about user attitudes towards each movie. Dataset collected from a particular video streaming platform are often incomplete and biased. Users have multiple ways to view the same movie, including online video systems (e.g., YouTube and Tencent Video), the traditional cinemas and TVs. The dataset collected from a single system will be severely biased due to user view behaviors in that particular system. Again, Douban is orthogonal to any particular system, thus it can provide us unbiased user feedbacks.

Finally, in addition to movies, a similar approach can be developed to analyze other commercial video categories, e.g., TV series. However, it needs considerable efforts to adapt our approach for some special video categories, e.g., user generated videos. On one hand, there may be a lack of statistical information of video makers and actors for videos made by users. On the other hand, videos with a slight difference could be uploaded by different users separately to the same online video system, resulting in the difficulty to analyze them aggregately. A possible solution is to group videos into different clone sets, and there is only a slight difference for videos in the same set [49], [50]. Movie is the category with the most commercial value, which is chosen to be first studied by our work.

[4] However, in fact they could have a significant overlap since a rater is normally supposed to comment on a video only if he (or she) is a potential viewer.

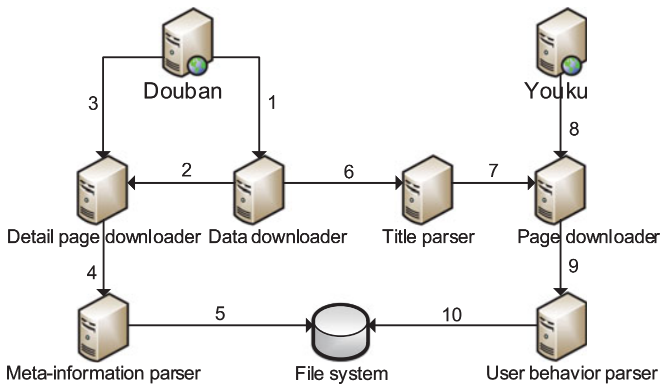



Fig. 1. The architecture of our data collection platform showing steps to crawl data.

## C. Data Collection

In order to collect all the above mentioned information about a movie, we build a dedicated data collection platform as illustrated in Fig. 1. There are two crawlers to collect data from Youku and Douban respectively. The first crawler obtains the title and meta information of all movies posted in Douban. The data downloader of the first crawler obtains JSON (JavaScript Object Notation) data from Douban, and then passes them to the title parser and the detail page downloader. The title parser then retrieves the movie titles, and passes them to the second crawler; while the detail page downloader is responsible for downloading each detail page. Then, meta-information parser retrieves movie meta information, Douban ratings and comments from detail pages. Movies with very few user ratings will be removed from the dataset due to the lack of statistical significance, which results in a total number of 15,931 movies.

After movie titles have been generated by the title parser of the first crawler, the second crawler collects Youku information for each discovered movie in Youku. Firstly, a movie searcher searches each movie title in Youku to acquire the link of a movie's play page. Then, page downloader crawls video play pages and passes crawled pages to user behavior parser, which starts to retrieve the information of view count, like/dislike numbers and user comments from movie play pages. There are only 4,714 movies found in Youku. Again, we remove these cold movies with very few user views, and finally we obtain 4,635 movies, which make up the *Youku&Douban* movie set. We define other movies that are not included by Youku as the *Douban-only* movie set.

For clarity, the entire crawling process has been marked with numbers in execution order in Fig. 1. To avoid interrupting Youku or Douban system, data collection is operated over a period of four days (i.e., from October 12, 2016 to October 15, 2016), executed in a distributed manner with multiple IP addresses. Overview of all datasets is shown in Table I.

Table I shows that Youku only provides not more than 30% movies if the Douban movie set is taken as the complete set. This is not surprising since video streaming providers have to be scrupulous with provided movies due to limited resources, e.g., bandwidth resource and the budget to purchase movies.

TABLE I
STATISTICS OF OUR DATASETS

| | Douban | Youku |
|---|---|---|
| # of movies | 15931 | 4635 |
| machine | CPU and 8G RAM | CPU and 8G RAM |
| Crawling period | 2016/10/12-2016/10/14 | 2016/10/14-2016/10/15 |

## D. Methodology of Data Processing

The Pearson correlation coefficient is an efficient measure to evaluate the correlation relationship between a pair of variables. Thus, we use the Pearson correlation to analyze each factor. With all significant factors in hand, a multiple correlation analysis can be conducted to amplify our ability to predict movie view counts.

We first describe our methodologies as follows. For variable $X$ and $Y$, the Pearson correlation coefficient is defined as:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}, \tag{1}$$

where $\mathrm{cov}(X,Y)$ is the covariance of $X$ and $Y$, $\sigma_X$ and $\sigma_Y$ are the standard deviations of $X$ and $Y$, respectively..

If we only have $n$ samples of $X$ and $Y$, then $\rho_{X,Y}$ can be represented by $r_{X,Y}$ as

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{2}$$

where $x_i$ and $y_i$ are the samples with index $i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and similarly for $\bar{y}$.

The value of $r_{X,Y}$ ranges from $-1$ to 1. It can be interpreted as follows: a larger coefficient $r_{X,Y}$ implies a stronger correlation between $X$ and $Y$. A negative coefficient implies a negative correlation between $X$ and $Y$.

Significance test should be executed before we can finally determine the correlation coefficient, which is used to exclude the possibility that the correlation relationship tested by us occurs by chance. Based on $r_{X,Y}$, the null hypothesis is that the true correlation coefficient $\rho_{X,Y}$ is 0. The significance test will return p-value as the probability that null hypothesis is true. Thus, the p-value must be very close to 0 if there really exists any correlation between $X$ and $Y$.

With all these significant factors, we can conduct a multiple correlation analysis. The coefficient of the multiple correlation $R$ ranges from 0 to 1. A higher $R$ indicates a better predictability of the dependent variable from the independent variables, otherwise it indicates a worse predictability. With $m$ independent variables ( denoted by $X_1, X_2, \ldots, X_m$) and the dependent variable ( denoted by $Y$), the multiple coefficient is the square root of the coefficient of determination, which can be derived as below:

$$R^2 = \mathbf{c}^\top \mathbf{R}_{XX}^{-1} \mathbf{c}, \tag{3}$$

where $\mathbf{c} = (r_{X_1,Y}, r_{X_2,Y}, \ldots, r_{X_m,Y})^\top$, $r_{X_j,Y}$ is the correlation coefficient between $X_j$ and $Y$, and $\mathbf{R}_{XX}^{-1}$ is the inverse matrix of the covariance matrix $\mathbf{R}_{XX}$. In $\mathbf{R}_{XX}$, the element on $i$th row and $j$th column is $r_{X_i,X_j}$. If all $m$ factors are

TABLE II
MEAN VALUE STATISTICS OF YOUKU AND DOUBAN DATA

| Youku Statistics | | Douban Statistics | |
|---|---|---|---|
| # of Youku views | 5532594 | # of ratings | 25144.35 |
| # of Youku likes | 13750.87 | Douban score | 6.73 |
| # of Youku dislikes | 1343.78 | # of HS comments | 7253.9 |
| # of Youku comments | 1073.46 | # of WTS comments | 245.21 |
| | | # of review comments | 144.63 |

TABLE III
PEARSON CORRELATION ANALYSIS OF YOUKU DATA

| statistic terms | corr. | p-value |
|---|---|---|
| # of Youku views v.s. # of Youku likes | 0.8632 | 0 |
| # of Youku views v.s. # of Youku comments | 0.8176 | 0 |
| # of Youku views v.s. # of Youku dislikes | 0.4344 | 6.9E-210 |

independent, $r_{X_i,X_j}$ is equal to 1 when $i = j$, otherwise 0. Then $R_{XX}$ is an identity matrix with $R^2 = \mathbf{c}^\top \mathbf{c}$, and the multiple coefficient is simply the sum of correlation coefficients of all factors.

