# On Migratory Behavior in Video Consumption

Huan Yan[1], Tzu-Heng Lin[1], Gang Wang[2], Yong Li[1], Haitao Zheng[3], Depeng Jin[1], Ben Y. Zhao[3]
[1]Tsinghua National Laboratory for Information Science and Technology,
Department of Electronic Engineering, Tsinghua University
[2]Department of Computer Science, Virginia Tech
[3]Department of Computer Science, University of Chicago
liyong07@tsinghua.edu.cn

## 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, and discuss the implications of our results to CPs.

## CCS CONCEPTS

• **Information systems** → *Web log analysis*;

## KEYWORDS

Video Consumption, Migratory Behavior

## 1 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 [16, 28]. Recently, the adoption of mobile devices and social networks further promotes the wide consumption of user-uploaded videos (YouTube, Vine) [31] and personal live streaming content (Periscope, Meerkat) [35].

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 [3, 25, 29, 30, 32]. 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 [7, 10, 12, 14, 16, 22–24, 39, 40]. 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 [38], 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. 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. Our analysis results lead to machine learning models to predict how likely users would migrate across CPs.


*CIKM'17 , November 6–10, 2017, Singapore, Singapore*

ACM ISBN 978-1-4503-4918-5/17/11…$15.00
https://doi.org/10.1145/3132847.3132884

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 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, 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.
- Fourth, user migration 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).

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.

## 2 RELATED WORK

**Video Access Behavior.** User behaviors in online video systems have been examined in the context of Video on Demand (VoD) [10, 22, 24, 27, 39, 40], International Protocol Television (IPTV) [7, 12], peer-to-peer (P2P) VoD [14] and live streaming [23], in different video services such as YouTube[39], 2008 Olympics [39], PPLive [22–24]. 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. Krishnan et al. analyze the influence of video stream quality on user behavior from multiple CPs using quasi-experimental designs [18], but they do not have an in-depth study of characterizing migration behaviors.

**Temporal Patterns.** For temporal analysis, Yu et al. propose a model for user arrival rate and video popularity [40]. Li et al. have reported their observations on daily and weekly patterns in a mobile VoD system [22]. Yin et al. focus on how the temporal dynamic nature of the system impacted user behavior [39]. Guo et al. model the video access patterns with stretched exponential distributions [13]. Instead of modeling daily/weekly patterns [14, 21, 22], 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 [5, 15, 24, 31]. For example, Cha et al. examine the geographical locality of an IPTV system [7]. Scellato et al. propose to use the location information in Twitter to predict the geographic popularity of YouTube videos [31]. Our work focuses on video viewing behaviors in a metropolis city.

**Migration Behavior in Social Network.** There are some works about migration behavior in social network [19, 26, 41, 42]. For instance, Kumar et al. explore user migration patterns between social media sites [19]. Newell et al. investigate how and why users in Reddit migrate to other Reddit-like alternative platforms [26]. Unlike them, we study user migration behavior across different CPs in video consumption.

## 3 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.

### 3.1 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 [6]. 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 [8], 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 the six CPs based the domain name of requested URLs. Note that all six 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 [37], we obtain data via collaborations with the ISP who carefully removed personally identifiable information (*e.g.*, IP), and anonymized user

| 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 1: Number of videos and views per category (the numbers are displayed in millions).**

| 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% |

**Table 2: Statistics of the 6 content providers (the numbers are displayed in millions). The penetration ratio is based on Internet Development Report of China.**

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).

## 3.2 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 1 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 2, YK is significantly larger than the other five with more videos (6.5 million), views (131 million) and users (3.1 million). This is consistent with the 2015 Internet report in China [6] 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. 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.

