

Just One More: Modeling Binge Watching Behavior

William Trouleau*
EPFL

Azin Ashkan
Technicolor Research

Weicong Ding
Technicolor Research

Brian Eriksson
Technicolor Research

ABSTRACT

Easy accessibility can often lead to over-consumption, as seen in food and alcohol habits. On video on-demand (VOD) services, this has recently been referred to as “binge watching”, where potentially entire seasons of TV shows are consumed in a single viewing session. While a user viewership model may reveal this binge watching behavior, creating an accurate model has several challenges, including censored data, deviations in the population, and the need to consider external influences on consumption habits. In this paper, we introduce a novel statistical mixture model that incorporates these factors and presents a “first of its kind” characterization of viewer consumption behavior using a real-world dataset that includes playback data from a VOD service. From our modeling, we tackle various predictive tasks to infer the consumption decisions of a user in a viewing session, including estimating the number of episodes they watch and classifying if they continue watching another episode. Using these insights, we then identify binge watching sessions based on deviation from normal viewing behavior. We observe different types of binge watching behavior, that binge watchers often view certain content out-of-order, and that binge watching is not a consistent behavior among our users. These insights and our findings have application in VOD revenue generation, consumer health applications, and customer retention analysis.

Keywords

Binge Watching, Censored Poisson Regression, Mixture Model

1. INTRODUCTION

In recent years, the ability to record user behavior has given rise to datasets containing consumption patterns of individuals for activities as diverse as alcohol consumption [1] and gambling [2]. Of course, not all users consume at the same rate, with some users specifically over-consuming during a single session. This is a compelling problem, as the ability to identify the over consumption of food, alcohol, or gambling behavior has applications in public health [18]. The focus of this paper is on modeling and inferring

*This work was completed while at Technicolor Research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '16, August 13 - 17, 2016, San Francisco, CA, USA

© 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM. ISBN 978-1-4503-4232-2/16/08...\$15.00

DOI: <http://dx.doi.org/10.1145/2939672.2939792>

the user consumption behavior in order to identify abnormal over-consumption, or “binging” behavior from real-world consumption data. Specifically, we focus on user binge watching of video content, or “binge watching”.

For years television episodes were consumed in a standard network broadcast format, where each new episode was released once a week and then consumed one-at-a-time. In recent years, with the advent of video on-demand (VOD) services, consumers now have the ability to access full seasons of television episodes at once. This has led to the rise of binge watching, where multiple TV episodes, and potentially entire seasons, are consumed in a single viewing session. Recent market research studies have shown that over 60% of consumers binge watch on a regular basis [17]. This change in watching habits, and consumers demands to continue binge watching, has resulted in modified viewership numbers, content release schedules, and consumer expectations.

Unfortunately, the precise modeling of user binge watching behavior is difficult. While the popular press usually defines binge watching with respect to a specific number of episodes consumed for all scenarios (*e.g.*, between two and six episodes in one sitting [3]), this ignores factors such as behavior changes due to the type of content or the day of the week. To create an accurate viewership model from real-world VOD playback data, several factors must be considered. The first factor is incorporating when the user has consumed all their available items, resulting in “censored” data, where the user cannot possibly consume more but may desire to. The second is the lack of homogeneous behavior among all users, where the binge users may behave considerably different than those that consume only a handful of episodes. Finally, external factors, such as the time of day, or details about the specific object being consumed – such as, what specific show is being watched – can determine different behavioral patterns.

In this paper, we model, characterize, and infer consumption behavior of users by accommodating all the factors mentioned above using real-world playback observations. We reject a one-size-fits-all approach and create a statistical mixture model that adapts to the many sub-populations of the observed user base. Additionally, we account for the possibility of censored data and learn the contribution of covariates (*i.e.*, external factors) on the number of items consumed. The result is framing our user viewership problem as *censored Poisson regression with latent factors*. To estimate the relevant parameters of this model, we derive a novel Expectation-Maximization (EM) algorithm that incorporates these features.

