On Distribution of User Movie Watching Time in a Large-scale Video Streaming System

Yishuai Chen  
School of Electrical and Information Engineering  
Beijing Jiaotong University  
Beijing, 100044, China  
yschen@bjtu.edu.cn

Yong Liu  
Department of Electrical and Computer Engineering  
Polytechnic Institute of New York University  
Brooklyn, NY, 11201.  
yongliu@poly.edu

Baoxian Zhang  
College of Comp. & Commun. Eng.  
University of Chinese Academy of Sciences  
Beijing, 100049, China  
bxzhang@ucas.ac.cn

Wei Zhu  
PPLive Inc.  
Shanghai, China  
stevenzhu@pplive.com

Abstract—Video watching time is a crucial measure for studying user watching behavior in online Internet on-demand (VoD) systems. It is important for system planning, user engagement study, and service quality evaluation. However, due to limited access to large-scale VoD systems, there is still a lack of accurate model for characterizing the distribution of user watching time on a per video basis. In this paper, we measure PPLive, one of the most popular commercial Internet VoD systems in China, over a three week period, and characterize user watching time distributions of 1,000 most popular movies. We find that a video's watching time can be modeled by a concatenation of exponential distribution (in the first several minutes of the video) and truncated power law distribution (in the remaining time of the video), when users watch the video without interruptions. For comparison, user watching time with user interactions such as seeking and/or pause operations does not follow such a distribution. We further reveal interesting characteristics regarding the relation between video’s watching time distribution and various watching/video-related features (including time-of-day, user ratings, and movie genres). Our measurement and modeling results bring forth important insights for design, deployment, and evaluation of Internet VoD systems.

Keywords—user watching time, online video, video-on-demand.

I. INTRODUCTION

For a video streaming service provider, it is crucial to understand the characteristics of videos’ watching time distribution for user engagement study and system planning and optimization. First, the system workload and network traffic that a video produces is determined by how much time users spend on watching the video. Second, video watching time is a good indicator of user engagement [1], which has attracted a lot of attention as Internet services are becoming more and more user-centric. Third, the performance of various system optimization algorithms (e.g., Peer-to-Peer streaming algorithms [2], prefix caching at proxy server [3], and pre-fetching at the client side [4]) is directly affected by users’ watching time distribution.

Despite of the importance of user watching time distributions, there is still a lack of accurate model in this aspect. The existing user watching time models were based on measurement data obtained by monitoring clients’ video traffic and signaling messages on media servers (e.g., [5]) or network routers (e.g., [6]). However, a large amount of downloaded video (e.g., 35–48 percent in Youtube [6]) are not watched by users due to pre-fetching and users’ early departures. Thus, a user's downloaded video is usually more than her watched video in a session. Moreover, a user's online session often includes many non-watching intervals (e.g., initial startup latency, re-buffering time, user pause time, etc.). Thus, a user's online time usually does not equal to her actual watching time.

In this paper, based on a large-scale and highly-accurate trace provided by PPLive, we study the distribution of user movie watching time and its relation to various watching/video-related features. PPLive is one of the most popular Internet video streaming systems in China. The latest industry report [7] shows that PPLive's daily unique users have reached 34 million. It has also been studied in a lot of research work (e.g., [3,8]). The trace is collected by the native PPLive client program, which is installed by users on their computers for video watching and accurately records users' video watching time. In this paper, we focus on studying movies as they are a major type of videos in online video-on-demand (VoD) systems and are relatively long (e.g., the mean length of the top 1,000 most popular movies in our measurement is 103.6 minutes), which makes the modeling of their user watching time distributions interesting and important. Our major findings are summarized as follows.

• For users who leave during the initial several minutes of movie opening credit screening period, their watching time follows exponential distribution, meaning that user leave rate in the initial screening period is constant, independent of how long she has stayed in this period.
For users who continue watching the movie after the initial screening period, their watching time follows a truncated power law distribution, meaning that user leave rate after the initial screening period is inversely proportional to the time that she has watched the movie.

User leave rate after the initial screening period varies diurnally and is highly correlated to movie genres and user ratings, meaning that users' engagement in videos varies diurnally and is highly correlated to these video features.