The above-mentioned methodology can be applied to explore correlations among multiple variables. It is not particularly developed for the dataset used in this paper. We can also apply such an approach to study the data collected from other platforms (e.g., IMDB, YouTube, Tencent Video).

## IV. COMPARATIVE ANALYSIS

In this section, we conduct a comparative study of the different characteristics of each dataset.

Table II shows the mean value of the collected data for each statistical term. In this table, we find that the mean Youku view count is much larger than the mean number of Douban ratings, however the mean number of likes/dislikes is smaller than the mean number of Douban ratings. The mean ratio of Youku likes is more than 0.9; while the mean Douban score is just 6.73. Therefore, Douban has a set of more reasonable user ratings reflecting more comprehensive user opinions. Given the high ratio of likes in Youku, it is difficult to evaluate movie quality. In addition, the Douban dataset is more comprehensive by providing us statistics for three kinds of comments from registered Douban users. Thus, it is feasible to explore user preferences based on the Douban dataset, which reflectst user opinions with more facets.

In Table III, we use the Pearson correlation coefficient as a metric to evaluate how two statistic measures correlate with each other in the Youku system. We also show the correlation results with corresponding p-values from significance tests in Table IV, in which all three pairs exhibit positive correlation. From the results, we can make the following two conclusions.

1) A video's Youku view count is highly positive correlated with the number of Youku likes and Youku comments, and moderately correlated with the number of Youku dislikes – It confirms that the process a user evaluates movie quality in Youku is closely correlated with the process to view movies of that user.

TABLE IV
PEARSON CORRELATION ANALYSIS OF DIFFERENT VARIABLES

| statistic terms | corr. | p-value |
|---|---|---|
| # of Youku views v.s. # of ratings with star 1 | 0.0509 | 0.0005 |
| # of Youku views v.s. # of ratings with star 2 | 0.1441 | 0 |
| # of Youku views v.s. # of ratings with star 3 | 0.3203 | 0 |
| # of Youku views v.s. # of ratings with star 4 | 0.4768 | 0 |
| # of Youku views v.s. # of ratings with star 5 | 0.4108 | 0 |
| # of Youku views v.s. Douban score | −0.1939 | 0.0 |

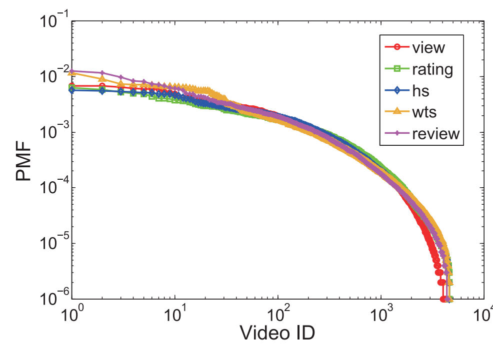



Fig. 2. CDF curves of number of Youku views or Douban ratings.

2) The extremely high ratio of likes in Table II further indicates that users prefer to view movies that look interesting, and are happy to vote likes if they appreciate them, but become reluctant to vote dislikes if these movies are disappointing.

In Fig. 2, we plot the probability mass function (PMF) for the fraction of view count taken by videos covered by Youku. Videos are ranked in a descending order by view counts. $x$ axis represents the ranked video ID, while $y$ axis represents the fraction of the total Youku view count taken by this video. Then, we extract these videos from Douban and plot the PMF curves for ratings (or each kind of Douban comments) similarly but reorder videos in the descending order according to the number of ratings (or the number of comments).

From Fig. 2, it is surprising to find that all PMF curves are so close and all of them drop sharply when video ID exceeds 1000. This result implies that there is a common underlying law guiding user attentions to mainly focus on a small fraction of popular videos though these videos are maintained separately by different platforms.

The result in Fig. 2 also implies that Douban ratings and Douban comments are supposed to be highly positive correlated with Youku view count if videos are ordered similarly by different curves. In this case, it will be quite easy to conduct cross-platform data analytics. However, as shown latter, the Pearson correlation coefficients between them are quite weak, about 0.2–0.3 for most cases. This means that a highly viewed video by Youku users does not necessarily attract intense Douban user rating and commenting, and vice versa. The underlying reason could be complicated, e.g., Youku views could be heavily affected by video recommendation and promotion strategies implemented by Youku. Thus, a joint analysis with such cross-platform information is not a trivial problem.

TABLE V
PEARSON CORRELATION ANALYSIS OF DOUBAN COMMENTS

| statistic terms | corr. | p-value |
|---|---|---|
| # of Youku views v.s. # of HS short comments | 0.2797 | 0 |
| # of Youku views v.s. # of WTS short comments | 0.2492 | 0 |
| # of Youku views v.s. # of reviews | 0.2019 | 0 |
| Douban score v.s. # of HS short comments | 0.2294 | 0 |
| Douban score v.s. # of WTS short comments | 0.0981 | 0 |
| Douban score v.s. # of reviews | 0.1801 | 0 |

## V. STATISTICAL ANALYSIS

In this section, we focus on exploring factors that correlate with Youku view counts. More specifically, we analyze how Douban ratings, Douban comments, release dates, online ages and movie makers relate with user view preferences. Based on the correlation results, we are able to unveil the latent user view behaviors.

### *A. Analysis of Douban Rating*

The dataset from Douban contains information about how raters rate movies. We conduct a Pearson correlation analysis between rater behavior and viewer behavior in Table IV.

Recall that Douban score ranges from 1 star to 5 stars, thus "# of ratings with star 1" represents the number of raters who vote 1 star for a movie and the rule is the same for other scores. As the results shown in Table IV, the low p-values in the last column indicate statistical significance for these factors. It is easy to understand the weak positive correlation between rating times and view count, but it is quite counterintuitive that there is a weak negative correlation between Douban score, which can be regarded as a measure of movie quality, and the number of Youku views. We summarize our observations as follows:

1) The number of raters voting 4-star or 5-star is more effective for predicting user view preferences than the number of raters who vote with other Douban scores. This is evidenced by the correlation results of the first five pairs of terms.
2) There is a negative correlation between view count and Douban score. This result implies that a high-quality movie does not necessarily result in a high view count. In addition to movie quality, there exist many other factors to influence view count (e.g., movie genre, actors, release date).

The above observations suggest that how the movie quality affects its view count is a complicated problem worth further exploration. In the next section, we particularly focus on investigating this relationship.

### *B. Analysis of Douban Comments*

Recall that each Douban user can post three kinds of comments, *want-to-see (WTS) comments, has-seen (HS) comments*, and *review comments*. We conduct a Pearson correlation study between Youku view count (or Douban score) with the numbers of three kinds of Douban comments in Table V.