Given our model, we infer various aspects of the user consumption behavior. This includes predicting the total number of episodes a user will consume from the start of a viewing session. Our technique results in an improvement of 4% in RMSE and 9% in MAE over other tested techniques. Additionally, we show how this model

can predict if the user will continue watching another episode during the session. For this application, our model results in classification AUC over 7% better than off-the-shelf techniques. These consistent improvements demonstrate the accuracy of our viewing model.

Finally, we demonstrate how this model can be used to detect binge watching sessions. We determine what constitutes “binge watching”, identify sessions that involve binge watching, and show how behavior of users in such sessions deviates from normal behavior. We describe the particular television show titles and genres that cause binge watching, with the observation that comedies (like “The Big Bang Theory” and “How I Met Your Mother”) result in more and longer binge watching sessions than dramatic shows (like “Homeland” and “The Walking Dead”). We also discover that users binge differently dependent on the content. On story-driven television shows (such as many dramas or “How I Met Your Mother”), a vast majority of users binge the episodes sequentially, allowing them to complete, or catch-up on, episodes quickly. On shows without a dominant storyline across episodes (such as “The Big Bang Theory” or “Modern Family”), almost half the binge sessions view episodes out-of-order, possibly as a background activity. Our insights demonstrate how binge watching is a distinctly different activity compared with regular viewing behavior.

2. RELATED WORK

Characterizing aggregate traffic on Video On-Demand (VOD) services has been a well studied area. Prior work often focused on detailing the traffic patterns from a specific service, such as Teliasonera [4], China Telecom [22], or YouTube [7]. These specific papers focused on content popularity and viewing trends for the purposes of service improvements. This has direct application on content caching and availability, but is separate from our focus on individual user sessions. The modeling of individual user sessions in VOD was the focus of [6]. This study analyzed content popularity and how it changes as a result of individual user behavior. To the best of our knowledge, no prior VOD study targets viewer consumption patterns with the consideration of the application of binge-watching characterization and inference. To this date, the primary source of binge watching characterization has come from explicit market research surveys.

Netflix conducted a survey on the binge watching behavior of 3,078 adults aged 18 and older, of whom 1,496 people (48.6% of respondents) stream TV shows at least once a week [3]. Among these participants, 61% reported they binge watch regularly, where binge watching is defined as watching two to three episodes of a single TV series in one session. More recently, TiVo conducted a similar survey on a group of 15,196 users [17]. They define binge watching as watching 3 or more episodes in one day. The results of this survey indicate that 91% of users report binge watching as a common behavior, of whom 40% and 69% reporting they had at least one binge watching session within a week and within a month of the survey, respectively. According to the responses from both surveys, factors driving binge watching include catching up on TV shows and compensating for the delay in learning about a show since the first time it was aired. We note that all of these prior research has focused on qualitative user responses, and not on actual user playback information.

Our focus on modeling user episode playback takes the form of building regression models on event counts. Parametric regression models, such as Poisson regression model and its variants [15], have been extensively used for modeling event data. Extensions, such as censored Poisson regression model in medical and biological studies [21], and mixtures of Poisson regression has also

been applied in various fields [10, 11, 16]. These extensions can be viewed as a limited case of our proposed model of censored Poisson regression with latent factors. Recently Karlis et al. [13] proposed a censored mixture Poisson regression model which is similar to our model. In contrast, we focus on session dependent censorship while this prior work considered the traditional setting where censorship threshold is fixed for all observations. Additionally, Karlis et al. simplifies their model parameter estimation, whereas our algorithm optimizes the exact likelihood. Finally, our regression model is a more general form of survival analysis, which has been commonly used in various other time-to-event domains, such as user engagement modeling and churn prediction [12], patient treatment studies [14], and auction price inference [20].

