On Migratory Behavior in Video Consumption

Huan Yan, Haohao Fu, Yong Li, *Senior Member, IEEE,* Tzu-Heng Lin, Gang Wang,
Haitao Zheng, Depeng Jin, Ben Y. Zhao

***Abstract*—Today's video streaming market is crowded with various content providers (CPs). For individual CPs, understanding user behavior, in particular how users migrate among different CPs, is crucial for improving users' on-site experience and the CP's chance of success. In this paper, we take a data-driven approach to analyze and model user migration behavior in video streaming, *i.e.*, users switching content provider during active sessions. Based on a large ISP dataset over two months (6 major content providers, 3.8 million users, and 315 million video requests), we study common migration patterns and reasons of migration. We find that migratory behavior is prevalent: 66% of users switch CPs with an average switching frequency of 13%. In addition, migration behaviors are highly diverse: regardless large or small CPs, they all have dedicated groups of users who like to switch to them for certain types of videos. Regarding reasons of migration, we find CP service quality rarely causes migration, while a few popular videos play a bigger role. Nearly 60% of cross-site migrations are landed to 0.14% top videos. Finally, we validate our findings by building an accurate regression model to predict user migration frequency as well as user survey, and discuss the implications of our results to CPs.**

***Index Terms*—Video Consumption, Migratory Behavior**

I. Introduction

Video streaming has become one of the most popular online activities, which creates an enormous market with various content providers (CPs). Video streaming services for movies and TV shows (Netflix, Hulu, Amazon Video) already take over more than 70% of the peak time traffic in North America [1], [2]. Recently, the adoption of mobile devices and social networks further promotes the wide consumption of user-uploaded videos (YouTube, Vine) [3] and personal live streaming content (Periscope, Meerkat) [4].

Various video CPs have formed a giant ecosystem, where it is common for different providers to offer similar services and fiercely compete for users. In addition to a handful of highly successful CPs, many more have already failed in the competition such as Yahoo's Screen, Verizon's Redbox, Shomi and Foxtel [5], [6], [7], [8], [9]. To succeed or even survive in this ecosystem, each CP strives to provide the best user experience, *i.e.*, with more intelligent video recommendation mechanisms and faster content delivery infrastructures.

For CPs, retaining user engagement is critical and yet challenging. It not only requires a deep understanding of user behavior on their own services, but also how and why users leave them to a competitor. In recent years, various studies have examined user behavior and video consumption patterns by focusing on *individual CPs* and specific contexts [2], [10], [11], [12], [13], [14], [15], [16], [17], [18]. Given the broad differences of the video content and features of different CPs, it is critical to look at user video consumption by putting different providers in the same picture. We have taken a very tentative analysis on them in [19], [20], but what we learned is rather limited.

In this paper, we take a data-driven approach to understand video consumption *across multiple CPs*. In particular, we focus on user migration, *i.e.*, switching CPs during active video viewing sessions. Our goal is to measure the prevalence of user migration across providers and extract common migration patterns. In addition, we seek to explore possible reasons that cause users to migrate, and eventually build models to predict user migratory frequency.

We achieve these goals by analyzing a large-scale ISP dataset which covers video viewing sessions of 3,870,858 users in Shanghai city over two months from November 1 to December 31 in 2015. The dataset contains in total 315 million video requests to 6 most popular video CPs in China including Youku, IQiyi, Sohu, Kankan, LeTV and Tencent Video. We obtain this dataset via our collaboration with a major ISP in China. Both parties have taken careful steps to protect and anonymize user information in this dataset (details in *Data* Section).

To understand how and why users migrate from one provider to another, we analyze different possible factors such as temporal characteristics of video viewing sessions, video categories, popularity of the providers, video refreshing, and even users' device types (Section IV). Based on our observations, we build a video sequence model to characterize cross-site migration. By clustering users' video viewing sequences, we are able to identify different user groups where they exhibit unique migration patterns (Section V). Our analysis results lead to machine learning models to predict how likely users would migrate across CPs (Section VI-B). Meanwhile, we conduct a user survey about user migratory behavior to examine our findings, which is discussed in Section VI-C.

Results from empirical analysis and modeling show user migratory behavior is highly dependent on device type, content categories and video popularity. Our high-level findings can be summarized as the follows:

- First, user migration across CPs is highly prevalent. 66% of users are likely to migrate across multiple providers during video watching. This is especially true for the

H. Yan, Y. Li, T. Lin and D. Jin are with State Key Laboratory on Microwave and Digital Communications, Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (E-mail: liyong07@tsinghua.edu.cn).
H. Fu is with the College of Letters and Science, UC Berkeley.
G. Wang is with Department of Computer Science at University of Illinois at Urbana-Champaign (UIUC).
H. Zheng and B. Zhao is with Department of Computer Science of University of Chicago.

active users with 100+ views, where 96% of them switch providers.

- Second, user migratory behaviors are highly diverse. Regardless how big or small the CPs are, they all have their dedicated groups of users who like to switch to them for certain types of videos.
- Third, for mobile video service having more diverse types, there also exist diverse migratory patterns. To be specific, there are specified groups of users who are likely to migrate not only across different CPs with the same type of video service, but also across different types of services. This provides an important insight for CPs to offer diverse video services to compete users.
- Fourth, CP service quality does not have a significant impact on user migration. Instead, a small number of highly popular videos play an important role: 0.14% top videos are associated to nearly 60% of cross-site migration events.
- Fifth, user migration frequency as a key reflection of user migratory behavior is predictable, particularly on active users. The best performing regression model (Random Forest) achieves 0.83 correlation coefficient between the predicted and actual migratory frequency (for users with 1000+ views).
- Sixth, according to our user survey, we find that it is significantly important of CPs to provide exclusive videos to retain user engagement, which is not explicitly revealed in our data analysis.

To the best of our knowledge, our study is the first to systematically analyze user video consumption and migration across different CPs. Results from large-scale empirical data reveals new insights about the complex interactions between users and content providers, providing guidelines for CPs to retain user loyalty and on-site engagement.

This paper is an extension from our previous work [21] to take a more valuable analysis. The main differences can be summarized as follows. First, to further illustrate our motivation on studying the migratory behaviors, we add the analysis of how CPs compete for users through uploading the same videos. Second, CP residence time is newly defined to explore the user engagement among different CPs. Third, we use another dataset containing the viewing logs from multiple video apps to further illustrate migratory patterns in mobile video service. Importantly, to more thorough analysis of the effect of video popularity on migration, we study the fraction of number of views on popular videos against the number of sites visited, and identify its positive correlation. Meanwhile, we add the analysis of two important features used in the prediction of migratory frequency, *i.e.*, the average number of daily views per user accessing different CPs, and the viewing time per videos. Lastly, we take a user survey about user migratory behavior to further examine our findings.

II. Related Work

Video Access Behavior. User behaviors in online video systems have been examined in the context of Video on Demand (VoD) [10], [11], [12], [13], [14], [22], International Protocol Television (IPTV) [15], [16], [23], peer-to-peer (P2P) VoD [17] and live streaming [18], in different video services such as YouTube[13], 2008 Olympics [13], PPLive [11], [12], [18]. For example, Hoiles *et al.* utilize YouTube meta-data features and social features to understand user engagement and video popularity dynamics [24]. Authors in [25] characterize YouTube viewers by introducing a View Count Modeling (VCM) framework in order to understand the viewership behavior. Wu *et al.* study the video view patterns on Tecent Video and model them to predict the popularity [26]. Lin *et al.* analyze mobile video viewing patterns in PPTV system and provide the important guidelines for the design of peer-assisted video delivery networks [27]. Ma *et al.* explore the relationships between viewers and broadcasters on mobile personal livecast services [28]. Xie *et al.* study the effect of access types on video consumption patterns with the data from a large-scale VoD system [29]. Huang *et al.* employ Auto Regressive Moving Average (ARMA) model to predict video popularity on VoD system [30]. These works usually focus on one single system, and their analysis is also limited to a specific context. In contrast, our work collects a dataset of video viewing behaviors across six major video providers with contexts such as VoD, IPTV, live streaming. This allows us to understand user migratory patterns across CPs. Although Krishnan *et al.* analyze the influence of video stream quality on user behavior from multiple CPs using quasi-experimental designs [31], they do not have an in-depth study of characterizing migratory behaviors.

Temporal Patterns. For temporal analysis, Yu *et al.* propose a model for user arrival rate and video popularity [14]. Li et al. have reported their observations on daily and weekly patterns in a mobile VoD system [11]. Yin *et al.* focus on how the temporal dynamic nature of the system impacted user behavior [13]. Guo *et al.* model the video access patterns with stretched exponential distributions [32]. Jiang *et al.* focus on understanding the temporal patterns of viral videos and then predict their peak day based on the observations [33]. Pinto *et al.* study the influence of early views on video popularity, and use daily views during its early days to more accurately predict popularity [34]. Lian *et al.* apply clustering approaches to reveal the daily patterns of online video viewing behavior of smart TV viewers [35]. Kamiyama *et al.* find that the daily view count of YouTube videos conforms to a lognormal distribution and propose an extended multiplicative process (MPP) method to model it [36]. Instead of modeling daily/weekly patterns [11], [17], [37], we focus on more fine-gain video switching patterns in the scales of hours or even minutes across different CPs.

Geographic Patterns. Others researchers have studied the location diversity of video consumption [3], [12], [38], [39], [40]. For example, Cha *et al.* examine the geographical locality of an IPTV system [16]. Scellato *et al.* propose to use the location information in Twitter to predict the geographic popularity of YouTube videos [3]. Brodersen *et al.* study the popularity distribution of YouTube videos across different geographic regions, and identify its strong geographic locality [38]. Li

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2020.3043467, IEEE Transactions on Network and Service Management

TABLE I
NUMBER OF VIDEOS AND VIEWS PER CATEGORY (THE NUMBERS ARE DISPLAYED IN MILLIONS).