TABLE VI
PEARSON CORRELATION ANALYSIS OF DOUBAN COMMENTS BY ELIMINATING OLD MOVIES

| statistic terms | corr. | p-value |
|---|---|---|
| # of Youku views v.s. # of HS short comments | 0.2959 | 0 |
| # of Youku views v.s. # of WTS short comments | 0.2606 | 0 |
| # of Youku views v.s. # of reviews | 0.2049 | 0 |
| Douban score v.s. # of HS short comments | 0.3017 | 0 |
| Douban score v.s. # of WTS short comments | 0.1297 | 0 |
| Douban score v.s. # of reviews | 0.2400 | 0 |

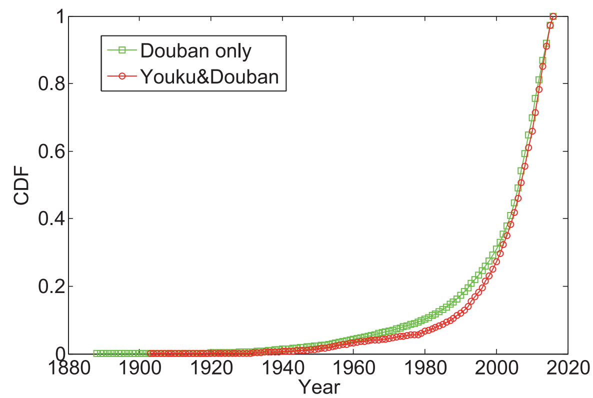



Fig. 3. CDF curves of movie production date.

Again, the p-values in the last column indicate statistical significance of this test, and correlations with Youku view count is more significant than correlations with Douban score. This result indicates that movies receiving high Youku view count tend to somehow introduce more intense rating and commenting from raters.

However, a question is that some old classical movies probably produced decades of years ago may only attract user ratings and comments, but fail to attract views. In Table VI, we repeat the correlation study of Table V by eliminating movies released before Jan. 1, 2006 since both Youku and Douban are established in about 2006 respectively. As observed, all correlation values are lifted a little bit, hence viewers' behavior tends to have higher correlation with raters' behavior for more recent movies. However, how exactly the time factor affects user behavior is a complicated problem, and we further explore it in the next subsection.

### *C. Effects of Release Date and Online Date*

Time effect is in fact rather complicated for video view count. We first introduce a few concepts to facilitate our presentation. The *release date* of a movie is the day when a movie is published to the public, while a movie's online date is the day when a movie is uploaded to the Youku system. We define *online delay* as the time difference between a movie's release date and its online date in Youku, and *movie age* as the time duration between a movie's online date and the data crawling date.

We first study the effect of release dates by grouping movies based on the year movies were released. Fig. 3 plots the CDF curves of the fraction of movies released per year for movies

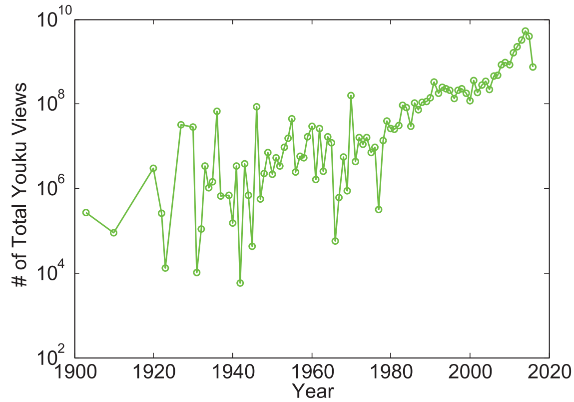



Fig. 4. The distribution of the number of total Youku views attracted by movies produced per year.

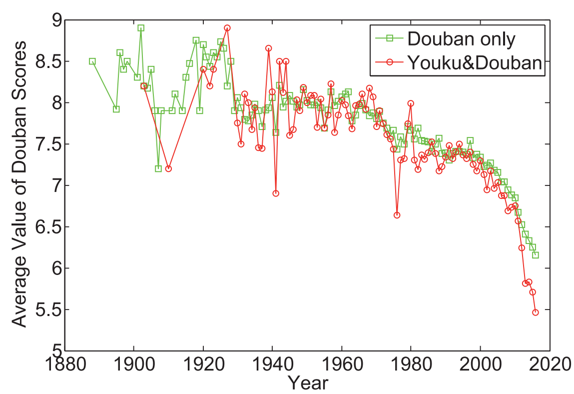



Fig. 5. The average Douban score for movies released in each year.

in Douban only (i.e., those movies absent from Youku system) and movies in both Youku and Douban respectively. As we can see from the figure, the number of movies released per year increases faster and faster with time, which indicates the flourish of film industry. By comparing these two curves, we find that there is no apparent trend that Youku opts to select more recent movies by also providing a certain number of old movies.

Fig. 4 presents the number of total Youku views generated by the movies released within the same year. By comparing Figs. 4 and 3, one can find two interesting phenomenons: 1) movies released before 1980 only attract a very small amount of Youku views; 2) there is a sudden drop caused by the view counts generated by movies released in 2015 and 2016.

It is not difficult to understand the observed phenomenons. Some old fashioned movies cannot attain user interests for viewing. People are often more attracted by new movies (which often have new stories with better pictures and sound quality, and are more heavily promoted by various advertisement). The sudden drop of views in 2015 and 2016 indicates that newly produced movies remain to be of interests to users for at least 1–2 years. The final view count is cumulated over a period and is much larger than the current value, but our data haven't included these views generated beyond the crawling date.

Fig. 5 plots the average Douban score for movies released per year. There is a clear trend that the average Douban score decreases with the release time for both Youku&Douban movies and Douban-only movies. It is observed that Youku did not

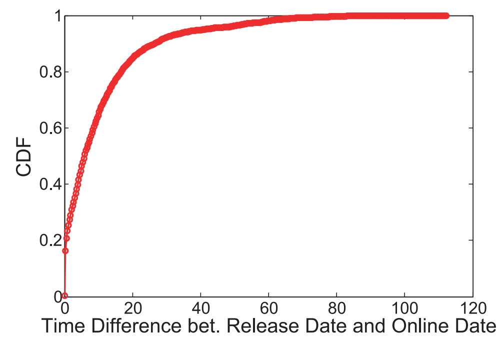



Fig. 6. CDF curve of online delays for all Youku movies.

TABLE VII
PEARSON CORRELATION ANALYSIS OF MOVIE AGE AND ONLINE DELAY

| statistic terms | corr. | p-value |
|---|---|---|
| # of Youku views v.s. Movie age | $-0.0097$ | 0.5139 |
| # of Youku views v.s. Online delay | $-0.1507$ | 0 |

particularly choose these high-score movies for their services. Movies released earlier than 1980 achieve very high Douban scores though they cannot attain a high volume of user views. This result also indicates that the flourish of movie making market mainly expands the capacity to produce movies instead of improving movie quality. This phenomena may also be caused by the fact that users are becoming more critical than before.

Based on the results of Figs. 3–5, we summarize our observations as follows. The release date indeed has a significant influence on the Youku view count, but little influence on the Douban score. Users appreciate old classical movies, but they may have viewed these old movies from TVs or cinemas, and are reluctant to watch them again in Youku. User eyeballs are mainly attracted by newly produced movies or movies strongly promoted by providers or movie producers though these movies may disappoint users resulting in very low Douban scores.

These insights are essential for online movie providers to select and promote movies. We infer that a movie's Douban score not only reflects the movie's quality, but also reflects how users are disappointed with it. Old classical movies mainly attract users to replay them which only takes a small portion of total views. For the whole system, most views are contributed by newly produced movies instead of high-score movies. This result can also explain the negative correlation between the Douban score and the movie view count presented in Table IV since user attentions mainly focus on new movies though the quality of these movies deteriorates with time .