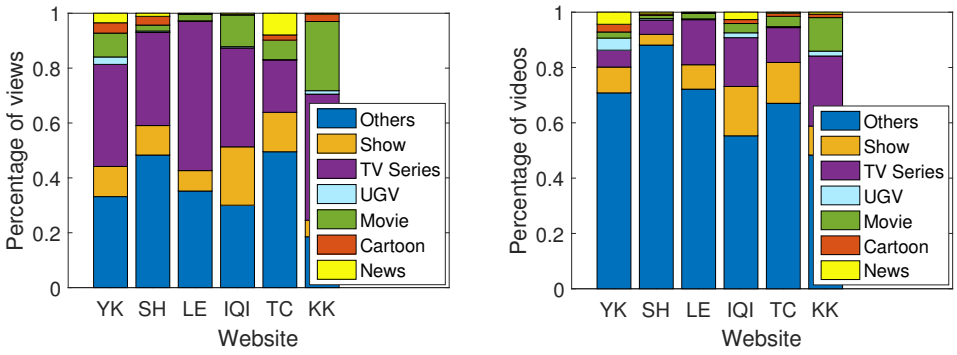
# 4 MIGRATORY BEHAVIOR ANALYSIS

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.

## 4.1 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 $V_i = \sum_{k=1}^{K} v_i^k$ $(1 \le i \le M)$ 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 $t_{i_j, i_{j+1}} = t_{i_{j+1}} - t_{i_j}$ $(1 \le j < N)$ as the time gap. 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. 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 CPs. As shown Figure 2, 99% of videos have a duration less than 100 minutes. Following existing




(a) Views (b) Videos

**Figure 1: Distribution of videos and views by categories in six CPs.**

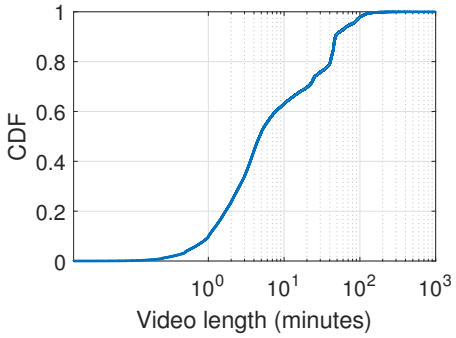



**Figure 2: CDF of video length.**

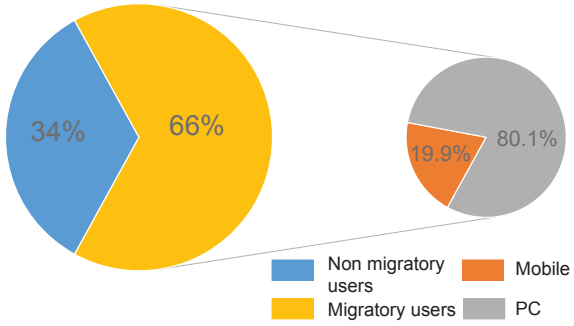



**Figure 3: Basic information about migratory users.**

work [1], 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 two metrics to drive our analysis:

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

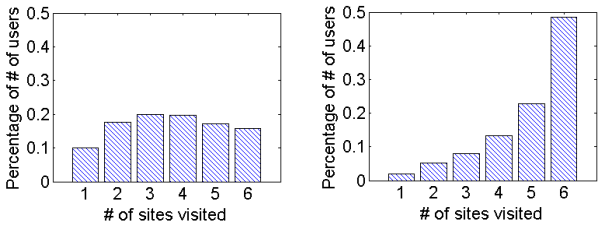
$$\overline{F} = \frac{\sum_{i=1}^{M}\sum_{k=1}^{K}\sum_{k'=1,k'\neq k}^{K} s_i^{k,k'}}{\sum_{i=1}^{M}(N_i-1)}. \quad (1)$$

  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,

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

- **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:

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




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

**Figure 4: Distribution of users visiting different number of CPs.**

## 4.2 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 3 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%.

| | 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% |

**Table 3: Migratory probability between different CPs. The column (row) represents the origin (target) CP.**

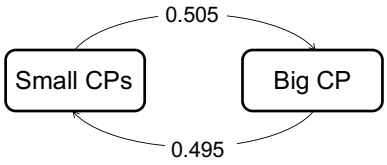



**Figure 5: Migratory probability between the big and small CPs.**

Figure 4 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, 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.

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.*

## 4.3 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 3. 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 5).