Category	# Views (10^6)	# Videos (10^6)
TV Series	115.5 (36.7%)	0.7 (7.4%)
Show	37.1 (11.8%)	0.8 (9.0%)
Movie	22.4 (7.1%)	0.2 (2.3%)
Cartoon	8.2 (2.6%)	0.2 (2.1%)
News	8.2 (2.6%)	0.3 (3.3%)
UGV	4.3 (1.4%)	0.3 (3.3%)
Others	119.3 (37.9%)	6.8 (72.7%)

TABLE II
STATISTICS OF THE 6 CONTENT PROVIDERS (THE NUMBERS ARE DISPLAYED IN MILLIONS). THE PENETRATION RATIO IS BASED ON INTERNET DEVELOPMENT REPORT OF CHINA.

Content Provider	YK	SH	LE	TC	IQI	KK
# Views (10^6)	131	74	35	33	32	10
# Users (10^6)	3.1	2.8	2.2	2.3	2.4	1.2
# Videos (10^6)	6.5	1.6	0.4	0.4	0.4	0.1
P2P service	N	N	N	N	N	Y
Social networks	N	N	N	Y	N	N
Penetr. Ratio	63%	46%	39%	54%	56%	33%

et al. investigate the uniformity of video requests across 33 geolocations in China, and find distinct popularity treads for popular and non-popular videos [12]. Our work focuses on migratory viewing behaviors in a metropolis city.

Migration Behavior in Social Network. There are some works about migration behavior in social network [41], [42], [43], [44]. For instance, Zhu *et al.* focus on the effects of overlapped membership on the survival of online communities [43]. Kumar *et al.* explore user migration patterns between social media sites [41]. Newell *et al.* investigate how and why users in Reddit migrate to other Reddit-like alternative platforms [42]. Unlike them, we study user migration behavior across different CPs in video consumption.

III. DATA

To study migration behavior in video consumption, we obtain a large-scale video viewing dataset from a major ISP in China via our collaboration. In the following we briefly describe our dataset and perform preliminary analysis.

A. Video Viewing Dataset

We obtain our dataset by focusing on 6 largest video content providers in China including Youku (YK), IQiyi (IQI), Sohu (SH), Kankan (KK), LeTV (LE), and Tencent Video (TC). They have the highest penetration rate in the market [45]. Note that all of them are Chinese domestic services, and most of their videos are free to watch. Because of the Great Firewall of China [46], large international video services like YouTube are not accessible in China and thus are neglected in our study. With the CP list, our collaborators at the ISP help to filter HTTP traffic to 6 CPs based the domain name of requested URLs. Note that all 6 CPs use HTTP protocol to deliver video content, which makes the filtering possible.

The resulting dataset contains the video viewing logs of 3,870,858 users in Shanghai city spanning over two months from November 1 to December 31 of 2015. This includes 315,069,400 viewing requests on 9,342,430 videos at the 6 CPs. Each viewing request is characterized by user ID, timestamp, device type and request URL. To obtain the detailed information about the video (*e.g.*, video category), we then use a web crawler to fetch the video URLs. This ISP network has an 85% of market share for the broadband access in China, which makes sure that our dataset provides a comprehensive view of video consumption across major CPs.

The user ID in our data is generated by the ISP, which is mapped to a device (*e.g.*, a smart phone, tablet or PC) instead of an IP address. We map users at the device-level, primarily considering different devices may lead to different video streaming experience for their screen size, network capacity and battery life. Using device-level ID helps to capture the fine-grained differences in video consumption. To protect user privacy, the user ID has been anonymized by the ISP (as a hashed bit string) before handling to us.

Ethics. Our study seeks to provide a better understanding of user video consumption and migration behaviors across content providers. The high-level goal is to help CPs to improve service quality for better user experience. Like existing studies [47], we obtain data via collaborations with the ISP who carefully removed personally identifiable information (*e.g.*, IP), and anonymized user ID before handing the data to us. Our study has received the approval from our university.

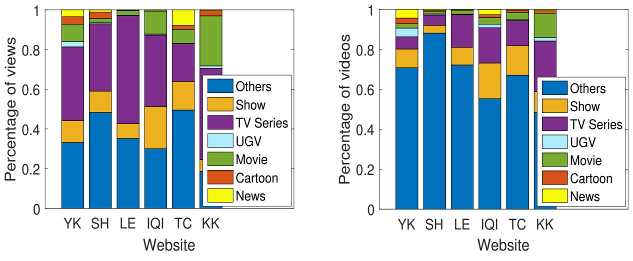
Impact of HTTPS. Since all six CPs use HTTP at the time of data collection, our study is not affected by HTTPS protocol. In the future, CPs may start to use HTTPS to encrypt the data. We believe most of our analysis metrics and methods can still be applicable in case HTTPS is used, *e.g.*, the timing and the sequence of the requests to different sites (which can still be identified by IP).

B. Preliminary Analysis

Next, we provide some preliminary analysis on video consumption across multiple providers. We seek to provide basic contexts for our later in-depth analysis.

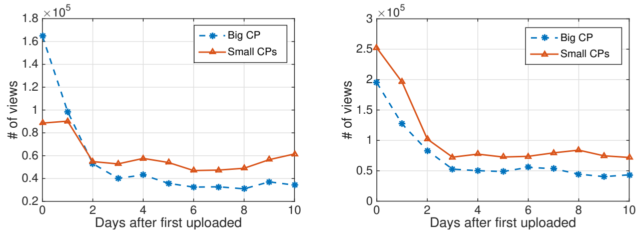
Video Category. Generally, video categories are labeled by CPs or video uploaders for convenient video search. Common video categories include “TV series”, “Show”, “Movie”, “Cartoon”, “News”, “User-generated videos (UGV)”. By resolving the video URLs, we collect these meta data labels from the respective CPs and classify videos into these 6 categories. Some videos have defunct URLs or have no category information, and we put them under “Others”. Table I shows the number of videos and views in each category. The most popular category is TV Series which has attracted 36.7% of the views with only 7.4% of videos. Note that the “Others” category, even though takes more than 70% of the videos, only attracts 30% of the views. Thus it should not impact our later investigations.

Differences and Similarities of Content Providers. Different CPs have their own emphasis and features. As shown in Table II, YK is significantly larger than the other five with more videos (6.5 million), views (131 million) and users (3.1



(a) Views (b) Videos

Fig. 1. Distribution of videos and views by categories in six CPs.



(a) Videos first uploaded by big CP (b) Videos first uploaded by small CPs

Fig. 2. Temporal distribution of views of the same videos since uploaded by big or small CPs.

million). This is consistent with the 2015 Internet report in China [45] where YK has the highest penetration ratio among all video services. In the rest of the paper, we regard YK as *big CP* and other CPs as *small CPs*.

The five small CPs are more specialized in providing certain types of videos (Figure 1). For instance, SH, LE and IQI are well known for movies, dramas and variety shows. TC's unique feature is the connection to a large social network Tencent QQ. TC also serves as a news portal where news are pushed through the social network. The impact is clear: even though "News" videos only take 0.5% of all TC videos, it has successfully drawn 8% of total views (Figure 1(b)). A counter example is IQI (with no social network): it also provides "News" (3%), but only draws less than 0.5% views. KK started its business for P2P downloading, and later expanded as a video streaming service specialized in providing "Movie" content.

Meanwhile, we find it is common that different CPs host the same video contents. By matching the video titles, we identify 102,297 videos hosted by more than one CP, which count for 19.76% of total views. This suggests intensive competitions among these CPs to attract users. To explore how CPs compete for users by uploading the same videos, we compute the number of views of the same videos per day since uploaded. Figure 2(a) shows the temporal distribution of views when these videos are first uploaded by big CP. For small CPs, the x-axis in this figure represents the uploading time gap with the granularity of day, which indicates that the same videos are uploaded after a few days. When the uploading time gap is 0, it means that small CPs may offer the same videos after a few minutes or hours. We can observe that at first two days big CP would attract more views, but later the view number on small CPs exceeds the big one. This shows that big CP may lose users when the same videos are also available on small CPs. Figure 2(b) is plotted when small CPs first upload the same videos. It can be observed that the view number on small CPs is larger than that on big one all the time. These results indicate that regardless of big or small CPs who first publish the same videos, small CPs can obtain higher user engagement. A possible reason is that there are a number of users who prefer to watching videos in small CPs. If they find the same videos available on small CPs, they would stay at their liked CPs for views. Thus, it is important of big CPs to provide more exclusive videos to attract views.

Given the above differences and similarities, user migration behavior would be highly complicated and also diverse for different CPs.

Mobile vs. PC. Video consumption from PC and mobile devices can be identified based on the device type. We find 30.4% of user IDs are associated to mobile devices, which contributes to only 13% of total video views. This suggests that PC is still the major platform for video viewing in China.

IV. Migratory Behavior Analysis

Our goal is to understand user video consumption and migratory behaviors across different CPs. We seek to answer two lines of questions. First, do users stick to one site or prefer to viewing at multiple sites? How often do users migrate across different providers? Second, what are the key factors that determine user migration patterns (*e.g.*, device types, popularity of CP, video categories, etc.)? To answer above questions, we first design a series of metrics to quantify user video viewing and migration, and then analyze the overall migratory behavior.

A. Metrics: Video Viewing and Migration

To measure users' video viewing, for a given user i, we model it as a sequence of viewing events: $Q_i = \{q_{i_1}, q_{i_2}, ..., q_{i_j}, ...\}$ with the corresponding timestamp $T_i = \{t_{i_1}, t_{i_2}, ..., t_{i_j}, ...\}$. We denote v_i^k as the total views in CP k from user i. The total number of views is defined as $V_i = \sum_{k=1}^{K} v_i^k$ $(1 \leq i \leq M)$ with M as the total number of users and K as the total number of CPs. The length of viewing sequence is N_i. Between two consecutive view events j and $j+1$, we denote $t_{i_j,i_{j+1}} = t_{i_{j+1}} - t_{i_j}$ $(1 \leq j < N)$ as the time gap. We add up the time gaps of n consecutive views at the k-th CP denoted by $d^k(n)$. $D^k(n)$ is the total number of such consecutive viewing "sessions". Finally, we denote $s_i^{k,k'}$ as the total number of times when user i migrates from CP k to CP k'.