Now, we study how the Youku online date (in the Youku system) and online delay relate with Youku view count. In Fig. 6, we plot the CDF curve of online delays for all Youku movies to have an overview of the distribution. There are about 60% movies with online delay less than 10 years but about 40% movies with online delay over 10 years. For a small fraction of movies, online delays could be over several decades.

In Table VII, we conduct a correlation analysis between Youku view count and movie age (or online delay). Surprisingly, there is no correlation between Youku view count and

TABLE VIII
PEARSON CORRELATION ANALYSIS WITH ANCIENT MOVIE GROUP AND RECENT MOVIE GROUP

| statistic terms | corr. | p-value |
|---|---|---|
| # of Youku views v.s. Online delay (Recent Group) | −0.2747 | 0 |
| # of Youku views v.s. Online delay (Ancient Group) | 0.0144 | 0.5674 |
| # of Youku views v.s. Douban score (Recent Group) | −0.1673 | 0 |
| # of Youku views v.s. Douban score (Ancient Group) | −0.0499 | 0.0470 |
| # of Youku views v.s. # of Douban ratings (Recent Group) | 0.2347 | 0 |
| # of Youku views v.s. # of Douban ratings (Ancient Group) | 0.0722 | 0 |

movie age, but a weak negative correlation between Youku view count and online delay. This result can be interpreted as that user interests to view a video probably diminish with time. Thus, a longer delay leads to less view interests for users. But a movie with a longer period being hosted on Youku does not necessarily have more views.

From Figs. 3 and 6, we observe that there are a number of videos released several decades ago with very large online delays. We speculate that the correlation should be weaker for these older movies. Thus, we heuristically divide all movies into two groups, *recent group* in which delay of each movie is less than 10 years,[5] and *ancient group* in which online delay of each movie is over 10 years. The result is presented in Table VIII, in which we observe that all correlations become pronounced if we only consider movies in the recent group, while the correlations are either statistical insignificant or very weak if we only consider movies in the ancient group. This study shows that users exhibit quite different behaviors when viewing (or rating) ancient and recent movies. We conjecture that users may be reluctant to view ancient movies repeatedly but would like to rate and comment them because of their high quality.

### D. Analysis of Movie Makers

In this subsection, we investigate how view count correlates with movie markers, i.e., casts and directors. Intuitively, more productive actors or directors tend to make more popular movies. Thus, to some extent, the popularity of a movie can be inferred from the eminence of its cast and director.

In Fig. 7, we plot the CDF curves for the number of movies involved by each actor and director. Surprisingly, there are a large portion of actors and directors who are only be involved with a single movie. About 80% of actors or directors produce no more than 5 movies.

To visualize how the activity of actors and directors affects Youku view count or Douban score, we conduct a correlation study in Table IX

From the results in Table IX, we found that the Youku view count is only slightly positive correlated with the productivity

[5]Given that Youku is established in 2006, the movies in recent group must be produced later than 1996.

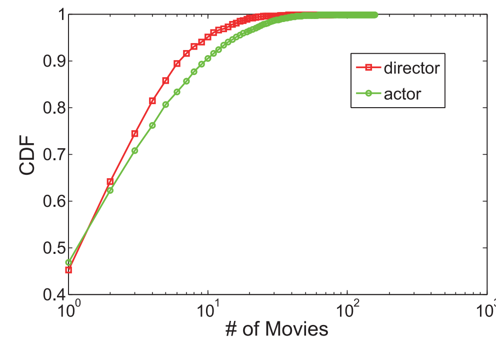



Fig. 7. CDF curves of the number of movies involved by an actor or director.

TABLE IX
PEARSON CORRELATION ANALYSIS OF THE ACTIVITY OF ACTORS AND DIRECTORS

| statistic terms | corr. | p-value |
|---|---|---|
| # of Youku views v.s. # of movies directed by directors | 0.0672 | 5E-6 |
| # of Youku views v.s. # of movies actors starred in | 0.1159 | 0 |
| # of Youku views v.s. # of actors | 0.1906 | 0 |
| Douban score v.s. # of movies directed by directors | 0.0774 | 0 |
| Douban score v.s. # of movies actors starred in | 0.0277 | 0.06 |
| Douban score v.s. # of actors | −0.0227 | 0.125 |

TABLE X
PEARSON CORRELATION ANALYSIS OF ACTORS' AND DIRECTORS' ACHIEVEMENT

| statistic terms | corr. | p-value |
|---|---|---|
| # of Youku views v.s. Avg. # of views of the cast | 0.3210 | 0 |
| # of Youku views v.s. Avg. # of view of directors | 0.1371 | 0 |
| # of Youku views v.s. Avg. Douban score of directors | −0.1189 | 0 |
| # of Youku views v.s. Avg. Douban score of the cast | −0.2817 | 0 |

of directors (measured by the number of movies directed), but is moderately positive correlated with actors' activity (measured by the number of movies in which actors are starred) and the actor population (which could be related with movie budget). Users' view decisions seem to be more heavily affected by the fame of actors instead of the fame of directors. However, the actor-related statistics have almost no correlation with the Douban score, from which we conjecture that users opt to rate movie quality based on movie content instead of actors or directors.

If an actor or director can persistently produce popular movies, view counts of an actor (or director)'s movies should be related with each other. To conduct this study, we need to first exclude those inactive actors and directors because the test result is likely statistical insignificant if there are too few movies involving the actor or director. Thus, in Table X, we conduct a correlation study between the Youku view count of a particular movie and the average view count of its actors' (or directors') other movies. If an actor (or director) is too inactive with no more than 4 movies, we will use the average value of all actors (or directors) by excluding the movie itself for the actor ( or director).

In Table X, we can find that there is moderate positive correlation between Youku view count and the average number of

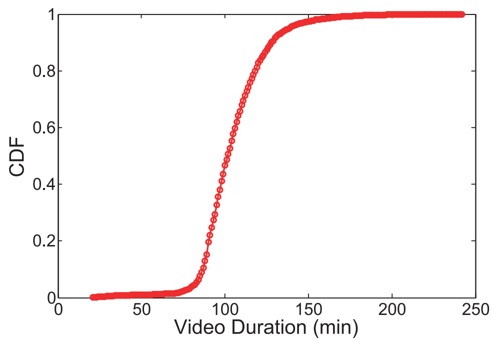



Fig. 8. CDF curves of movie duration.

TABLE XI
PEARSON CORRELATION ANALYSIS OF DIFFERENT VARIABLES

| statistic terms | corr. | p-value |
|---|---|---|
| # of Youku views v.s. Movie duration | −0.0133 | 0.3754 |
| Douban score v.s. Movie duration | 0.2816 | 0 |

views of the cast (or the average number of view of directors). Again, the Youku view count has moderate negative correlation with the average score of directors (or the average score of actors), which is consistent with the previous result that Youku view count has a negative correlation with the Douban score.

Finally, we evaluate the influence of movie duration. Fig. 8 plots the CDF curve of the durations for all Youku movies. The durations of most movies are in the range from 90 minutes to 140 minutes. We conduct a Pearson correlation analysis in Table XI, which shows that movie duration has a moderate positive correlation with the Douban score, but has no correlation with the view count. This result shows the complicated relationship between movie quality and its attractiveness again, which motivates us to further explore this problem in the next section.