We seek to further understand whether video category influences the migratory behaviors by showing the results in Table 4. 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.

| | 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 4: Migratory probability on different video categories. The column (row) represents the origin (target) categories.**

| | 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 |

**Table 5: Video categories that users migrate to at the target CP. The column (row) represents the target CP (categories). (The numbers are displayed in thousands)**

Finally, we examine which categories users migrate to when switching to a particular site. As shown in Table 5, for "TV", "Movie" and "Show", 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.*

# 5 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.

## 5.1 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.

| 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) |

**Table 6: Possible user behaviors that correspond to different time gaps.**

**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 $\{q_{i_1}, t_{i_1,i_2}, q_{i_2}, t_{i_2,i_3}, q_{i_3} \ldots q_{i_N}\}$, 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 6). 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 4.1.

**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 [36] 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 $[e_{v,1}, e_{v,2}, ..., e_{v,T}]$ ($v = i, j$, $T = |C_k(Y_i, Y_j)|$). The sequence similarity metric is computed by Tanimoto coefficient:

$$Z_k(Y_i, Y_j) = \frac{\sum_{m=1}^{T} e_{i,m} e_{j,m}}{\sum_{m=1}^{T} e_{i,m}^2 + \sum_{m=1}^{T} e_{j,m}^2 - \sum_{m=1}^{T} e_{i,m} e_{j,m}}, \quad (4)$$

which considers both the direction and magnitude of two vectors. We set $k = 5$ for our analysis following the settings in [36]. To identify clusters in the graph, we use the *Divisive Hierarchical Clustering Algorithm* [9], which is suitable for finding arbitrary cluster shapes.

## 5.2 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 [2]). For our analysis, we focus on the largest 12 clusters shown in Figure 6, which covers 99% of the selected users.

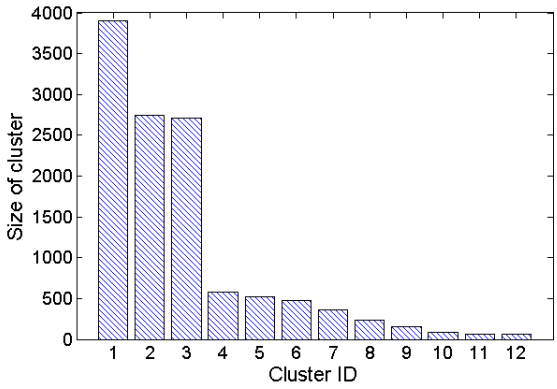



**Figure 6: Number of users in top 12 clusters.**

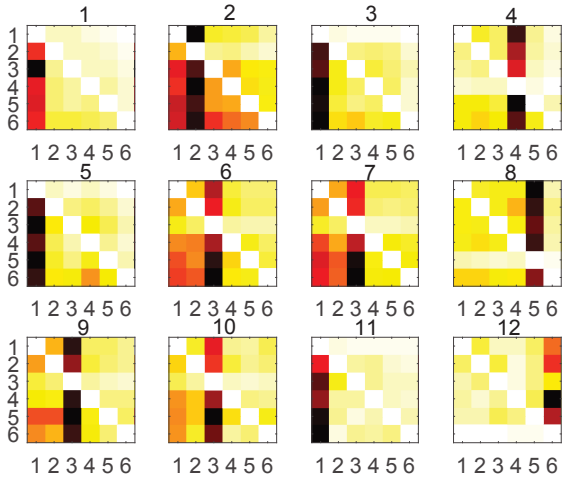



**Figure 7: CP migratory probability for the top 12 clusters. Each heatmap represents one cluster. The column (row) represents the origin (target) CP where numbers 1–6 represent YK, SH, LE, IQI, TC, KK respectively.**

**Cluster Analysis.** With a focus on users' cross-site migratory patterns, we first plot a heatmap in Figure 7. It shows the probability of migrating from one CP (row) to another CP (column) 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 8. 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