We define migratory behavior as users switching CP during active video viewing sessions. Table III summarize the main definitions in our paper. To identify migration, we first need to determine if a session is still alive. This is decided by setting a threshold: if a user has not issued any request for a duration (x minutes), he/she is offline. To pick a reliable threshold, we need to first analyze the video length. We do so by crawling a random sample of 439,673 videos from six

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2020.3043467, IEEE Transactions on Network and Service Management

TABLE III
THE MAIN DEFINITIONS IN THE PAPER.

Term	Definition
User	Each user is identified by both IP address and device.
View	Each view is defined as a user requesting a video URL.
Event	Each event is denoted as a tuple of CP and video category.
Session	Each session is defined as a user consecutively viewing videos until he/she is offline.

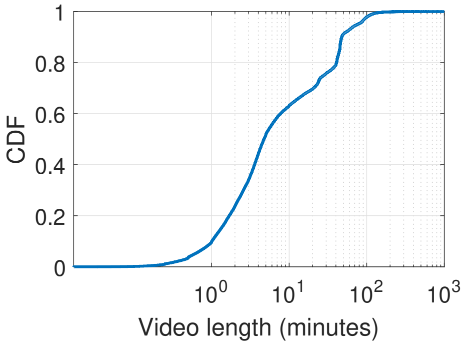


Fig. 3. CDF of video length.

CPs. As shown Figure 3, 99% of videos have a duration less than 100 minutes. Following existing work [48], we add an extra 20 minutes of inactive time for each active session. This produces a threshold of 120 minutes — if a user does not send any video requests for 120 minutes, she/he is offline, which is a relatively conservative threshold to capture most of migration behaviors.

To understand the overall user migratory behavior, we define three metrics to drive our analysis:

- **Migratory Frequency** measures how frequently users migrate between different CPs, defined as

$$\overline{F} = \frac{\sum_{i=1}^{M}\sum_{k=1}^{K}\sum_{k^{'}=1,k^{'}\neq k}^{K} s_i^{k,k^{'}}}{\sum_{i=1}^{M}(N_i-1)}. \tag{1}$$

Its value ranges from 0 to 1. In particular, $\overline{F}=0$ indicates all users only watch videos in a single CP. For user i, the migratory frequency can be computed as follows,

$$F_i = \frac{\sum_{k=1}^{K}\sum_{k^{'}=1,k^{'}\neq k}^{K} s_i^{k,k^{'}}}{N_i-1}. \tag{2}$$

- **CP Migratory Probability** measures how likely a user migrates from one CP to another. The probability of user i to migrate from CP k to $k^{'}$ is defined as:

$$P_{k,k^{'}} = \frac{\sum_{i=1}^{M} s_i^{k,k^{'}}}{\sum_{k^{'}=1}^{K}\sum_{i=1}^{M} s_i^{k,k^{'}}}. \tag{3}$$

- **CP Residence Time** measures the time a user spends on consecutively viewing videos on a CP. For CP k, residence time is

$$R_k = \frac{\sum_{n=1}^{max(N_i)-1} d^k(n)}{\sum_{n=1}^{max(N_i)-1} D^k(n)}. \tag{4}$$

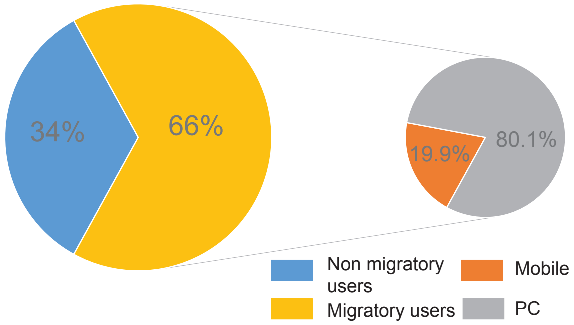
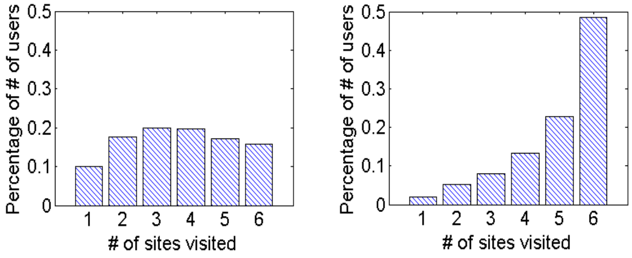


Fig. 4. Basic information about migratory users.



(a) All Users (b) Users with 100+ Views

Fig. 5. Distribution of users visiting different number of CPs.

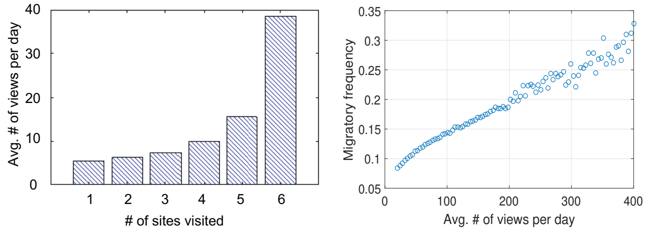
B. Prevalence of Migration Behavior

Since most videos on 6 CPs are free to watch, the impacts of pricing on the migratory behaviors among CPs can be neglected. First, we examine how often users visit multiple CPs. Figure 4 shows 66% of users visit more than one CP. According to (1), their average migratory frequency reaches 13%. Among these migratory users, the number of PC users is four times of that of mobile users. Further, the migratory ratio of users using PCs and mobile devices are 75.6% and 43.8% respectively, indicating that PC users prefer to switching CPs. In addition, for the active users with 100+ views (19.8% of users), the proportion of migratory users reaches 96%.

Figure 5 shows that the distribution of users accessing multiple CPs to watch videos. Most users visit more than one CP; while for users who have 100+ views recorded in 2 months, nearly 50% of them visit 6 CPs and only 2% stick to one single CP, which suggests that users do use multiple CPs to access video content.

To take a preliminary analysis of users who access multiple CPs, we compute average number of daily views per user visiting different number of websites. As shown in Figure 6 (a), we observe that average number of views per day exhibits a positive correlation with the number of websites. Then, we explore the relation between migratory frequency and average views per day in terms of users. From Figure 6 (b), we find that as average views increase migratory frequency become higher. These suggest that users who watch more videos during a time period of one day prone to switch CPs.

In summary, we observe that *migratory behavior is prevalent when users watch videos across different CPs, and PC users are prone to cross-site migration compared with mobile users.*



(a) Users visiting different number of websites

(b) Migratory frequency

Fig. 6. Average number of daily views per user against (a) different number of CPs and (b) migratory frequency.

TABLE IV
MIGRATORY PROBABILITY BETWEEN DIFFERENT CPS. CPS ARE LISTED IN THE COLUMN (ORIGIN) AND THE ROW (TARGET).

	YK	SH	LE	IQI	TC	KK
YK	89.8%	3.8%	2.1%	1.8%	1.7%	0.7%
SH	6.5%	84.9%	2.7%	2.6%	2.2%	1.1%
LE	8.7%	5.8%	76.9%	3.6%	3.1%	2.0%
IQI	9.2%	7.4%	4.7%	72.8%	4.0%	1.9%
TC	8.7%	6.0%	3.8%	3.9%	75.9%	1.7%
KK	12.7%	10.00%	9.1%	7.0%	5.9%	55.4%

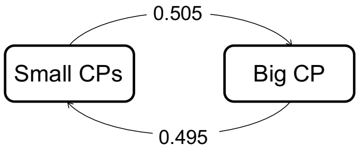


Fig. 7. Migratory probability between the big and small CPs.

TABLE V
MIGRATORY PROBABILITY ON DIFFERENT VIDEO CATEGORIES. THE CATEGORIES ARE LISTED IN THE COLUMN (ORIGIN) AND THE ROW (TARGET).

	TV Series	Show	Movie	News	Cartoon	UGV
TV Series	67.3%	16.2%	11.2%	2.6%	2.1%	0.8%
Show	47.8%	33.1%	12.3%	3.1%	2.5%	1.2%
Movie	37.3%	14.0%	43.5%	2.2%	2.2%	0.9%
News	54.4%	21.8%	14.0%	5.0%	3.0%	1.8%
Cartoon	51.1%	21.1%	15.5%	3.5%	7.1%	1.7%
UGV	48.3%	23.9%	16.9%	4.7%	4.0%	2.3%

TABLE VI
THE NUMBER OF VIEWS ON VIDEO CATEGORIES THAT USERS MIGRATE TO AT THE TARGET CP. THE COLUMN (ROW) REPRESENTS THE TARGET CP (CATEGORIES). (THE NUMBERS ARE DISPLAYED IN THOUSANDS)

	TV Series	Show	Movie	News	Cartoon	UGV
YK	4,445.2	1,135.6	1,570.3	237.6	268.7	176.4
SH	2,744.0	1,072.9	346.9	98.3	289.9	47.0
LE	2,770.3	645.5	237.9	14.2	1.9	3.3
TC	1,053.9	700.8	553.5	359.7	95.1	4.0
IQI	1,647.6	1,260.0	895.0	39.2	10.0	29.4
KK	1,055.6	162.0	799.4	9.7	60.0	34.0

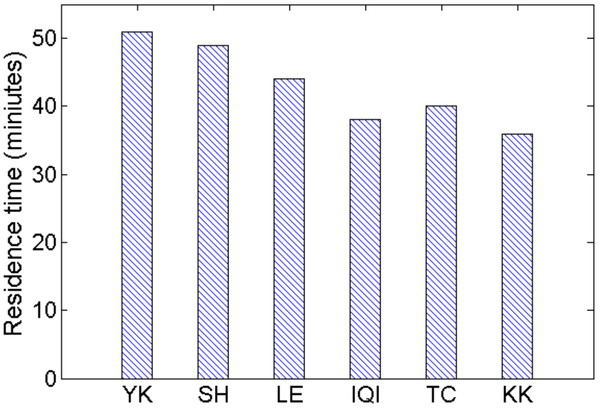


Fig. 8. Average estimated residence time.