The rest of this paper is organized as follows. Section II briefly reviews related work. Section III introduces our measurement and analysis methodologies. Sections IV and V present our measurement and modeling results. In Section VI, we establish the relation between user leaving rate and key parameters in the model. Section VII concludes this paper.

II. RELATED WORK

Some existing work measured and characterized user watching time in Internet-based VoD systems [1,3,6,9,10]. They mainly focused on observing the impact of service quality [1], video length [3,9,10], video type [10], video popularity [10], access method [6], and pricing [9] on user watching time. They did not model user watching time distribution on a per video basis as we shall do in this paper. Some existing work (e.g., [11]) reported aggregated watching time distribution of all videos’ or channels’ sessions. In contrast, in this paper, we characterize user watching time distribution on a per movie basis.

References [8,12] reported user watching time distribution on a per channel basis in live streaming systems. However, they observed aggregated watching time distribution of a channels’ all sessions (e.g., sessions with and without playback interruptions) and did not give concise mathematical expression of users’ watching time distribution. In contrast, in this paper, we separate users’ watching sessions in different cases to analyze and obtain an accurate mathematical model of users’ video watching time. Moreover, users' watching behavior in live streaming systems is quite different from that in VoD systems. Specifically, after a user selects a live streaming channel, she watches the channel's content from the channel's current playback point. As the user has to stop when the event ends, her watching time is highly dependent on her arriving time. In contrast, in VoD, after a user clicks a video, she always watches from the beginning of the video, and her watching time is mainly decided by her interest in the video.

III. MEASUREMENT AND ANALYSIS METHODOLOGIES

A. Trace from PPLive

The system we measured is PPLive, which provides free VoD services on the Internet. A user can access its videos through a web browser or a client program with a user interface similar to most other popular online video streaming systems. When playing a video, users can pause, resume, and seek to any position of the video. Video are mainly in Chinese (or foreign languages with Chinese captions), and the users are mainly Chinese, in either China or other countries around the world.

We believe that the measurement and modeling results obtained in this paper are applicable for other PC-based Internet VoD systems due to the following reasons. 1) PPLive's client program has a user interface similar to most other popular online video streaming systems. As a result, users’ watching behavior in these systems are similar. 2) In this paper, we focus on characterizing users’ natural watching behavior, i.e., watching with neither playback interruptions nor user interactions. In this case, our measurement and modeling results are not affected by the system architecture (e.g., client/server or peer-to-peer) and streaming quality (e.g., playback interruptions) of the underlying streaming system and exactly reflects users’ nature of video watching, which we believe is consistent across different video streaming systems. 3) The subscriber base of PPLive is huge and geographically widely distributed all over the world, which ensures our results can reflect the general characteristics of global Internet users rather than local or regional users.

The measurement data used in this work were provided by PPLive from their log server since this work is jointly carried out with PPLive. The data were collected by the native PPLive client program, which is installed by users on their computers for video watching. The program records information as listed in Table I for each video session and sends the logged information back to a PPLive log collection server. The trace period is three weeks, from June 27th to July 16th of 2010, without logs on July 4th and 7th. During the measurement period, PPLive provided streaming services for more than 100k videos to users and these videos led to more than 300 million sessions. To protect user privacy, PPLive does not keep any information that could be used to directly or indirectly identify individual users. We also use PPLive's system log which records metadata of video files, including video length, genre, and user rating.

B. Session Model

We characterize a user’s video watching behavior at the session level. After a user starts watching a video, her video session ends either when the video is played to its end, or she quits early. Previous studies (e.g., [6,11]) have showed that a large number of users leave during the initial several-minutes of opening credit period and thus treat this several minutes period as the initial screening phase, in which a user gets an initial impression of the video and decides whether to continue watching. In this paper, we adopt a similar method. Specifically, a session consists of the following two phases (see Fig. 1).

<table>
<thead>
<tr>
<th>Type</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session-related</td>
<td>Video ID</td>
</tr>
<tr>
<td></td>
<td>Duration of the session, from the time when a user clicks a video to the time when the user leaves</td>
</tr>
<tr>
<td>User Interaction</td>
<td>Number of seeking operations</td>
</tr>
<tr>
<td></td>
<td>Total seeking latency</td>
</tr>
<tr>
<td></td>
<td>Total pause time</td>
</tr>
<tr>
<td>Network Interruption</td>
<td>Number of playback interruptions</td>
</tr>
<tr>
<td></td>
<td>Total playback interruption latency</td>
</tr>
</tbody>
</table>

Fig. 1. Illustration of two possible phases in a video watching session.