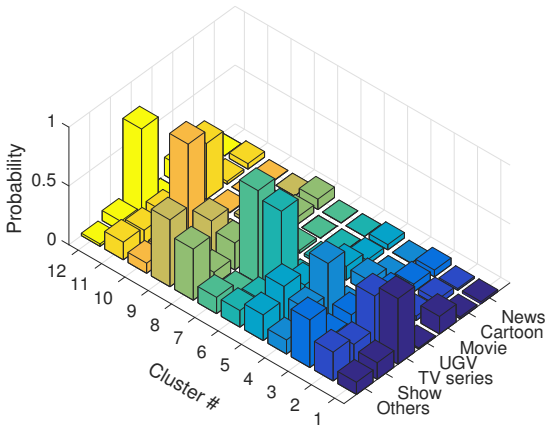



**Figure 8: Probability distribution of video categories that users migrate to in each cluster.**

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.*

## 6 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.

### 6.1 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.

*6.1.1 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

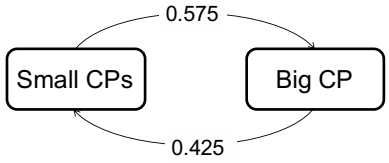



**Figure 9: Migratory ratio between the big and small CPs on migration events caused by refreshing failure.**

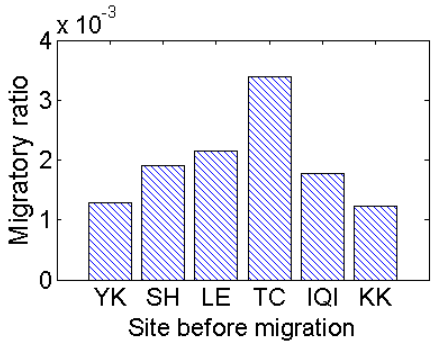



**Figure 10: Ratio of migratory due to refreshing failure (threshold=30s).**

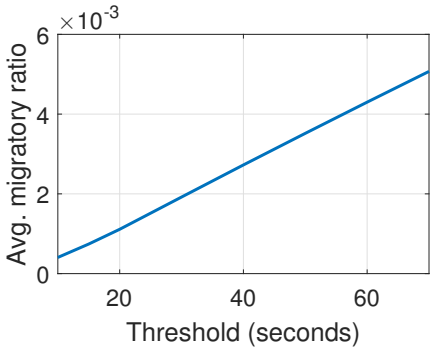



**Figure 11: Migratory ratio with different time threshold.**

*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 9). 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 10, the migratory ratio is smaller than 0.4% at all six sites, and average ratio is only 0.19%, which suggests refreshing failures rarely 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 11. We find that the

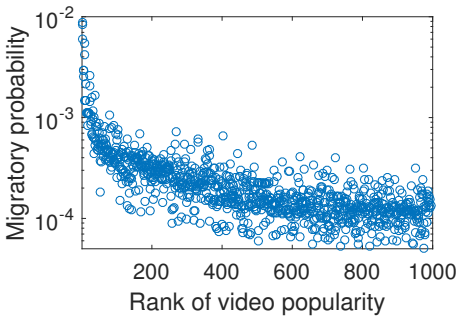



**Figure 12: Migratory probability for watching popular videos. Popular videos are ranked by # of views.**

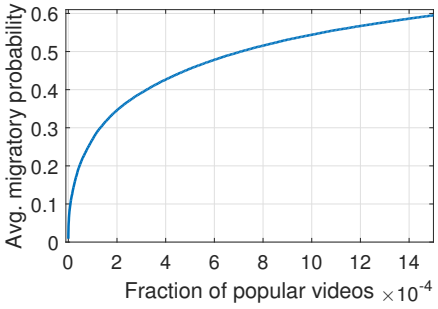



**Figure 13: Ratio of migration events to watch popular videos over all migration events.**

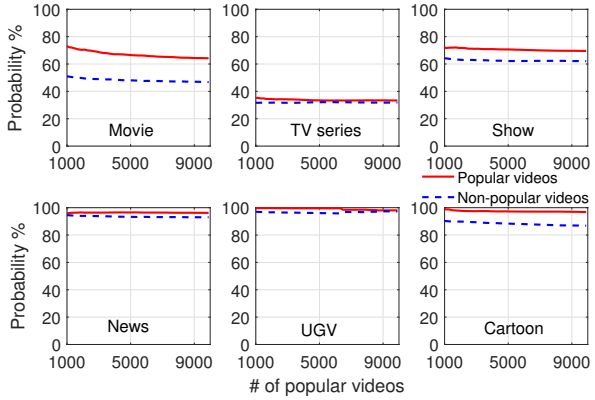



**Figure 14: Probability that a user migrates to watch a non-popular/popular video of a different category from the video she watched before migration.**

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.