Discussion: Our work has considered two kinds of statistical variables. The first kind is from the Youku system, e.g., online date and view count; while the second kind is from the Douban platform, e.g., ratings, comments, metadata. By combining them together, we can conduct a joint analysis to reveal latent user behaviors and quantify effective factors correlating with user view preferences, which cannot be observed if we only have Youku dataset. This pioneering work can guide online video providers to refine their system operations by synthetically considering more different facets of individual movies.

## VI. GROUP BASED ANALYSIS

The most counterintuitive result we have obtained so far is the weak negative correlation between the Douban score and Youku view count. This section further investigates the relationship between movie quality (evaluated in terms of Douban scores) and user view interests (evaluated in terms of Youku views).

To clearly visualize this relationship, we explore this problem by grouping all movies according to Douban score and view count. More specifically, we classify all movies into four groups:*high-view high-score (hv-hs), high-view low- score (hv-ls), low-view high- score (lv-hs) and low-view low-score (lv-ls)*. Fig. 9 presents how the movies are classified into four clusters

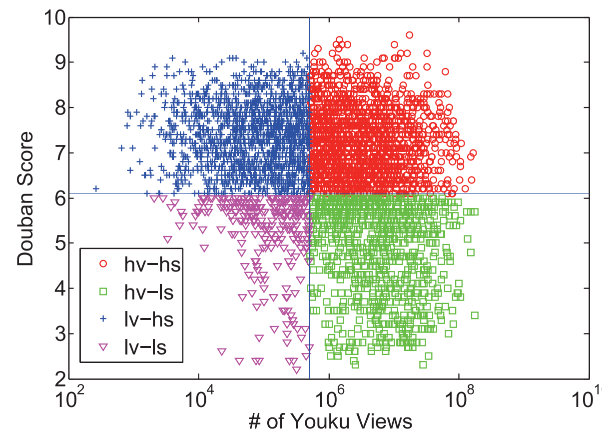



Fig. 9. The scatter plot showing how to group movies.

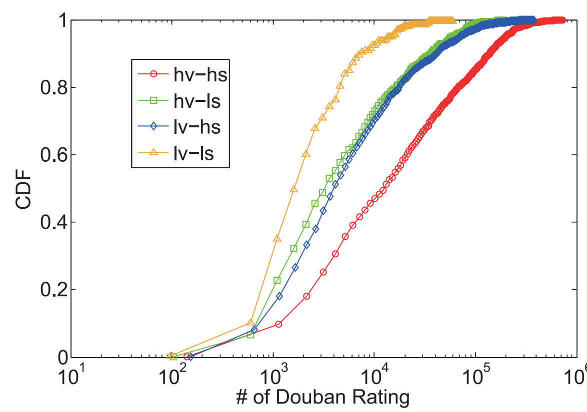



Fig. 10. CDF curves of the number of Douban ratings for each group.

by running the K-means algorithm using the Douban score and the logarithm value of Youku view count. Note that K-means algorithm is not meant to detect concealing patterns. Instead, it is applied to determine the boundary for movie grouping. In Fig. 9, each point represents a movie, and the boundary conditions to classify movies are indicated by two straight lines in Fig. 9. The number of movies in each group is 1,761 for hv-hs, 852 for hv-ls, 1,746 for lv-hs and 276 for lv-ls respectively.

As shown in Fig. 9, there exist a large number of movies belonging to either hv-ls group or lv-hs group. The reasons for this phenomenon are multifold: one possibility is the movie makers can invest a lot of money to promote their movies even though the movie quality is not very good; Another possibility is that a high-quality literary movie may only be appreciated by a small population resulting in low views. This section focuses on mining the features of the two special movie groups: hv-ls and lv-hs because it is ineffective to infer view preferences of users for these movies through their Douban scores.

Fig. 10 plots CDF curves of the number of Douban ratings for each movie group. It is easy to separate the hv-hs group from the lv-ls group because the former one received the most number of user ratings but the latter one received the least. The curves of lv-hs group and hv-ls group are in the region between hv-hs group and lv-ls group, and the gap between them is very small. Recall that old movies tend to achieve good ratings but new movies tend to capture user eyeballs as we have shown in

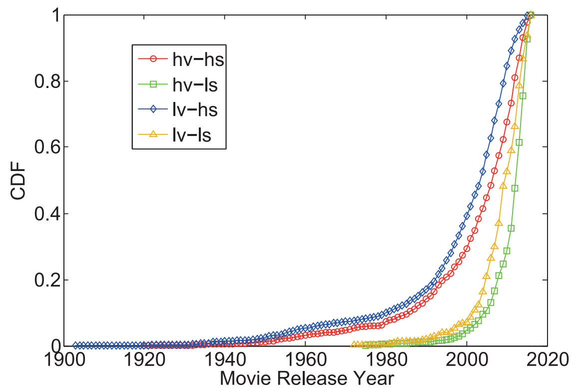



Fig. 11. CDF curves of movies' release date for each group.

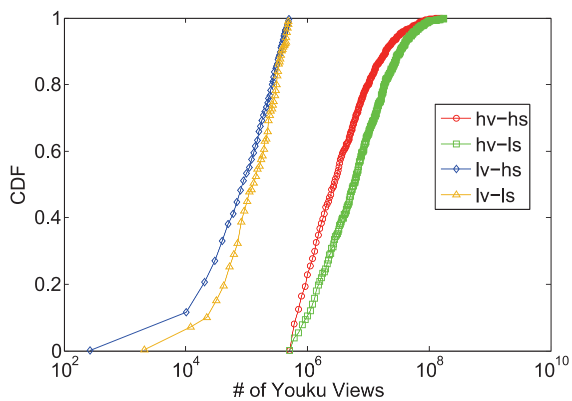



Fig. 12. CDF curves of Youku views for each movie group.

Figs. 4 and 5, we conjecture that there is a large fraction of old movies in the lv-hs group and a large number of new movies in the hv-ls group. By observing the curves of hv-hs group and hv-ls group, it seems that users are reluctant to rate low score movies, based on which we infer that users tend to keep silent with movies they do not appreciate.

To verify our conjecture about the movie age of each group, we plot the CDF curves of movies' release dates for each group in Fig. 11. The lv-hs group has the largest fraction of old fashioned movies, and the fraction of movies released earlier than the year of 2000 exceeds 50% in both lv-hs and hv-hs. From Fig. 3, we have shown that movies released before the year of 2000 only take about 20 percent. It is surprising to find that there is very few old movies included in lv-ls and hv-ls groups, which indicates that new movies tend to attain low scores. Based on this result, we can deduce that some old movies with high quality mainly included in the hv-hs group can keep attracting user eyeballs, but the other old movies mainly included in the lv-hs group have lost users' view interests though they can attain high Douban scores. The hv-ls and lv-ls curves also indicate that too many low-score movies are produced recently, even though some of them can attract users' view interests.