3. VIDEO ON-DEMAND DATASET

We received a sampled set of anonymized users from a US-based streaming video on-demand service across a timespan of 16 months from January 2014 to April 2015. This service is pay-per-content and consists of user interactions (*e.g.*, content play actions, pause actions, etc.) from TV, tablet, and phone devices. Our focus throughout the remainder of the paper is on modeling television consumption habits via this dataset, specifically with assessing abnormal binge consumption patterns. In order to assess viewing consumption patterns, we must first define a discrete television viewing session. Informed by prior work, we use the following definition,

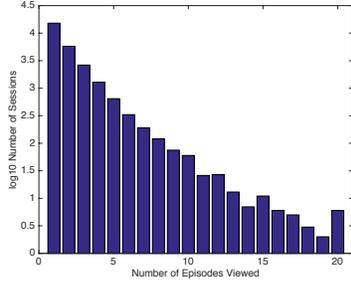
DEFINITION 1. *A watching session consists of all interactions of a user containing watching at least one episode of television and with less than an hour between interactions.*

As a result, we created user viewing sessions over our sample of VOD data. We then removed the least popular shows and kept television series that have been watched in more than 100 sessions. In order to filter out infrequent users, we only kept sessions of users who purchased five or more episodes. The resulting dataset contains 65 popular television titles, 3,488 users, and 26,404 total viewing sessions across these users, with each session lasting an average of 91.8 minutes and a median of 62.0 minutes.

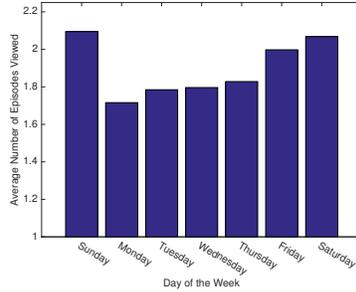
Considering just the number of episodes consumed in each session, the histogram can be seen in Figure 1a presented in log scale. As expected, we find a majority of users consume only a handful of episodes during each session. We also find that this distribution has a relatively heavy tail, indicating potential abnormal over-consumption behavior in some sessions.

Using this actual VOD playback dataset, we find that viewing behavior is not a consistent phenomenon. For example, in Figure 1b, we show the average number of episodes viewed per session given the day of the week. This plot indicates that viewers have shorter viewing sessions on weekdays compared with the weekend (*i.e.*, Friday, Saturday, and Sunday). Additionally, we show the average number of episodes consumed relative to the start hour of the session in Figure 1c for both weekday and weekend. Again, we find that the time context changes user behavior. In particular, there are more viewing activities during weekend nights as opposed to weekend days, and the difference between weekend and weekday views appears to be higher later in the day. Finally, our dataset also reveals different behaviors relative to the mechanism the viewer uses to consume content. We find that on mobile devices users consume on-average 1.58 episodes per session, while on televisions they consume 2.00 episodes on average.

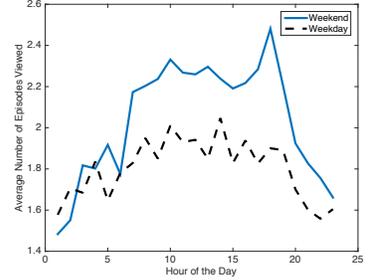
In addition to the user context of time and device, viewing behavior is also modified by the content itself. We find that TV shows generally fall in two categories in terms of the length of their



(a) Histogram with respect to the number of episodes viewed per session in log scale.



(b) Average number of episodes viewed per session relative to the day of the week.



(c) Average number of episodes viewed per session relative to the hour of day.

Figure 1: Viewing statistics.

individual episodes; with episodes commonly either 22 minutes or 44 minutes in length, such that with commercials the episodes would be either a half-hour or one hour in length, respectively. In our VOD data, the average number of episodes per session is reported to be 1.79 for longer episodes whereas it is 2.54 for shorter episodes.

In Table 1, we also show the average number of episodes consumed for specific television shows. We find that this changes dramatically between different titles and different genres. For example, genres of drama/horror/action tended to have fewer episodes viewed per session compared with television shows in the comedy genre. Additionally, we find that even shows in the same genre can have different behavior, such as the more narrative driven comedy show “How I Met Your Mother” resulting in viewers watching on-average half an episode more per session compared with the less story-focused comedy of “The Big Bang Theory”.