*6.1.2 Video Popularity.* Popular videos attract more user viewing, and are likely to trigger user migration. To investigate it, on one hand, 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 12). 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. Further, we compute the average probability among top popular videos and show the results in Figure 13. 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 14, 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 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.*

| 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 |

**Table 7: The feature list for prediction of user migratory frequency.**

| 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 |

**Table 8: The correlation coefficient between predicted and real value from three prediction methods.**

## 6.2 Migration Prediction

So far, we analyzed user migration behavior and possible reasons. We next build a prediction model to validate our findings. More

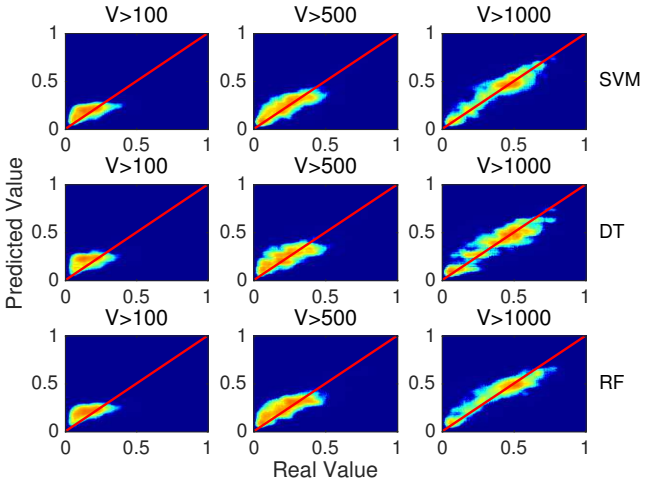



**Figure 15: Correlation between prediction result and real value for users with different number of video viewing events (V denotes # of views per user).**

| 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 |

**Table 9: The weights of features of three methods.**

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 7, these features include: the fraction views on opular 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) [11], Decision Tree (DT) [4] and Random Forest (RF) [33]. In our experiment, we select 10,000 users with 100+, 1000+, 1000+ views respectively and run 5 fold cross-validation. We use heatmap to intuitively illustrate the correlation between the predicted and real value (Figure 15). 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 8. We observe that more active users are more predicable. 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).

To explore the importance of features, we compute the feature weights of three methods in Table 9, where a higher weight indicates more important feature. Note that the three models compute weights differently: SVM uses sensitivity analysis on features [17], while DT measures the goodness of each split inside the tree [4]. RF measures the decrease in node impurities on features [33]. 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 viewing time per video* (feature 7), which is also highly indicative of migration (missed by previous sections). 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 7, 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.* viewing time per video).

## 6.3 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) CPs should pay more attention to their contents, rather than the networking service quality 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.

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.*, viewing time per video). This can help individual CPs to estimate users' likelihood of migration, and deploy targeted engagement mechanisms.

## 6.4 Limitations

There are a few limitations in our study. First, our analysis on the possible reasons of migration is by no means complete. Certain factors such as social influence from friends [20, 34] and user demographics cannot be captured by our data. Our future work will explore a qualitative approach to examine user motivations for

switching CPs and cross-examine the results with our empirical study. Second, our study primarily focuses on Chinese video streaming market. Future research is needed to expand the analysis scope (when related data becomes available).

## 7 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. In addition, we study the potential reasons about user migration which leads to an accurate prediction model for migration frequency. 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.

## 8 ACKNOWLEDGMENT

This work is supported by research fund of Tsinghua University-Tencent Joint Laboratory for Internet Innovation Technology and NSF grant CNS-1717028.