C. CP Migration Probability

A series of critical questions for CPs are, when users leave their site to a new one, where do users go, and why they leave. We analyze the *migratory probability* across CPs in Table IV. We observe the highest values at the diagonal of the matrix. After watching each video, users generally have a higher probability of staying on the same site than migrating to a different CP. Among different CPs, YK (the largest site) has the highest chance to keep their users, while KK (the smallest site) is mostly likely to lose their users to other CPs. When migration happens, YK is also the most probable destination. This indicates the size/popularity of the CP matters. However, comparing YK (big) with the all other 5 sites together (small), the difference becomes less significant (Figure 7).

We seek to further understand whether video category influences the migratory behaviors by showing the results in Table V. The most dominating trend is that users would migrate to more popular video categories such as “TV” or “Show” during the migration. Besides, we also observe some users would switch site for the same category of the videos (the numbers along diagonal are slightly higher than the nearby numbers *e.g.*, “Movie” and “Cartoon”). Further, we consider whether the categories of videos viewed before and after migration are the same or not and compute the average migratory probability. The obtained results show that the probability of migrating to the same category is relatively higher (52.56%), which suggests that users are more likely to watch the same video categories during the migration.

Then, we examine which categories users migrate to when switching to a particular site. As shown in Table VI, for “TV” and “Movie” , the largest site YK has a dominating influence over the other sites. However, we do observe that smaller CPs’ unique features help them to draw users during migration. For instance, TC, with the help of its social network to push “News” videos, has received most views when users switch site to watch “News”; while SH is currently doing well in “Cartoon” category. Although IQI and KK are two smallest CP, they attract more “Movie” views from migration right after YK due to their emphasis on Movie content.

Overall, we conclude that *regardless of big or small CPs, certain users prone to migrate to them. Further, during the migration, users are more likely to switch to the same video category.*

D. CP Residence Time

Although we have analyzed the overall migratory behavior at different CPs, it does not consider the time factor involving user engagement on different CPs. Thus, we introduce *CP Residence Time* that estimates how a CP can retain users for consecutive views. We examine the residence time of users

TABLE VII
POSSIBLE USER BEHAVIORS THAT CORRESPOND TO DIFFERENT TIME GAPS.

Time Gap	Possible Behavior
$(0, 1min]$	Scanning through video pages quickly
$(1min, 30min]$	Watching short video clips
$(30min, 1h]$	Watching TV series
$(1h, 2h]$	Watching movies
$(2h, +\infty]$	Taking a break (offline)

on each CP in Figure 8, as a proxy for the level of user-engagement. The results show that the estimated residence time on YK is longest (51 mins), but not significantly longer compared to the rest CPs (*e.g.*, the shortest one is KK of 36 mins). This is consistent with our earlier observation: even though YK has a clear advantage in attracting users, other smaller CPs can still maintain a similar level of user engagement once users get on their sites. The above analysis further illustrates the importance of our study on the migratory behavior.

V. CLUSTERING MIGRATORY PATTERNS

Thus far, we have analyzed users' video consumption and migratory behaviors by treating users as a single population. However, there could be different user behaviors within this population. Now, we explore what are the major types of user migratory behavior over multiple providers? How can providers retain user engagement for different user types? To answer these questions, we apply an unsupervised mining method to cluster users' video viewing sequences.

A. Viewing Sequence Clustering

Now, we build an unsupervised model to identify groups of prevalent behaviors among users by clustering the user viewing sequences. This is done by building a similarity graph for viewing sequences, where each node in the graph represents a user and the edges are weighed based on the "similarity" of sequences. Partitioning the graph produces clusters of users with similar activities.

Viewing Sequences. Each user's viewing sequence is a sequence of video viewing events with the time gaps between events. We model the sequence of user i as $\{q_{i_1}, t_{i_1,i_2}, q_{i_2}, t_{i_2,i_3}, q_{i_3}...q_{i_N}\}$, where q_{i_j} is the j_{th} event of the user, t_{i_j-1,i_j} is the time gap between two click events. To capture both CP and video category, each event q is denoted as a tuple of them, *e.g.*, $(YK, Movie)$. For easy comparison of sequences, we also discretize the time gaps as events. In this paper, we classify the time intervals as $(0, 1min]$, $(1min, 30min]$, $(30min, 1h]$, $(1h, 2h]$, $(2h, +\infty)$ that correspond to possible viewing behaviors (Table VII). This classification is based on the estimated video length of each category as well as the threshold of active video viewing sessions discussed in Section IV-A.

Similarity Graph and Partitioning. Our high-level intuition is that user behaviors would form clusters, *i.e.,* users behave similarly at certain aspects. To capture such clusters, we map user's viewing sequences into a similarity graph [49] and partition the graph to produce groups of users with similar activities. In this graph, each node is a user, and the edges measure the similarity of any two sequences. Our similarity metric considers the visited CP, video category and time gap at the same time. For a given sequence $Y = (y_1, y_2, ..., y_n)$, we compute all possible subsequences (or k-grams) as $\Phi_k(Y) = \{y(k)|y(k) = (y_{2i-1}, y_{2i}, ..., y_{2i+k-2}), i \in [1, \frac{n+2-k}{2}]\}$. Then, given two sequences, we measure their similarity based on common subsequences $C_k(Y_i, Y_j) = \Phi_k(Y_i) \cup \Phi_k(Y_j)$, and the frequency of each subsequence $[e_{\nu,1}, e_{\nu,2}, ..., e_{\nu,T}]$ ($\nu = i, j$, $T = |C_k(Y_i, Y_j)|$). The sequence similarity metric is computed by Tanimoto coefficient:

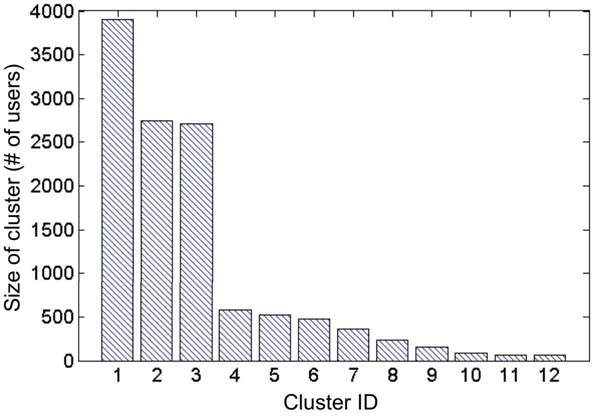


Fig. 9. Number of users in top 12 clusters.

$$Z_k(Y_i, Y_j) = \frac{\sum_{m=1}^{T} e_{i,m} e_{j,m}}{\sum_{m=1}^{T} e_{i,m}^2 + \sum_{m=1}^{T} e_{j,m}^2 - \sum_{m=1}^{T} e_{i,m} e_{j,m}}, \quad (5)$$

which considers both the direction and magnitude of two vectors. We set $k = 5$ for our analysis following the settings in [49]. To identify clusters in the graph, we use the *Divisive Hierarchical Clustering Algorithm* [50], which is suitable for finding arbitrary cluster shapes.

B. Sequence Clustering Results

Data Clustering. Building a complete similarity graph is too costly given the size of our dataset ($O(n^2)$). Thus, we rely on sampling to build similarity graph by seeking to give a fair consideration for users who visit different number of CPs. More specifically, we randomly select 2000 users from those who visit x sites, where $x = 1, 2, .., 6$. In total, this gives us 12,000 users to build a similarity graph. After clustering, we obtain in total 24 clusters (the number of clusters is determined by clustering quality metric: modularity [51]). For our analysis, we focus on the largest 12 clusters shown in Figure 9, which covers 99% of the selected users.

Cluster Analysis. To help the understanding the behaviors captured by each cluster, we select key features (*i.e.*, subsequences) that are strongly associated to each cluster. More specifically, we use Chi-square statistics (χ^2) [52] to measure the correlation of a feature and a cluster.

It can be computed as

$$\chi^2(t, c) = \frac{M(AD - CB)^2}{(A + C)(A + B)(C + D)(B + D)}, \quad (6)$$

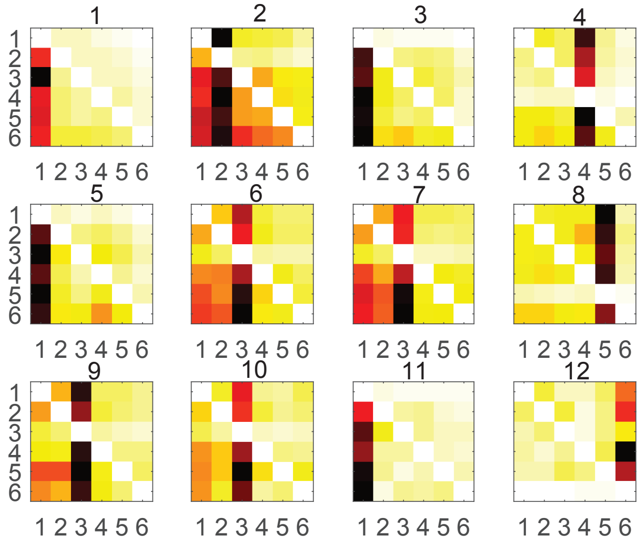


Fig. 10. CP migratory probability for the top 12 clusters. Each heatmap represents one cluster. The column (row) represents the origin (target) CP where numbers 1–6 represent YK, SH, LE, IQI, TC, KK respectively.