Table 1: Average Number of episodes consumed per session for given television shows.

Title	Genre	Mean Episodes per Session	Running Time per Episode
Walking Dead	Horror	1.72	44 min.
Breaking Bad	Drama	1.78	49 min.
Arrow	Action	2.01	44 min.
Big Bang Theory	Comedy	2.51	22 min.
How I Met Your Mother	Comedy	3.07	22 min.

We also need to consider the availability of content for particular television shows. For example, consider there are only three episodes of a particular television series available for the user, and the user consumes all three episodes. This is *censorship* behavior. In our dataset, we find that 20.9% of all sessions are censored, indicating that the termination of a session may not have been by user choice. We find this phenomena is not uniformly distributed across television shows, with 41.1% of “The Walking Dead” sessions censored, only 6.04% of “The Big Bang Theory” censored, and 17.4% of “How I Met Your Mother” sessions censored.

Our observations from this dataset indicate considerable differences in viewing behavior due to user context (*e.g.*, time of the day, day of the week) and the television show content. In addition, the frequent occurrence of session censorship indicates that many of the session terminations may be forced by content availability. These insights indicate the need to consider these multitude of factors in order to accurately model user viewing behavior.

4. CONSUMPTION BEHAVIOR MODEL

The characterization of our playback dataset indicates pre-

dictable patterns in user behavior. We find that certain timing, content, and metadata all result in distinct differences in viewing patterns. This points to the ability to model user behavior by incorporating these factors. An accurate viewing model allows for user behavior inference, such as, prediction of consumption, estimation of session length, and the extraction of binge watching events.

In this section, we present the assumptions and details of the proposed model of the consumption behavior of users. As mentioned earlier, a key point in our model is to account for the possibility of censorship for sessions where the user cannot consume more episodes but may, in fact, desire to. We also consider the impact of external factors, such as content and context features, on user consumption behavior. Finally, to account for the variability of behavior among user population, the proposed model consists of a mixture of components. A novel Expectation-Maximization (EM) algorithm is proposed to estimate the parameters that correspond to these factors.

4.1 Distribution Assumptions

Let v_i denote the number of episode views in the session i ¹. As defined previously, a session is the unit of interaction time of the user with the VOD in a single sitting, and the number of views is the discrete value of interest on which our model focuses. An initial assumption is that this value follows a Poisson distribution with parameter λ , whose probability mass function is,

$$f(v_i) = \Pr(v_i; \lambda) = \frac{\lambda^{v_i}}{e^{-\lambda} v_i!} \quad (1)$$

We choose the Poisson distribution for this purpose mainly because it is suitable to express the probability of counting events, such as the number of episodes viewed. While different discrete distributions could be used here, such as a geometric distribution, we later find in Section 5 that this offers no improvements.

4.2 Censorship in Consumption Behavior

In the VOD domain of our interest, we consider censorship to occur if the last episode watched in a session is the latest episode of the corresponding television series that is aired and available on the VOD service at the time of viewing². Therefore, the act of censorship in our setting varies and depends on the current status of the show at the time of a session. For ease of notation, we introduce

¹For ease of notation, we consider v_i to be the number of episodes consumed in session i , minus one (*e.g.*, viewing one episode would result in $v_i = 0$.)

²It is noted this concept of censorship is different from that of used in media content which is known as the act of suppressing unwanted parts from the content.

the value h_i to denote the number of episodes available to the user during session i . This *copyright threshold* depends on the content that is viewed in the session; hence, it is session dependent. This is different from the standard censorship setting in which a fixed censorship threshold across all observations is induced by, for instance, the ending of a clinical trial [14] or user inactivity after a certain point on the study [9].

4.3 Censored Poisson Regression with Latent Factors

We now define our consumption model that incorporates (1) - session censorship, (2) - the contribution of user context in the form of covariates, and (3) - the heterogeneous nature of the user population.