TABLE I. LOGGED INFORMATION FOR EACH SESSION.
• Video Screening Phase. If a user quits during the initial screening phase, we call her video watching time as the user's video screening time and denote it by $T_{scr}$.

• Video Watching Phase. After the video screening phase, if a user chooses to continue watching, then she starts watching the actual video content. We call her watching time (counting from the beginning of the session) as the user's video watching time and denote it by $T_{watch}$.

C. Statistical Model Fitting Method

We use both visual and quantitative methods to study the distribution of $T_{scr}$ and $T_{watch}$ and obtain their border. If the measured data follow a hypothetical exponential distribution, i.e., its PDF (Probability Density Function) has the form of \( f(t) = Ce^{-\lambda t} \), where \( C \) is a constant and \( \lambda \) is the rate parameter, the PDF curve will follow a straight line with slope \(-\lambda\) with \( x\)-axis on linear scale and \( y\)-axis on log scale. If the measured data follow a hypothetical power law distribution, i.e., its PDF has the form of \( f(t) = Ct^{-\alpha} \), where \( C \) is a constant and \( \alpha \) is the power law exponent, its CCDF (Complementary Cumulative Distribution Function) on the log-log scale is a straight line with slope \(-(\alpha-1)\), as its CCDF is \( 1-F(t) = Dt^{-(\alpha-1)} \), where \( D \) is a constant and equal to \( C(\alpha-1) \), and \( \log(1-F(t)) = \log(D) - (\alpha-1)\log(t) \). We further apply linear regression on the distribution to obtain the slope of the line, which reflects the model parameter \( \alpha \) or \( \alpha-1 \). Then we evaluate the goodness-of-fit of the fitting by R-squared (\( R^2 \)), which is a commonly used metric to evaluate how well a regression line approximates real data points, and Kolmogorov-Smirnov (K-S) distance [13]. Given a set of random samples, \( x_1, x_2, ..., x_n \), whose empirical CDF (Cumulative Distribution Function) is denoted by \( S(x) \), denoting the CDF of the hypothesis distribution by \( F(x) \), the K-S distance between the empirical and hypothesis distribution is defined as \( KS(F, S) = \sup_{x}[F(x) - S(x)] \). In general, a \( R^2 \) closer to 1 and a K-S distance closer to 0 suggest a better fit.

D. Analysis Methodology

We categorize users’ video watching sessions into the following four types and characterize the distributions of user watching time in these four patterns, respectively. We will demonstrate how the above classification of sessions can improve the accuracy of our modeling results and bring forth new findings in later sections.

• Natural leave: In natural leave pattern, a user watches a video continuously throughout the whole session, without playback interruptions or user interactions (i.e., seeking or pause). In this case, the user's watching time is a natural reflection of her attention span on the video.

• Interruption/seek/pause-only. In each of these three patterns, a user experiences one and only one type of events during the session. For instance, a session in the interruption-only category only experiences playback interruption(s). Seek-only and pause-only patterns are similarly defined. By comparing the user watching time distribution for each of these patterns with that for the natural leave pattern, we can examine the impact of each type of events on users’ watching time distribution.

Moreover, to accurately measure a user's engagement time in a session, we only focus on the user's actual video watching time by excluding all those non-watching time periods, including user pause and seeking time, and playback interruption latency caused by network/system-related reasons.

IV. SINGLE MOVIE ANALYSIS

To observe the user watching time distribution on a per-video basis, in this section, we randomly select a popular movie available in our trace. By using its watching data (after necessary preprocessing and classification as described in the preceding section), we show the video’s watching time distributions in different watching phases and model them mathematically. In the next section, we will further examine the validity of the models obtained here for each of the 1,000 most popular movies.