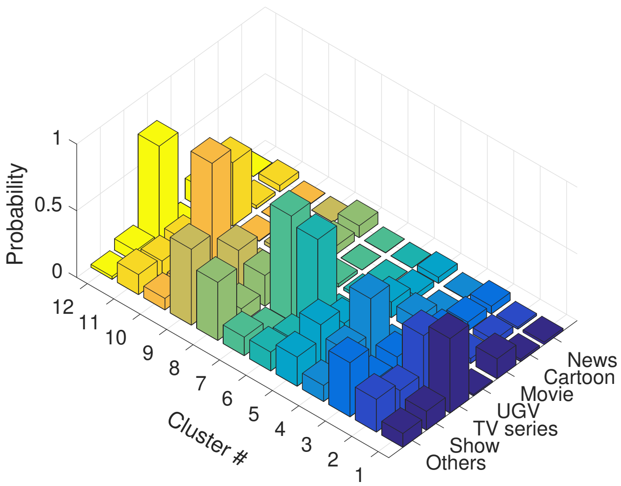


Fig. 11. Probability distribution of video categories that users migrate to in each cluster.

where t and c represent the selected feature (subsequences) and cluster respectively. A is the frequency of feature t occurring in cluster c, B is the frequency of feature t occurring outside cluster c, C is the frequency of other features occurring in cluster c and D is frequency of other features occurring outside cluster c. $\chi^2 = 0$ if the feature and the cluster are independent; A higher χ^2 score indicates a stronger association for the feature and the cluster.

To obtain the top feature of each cluster, we select the highest χ^2 score of each feature based on the distribution of users in the cluster and outside this cluster. Since we focus on users' cross-site migratory patterns, we choose the top feature satisfying such patterns in each cluster. To intuitively show the result, we first plot a heatmap in Figure 10. It shows the probability of migrating from one CP (column) to another CP (row) with darker color represents a higher migratory probability. In the meantime, we also examine what video categories users in each cluster are likely to migrate to in Figure 11. Our results confirm our intuition that users do have very different migratory behaviors. For instance, cluster 1, 3, 5 and 11 have users who are likely to migrate to YK, but the target video categories are different. For example, users in cluster 1 are likely to migrate to YK to watch TV series, while users in cluster 11 are likely to migrate to YK to watch movies. Cluster 2 has users who often migrate to SH to watch TV series; Even for the smallest CPs such as TC and KK, there are dedicated clusters of users who are likely to migrate to them (cluster 8 and 12). These results confirm that even smaller CPs can still receive preferences from certain types of users and the migration behavior exhibits great differences for different groups of users. For service providers, understanding such migration behavior patterns can help to engage their users. By classifying users into different migration patterns, CPs can make better recommendations on intended videos on the same site to keep users from migrating to other CPs.

Finally, we analyze the time intervals for users to migrate from one CP to another. For different clusters, we do not observe significant differences. This suggests that temporal features are less important in identifying migratory behaviors compared to video categories and CP preferences.

In summary, our results further validate that *regardless of big or small CPs, users all have their dedicated groups, where they like to switch for certain video categories.*

C. Migratory Pattern Analysis on Mobile Video Service

In recent years, mobile video services are rapidly changing, *e.g.,* a growing number of people prefer to visiting the mobile video apps for video viewing by their mobile devices. Thus, for users accessing to mobile video service, a key question is what are the major patterns of migratory behavior over different providers. To investigate it, we use the dataset from [53] to model the users' video viewing sequences. This dataset records the logs of user clicks from 34 popular mobile video apps[1] classified into three kinds of mobile video service, including traditional video portals, user-generated video service and personalized livestreaming [53]. With it, we adopt our proposed clustering method to characterize the usage patterns of mobile video service. As previously mentioned, it is very time-consuming to construct a complete similarity graph. Thus, we randomly choose 10,000 users and filter the users with smaller than 3 clicks. Since we do not have the information of video category due to the limitation of dataset, we denote each event q as the app where the user clicks when modeling user viewing sequences.

Finally, we obtain 8 main clusters containing 98.2% of selected users. Table VIII shows the selected top features and their scores in 8 clusters. We observe that these top features corresponding to user behaviors are mainly intra-CP switching.

[1] The three types and their corresponding apps are traditional video portals (Tecent Video, Sohu Video, IQiyi, 56 Video, Youku, LeTV, Baofeng, PPTV, Tudou, CNTV, Kankan, Baomihua, Ku6, Migu Video, TV189, 360 Video, PPStream, Mangguo TV), personalized livestreaming (YY, Douyu, Inke, Huya, Huajiao, Fanxing Chushou TV, Panda TV, Quanmin TV, KKTV), and user-generated video service (Bilibili, 51 TV, AcFun, Miaopai, Meipai, Xiaojiaxiu).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2020.3043467, IEEE Transactions on Network and Service Management

TABLE VIII
TOP 8 CLUSTERS WHERE WE SELECT TOP FEATURES PER CLUSTER FOR INTRA-SITE SWITCHING AND CROSS-SITE SWITCHING. FOR BREVITY, WE USE 1M, 30M, 1H, 2H AND 2H+ TO REPRESENT EACH TIME BUCKET RESPECTIVELY.

ID	Users	Chi^2	Top Feature (Intra & Cross-site)

1	1625	17870.9
		88.5
			{Tecent Video, 1m, Tecent Video, 1m, Tecent Video}
			{Miaopai, 2h+, Tecent Video, 1m, Tecent Video}

2	1465	27494.5
		987.7
			{IQiyi, 1m, IQiyi, 1m, IQiyi}
			{Tecent Video, 2h+, IQiyi, 1m, IQiyi}

3	1099	26265.0
		1187.7
			{Youku, 2h+, Youku, 2h+, Youku}
			{Tecent Video, 2h+, Youku, 1m, Youku}

4	997	72986.3
		1588.3
			{360 Video, 2h+, 360 Video, 2h+, 360 Video}
			{360 Video, 2h+, 360 Video, 2h+, Tecent Video}

5	981	14849.4
		1740.6
			{Miaopai, 1m, Miaopai, 2h+, Miaopai}
			{Bilibili, 1m, Bilibili, 2h+, Miaopai}

6	432	19228.5 1342.0	{Douyu, 1m, Douyu, 1m, Douyu} {Youku, 30m, Tudou, 1m, Tudou}
7	260	915.0 129.0	{PPStream, 1m, PPStream, 1m, PPStream} {IQiyi, 1m, PPStream, 1m, PPStream}

8	239	44106.3
		179.9
			{PPTV, 1m, PPStream, 1m, PPStream}
			{Miaopai, 2h+, Miaopai, 2h+, Tecent Video}

Further, the usage patterns in Cluster 1, 2, 6, 7 and 8 are the quick clicks within the apps, with the reason that it is more convenient for users to use mobile apps to watch online videos. Meanwhile, some apps that provide the video service different from the traditional video portals (*e.g.,* Youku) also attract a group of users to view. For example, the users in Cluster 6 quickly click the app of Douyu that offers the platform for personalized livestreaming. With a focus on the cross-site migratory patterns, different groups of users who behave distinctively in switching apps for views. For example, the users in Cluster 1 migrate to IQiyi after 2 hours while Cluster 6 has users who switch from Youku to Tudou in less than 30 minutes. In particular, it is observed that a group of users are likely to migrate to the app that offers a different kind of video service, which is unexplored in previous analysis. For example, Cluster 1 has the users who switch from Miaopai to Tecent Video, where the former has expertise on making the user-generated videos while the latter belongs to traditional video portals. This indicates that users may leave to a competitor since different kinds of video services are required, which provides a valuable reference for CPs to offer diverse video services to retain and compete users. Based on above analysis, we conclude that there are diverse migratory behaviors in mobile video service.

VI. MIGRATION REASONS & PREDICTION

Thus far, our results suggest users prone to migrate from one site to another for video consumption. For individual CPs, it is crucial to understand the reasons why users leave their sites and migrate to competitors. This allows CPs to develop more targeted mechanisms to retain user engagement and loyalty. In this section, we first analyze the possible reasons behind migration. Then, we validate our findings with a prediction model for user migration.

A. Migration Reasoning

We explore migration reasons from two aspects: CP and video. First, CP's (poor) service quality may be an important factor that triggers user migration; Second, for videos, the popularity of video may influence the users' viewing and migration behavior.

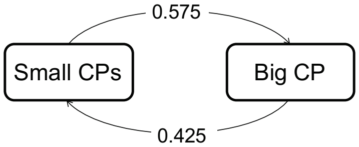


Fig. 12. Migratory ratio between the big and small CPs on migration events caused by refreshing failure.

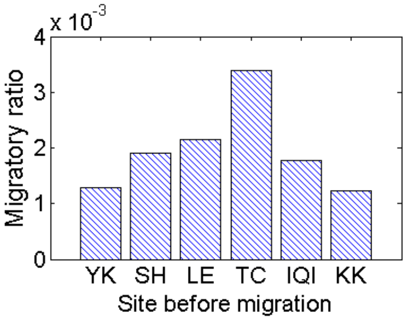


Fig. 13. Ratio of migratory due to refreshing failure (threshold=30s).

1) CP Service Quality: If a user sends multiple consecutive requests on the same video but fails to view it due to long startup or rebuffer delay, she/he may switch to a new site for videos in a short time. To better investigate such phenomenon, we detect *video refreshing* events if a user sends at least two consecutive requests on the same video within a small time threshold. Further, we define *refresh failure* if a user immediately starts to watch videos at a different site after refreshing. We set the time threshold as 30 seconds and evaluate the sensitivity of the threshold later.

To quantify the influence of CP service quality, we count the migratory ratio of refresh failure considering the switching direction of the big and small CPs (Figure 12). We observe that there is no significant difference for the ratio between these two types of migrations. Further, we compute migratory ratio as the number of refresh failures divided by the number of migrations in the same site. As shown in Figure 13, the migratory ratio is smaller than 0.4% at all six sites, and average ratio is only 0.19%, which suggests refreshing failures rarely

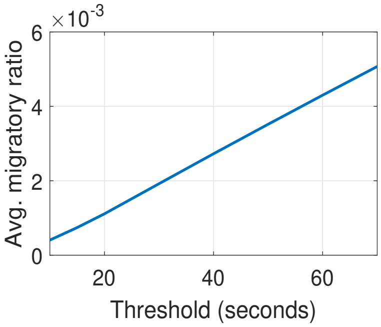


Fig. 14. Migratory ratio with different time threshold.