Censorship – We define $c_i = \mathbb{I}(v_i = h_i)$ as a binary variable indicating whether the i^{th} session is censored, where the user has consumed the latest available episodes in the session. By assuming independence across N observed sessions, the likelihood of single Poisson distribution with censorship is obtained as:

$$\Pr(\mathbf{v}; \lambda) = \prod_{i=1}^N \left[\frac{\lambda^{v_i}}{e^{\lambda} v_i!} \right]^{1-c_i} \left[\Pr(v_i \geq h_i) \right]^{c_i} \quad (2)$$

Note that in Equation (2), the likelihood of viewership for a censored observation is considered as $\Pr(v_i \geq h_i)$ to take into account the chance that the user could continue watching any further episode(s) of the series, had the episode(s) been available for viewing at the time.

External Factors – Each session i can be represented by a group of external factors, such as the viewing device used, the time of day, etc. These constitute the vector of d covariates $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,d})^\top$. Instead of assuming a fixed λ parameter for all sessions (as in Eq. (1) and (2)), a session dependent parameter λ_i is introduced to capture the dependency of consumption rate on these factors as discussed in Section 3. We use the log-linear function commonly used as the link function in Poisson regressions,

$$\log \lambda_i = \mathbf{x}_i^\top \boldsymbol{\beta} \quad (3)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^\top$ is the corresponding coefficient vector. This facilitates the derivation of our EM algorithm presented later.

Mixture Components – Finally, to account for the variability in the behavior of the user population, we consider a mixture of Poisson distributions to model the heterogeneous consumption behavior between sessions.

Let K denote the number of mixture components, and $k \in \{1 \dots K\}$ denote the index of each component. The probability mass for v_i can be written as follows:

$$\Pr(v_i; \boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{k=1}^K \pi_k \left[\frac{\lambda_{i,k}^{v_i}}{e^{\lambda_{i,k}} v_i!} \right]^{1-c_i} \left[\Pr_k(v_i \geq h_i) \right]^{c_i} \quad (4)$$

where π_k is the weight of the mixture component k , $\pi_k \geq 0$, $\sum_{k=1}^K \pi_k = 1$, and $\Pr_k(v_i \geq h_i)$ is the likelihood of watching at least h_i episodes in the session according to component k of the mixture model. Finally, the likelihood of all N sessions is:

$$\Pr(\mathbf{v}; \boldsymbol{\beta}, \boldsymbol{\pi}) = \prod_{i=1}^N \sum_{k=1}^K \pi_k \left[\frac{\lambda_{i,k}^{v_i}}{e^{\lambda_{i,k}} v_i!} \right]^{1-c_i} \left[\Pr_k(v_i \geq h_i) \right]^{c_i} \quad (5)$$

Note that, the session dependent Poisson parameter and the coefficient vector, respectively denoted by $\lambda_{i,k}$ and $\boldsymbol{\beta}_k$, both depend on the component k . In other words, the heterogeneity of the user population results in various behavioral patterns that are captured

Table 2: List of notations.

Notation	Definition
\mathbf{x}_i	Covariate vector of each session i
v_i	The number of views in each session i
h_i	The number of episodes available for session i
c_i	Censorship indicator for each session i
$\lambda_{i,k}$	Poisson parameter for component $k \in \{1, \dots, K\}$ and corresponding to session i
$\boldsymbol{\beta}_k$	Coefficient vector for component $k \in \{1, \dots, K\}$
π_k	Mixture weight for component $k \in \{1, \dots, K\}$

by different variants of these parameters. Similar to Equation (3), the log-linear relation holds between these two parameters in each component k : $\log(\lambda_{i,k}) = \mathbf{x}_i^\top \boldsymbol{\beta}_k$. It is noted that the censored likelihood $\Pr_k(v_i \geq h_i)$ is calculated using the session dependent consumption rate $\lambda_{k,i}$, hence, it is also a function of $\boldsymbol{\beta}_k$. A summary of notations is provided in Table 2.