## REFERENCES

[1] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. 2009. Characterizing User Behavior in Online Social Networks. In *Proceedings of IMC*. 49–62.
[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (2008), P10008.
[3] James bradshaw. 2016. Video-streaming Service Shomi To Shut Down at End of November. The Globe and Mail. (2016).
[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Wadsworth International Group.
[5] Anders Brodersen, Salvatore Scellato, and Mirjam Wattenhofer. 2012. Youtube Around the World: Geographic Popularity of Videos. In *Proceedings of WWW*. 241–250.
[6] China Internet Network Information Center. 2015. *Statistical Report on Internet Development in China*.
[7] Meeyoung Cha, Pablo Rodriguez, Jon Crowcroft, Sue Moon, and Xavier Amatriain. 2008. Watching Television Over an IP Network. In *Proceedings of IMC*. 71–84.
[8] Richard Clayton, Steven J. Murdoch, and Robert N. M. Watson. 2006. Ignoring the Great Firewall of China. In *Proceedings of PETS*.
[9] Inderjit S Dhillon, Subramanyam Mallela, and Rahul Kumar. 2003. A Divisive Information Theoretic Feature Clustering Algorithm for Text Classification. *The Journal of Machine Learning Research* 3 (2003), 1265–1287.
[10] Florin Dobrian, Vyas Sekar, Asad Awan, Ion Stoica, Dilip Joseph, Aditya Ganjam, Jibin Zhan, and Hui Zhang. 2011. Understanding the Impact of Video Quality on User Engagement. *ACM SIGCOMM Computer Communication Review* 41, 4 (2011), 362–373.
[11] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex J. Smola, and Vladimir Vapnik. 1997. Support Vector Regression Machines. In *Advances in Neural Information Processing Systems 9*. MIT Press, 155–161.
[12] Vijay Gopalakrishnan, Rittwik Jana, KK Ramakrishnan, Deborah F Swayne, and Vinay A Vaishampayan. 2011. Understanding Couch Potatoes: Measurement and Modeling of Interactive Usage of IPTV At Large Scale. In *Proceedings of IMC*. 225–242.
[13] Lei Guo, Enhua Tan, Songqing Chen, Zhen Xiao, and Xiaodong Zhang. 2008. The Stretched Exponential Distribution of Internet Media Access Patterns. In *Proceedings of ACM PODC*. 283–294.
[14] Yan Huang, Tom ZJ Fu, Dah-Ming Chiu, John Lui, and Cheng Huang. 2008. Challenges, Design and Analysis of a Large-scale p2p-vod System. In *Proceedings of ACM SIGCOMM Computer Communication Review*. 375–388.
[15] Kévin Huguenin, Anne-Marie Kermarrec, Konstantinos Kloudas, and François Taïani. 2012. Content and Geographical Locality in User-generated Content Sharing Systems. In *Proceedings of NOSSDAV*. 77–82.
[16] Dmytro Karamshuk, Nishanth Sastry, Andrew Secker, and Jigna Chandaria. 2015. On Factors Affecting the Usage and Adoption of a Nation-wide TV Streaming Service. In *Proceedings of INFOCOM*. 837–845.
[17] R. H. Kewley, M. J. Embrechts, and C. Breneman. 2000. Data Strip Mining for the Virtual Design of Pharmaceuticals with Neural Networks. *IEEE Transactions on Neural Networks* 11, 3 (2000), 668–679.
[18] S. Shunmuga Krishnan and Ramesh K. Sitaraman. 2012. Video Stream Quality Impacts Viewer Behavior: Inferring Causality Using Quasi-experimental Designs. In *Proceedings of IMC*. 211–224.
[19] Shamanth Kumar, Reza Zafarani, and Huan Liu. 2011. Understanding User Migration Patterns in Social Media. In *Proceedings of AAAI*. 1204–1209.
[20] Haitao Li, Xiaoqiang Ma, Feng Wang, Jiangchuan Liu, and Ke Xu. 2013. On Popularity Prediction of Videos Shared in Online Social Networks. In *Proceedings of CIKM*. 169–178.
[21] Yuheng Li, Yiping Zhang, and Ruixi Yuan. 2011. Measurement and Analysis of a Large Scale Commercial Mobile Internet TV System. In *Proceedings of IMC*. 209–224.