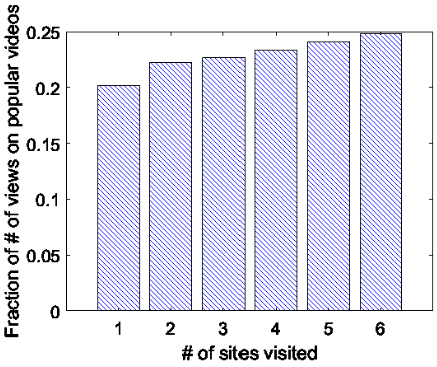


Fig. 15. Fraction of number of views on top 1000 videos against the number of websites visited.

cause migration regardless of big or small CPs. To check the sensitivity of the threshold, we show the migratory ratio with different thresholds in Figure 14. We find that the average migratory ratio increases with the threshold but is still smaller than 0.06% even when it increases to 70 seconds. Based on these two observations, we conclude that CP service quality has minor impacts on user migration.

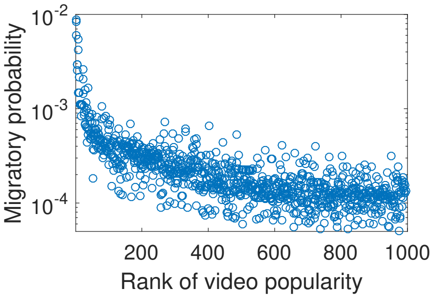


Fig. 16. Migratory probability for watching popular videos. Popular videos are ranked by # of views.

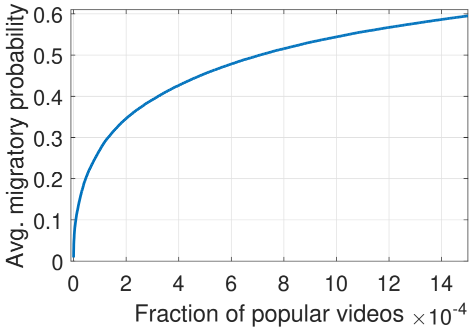


Fig. 17. Ratio of migration events to watch popular videos over all migration events.

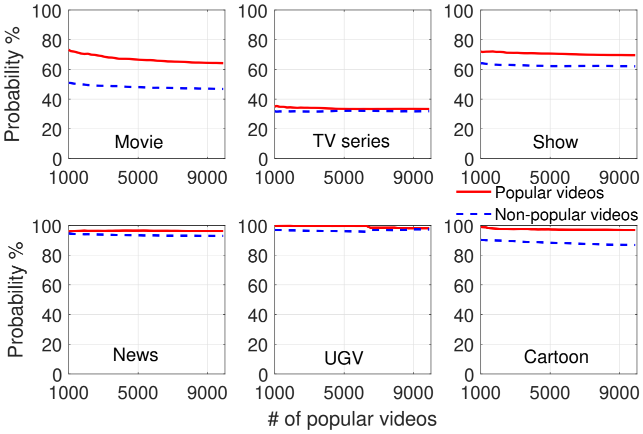


Fig. 18. Probability that a user migrates to watch a non-popular/popular video of a different category from the video she watched before migration.

2) Video Popularity: Popular videos attract more user viewing, and are likely to trigger user migration. To investigate it, we calculate the fraction of views on top 1000 videos against the number of sites visited. In Figure 15, we observe that the fraction of views increases as the number of sites visited. This suggests that users who visit more sites prefer more popular videos. Further, we rank the video popularity by number of views, and plot the probability of migrating to popular videos over the total number of migrations (Figure 16). From the results, we observe that migratory probability exhibits a negative correlation with the rank of video popularity, which suggests that more popular videos cause more migrations. On the other hand, we compute the average probability among top popular videos and show the results in Figure 17. We find that a very small fraction of popular videos counts for most of the migrations. Nearly 60% of cross-site migration is landed to 0.14% top videos. Even for the top 1000 extremely popular videos (0.011% of all videos), they trigger 27.73% of all the cross-site migration. In comparison, the probability that users watch these popular videos within the same site is 23.5%, which is smaller than that of user migration.

A key question is, are users intentionally looking for these videos or do they reach popular videos due to the recommendations of destination CP. To explore it, we treat users *watching popular videos in a different category* after migration, as a signal of being distracted by the destination CP's recommendation. We compute the probability of migrating to *different categories* of popular videos after migration, and use non-popular videos as comparison. As shown in Figure 18, except for "Movie", there is no significant difference ($< 10\%$) in the probability of changing categories between popular or non-popular videos. There is a significant 20% difference for the movie category. It is possible that after watching a long movie, users are more likely to migrate to another site to watch recommended videos in other categories.

As a brief summary, we obtain two enlightened findings: 1) *CP service quality does rarely cause the user migration from*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2020.3043467, IEEE Transactions on Network and Service Management

TABLE IX
THE FEATURE LIST FOR PREDICTION OF USER MIGRATORY FREQUENCY.

ID	Feature
1	Fraction of # of views on popular videos
2	Device type used for watching videos
3	# of video categories viewed
4	Avg. # of views per day
5	# of offline
6	Avg. # of views during consecutive sessions
7	Avg. estimated viewing time per video

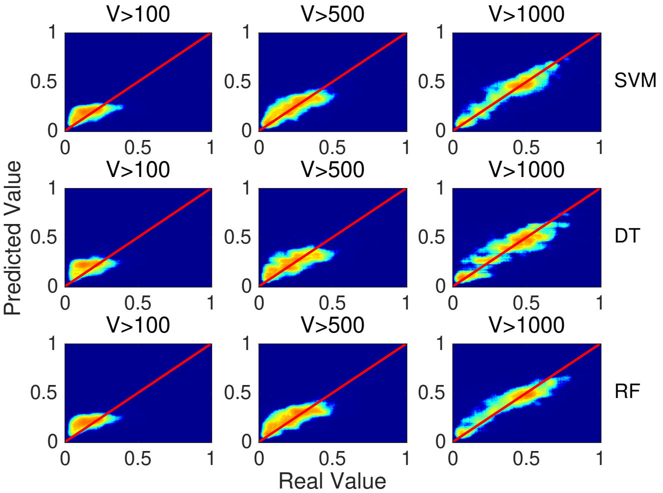


Fig. 19. Correlation between prediction result and real value for users with different number of video viewing events (V denotes # of views per user).

one CP to another; 2) *users prefer to migrating to another site for popular videos. Especially, when watching movies and doing the migration, they are more likely to watch popular videos with other categories*.

B. Migration Prediction

So far, we analyzed user migration behavior and possible reasons. We next build a prediction model to validate our findings. More specifically, we seek to predict migratory frequency, *i.e.*, how frequently a user would switch CPs (see Equation 2). This metric can be useful for CPs to estimate user loyalty and re-engage unsatisfied users.

Our prediction is based on regression models to predict migratory frequency. We first select key features for each user based on early obtained insights. As shown in Table IX, these features include: the fraction of views on popular videos over all the videos the user watched, the user's device type, number of video categories previously watched, average number of views per day. We also include features to characterize the video streaming sessions such as number of offline events, number of views per session, and estimated viewing time per video.

Based on these features, we build regression models using three widely used machine learning methods: Support Vector Machine (SVM) [54], Decision Tree (DT) [55] and Random Forest (RF) [56]. In our experiment, we select 10,000 users with 100+, 1000+, 1000+ views respectively and run 5 fold cross-validation. We use Root Mean Square Error (RMSE) to evaluate the accuracy between predicted and real value. Table X lists RMSE value of three models. From the results, we find that their values are smaller than 0.1, and RF can achieve most accurate one. Further, we use heatmap to intuitively illustrate the correlation between the predicted and real value (Figure 19). If each predicted value matches real value perfectly, all the dots would be distributed along with the diagonal. The results show that our prediction models are effective. The correlation coefficients between predicted and real values are listed in Table XI. For users with 1000+ views, our models predict migratory frequency with a correlation over 0.79 (regardless SVM, DT or RF). Among different models, RF is the most accurate one (0.83).

TABLE X
THE RMSE OF PREDICTION AMONG THREE PREDICTION METHODS.

Method	Views>100	Views>500	Views>1000
SVM	0.06	0.0877	0.0958
DT	0.0641	0.0902	0.0996
RF	0.0602	0.0839	0.0913

TABLE XI
THE CORRELATION COEFFICIENT BETWEEN PREDICTED AND REAL VALUE FROM THREE PREDICTION METHODS.

Method	Views>100	Views>500	Views>1000
SVM	0.66	0.73	0.8
DT	0.6	0.71	0.79
RF	0.66	0.76	0.83

To explore the importance of features, we compute the feature weights of three methods in Table XII, where a higher weight indicates more important feature. Note that the three models compute weights differently: SVM uses sensitivity analysis on features [57], while DT measures the goodness of each split inside the tree [55]. RF measures the decrease in node impurities on features [56]. Despite the differences, the top feature across all three models is consistent. The fraction of views on popular videos is still the strongest indicator of user migration. In addition, we identified a new feature *the average viewing time per video* (feature 7), which is also highly indicative of migration. We examine the relationship between migratory frequency and average viewing time per video, and plot the result in Figure 20. We observe that when users watch short videos (< 20 minutes) migratory frequency increase with average viewing time of each video. In contrast, when watching videos longer than 20 minutes, users do not frequently switch CPs. Intuitively, if a user constantly closes videos before finishing, it indicates unsatisfying experience and a tendency to migrate.

Note that our prediction experiments are not intended to provide off-the-shelf prediction tools for individual CPs. As shown in Table IX, certain features require a global view of user traffic data. Instead, we use the prediction model for inquiry and validation on our early findings, and identify new signals to predict migration (*e.g.* average viewing time per video).

C. User Study

To further evaluate our findings, we develop an online survey about users' viewing behaviors. We design a questionnaire containing 6 questions in the survey[2]. To avoid multiple

[2] The full question list is available at https://www.dropbox.com/s/7w9q1zvbkiwe814/Questionnaire.pdf?dl=0

TABLE XII
THE WEIGHTS OF FEATURES OF THREE METHODS.