4.4 Parameter Estimation

We consider Maximum Likelihood Estimation (MLE) for the model parameters $\Theta = \{\boldsymbol{\beta}_k, \pi_k\}_{k=1}^K$ given N observed sessions, their covariates, and the session dependent censorship thresholds.

A closed-form MLE is intractable in our mixture model, hence we instead use an Expectation-Maximization (EM) based approximation [5, 8]. We note that the introduction of covariates and session dependent censorship requires the derivation of a specific EM algorithm for our model.

Let $z_i \in \{1, \dots, K\}$ be the mixture assignment of session i . Our goal is to lower bound the data likelihood function by,

$$\log \Pr(v_i; \Theta) \geq \sum_{k=1}^K q(z_i = k) \log \frac{\Pr(v_i, z_i = k; \Theta)}{q(z_i = k)} \quad (6)$$

and maximize this lower-bound. In the E-Step of the iteration t of the algorithm, we fix $\Theta = \Theta^{(t-1)}$ from M-Step of the previous iteration and maximize the lower-bound with respect to all distributions $q(z)$; in the M-Step, we fix $q(z)$ and find the optimum $\Theta^{(t)}$ that maximizes the lower-bound. The E-Step and M-Step are conducted alternatively.

E-Step: By Jensen's inequality, the optimal solution for $q(z)$ is $q(z_i = k) = \Pr(z_i = k | v_i; \Theta^{(t-1)})$ with $\Theta^{(t-1)}$ from previous iteration. This can be calculated explicitly,

$$\begin{aligned} \tau_{i,k}^{(t-1)} &:= \Pr(z_i = k | v_i; \Theta^{(t-1)}) \\ &= \frac{\pi_k^{(t-1)} \left[\frac{\lambda_{i,k}^{(t-1) v_i}}{e^{\lambda_{i,k}^{(t-1)}} v_i!} \right]^{1-c_i} \left[\Pr_k^{(t-1)}(v_i \geq h_i) \right]^{c_i}}{\sum_{j=1}^K \pi_j^{(t-1)} \left[\frac{\lambda_{i,j}^{(t-1) v_i}}{e^{\lambda_{i,j}^{(t-1)}} v_i!} \right]^{1-c_i} \left[\Pr_k^{(t-1)}(v_i \geq h_i) \right]^{c_i}} \end{aligned} \quad (7)$$

where $\Pr_k^{(t-1)}$ is calculated using $\Theta^{(t-1)}$. Note that $\tau_{i,k}$ is the posterior probability of the session i being sampled from mixture component k .

M-Step: Using Eq (7), the lower-bound in Eq. (6) after E-Step can be written as:

$$\begin{aligned} Q(\Theta; \Theta^{(t-1)}) &= \sum_{i=1}^N \sum_{k=1}^K \tau_{i,k}^{(t-1)} \left[\log(\pi_k) + c_i \log(\Pr_k(v_i \geq h_i)) \right. \\ &\quad \left. + (1 - c_i)(v_i \log(\lambda_{i,k}) - \log(v_i!) - \lambda_{i,k}) \right] \end{aligned} \quad (8)$$

Now we optimize $\Theta^{(t)} = \arg \max Q(\Theta; \Theta^{(t-1)})$. For π_k :

$$\pi_k^{(t)} = \arg \max \sum_{i=1}^N \tau_{i,k}^{(t-1)} \log(\pi_k) + \gamma(1 - \sum_j \pi_j) \quad (9)$$

where γ is the Lagrange multiplier to take into account the constraint that π_j values must sum to one. For β_k , we numerically solve the following problem:

$$\beta_k^{(t)} = \arg \max \sum_{i=1}^N \tau_{i,k}^{(t-1)} \left[(1 - c_i)(v_i \log(\lambda_{i,k}) - \lambda_{i,k}) + c_i \log(\text{Pr}_k(v_i \geq h_i)) \right] \quad (10)$$

The detailed derivations can be found in Appendix.

Algorithm 1 Censored EM Algorithm: EM-fit

Input: Session covariates $\{x_i\}$, consumption observations $\{v_i\}$, censorship thresholds $\{h_i\}$, number of iterations T , number of components K .