[22] Zhenyu Li, Jiali Lin, Marc-Ismael Akodjenou, Gaogang Xie, Mohamed Ali Kaafar, Yun Jin, and Gang Peng. 2012. Watching Videos From Everywhere: A Study of the PPTV Mobile Vod System. In *Proceedings of IMC*. 185–198.
[23] Zhenyu Li, Gaogang Xie, Mohamed Ali Kaafar, and Kave Salamatian. 2015. User Behavior Characterization of a Large-scale Mobile Live Streaming System. In *Proceedings of WWW*. 307–313.
[24] Zhenyu Li, Gaogang Xie, Jiali Lin, Yun Jin, Mohamed-Ali Kaafar, et al. 2014. On the Geographic Patterns of A Large-scale Mobile Video-on-demand System. In *Proceedings of INFOCOM*. 397–405.
[25] Nichole McNiel. 2016. Yahoo! Shutters Their Streaming Video Service. Yahoo! Had A Streaming Video Service? TheAmericanGenius. (2016).
[26] Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. In *Proceedings of ICWSM*. 297–288.
[27] Minsu Park, Mor Naaman, and Jonah Berger. 2016. A Data-Driven Study of View Duration on YouTube. In *Proceedings of ICWSM*.
[28] Emil Protalinski. 2016. Streaming Services Now Account For Over 70% of Peak Traffic in North America, Netflix Dominates with 37%. Venture Beat. (2016).
[29] Claire Reilly. 2015. EzyFlix Goes Dark As Streaming Battle Claims Its First Casualty. CNet News. (2015).
[30] Claire Reilly. 2016. Goodbye Presto: Foxtel's Streaming Service to Shut Down. CNet News. (2016).
[31] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, and Jon Crowcroft. 2011. Track Globally, Deliver Locally: Improving Content Delivery Networks By Tracking Geographic Social Cascades. In *Proceedings of WWW*. 457–466.
[32] Todd Spangler. 2014. Verizon, Redbox to Pull Plug on Video-Streaming Service. Variety. (2014).
[33] Leo Breiman Statistics and Leo Breiman. 2001. *Random Forests*. 5–32 pages.
[34] David Vallet, Shlomo Berkovsky, Sebastien Ardon, Anirban Mahanti, and Mohamed Ali Kafaar. 2015. Characterizing and Predicting Viral-and-Popular Video Content. In *Proceedings of CIKM*. 1591–1600.
[35] Bolun Wang, Xinyi Zhang, Gang Wang, Haitao Zheng, and Ben Y. Zhao. 2016. Anatomy of a Personalized Livestreaming System. In *Proceedings of IMC*.
[36] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, and Ben Y. Zhao. 2016. Unsupervised Clickstream Clustering for User Behavior Analysis. In *Proceedings of CHI*. 225–236.
[37] Ning Xia, Han Hee Song, Yong Liao, Marios Iliofotou, Antonio Nucci, Zhi-Li Zhang, and Aleksandar Kuzmanovic. 2013. Mosaic: Quantifying Privacy Leakage in Mobile Networks. In *Proceedings of SIGCOMM*. 279–290.
[38] Huan Yan, Tzu-Heng Lin, Gang Wang, Yong Li, Haitao Zheng, Depeng Jin, and Ben Zhao. 2017. A First Look at User Switching Behaviors Over Multiple Video Content Providers. In *Proceedings of ICWSM*. 700–703.
[39] Hao Yin, Xuening Liu, Feng Qiu, Ning Xia, Chuang Lin, Hui Zhang, Vyas Sekar, and Geyong Min. 2009. Inside the Bird's Nest: Measurements of Large-scale Live VoD from the 2008 Olympics. In *Proceedings of IMC*. 442–455.
[40] Hongliang Yu, Dongdong Zheng, Ben Y Zhao, and Weimin Zheng. 2006. Understanding User Behavior in Large-scale Video-on-demand Systems. In *Proceedings of ACM SIGOPS Operating Systems Review*. 333–344.
[41] Haiyi Zhu, Jilin Chen, Tara Matthews, Aditya Pal, Hernan Badenes, and Robert E. Kraut. 2014. Selecting an Effective Niche: An Ecological View of the Success of Online Communities. In *Proceedings of CHI*. 301–310.
[42] Haiyi Zhu, Robert E. Kraut, and Aniket Kittur. 2014. The Impact of Membership Overlap on the Survival of Online Communities. In *Proceedings of CHI*. 281–290.