ID	RF Weight	SVM Weight	DT Weight
1	0.50	0.26	0.56
7	0.13	0.15	0.12
4	0.13	0.04	0.09
6	0.08	0.13	0.09
5	0.07	0.09	0.05
2	0.05	0.19	0.06
3	0.04	0.14	0.03

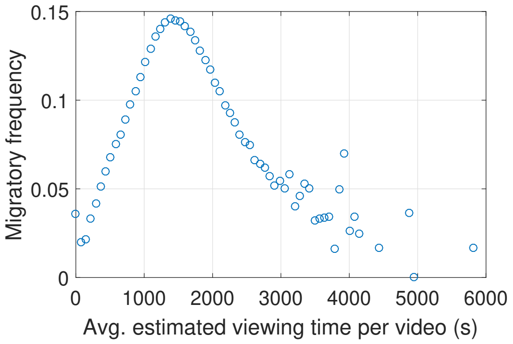


Fig. 20. Migratory frequency against average estimated viewing time per video.

submissions from the same respondent, we set up a tool that can examine the browser cookies.

We invite the people in the university to fill in this questionnaire since they are active in watching online videos. Finally, 107 respondents complete the survey in total. Table XIII shows the specific reasons why users choose a specified CP for video views, where the number and percentage of respondents are ranked in descending order. From the result, we find that providing exclusive videos is a most important factor influencing user engagement on different CPs. Besides, 28.97% of respondents claim popular videos also attract them, and its rank is right after the exclusive videos. Further, we observe that 34.58% of respondents would switch CPs during active viewing sessions. Among these respondents, 67.57% of them switch CPs for popular videos, while the percentage of respondents who choose service quality only occupies 29.73%. This is consistent with our previous analysis (Section VI-A). In addition, from Table XIV, we observe that different respondents have their CP preferences (big or small CPs) when switching. Moreover, in our survey, we find that 78.38% of respondents prefer to switching for certain video categories. These further validate our findings in Section V.

D. Practical Implications

Our results have a number of practical implications to CPs.

First, we identified a number of factors that contribute to user migration across CPs. This provides guidelines for CPs to optimize their services. 1) In comparison with the networking service quality, video contents should be paid more attention, which is beneficial of CPs to compete with their competitors. In particular, identifying and indexing trending videos across the Internet can help to engage their users. Also, developing their uniquely featured video categories (*e.g.*, News for TC, Movies for IQI and KK) helps to attract incoming migrants, even from larger sites. 2) CPs should pay attention to video recommendation in the same video categories — users often migrate to other CPs to watch (trending) videos of the same category as they watched before migration.

TABLE XIII
THE REASONS FOR CHOOSING A SPECIFIED SITE TO WATCH VIDEOS.

ID	Reasons	Number	Percentage
2	Exclusive videos	48	44.86%
1	Popular videos	31	28.97%
3	Good service quality	18	16.82%
4	Good video recommendation	10	9.35%

TABLE XIV
THE PERCENTAGE OF RESPONDENTS WHO PREFER TO SWITCHING TO A SPECIFIED CP.

	YK	SH	LE	TC	IQI	KK
Percentage	18.92%	0	0	37.84%	40.54%	2.7%

Second, our experiments above show that migration behavior is predictable. However, in practice, there are challenges to make the prediction tool directly available to individual CPs since certain features require a global view of the network traffic. This gives ISPs the opportunity to provide services to CPs, to compute global features on their behalf. Future research will be needed to guarantee CPs cannot reverse-engineer a user's detailed browsing traces from these statistical features. In addition, we find other signals that do not require global statistics (*e.g.*, average viewing time per video). This can help individual CPs to estimate users' likelihood of migration, and deploy targeted engagement mechanisms.

E. Limitations

There are a few limitations in our study. First, our study primarily focuses on Chinese video streaming market. Future research is needed to expand the analysis scope (when related data becomes available). Second, we study user migratory behavior based on the video viewing records collected 4 years ago. Although today's user viewing behavior may be different due to the proliferation of smart devices and the appearance of new video services (*i.e.,* live streaming), our analysis can still provide valuable insights. On one hand, our metrics and methods are defined based on some basic information including user ID, timestamp, and request URL, and can be applied to other similar dataset. For example, the proposed sequencing model only needs the above information as the input, thus it can be generalized to predict the migratory patterns. On the other hand, the analysis results can offer some insightful guidelines for CPs to improve their services. For example, we find that it is helpful of CPs to provide more popular video contents and develop featured video categories to retain users.

VII. Conclusion

To the best of our knowledge, this is the first study to systematically analyze user video consumption and migratory behaviors across different content providers. We not only uncover the overall patterns of how users migrate from one CP to another, but identify distinct groups of users with highly different migratory behaviors. Moreover, we study the potential reasons about user migration, which leads to an

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2020.3043467, IEEE Transactions on Network and Service Management

accurate prediction model for migration frequency. In addition, we conduct a user survey to further validate our analysis, and obtain new findings. CPs can utilize these findings to improve their services and better engage users. As future work, we plan to investigate long-term migration behavior across CPs.

References

[1] E. Protalinski, “Streaming services now account for over 70% of peak traffic in north america, netflix dominates with 37%,” Venture Beat, 2016.

[2] D. Karamshuk, N. Sastry, A. Secker, and J. Chandaria, “On factors affecting the usage and adoption of a nation-wide tv streaming service,” in *Proceedings of INFOCOM*, 2015, pp. 837–845.

[3] S. Scellato, C. Mascolo, M. Musolesi, and J. Crowcroft, “Track globally, deliver locally: Improving content delivery networks by tracking geographic social cascades,” in *Proceedings of WWW*, 2011, pp. 457–466.

[4] B. Wang, X. Zhang, G. Wang, H. Zheng, and B. Y. Zhao, “Anatomy of a personalized livestreaming system,” in *Proceedings of IMC*, 2016.

[5] N. McNiel, “Yahoo! shutters their streaming video service. yahoo! had a streaming video service?” TheAmericanGenius, 2016.

[6] J. bradshaw, “Video-streaming service shomi to shut down at end of november,” The Globe and Mail, 2016.

[7] C. Reilly, “Goodbye presto: Foxtel’s streaming service to shut down,” CNet News, 2016.

[8] T. Spangler, “Verizon, redbox to pull plug on video-streaming service,” Variety, 2014.

[9] C. Reilly, “Ezyflix goes dark as streaming battle claims its first casualty,” CNet News, 2015.

[10] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang, “Understanding the impact of video quality on user engagement,” *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4, pp. 362–373, 2011.

[11] Z. Li, J. Lin, M.-I. Akodjenou, G. Xie, M. A. Kaafar, Y. Jin, and G. Peng, “Watching videos from everywhere: A study of the pptv mobile vod system,” in *Proceedings of IMC*, 2012, pp. 185–198.

[12] Z. Li, G. Xie, J. Lin, Y. Jin, M.-A. Kaafar *et al.*, “On the geographic patterns of a large-scale mobile video-on-demand system,” in *Proceedings of INFOCOM*, 2014, pp. 397–405.

[13] H. Yin, X. Liu, F. Qiu, N. Xia, C. Lin, H. Zhang, V. Sekar, and G. Min, “Inside the bird’s nest: Measurements of large-scale live vod from the 2008 olympics,” in *Proceedings of IMC*, 2009, pp. 442–455.

[14] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng, “Understanding user behavior in large-scale video-on-demand systems,” in *Proceedings of ACM SIGOPS Operating Systems Review*, 2006, pp. 333–344.

[15] V. Gopalakrishnan, R. Jana, K. Ramakrishnan, D. F. Swayne, and V. A. Vaishampayan, “Understanding couch potatoes: Measurement and modeling of interactive usage of iptv at large scale,” in *Proceedings of IMC*, 2011, pp. 225–242.

[16] M. Cha, P. Rodriguez, J. Crowcroft, S. Moon, and X. Amatriain, “Watching television over an ip network,” in *Proceedings of IMC*, 2008, pp. 71–84.

[17] Y. Huang, T. Z. Fu, D.-M. Chiu, J. Lui, and C. Huang, “Challenges, design and analysis of a large-scale p2p-vod system,” in *Proceedings of ACM SIGCOMM Computer Communication Review*, 2008, pp. 375–388.

[18] Z. Li, G. Xie, M. A. Kaafar, and K. Salamatian, “User behavior characterization of a large-scale mobile live streaming system,” in *Proceedings of WWW*, 2015, pp. 307–313.

[19] H. Yan, T.-H. Lin, G. Wang, Y. Li, H. Zheng, D. Jin, and B. Zhao, “A first look at user switching behaviors over multiple video content providers,” in *Proceedings of ICWSM*, 2017, pp. 700–703.

[20] H. Yan, T.-H. Lin, C. Gao, Y. Li, and D. Jin, “On the understanding of video streaming viewing behaviors across different content providers,” *IEEE Transactions on Network and Service Management*, vol. 15, no. 1, pp. 444–457, 2018.

[21] H. Yan, T.-H. Lin, G. Wang, Y. Li, H. Zheng, D. Jin, and B. Y. Zhao, “On migratory behavior in video consumption,” in *Proceedings of CIKM*, 2017, pp. 1109–1118.

[22] M. Park, M. Naaman, and J. Berger, “A data-driven study of view duration on youtube,” in *Proceedings of ICWSM*, 2016.

[23] M. Elhoseny, A. Shehab, and L. Osman, “An empirical analysis of user behavior for p2p iptv workloads,” in *International Conference on Advanced Machine Learning Technologies and Applications*. Springer, 2018, pp. 252–263.

[24] A. A. W. Hoiles and V. Krishnamurthy, “Engagement and popularity dynamics of youtube videos and sensitivity to meta-data,” in *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 7, 2017, pp. 1426–1437.

[25] N. Aggrawal, A. Arora, and A. Anand, “Modeling and characterizing viewers of you tube videos,” *International Journal of System Assurance Engineering and Management*, vol. 9, no. 2, pp. 539–546, 2018.