Initialize parameters $\Theta^{(0)} = (\pi^{(0)}, \beta^{(0)})$

for $t \in 1 \dots T$ **do**

E-Step: Compute $\tau_{i,k}^{(t)}$ using Equation 7, for $i = 1 \dots N$, $k = 1 \dots K$.

M-Step: Compute $\pi_k^{(t)}$ using Equation 9, and $\beta_k^{(t)}$ using Equation 10, for $k = 1 \dots K$

end for

Output: Estimated parameter values, $\hat{\Theta} = (\pi^{(T)}, \beta^{(T)})$

We present the pseudocode of the estimation algorithm, called EM-fit, in Algorithm 1. EM-fit receives three session dependent inputs: covariates \mathbf{x}_i , episode counts v_i , and the corresponding censorship criterion h_i for sessions $i = 1, \dots, N$, and the number of components K . EM-fit iterates over the E-Step and M-Step for T steps, where T is the index of the step in which the convergence criterion is satisfied. We use the common convergence criterion that the algorithm is deemed to have converged when the change in the data log-likelihood falls below some threshold, which we specify in Section 5.

5. EXPERIMENTAL STUDY

In this section, we present results on the real-world VOD viewership dataset described in Section 3. We first show that our model can fit better than simpler models and that each of our factors contribute to a more accurate fit to our observed data. We then show that our model can be used to predict user behaviors in two application scenarios: inferring the number of episodes viewed in a session and classifying if the user will continue to watch the next episode.

To infer our model parameters, we run Algorithm 1. For the specific choice of covariates, \mathbf{x}_i , we consider the following:

- The title of television series, e.g., *Arrow*, *Homeland*
- The hour in the day when a session starts, e.g., 7 AM, 10 PM
- The day of week when a session starts, e.g., Monday
- The device used in the session, e.g., iPad, iPhone, television

All these features are binarized and concatenated, resulting in a feature vector for each session of dimension 104.

We run Algorithm 1 until the change of the log-likelihood on training data is below 0.01%. For updating the coefficient vectors, β_k , we use gradient descent and stop when the frobenius norm of gradient is below 0.01. Parameter vector $\beta_k^{(0)}$ is randomly initialized and $\pi^{(0)}$ is uniformly initialized from the $K - 1$ dimensional

simplex. To avoid local minimas, we use 10 random initialization for Algorithm 1. Finally, all results in this section are presented with respect to 5-fold cross validation.

5.1 Model Selection

First the hypothesis of using the Poisson distribution is tested. We compare the Poisson distribution against another candidate, geometric distribution. The estimation for geometric distribution with covariates, censorship, and latent factors can be similarly derived as we performed for the Poisson case in Section 4. Figure 2 compares the empirical probabilities of the observed number of episodes per session with the fitted models for a single Poisson, single geometric distribution, mixture Poisson ($K = 3$) and mixture geometric ($K = 3$). We find the average log-likelihood to be -1.30 for single Geometric, -1.34 for single Poisson, -1.27 for mixture of Geometric, and -1.12 for mixture of Poisson. We find that although a single geometric distribution can fit better than a single Poisson distribution, the mixture of Poisson can fit best to the data. The figure demonstrates that this is especially true with respect to the tail of the distribution. It is also known that mixture of Poisson distributions often have thick tails which make them suitable for long-tailed data [19]. This reinforces our decision to use the Poisson distribution in our model.

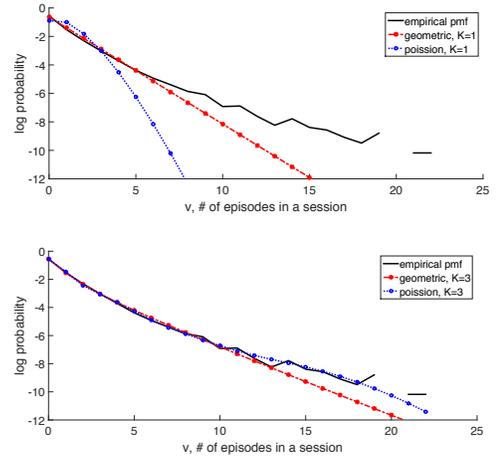


Figure 2: The learned model for single Poisson/Geometric (upper), and mixture Poisson/Geometric (bottom) on the VOD data.