Fig. 12 presents the CDF curves of Youku views for each movie group. As expected, hv-ls and hv-hs curves are apart from lv-lv and lv-hs curves because these groups are classified based on Youku views. The interesting issue is that movies in the hv-ls group have more average views than these in the hv-hs group, and the argument is the same for the lv-ls and

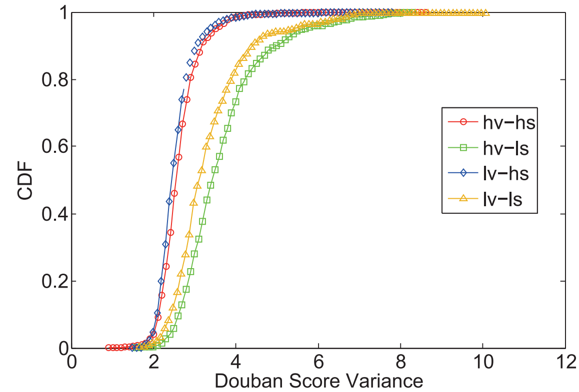



Fig. 13. CDF curves of score variances for four groups.

lv-hs groups. As shown in Table IV, there is a weak negative correlation between views and scores. This property holds in both high view groups and low view groups, based on which we infer that movies targeting more users tend to be rated with a lower score because it is more difficult to flatter users with diverse interests. By comparing Figs. 10 and 12, we find that movies in the hv-ls group with more views receives much fewer ratings than movies in the hv-hs group. The same argument also applies for movies in the lv-ls and lv-hs groups. This comparison confirms our inference that users prefer not to comment or rate on bad movies (or low-score movies).

We define score variance $v$ as the metric to evaluate the divergence of user opinions on a movie's quality. Let $r$ and $\mathcal{R}$ denote the value of the rating and the set of all possible values of ratings. The score and the score variance of a particular movie are calculated by Douban as $s = \frac{1}{N}\sum_{\forall r \in \mathcal{R}} N_r \times r$ and $v = \sum_{\forall r \in \mathcal{R}} \frac{N_r}{N}(r-s)^2$ respectively, where $N$ is the total number of ratings and $N_r$ is the number of ratings with value $r$.

We plot the CDF curve of Douban score variances for each group in Fig. 13. It is interesting to find that most movies with a low score have much larger score variance than movies with a high score. About 90% high-score movies have variances larger than 3; while more than 70% low-score movies have a variance larger than 3. In particular, the hv-ls group has the largest average score variance. This result indicates that user opinions are diverge on bad movies, but for good movies, user opinions are quite coherent.

A Pearson correlation coefficient between Douban score variance and Youku view count is 0.1977 with p-value as zero, which shows that a larger score variance implies a higher view count.

*Discussion:* For online video providers, the major goal is to maximize the view count of their movies since more views often imply higher advertisement revenue. However, based on our study, predicting movie view count is challenging since view count depends on many different factors. It is difficult to accurately predict a video's attractiveness hence its view count simply from one or two factors. On the other hand, movie quality should also be taken into account by online video providers for their services. Although users prefer to view fresh movies, users may leave the system if they are frustrated by too many low-quality movies (especially when only low-quality fresh movies are available). In addition, movies are probably produced for

TABLE XII
THE LIST OF ELEMENTS IN EACH FACTOR SET

| Set | Elements |
|---|---|
| $\mathcal{F}_1$ | Douban score, # of ratings with star 1, # of ratings with star 2, # of ratings with star 3, # of ratings with star 4, # of ratings with star 5, Douban score variance, # of HS short comments, # of WTS short comments, # of reviews, online delay, Avg. Douban score of the cast, Avg. Douban score of directors |
| $\mathcal{F}_2$ | Avg. # of views of the cast, Avg. # of views of directors, # of actors, # of movies actors starred in, # of movies directed by directors, online delay |

TABLE XIII
THE RESULT OF MULTIPLE CORRELATION ANALYSIS

| Used factors | Multiple correlation coefficient $R$ | p-value |
|---|---|---|
| $\mathcal{F}_1$ | 0.5379 | 0 |
| $\mathcal{F}_2$ | 0.3772 | 0 |
| $\mathcal{F}_1 + \mathcal{F}_2$ | 0.5473 | 0 |

different user groups. Online video service providers should also consider to provision some temporarily unpopular movies with high quality to attract users with different tastes.

## VII. MULTIPLE CORRELATION ANALYSIS

This section evaluates how well the movie view count can be predicted by the factors extracted from the Douban dataset. First of all, we classify all effective factors we have tested so far into two sets: $\mathcal{F}_1$ and $\mathcal{F}_2$. Our principle is to put all factors merely extracted from the Douban dataset into $\mathcal{F}_1$, and the others in $\mathcal{F}_2$. The elements in each set are presented in Table XII

As we can see from Table XII, the factors in $\mathcal{F}_2$ needs to acquire the view count information from Youku. In contrast, $\mathcal{F}_1$ only needs information from Douban.

We conduct correlation analysis using factors in $\mathcal{F}_1$, $\mathcal{F}_2$ and $\mathcal{F}_1 + \mathcal{F}_2$ as independent variables respectively. The result is presented in Table XIII, in which the multiple correlation coefficient $R$ indicates the ability to predict the Youku view count with these independent factors according to the preliminary knowledge introduced in Section III. The Douban dataset can significantly increase the prediction ability by increasing $R$ from 0.3772 to 0.5473. Although, multiple correlation analysis may not be the optimal model for this problem, it shows the potential benefit of the multimodal information provided by Douban.

We enumerate two examples to illustrate how online video providers can utilize our findings to improve their services.

1) Imagine that an online video provider intends to supply a video released two years ago. However, without any operation data, it is difficult for the provider to estimate the video's attractiveness. Fortunately, Douban collects the video rating and commenting information since its release date, which can be used to estimate the video's attractiveness by extracting factors introduced in our work. By only applying a simple multiple correlation analysis, the provider can estimate its attractiveness without significant deviation.
2) The information extracted from Douban can be incorporated into a provider's recommendation framework. Through analyzing the dataset from Douban, we can estimate a video's attractiveness more accurately based on feedbacks reported by users using different systems. By comparing with the video' actual view count in the system, the provider can adjust recommendation strategies by identifying videos that have not been recommended adequately.

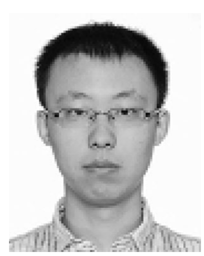
## VIII. CONCLUSION

In this paper, we investigate video view preferences of users by jointly analyzing data collected from two separately operated platforms: Youku (an online video provider) and Douban (an online video database). We find that a movie's view count is not only determined by the quality of the movie but also affected by many other factors such as release date, actors, director(s). This work also unveils latent online viewers' behaviors. For example, users are more reluctant to rate movies that they are not interested in. This work indicates the feasibility to estimate user view preferences by leveraging cross-platform information, supported by the results of correlation analysis. How to create an advanced model to estimate and even predict movie view counts with these independent factors will be an interesting future work.