[26] D. M. C. J. Wu, Y. Zhou and Z. Zhu, “Modeling dynamics of online video popularity,” in *IEEE Transactions on Multimedia*, vol. 18, no. 9, 2016, pp. 1882–1895.

[27] G. X. Jiali Lin, Zhenyu Li, Y. Sun, K. Salamatian, and W. Wang, “Mobile video popularity distributions and the potential of peer-assisted video delivery,” in *IEEE Communications Magazine*, vol. 51, no. 11, 2013, pp. 120–126.

[28] M. Ma, L. Zhang, J. Liu, Z. Wang, H. Pang, L. Sun, W. Li, G. Hou, and K. Chu, “Characterizing user behaviors in mobile personal livecast: Towards an edge computing-assisted paradigm,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 3s, p. 66, 2018.

[29] G. Xie, Z. Li, M. A. Kaafar, and Q. Wu, “Access types effect on internet video services and its implications on cdn caching,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1183–1196, 2017.

[30] L. Huang, B. Ding, A. Wang, Y. Xu, Y. Zhou, and X. Li, “User behavior analysis and video popularity prediction on a large-scale vod system,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 3s, p. 67, 2018.

[31] S. S. Krishnan and R. K. Sitaraman, “Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs,” in *Proceedings of IMC*, 2012, pp. 211–224.

[32] L. Guo, E. Tan, S. Chen, Z. Xiao, and X. Zhang, “The stretched exponential distribution of internet media access patterns,” in *Proceedings of ACM PODC*, 2008, pp. 283–294.

[33] L. Jiang, Y. Miao, Y. Yang, Z. Lan, and A. G. Hauptmann, “Viral video style: A closer look at viral videos on youtube,” in *Proceedings of ICMR*, 2014, pp. 193–200.

[34] J. M. A. Henrique Pinto and M. A. Gonçalves, “Using early view patterns to predict the popularity of youtube videos,” in *Proceedings of WSDM*, 2013, pp. 365–374.

[35] T. Lian, Z. Chen, Y. Lin, and J. Ma, “Temporal patterns of the online video viewing behavior of smart tv viewers,” *Journal of the Association for Information Science and Technology*, vol. 69, no. 5, pp. 647–659, 2018.

[36] N. Kamiyama and M. Murata, “Reproducing popularity distribution of youtube videos,” *IEEE Transactions on Network and Service Management*, 2019.

[37] Y. Li, Y. Zhang, and R. Yuan, “Measurement and analysis of a large scale commercial mobile internet tv system,” in *Proceedings of IMC*, 2011, pp. 209–224.

[38] A. Brodersen, S. Scellato, and M. Wattenhofer, “Youtube around the world: Geographic popularity of videos,” in *Proceedings of WWW*, 2012, pp. 241–250.

[39] K. Huguenin, A.-M. Kermarrec, K. Kloudas, and F. Taïani, “Content and geographical locality in user-generated content sharing systems,” in *Proceedings of NOSSDAV*, 2012, pp. 77–82.

[40] H. Yan, J. Liu, Y. Li, D. Jin, and S. Chen, “Spatial popularity and similarity of watching videos in a large city,” in *Proceedings of IEEE GLOBECOM*, 2016, pp. 1–6.

[41] S. Kumar, R. Zafarani, and H. Liu, “Understanding user migration patterns in social media,” in *Proceedings of AAAI*, 2011, pp. 1204–1209.

[42] E. Newell, D. Jurgens, H. M. Saleem, H. Vala, J. Sassine, C. Armstrong, and D. Ruths, “User migration in online social networks: A case study on reddit during a period of community unrest,” in *Proceedings of ICWSM*, 2016, pp. 297–288.

[43] H. Zhu, R. E. Kraut, and A. Kittur, “The impact of membership overlap on the survival of online communities,” in *Proceedings of CHI*, 2014, pp. 281–290.

[44] H. Zhu, J. Chen, T. Matthews, A. Pal, H. Badenes, and R. E. Kraut, “Selecting an effective niche: An ecological view of the success of online communities,” in *Proceedings of CHI*, 2014, pp. 301–310.

[45] C. I. N. I. Center, *Statistical Report on Internet Development in China*, 2015.

[46] R. Clayton, S. J. Murdoch, and R. N. M. Watson, “Ignoring the great firewall of china,” in *Proceedings of PETS*, 2006.

[47] N. Xia, H. H. Song, Y. Liao, M. Iliofotou, A. Nucci, Z.-L. Zhang, and A. Kuzmanovic, “Mosaic: Quantifying privacy leakage in mobile networks,” in *Proceedings of SIGCOMM*, 2013, pp. 279–290.

[48] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proceedings of IMC*, 2009, pp. 49–62.

[49] G. Wang, X. Zhang, S. Tang, H. Zheng, and B. Y. Zhao, "Unsupervised clickstream clustering for user behavior analysis," in *Proceedings of CHI*, 2016, pp. 225–236.

[50] I. S. Dhillon, S. Mallela, and R. Kumar, "A divisive information theoretic feature clustering algorithm for text classification," *The Journal of Machine Learning Research*, vol. 3, pp. 1265–1287, 2003.

[51] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[52] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of ICML*, ser. ICML '97, San Francisco, CA, USA, 1997.

[53] H. Yan, T. Lin, M. Zeng, J. Wu, J. Huang, Y. Li, and D. Jin, "Characterizing the usage of mobile video service in cellular networks," in *IEEE Globecom*, 2017, pp. 1–6.

[54] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems 9*. MIT Press, 1997, pp. 155–161.

[55] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth International Group, 1984.

[56] L. B. Statistics and L. Breiman, *Random Forests*, 2001.

[57] R. H. Kewley, M. J. Embrechts, and C. Breneman, "Data strip mining for the virtual design of pharmaceuticals with neural networks," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 668–679, 2000.

Huan Yan received the B.S. degree in electronic information engineering from Nanchang University, Nanchang, China, in 2009, and M.S. degree in communication engineering from Hangzhou Dianzi University, Hangzhou, China, in 2012. He is currently working toward the Ph.D. degree in the department of electronic engineering of Tsinghua University, Beijing, China. His current research interests include network big data and software-defined network.

Haohao Fu is currently working toward the undergraduate degree with the College of Letters and Science, UC Berkeley, Berkeley, California.

Yong Li (M'09–SM'16) received the B.S. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2007 and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012. He is currently a Faculty Member of the Department of Electronic Engineering, Tsinghua University.

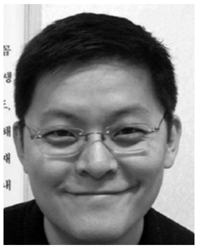
Dr. Li has served as General Chair, TPC Chair, TPC Member for several international workshops and conferences, and he is on the editorial board of two IEEE journals. His papers have total citations more than 3600 (six papers exceed 100 citations, Google Scholar). Among them, eight are ESI Highly Cited Papers in Computer Science, and four receive conference Best Paper (run-up) Awards. He received IEEE 2016 ComSoc Asia-Pacific Outstanding Young Researchers and Young Talent Program of China Association for Science and Technology.

Tzu-Heng Lin is currently an undergraduate student in the Department of Electronic Engineering, Tsinghua University, Beijing, China. His current research interests include big data mining and user behavior analysis.

Gang Wang Gang Wang received the B.E. degree in electrical engineering from Tsinghua University, Beijing, China, in 2010, and is currently pursuing the Ph.D. degree in computer science at the University of California, Santa Barbara, CA, USA. His research interests are security and privacy, mobile networks, online social networks, and crowdsourcing systems.

Haitao Zheng (M'99–SM'09) received the B.S. degree (with highest honor) in electrical engineering from Xian Jiaotong University, Xian, China, in 1995, and the M.S.E.E. and Ph.D .degrees in electrical and computer engineering from the University of Maryland, College Park, MD, USA, in 1998 and 1999, respectively. She joined the Wireless Research Lab, Bell-Labs, Lucent Technologies, Crawford Hill, NJ, USA, as a Member of Technical Staff in 1999 and moved to Microsoft Research Asia, Beijing, China, as a project lead and Researcher in 2004. Since 2005, she has been a faculty member with the Computer Science Department, University of California, Santa Barbara, where she is now a Professor. Her recent research interests include wireless systems and networking and social networks. Dr. Zheng was named as the 2005 MIT Technology Review Top 35 Innova- tors under the age of 35 for her work on cognitive radios. Her work was selected by MIT Technology Review as one of the 10 Emerging Technologies in 2006. She also received the 2002 Bell Laboratories President's Gold Award from Lucent Bell-Labs and the 1998–1999 George Harhalakis Outstanding Graduate Student Award from the Institute of System Research, University of Maryland.

Depeng Jin received the B.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 1995 and 1999 respectively both in electronics engineering. He is a professor at Tsinghua University and the chair of Department of Electronic Engineering. Dr. Jin was awarded National Scientific and Technological Innovation Prize (Second Class) in 2002. His research fields include telecommunications, high-speed networks, ASIC design and future Internet architecture.



Ben Y. Zhao received the B.S. degree from Yale Uni- versity, New Haven, CT, USA, in 1997, and the M.S. and Ph.D. degrees from the University of California (UC), Berkeley, CA, USA, in 2000 and 2004, respec- tively, all in computer science. He is a Professor with the Computer Science Department, UC, Santa Barbara, CA, USA. He has published over 100 publications in areas of security and privacy, networked and distributed systems, wireless networks, and data-intensive computing. Prof. Zhao has served as Program Chair for top conferences (WOSN, WWW 2013 OSN track, IPTPS, IEEE P2P) and is a co-founder and Steering Committee member of the ACM Conference on Online Social Networks (COSN). He is a recipient of the National Science Foundation's CAREER Award, MIT Technology Review's TR-35 Award (Young Innovators Under 35), and Computerworld's Top 40 Technology Innovators Award. His work has been covered by media outlets such as The New York Times, The Boston Globe, MIT Technology Review, and Slashdot.