The selected movie (also called exemplar movie hereafter) actually ranks the sixth in terms of popularity among all the movies in our trace and it has more than 400 thousand sessions. Its duration is 98.5 minutes. We only study those sessions in which users’ valid watching time is larger than zero. We call such sessions as valid sessions. Among all the sessions for the exemplar video, 87.11% are valid sessions. Among the valid sessions, 31.45%, 15.05%, 12.25% and 11.83% of sessions belong to natural leave, interruption-only, seek-only, and pause-only categories, respectively.

A. Exponential Distribution of $T_{scr}$

We first study the distribution of users’ screening time $T_{scr}$ if they leave in the initial screening phase. Fig. 2 plots the PDF of $T_{scr}$ for all valid sessions and valid sessions in natural leave pattern respectively, with \( x\)-axis on the linear scale and \( y\)-axis on the log scale. We selected 5 seconds as the bin width for plotting the PDF. We also tried other bin sizes and found the results are similar. As shown in Fig. 2, while fluctuating, the PDF curves have a trend showing that the probability of $T_{scr}$ decreases exponentially in $[0,220s]$, indicating that users’ video screening time follows an exponential distribution. We apply linear regression on the data from 0s to 220s to obtain the model parameter \( \lambda \) and corresponding \( R^2 \) score, which are 0.0098 and 0.8837, respectively. Such a \( R^2 \) is quite close to 1, meaning the fitting is very good. For interruption-only, pause-only, and seek-only sessions, we also find that their user watching time follows exponential distribution, with \( \lambda \) close to 0.01. We did not plot them in Fig. 2 as they overlap with the current curves.

B. Truncated Power-Law Distribution of $T_{watch}$

We now study the distribution of movie watching time of users who continue watching beyond the initial screening phase, i.e., user movie watching time $T_{watch}$. As shown in Fig. 2, the distribution of user screening time approximates exponential distribution during the initial phase $[0,220s]$. Thus, we choose 220s as the length of the initial screening phase for the exemplar movie. Accordingly, we filter out all those sessions that terminated in the initial screening phase and observe user watching time of the remaining sessions. Fig. 3 plots the CCDF of $T_{watch}$ of all the remaining sessions, and sessions in different patterns, on a log-log scale. The watching time distribution of interruption-only pattern is almost exactly the same as that of natural leave pattern. Thus, we do not plot them in Fig. 3. This is because 93.16 percent of interruption-only sessions experience less than five interruptions, which have no obvious impact on users’ watching time distributions.
As shown in Fig. 3, the CCDF of $T_{\text{watch}}$ for sessions in natural leave pattern is very straight on the log-log scale, indicating that $T_{\text{watch}}$ for this type of sessions follows power-law distribution. Moreover, the CCDF quickly drops to zero when $T_{\text{watch}}$ approaches $L$, which is users' maximum watching time in natural leave pattern. Such a distribution is the so-called "truncated power-law distribution" [12]. To obtain the power law exponent, we do linear regression on the measurement results and get the slope of the line $-\alpha = -0.4076$, with fitness score $R^2 = 0.988$. We also characterize the distribution of $T_{\text{watch}}$ for seeking-only or pause-only sessions. As shown in Fig. 3, the CCDFs of $T_{\text{watch}}$ for sessions of these two categories are considerably different from that of natural leave sessions and are no longer straight, meaning that $T_{\text{watch}}$ for seeking-only or pause-only sessions no longer follows truncated power law distribution. This result explains why previous studies failed to find the power law property of video watching time distribution: They mix sessions of different watching patterns and observe the aggregated watching time distribution of all sessions. The distribution, however, does not follow power law distribution.

C. Model for User Watching Time

The above results suggest that video watching time follows an "exponential + truncated power law" distribution when we focus on sessions in natural leave pattern. Based on this observation, we propose the following PDF function of user watching time $t$ on a per movie basis:

$$f(t) = \begin{cases} b \lambda e^{-\lambda t} & t \leq T_1 \\ c(\alpha - 1) T_1^{(\alpha - 1)} t^{-\alpha} & T_1 < t < L \\ c T_1^{(\alpha - 1)} L^{-(\alpha - 1)} & L \end{cases} \quad (1)$$

where $T_1$ is the border between $T_{\text{scr}}$ and $T_{\text{watch}}$, $\lambda$ and $\alpha$ are model parameters that can be obtained by linear regressions. For the exemplar movie, as shown in Sections IV.A and B, we have $T_1 = 220$ s, $\lambda = 0.0098$, and $\alpha = 1.4076$. $c$ is the probability that users watch more than $T_1$ time. For the exemplar movie, we have $c = 0.636$. As the cumulative probability of user leaving during the screening phase is $\int_0^{T_1} b \lambda e^{-\lambda t} dt$, which equals to $1 - c$, we have $b = (1 - c)/(1 - e^{-\lambda T_1}) = 0.5352$.