## REFERENCES

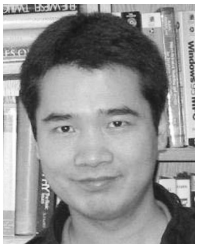
[1] "Imdb". 2017. [Online]. Available: http://www.imdb.com/
[2] Y. Xu *et al.*, "Modeling buffer starvations of video streaming in cellular networks with large-scale measurement of user behavior," *IEEE Trans. Moblie Comput.*, vol. 16, no. 8, pp. 2228–2245, Aug. 2017.
[3] Z. L, Y. Ji, X. Jiang, and Y. Tanaka, "User-behavior driven video caching in content centric network," in *Proc. 3rd ACM Conf. Inf.-Centric Netw.*, 2016, pp. 197–198.
[4] X. Gu, Y. Zhou, D. Wu, T. Chan, and M. Chen, "Unveiling latent behaviors of video viewers with cross-platform information," in *Proc. 27th ACM 27th Workshop Netw. Oper. Syst. Support Digit. Audio Video*, 2017, pp. 31–36.
[5] C. Zhang and J. Liu, "On crowdsourced interactive live streaming: A twitch. TV-based measurement study," in *Proc. 25th Workshop Netw. Oper. Syst. Support Digit. Audio Video*, 2015, pp. 55–60.
[6] X. Cheng, J. Liu, and C. Dale, "Understanding the characteristics of internet short video sharing: A Youtube-based measurement study," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1184–1194, Aug. 2013.
[7] S. Krishnan and R. Sitaraman, "Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 2001–2014, Dec. 2013.
[8] A. Ali-Eldin, M. Kihl, J. Tordsson, and E. Elmroth, "Analysis and characterization of a video-on-demand service workload," in *Proc. 6th ACM Multimedia Syst. Conf.*, 2015, pp. 189–200.
[9] S. Krishnan and R. Sitaraman, "Understanding the effectiveness of video ads: A measurement study," in *Proc. ACM Conf. Internet Meas. Conf.*, 2013, pp. 149–162.
[10] S. Sikdar *et al.*, "Identifying and characterizing sleeping beauties on Youtube," in *Proc. 19th ACM Conf. Comput. Supported Cooperative Work Soc. Comput. Companion*, 2016, pp. 405–408.

[11] J. Wu, Y. Zhou, D. Chiu, and Z. Zhu, "Modeling dynamics of online video popularity," in *Proc. 23rd IEEE/ACM Int. Symp. Qual. Service*, 2015, pp. 141–146.
[12] Z. Guo, J. Huang, J. He, X. Hei, and D. Wu, "Unveiling the patterns of video tweeting: A Sina Weibo-based measurement study," in *Proceedings of the 14th International Conference Passive and Active Network Measurement*. New York, NY, USA: Springer, 2013, pp. 166–175.
[13] M. Chen, Y. Hao, Y. Li, D. Wu, and D. Huang, "Demo: Lives: Learning through interactive video and emotion-aware system," in *Proc. 16th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2015, pp. 399–400.
[14] M. Chen, Y. Hao, S. Mao, D. Wu, and Y. Zhou, "User intent-oriented video QOE with emotion detection networking," in *Proc. IEEE Global Commun. Conf.*, 2016, pp. 1552–1559.
[15] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep features learning for medical image analysis with convolutional autoencoder neural network," *IEEE Trans. Big Data*, to be published.
[16] L. Cui, L. Dong, X. Fu, Z. Wen, N. Lu, and G. Zhang, "A video recommendation algorithm based on the combination of video content and social network," *Concurrency Comput.: Practice Experience*, vol. 29, no. 14, 2017, Art. no. e3900.
[17] P. Zhou, Y. Zhou, D. Wu, and H. Jin, "Differentially private online learning for cloud-based video recommendation with multimedia big data in social networks," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1217–1229, Jun. 2016.
[18] M. S. Hossain, M. F. Alhamid, and G. Muhammad, "Collaborative analysis model for trending images on social networks," *Future Gener. Comput. Syst.*, to be published.
[19] N. Liu, M. Zhao, E. Xiang, and Q. Yang, "Online evolutionary collaborative filtering," in *Proc. 4th ACM Conf. Recommender Syst.* 2010, pp. 95–102.
[20] Y. Koren, "Collaborative filtering with temporal dynamics," *Commun. ACM*, vol. 53, no. 4, pp. 89–97, 2010.
[21] Y. Li, D. Zhang, Z. Lan, and K. Tan, "Context-aware advertisement recommendation for high-speed social news feeding," in *Proc. 32nd IEEE Int. Conf. Data Eng.*, 2016, pp. 505–516.
[22] T. Han *et al.*, "Dancelets mining for video recommendation based on dance styles," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 712–724, Apr. 2017.
[23] L. Sun, X. Wang, Z. Wang, H. Zhao, and W. Zhu, "Social-aware video recommendation for online social groups," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 609–618, Mar. 2017.
[24] J. Zhang, Y. Zhou, D. Wu, and C. Yang, "Context-aware video recommendation by mining users view preferences based on access points," in *Proc. 27th Workshop Netw. Oper. Syst. Support Digit. Audio Video*, 2017, pp. 37–42.
[25] Y. Wen, X. Zhu, J. J. Rodrigues, and C. W. Chen, "Cloud mobile media: Reflections and outlook," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 885–902, Jun. 2014.
[26] H. Hu *et al.*, "Community based effective social video contents placement in cloud centric CDN network," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2014, pp. 1–6.
[27] H. Hu *et al.*, "Joint content replication and request routing for social video distribution over cloud CDN: A community clustering method," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 7, pp. 1320–1333, Jul. 2016.
[28] L. Zhou, "QOE-driven delay announcement for cloud mobile media," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 1, pp. 84–94, Jan. 2017.
[29] L. Zhou, "On data-driven delay estimation for media cloud," *IEEE Trans. Multimedia*, vol. 18, no. 5, pp. 905–915, May 2016.
[30] G. Gao *et al.*, "Cost optimal video transcoding in media cloud: Insights from user viewing pattern," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2014, pp. 1–6.
[31] G. Gao, W. Zhang, Y. Wen, Z. Wang, and W. Zhu, "Towards cost-efficient video transcoding in media cloud: Insights learned from user viewing patterns," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1286–1296, Aug. 2015.
[32] G. Gao, Y. Wen, W. Zhang, and H. Hu, "Cost-efficient and QOS-aware content management in media cloud: Implementation and evaluation," in *Proc. IEEE Int. Conf. Commun.*, 2015, pp. 6880–6886.
[33] H. Hu, Y. Wen, and D. Niyato, "Public cloud storage-assisted mobile social video sharing: A supermodular game approach," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 3, pp. 545–556, Mar. 2017.
[34] H. Hu, Y. Wen, and D. Niyato, "Spectrum allocation and bitrate adjustment for mobile social video sharing: Potential game with online QOS learning approach," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 4, pp. 935–948, Apr. 2017.
[35] H. Li, X. Ma, F. Wang, J. Liu, and K. Xu, "On popularity prediction of videos shared in online social networks," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2013, pp. 169–178.
[36] H. Yu, L. Xie, and S. Sanner, "Twitter-driven Youtube views: Beyond individual influencers," in *Proc. 22nd ACM Int. Conf. Multimedia,* 2014, pp. 869–872.
[37] Z. Deng, J. Sang, and C. Xu, "Personalized video recommendation based on cross-platform user modeling," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2013, pp. 1–6.
[38] T. Yoshida, G. Irie, H. Arai, and Y. Taniguchi, "Towards semantic and affective content-based video recommendation," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2013, pp. 1–6.
[39] W. Min, B.-K. Bao, C. Xu, and M. S. Hossain, "Cross-platform multimodal topic modeling for personalized inter-platform recommendation," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1787–1801, Oct. 2015.
[40] B.-K. Bao, C. Xu, W. Min, and M. S. Hossain, "Cross-platform emerging topic detection and elaboration from multimedia streams," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 11, no. 4, 2015, Art. no. 54.
[41] J. Davidson *et al.*, "The Youtube video recommendation system," in *Proc. 4th ACM Conf. Recommender Syst.*, 2010, pp. 293–296.
[42] A. B. Barragáns-Martínez *et al.*, "A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition," *Inf. Sci.*, vol. 180, no. 22, pp. 4290–4311, 2010.
[43] X. Zhao *et al.*, "Integrating rich information for video recommendation with multi-task rank aggregation," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 1521–1524.
[44] B. Yang *et al.*, "Online video recommendation based on multimodal fusion and relevance feedback," in *Proc. 6th ACM Int. Conf. Image Video Retrieval*, 2007, pp. 73–80.
[45] "Youku". 2017. [Online]. Available: http://www.youku.com/
[46] "Douban". 2017. [Online]. Available: https://www.douban.com/
[47] "Netflix". 2017. [Online]. Available: https://www.netflix.com/
[48] Z. Wang, L. Sun, C. Wu, and S. Yang, "Enhancing internet-scale video service deployment using microblog-based prediction," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 3, pp. 775–785, Mar. 2015.
[49] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," *IEEE/ACM Trans. Netw.*, vol. 17, no. 5, pp. 1357–1370, Oct. 2009.
[50] Y. Borghol, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "The untold story of the clones: Content-agnostic factors that impact youtube video popularity," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1186–1194.