Eq. (1) means that, when $t \leq T_1$, i.e., in the initial screening phase, $t$ follows exponential distribution. After $T_1$, it follows power law distribution before the movie ends (i.e., $t < L$). Then, when $t = L$, as the movie ends and users have nothing to watch further, all remaining users leave. Fig. 4 plots the measurement and modeling results of CDF of users' natural watching time of the exemplar movie. As shown in Fig. 4, the modeling result fits the measurement result quite well. The K-S distance between the empirical and modeling distribution is 0.021, which is very small and validates the goodness-of-fit of our model.

V. ANALYSIS OF TOP-1000 MOVIES

In this section, we give more quantitative results in a broader view to observe whether the "exponential + truncated power law" distribution proposed in Section IV is applicable to other movies. Specifically, we study the most popular 1,000 movies as each of them has more than 10,000 sessions in the tracing period and thus has enough number of sessions for us to do statistical analysis. Their video coding rates are all 400 kbps.

A. Analysis Method

Given the large number of movies we have, it is impractical to examine each movie's distribution visually as done in Section IV. To handle this issue, we directly fit the empirical distribution curve of each movie with the hypothesis distribution model in (1), using the model fitting method described in Section IV. We then check the goodness-of-fitting by the resulting R$^2$ and K-S distance. To determine $T_1$ for each movie, we tried different $T_1$ values, and for each $T_1$ value, we fit the measured data with (1) and obtain a K-S distance. We then select the $T_1$ value which produces the minimal K-S distance. Fig. 5a plots PDF of $T_1$ for all the 1,000 movies.

B. Goodness-of-fit of Our Proposed Model

This section evaluates the goodness-of-fit of our proposed distribution model in (1) for all the 1,000 movies. Fig. 5b plots the CDF of the K-S distances between the empirical distributions and our model for all the 1,000 movies. TABLE II reports the values of K-S distances at the 10th, 25th, 50th, and 75th percentile. As shown in TABLE II, 50 percent of all the movies have K-S distances < 0.05, and 75 percent of all the movies have K-S distances < 0.07, meaning good fits. We

<table>
<thead>
<tr>
<th>Metrics</th>
<th>10th</th>
<th>25th</th>
<th>50th</th>
<th>75th</th>
<th>90th</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$ (seconds)</td>
<td>140</td>
<td>207.5</td>
<td>310</td>
<td>410</td>
<td>510</td>
</tr>
<tr>
<td>K-S Distance</td>
<td>0.02</td>
<td>0.03</td>
<td>0.05</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>$R^2$ of $T_{\text{scr}}$</td>
<td>0.659</td>
<td>0.776</td>
<td>0.846</td>
<td>0.9</td>
<td>0.928</td>
</tr>
<tr>
<td>$R^2$ of $T_{\text{watch}}$</td>
<td>0.962</td>
<td>0.975</td>
<td>0.987</td>
<td>0.994</td>
<td>0.997</td>
</tr>
</tbody>
</table>
further evaluate the accuracy of the exponential distribution model for $T_{\text{watch}}$ and the truncated power-law distribution model for $T_{\text{watch}}$ for the 1,000 movies, respectively, by using $R^2$. Figs. 5c and 5d plot the CDF of their $R^2$ scores for all the movies and TABLE II reports the values of their $R^2$ at the 10th, 25th, 50th, and 75th percentile, respectively. As shown in TABLE II, the 25th percentile of $R^2$ is 0.776, meaning that 75 percent of the movies have $R^2 \geq 0.776$. Generally, due to the fluctuation of PDF curves, linear regression on the PDF curves lead to relatively small $R^2$ values. Thus, we conclude that $T_{\text{watch}}$ follows exponential distribution for a majority of the most popular movies. For $T_{\text{watch}}$ as shown in TABLE II, 90% movies have $R^2 \geq 0.962$, which is close to 1, meaning good fits. Thus, we conclude that $T_{\text{watch}}$ follows truncated power law distribution for nearly all the most popular movies in our trace period. We further grouped sessions based on their starting times and validated that, for each group of sessions started in each hour of a day, their $T_{\text{watch}}$ follows truncated power law distribution with high $R^2$ scores as well. In addition, we find the goodness-of-fit of model and videos’ session numbers are uncorrelated. For instance, the Pearson correlation coefficient between $R^2$ of $T_{\text{watch}}$ and videos’ session numbers is 0.0274, suggesting they are nearly uncorrelated.

We also characterize the distributions of the model parameters $\lambda$ and $\alpha$ and find that log-normal distribution can characterize their distributions very well (see Figs. 5c and 5f). Besides, normal distribution can characterize the distribution of $T_I$ (see Fig. 5a). Table III shows the model parameters obtained by using maximum likelihood estimates (MLEs) and K-S distances.

VI. CHARACTERIZING USER LEAVE RATE

Users’ watching time distribution reflects their leave rate during video watching. Similar to [14], we define user's instantaneous leave rate when her movie watching time reaches $t$ as $h(t) = f(t)/(1-F(t))$, where $f(t)$ is the PDF of users’ movie watching time, $F(t)$ is the CDF, and accordingly $1-F(t)$ is the CCDF. For exponential distribution, as $f(t) = \lambda e^{-\lambda t}$ and $1 - F(t) = e^{-\lambda t}$, we have $h(t) = \lambda$, meaning that user leave rate during the initial screening phase is constant, independent of how long she has distribution, meaning that user leave rate during the initial screening phase is constant, independent of how long she has watched (screened) a movie. In comparison, for power-law distribution, as $f(t) = Ct^{-s}$, $1 - F(t) = D(Ct)^{-(s-1)}$ with $D=C/(\alpha-1)$, we have $h(t)=\alpha(t-1)/t$, meaning that when a user keeps watching a movie after the screening phase, the longer she has watched, the less likely she would leave. This finding is consistent with the observation [15] that the power-law pattern in humans could be due to people having some perception of their past activity rate and thereby reacting by reducing their activity intensity.

We then study the correlation between user leave rate during the video watching phase and their session start time. As shown in Fig. 6a, $\alpha$-1 varies with the time of day of the session start time considerably. It has the minimum value 0.2443 at 4AM and the maximum value 0.4881 at 9AM, meaning that user leave rate at 9AM is 0.4881/0.2443=2 times of that at 4AM. Moreover, user leave rate is correlated to users’ work-and-rest schedules. For instance, from 0AM to 6AM, users are usually in rest state and thus they can keep watching movies for long time (i.e., low $\alpha$-1); from 8AM to 1PM, users are usually busy and thus they have a high leave rate (i.e., high $\alpha$-1). We further find that $\alpha$-1 is independent on day of week of the session start time, suggesting that it is appropriate to characterize a video’s user watching time based on one day’s measurement data.

We next study the correlation between user leave rate and user rating. User rating is commonly used in Internet-based streaming systems for users to express their levels of satisfaction about movies. Similar to other Internet-based streaming systems, in PPLive, a user can rate a movie with one of scores from 1, 2, 3, 4, and 5, and the system then averages all the user rating scores of the movie and doubles the resulting value to obtain an overall rating for the movie in the range from 0 to 10. We adopt the “binned” correlation analysis method proposed in [1] by binning movies by their user ratings and then calculating the mean $\alpha$-1 of movies in each bin. To ensure a calculated mean $\alpha$-1 to be meaningful, we only consider bins with at least 10 movies. Fig. 6b plots the mean $\alpha$-1 versus user rating. The bin width is 0.17. We also tried other bin widths and found that the result is similar. As shown in Fig. 6b, when user ratings are lower than 9.17, while fluctuating, there is a trend showing that $\alpha$-1 linearly decreases with user rating, meaning that users quit watching movies with higher ratings with lower probability. For movies with user ratings higher than 9.17, however, the distributions of $\alpha$-1 are relatively stable, meaning that, when user rating is high enough, movies have similar user leave rates.