**Yipeng Zhou** received the B.S. degree from Department of Computer Science, University of Science and Technology of China, Hefei, China, and the M.Phil. and Ph.D. degrees from the Department of Information Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong. He was an Assistant Professor with the College of Computer Science and Software Engineering, Shenzhen University, from September 2013 to August 2016. From 2012 to 2013, he was a Postdoctoral Research Fellow with the Institute of Network Coding, CUHK. He is currently a Research Fellow with the Institute for Telecommunications Research, University of South Australia, Adelaide, SA, Australia. His main research interests include modeling and analysis of large-scaled networking systems, analysis of user behaviors of online video systems, and crowdsourcing-based content distribution.

**Xuhong Gu** received the B.S. degree in network engineering in 2016 from Sun Yat-Sen University, Guangzhou, China, where he is currently working toward the Master's degree in the School of Data and Computer Science. His research interests mainly include multimedia networking, and analysis of user behaviors of online video systems.



**Di Wu** (M'06–SM'17) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2000, the M.S. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2003, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2007. He was a Postdoctoral Researcher with the Department of Computer Science and Engineering, Polytechnic Institute of New York University, Brooklyn, NY, USA, from 2007 to 2009, advised by Prof. K. W. Ross. He is currently a Professor and the Assistant Dean of the School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China. His research interests include cloud computing, multimedia communication, Internet measurement, and network security. Prof. Wu was a co-recipient of the IEEE INFOCOM 2009 Best Paper Award. He has served as an Editor of the *Journal of Telecommunication Systems* (Springer), *Journal of Communications and Networks*, *Peer-to-Peer Networking and Applications* (Springer), *Security and Communication Networks* (Wiley), and the *KSII Transactions on Internet and Information Systems*, and a Guest Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He has also served as the MSIG Chair of the Multimedia Communications Technical Committee in the IEEE Communications Society from 2014 to 2016. He served as the TPC Co-Chair of the IEEE Global Communications Conference—Cloud Computing Systems, and Networks, and Applications in 2014, the Chair of the CCF Young Computer Scientists and Engineers Forum—Guangzhou from 2014 to 2015, and a member of the Council of China Computer Federation.

**Min Chen** (M'08–SM'09) has been a Full Professor with the School of Computer Science and Technology, Huazhong University of Science and Technology (HUST), Wuhan, China, since February 2012. He is the Director of Embedded and Pervasive Computing Lab, HUST. He was an Assistant Professor with the School of Computer Science and Engineering, Seoul National University (SNU). He was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of British Columbia (UBC) for three years. Before joining UBC, he was a Postdoctoral Fellow at SNU for one and a half years. Prof. Chen received the Best Paper Award from QShine 2008, IEEE ICC 2012, ICST Industrial IoT 2016, and IEEE IWCMC 2016. He is the Chair of IEEE Computer Society Special Technical Communities on Big Data. He is as an Editor or Associate Editor of *Information Sciences*, *Information Fusion*, IEEE ACCESS, etc. He is a Guest Editor of IEEE NETWORK, IEEE WIRELESS COMMUNICATIONS, IEEE TRANSACTION SERVICE COMPUTING, etc. He is the Co-Chair of IEEE ICC 2012—Communications Theory Symposium, and the Co-Chair of the IEEE ICC 2013—Wireless Networks Symposium. He is the General Co-Chair of the IEEE CIT-2012, Tridentcom 2014, Mobimedia 2015, and Tridentcom 2017. He is a Keynote Speaker of CyberC 2012, Mobiquitous 2012, Cloudcomp 2015, IndustrialIoT 2016, and the 7th brainstorming workshop on 5G wireless. He has authored or coauthored more than 300 paper publications, including 200+ SCI papers, 80+ IEEE Transactions/Journal papers, 16 ISI highly cited papers, and 8 hot papers. He has published four books: *OPNET IoT Simulation* (HUST Press, 2015), *Big Data Inspiration* (HUST Press, 2015), *5G Software Defined Networks* (HUST Press, 2016), and *Introduction to Cognitive Computing* (HUST Press, 2017), a book on big data: *Big Data Related Technologies* (2014, Springer) and a book on 5G: *Cloud Based 5G Wireless Networks* (2016, Springer). His latest book (coauthored with Prof. Kai Hwang), is *Big Data Analytics for Cloud/IoT and Cognitive Computing* (Wiley, May 2017). His Google Scholars Citations reached 10 500+ with an h-index of 50. His top paper was cited 1000+ times. He received the IEEE Communications Society Fred W. Ellersick Prize in 2017. His research interests include cyber-physical systems, IoT Sensing, 5G networks, mobile cloud computing, SDN, healthcare big data, medical cloud privacy and security, body area networks, emotion communications and robotics, etc.

**Terence H. Chan** received the B.Sc. (Math), Master's, and Ph.D. degrees in information engineering from The Chinese University of Hong Kong (CUHK), Hong Kong, in 1996, 1998, and 2000, respectively. In 2001, he was a Visiting Assistant Professor with the Department of Information Engineering, CUHK. From February 2002 to June 2004, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Toronto. He was an Assistant Professor at the University of Regina from 2004 to 2006. He is currently an Associate Professor with the Institute for Telecommunications Research, University of South Australia, Adelaide, SA, Australia.

**Siu-Wai Ho** received the B.Eng., M.Phil., and Ph.D. degrees in information engineering from The Chinese University of Hong Kong, Hong Kong, in 2000, 2003, and 2006, respectively. During 2006–2008, he was a Postdoctoral Research Fellow with the Department of Electrical Engineering, Princeton University, Princeton, NJ, USA. Since 2009, he has been with the Institute for Telecommunications Research, University of South Australia, Adelaide, SA, Australia, where he is currently a Senior Research Fellow. He received the Croucher Foundation Fellowship for 2006–2008, the 2008 Young Scientist Award from the Hong Kong Institution of Science, and the Australian Research Council Australian Postdoctoral Fellowship for 2010–2013. His paper received the Best Student Paper Award in the 2016 Australian Communication Theory Workshop. His projects received the 2016 National iAward—Consumer Category from the Australian Information Industry Association and an honorary mention in the 2015 ComSoc Student Competition "Communications Technology Changing the World" organized by the IEEE Communications Society.