![Fig. 5. Distributions of model parameters and goodness-of-fit results. “Meas” and “Model” represent the curves for measurement results and fitting results using normal (see (a)) or log-normal distribution (see (e) and (f)), respectively.](image)

![Fig. 6. (a) Curve for $\alpha$-1 versus different hours of a day, from 0 AM to 24 PM. Each value shown here is the mean of $\alpha$-1 for all the 1000 movies. (b) Relation between $\alpha$-1 and user rating.](image)
We finally study the correlation between the user leave rate and movie genres. In PPLive, each movie has 1-4 genre tags among totally 22 tags. To ensure the analysis is meaningful, we only consider movie genres including at least 100 movies in the 1,000 most popular movies. Table IV shows the mean $\alpha$-1 of the selected movie genres. As shown in Table IV, $\alpha$-1 is dependent on movies’ genres. For instance, the mean $\alpha$-1 of love movies is 0.5591, which is 0.5591/0.4429=1.26 time that of action movies, meaning that user leave rate when they watch love movies is 1.26 times of that when they watch action movies. We had also studied the correlation between $\alpha$-1 and movie popularity and found no correlation. This might because all the movies under study are popular. We conjecture movie popularity will impact user leave rate if we include more long-tail movies.

<table>
<thead>
<tr>
<th>Movie Genres</th>
<th>Mean $\alpha$-1</th>
<th>Movie Genres</th>
<th>Mean $\alpha$-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action</td>
<td>0.4429</td>
<td>Thriller</td>
<td>0.5006</td>
</tr>
<tr>
<td>Horror</td>
<td>0.4644</td>
<td>Crime</td>
<td>0.5078</td>
</tr>
<tr>
<td>Comedy</td>
<td>0.4706</td>
<td>Drama</td>
<td>0.5314</td>
</tr>
<tr>
<td>Adventure</td>
<td>0.4969</td>
<td>Love</td>
<td>0.5591</td>
</tr>
</tbody>
</table>

VII. POTENTIAL APPLICATIONS OF OUR MODELING RESULTS

This section discusses some potential applications of our modeling results and findings in the preceding sections.

First, our model results and findings can be applied in a streaming workload generator (e.g., [5]) to properly generate sessions’ lengths and consequently system workload for simulating users’ early leaving behavior when they watch different types of videos at different time of day.

Second, our model can be applied in analysis of algorithms in video streaming systems. For instance, a user usually pre-fetches some media contents ahead of her playback position to ensure the playback continuity in the presence of packet losses, departure of source-peer, and network delivery jitters. When the user leaves early, however, the pre-fetched video contents are wasted. Our results are beneficial for modeling analysis of this problem and optimizing the amount of pre-fetched media to reduce the amount of contents downloaded but not watched and thus wasted [4]. At the proxy server side, it is also proposed that the proxy server should just cache the prefix of a video to maximize the effectiveness of its storage space [3]. Our modeling results are useful for properly setting the prefix size.

Finally, our results are beneficial for understanding users’ engagement [1], which have attracted a lot of attention as Internet services are becoming user-centric. Specifically, as users’ leaving rate reflects users’ engagement in a video, and our analysis in the preceding section establishes the relationship between a video's user leaving rate and the power law exponent of the video watching time distribution, our measurement results are beneficial for understanding users’ engagement. For instance, our measurement result that the power law exponent of video watching time distribution varies diurnally suggests that users’ engagement in watching video varies diurnally. Similarly, our measurement result that the power law exponents are highly correlated to video genres and user ratings reflects that users’ engagements are highly correlated to these video features.

| TABLE IV. LEAVE RATES OF DIFFERENT MOVIE GENRES. |

REFERENCES


VIII. CONCLUSION

Based on a large-scale measurement trace from a popular online video streaming system, we conducted a thorough study on user video watching time in different watching phases on a per-video basis. Our results show that exponential distribution can properly characterize user early departure time during the initial screening phase and truncated power law distribution can properly characterize user video watching time after the screening phase. Based on the built model, we further study user leave rate and show its dependency with session start time and movie features. Our results improve the understanding on user movie watching behavior and engagement and can be useful for the design and operation of online video streaming systems. In our future work, we will apply the modeling results in this paper to optimize the pre-fetching and prefix caching algorithms.