Next, we consider the choice of the number of mixture components K . The predictive log-likelihood is used as a performance metric and evaluated using 5-fold cross-validation. Specifically, we estimate the model parameters $\{\beta_k, \pi_k\}_{k=1}^K$ from the training set and then evaluate the log-likelihood of the hold-out validation set for various values of K . We also normalize the log-likelihood by the number of sessions in the validation set.

The average predictive log-likelihood across 5-fold cross-validation as well as the standard deviation of this value are depicted in Figure 3. We observe that $K = 3$ is the knee point of the curve: by increasing $K = 1$ to $K = 3$ the likelihood improves by a large margin. Meanwhile, for $K > 3$ there is no significant change in likelihood. This suggests that $K = 3$ is a good candidate. As a result, we fix $K = 3$ for the rest of the paper.

5.2 Model Component Performance

We next compare the performance of our proposed model against multiple baselines in fitting to the real-world data. These baselines can be viewed as different combinations of the three key factors - mixture, censorship, and covariates. They are summarized in Ta-

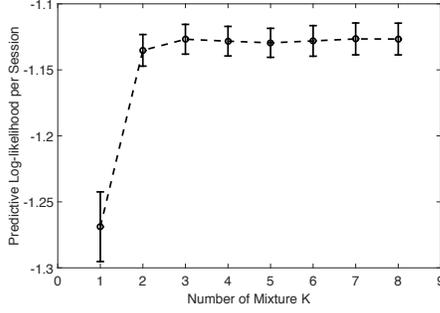


Figure 3: Predictive log-likelihood for various values of K .

Table 3: Settings of the proposed model and various baselines. ‘Y’ represents the existence of mixture, censorship or covariates in the model. ‘N’ represents their absence. Our model is ‘H’.

Description	ID	Mixture	Censoring	Covariate
Single Poisson	A	N	N	N
Poisson Regression	B	N	N	Y
Mixture Poisson	C	Y	N	N
Poisson Regression with Latent Factors	D	Y	N	Y
Censored Poisson	E	N	Y	N
Censored Single Poisson Regression	F	N	Y	Y
Censored Mixture Poisson	G	Y	Y	N
Proposed Model	H	Y	Y	Y

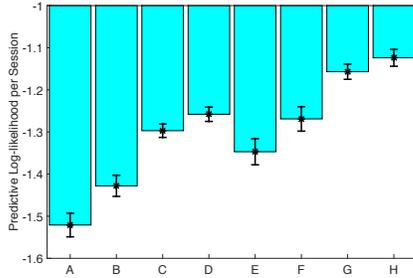


Figure 4: Predictive log-likelihood of the proposed model comparing to various baselines as detailed in Table 3. The predictive log-likelihood measures the goodness of fit of each model.

ble 3. For instance, the censored Poisson regression model (F) is a special case of our model when $K = 1$. As a consequence, we can understand the importance and significance of these components of our model. We note that EM-based estimation algorithms can similarly be derived for these models.

We use predictive log-likelihood on 5-fold cross-validation as the performance metric to evaluate the robustness of different settings as adding modeling complexity does not necessarily improve the predictive likelihood. The results are shown in Figure 4. The log-likelihoods are normalized by the number of sessions in each validation set and the standard deviations over 5-folds are reported.

Overall, our model achieves the best predictive accuracy compared to all the baselines. We find that introduction of latent factors, the censorship, and the covariates all contribute significantly to improving modeling accuracy. One can compare the proposed model (H) against the model without mixtures (F) and observe that adding these mixture components can improve the modeling accuracy by a large margin. Additionally, the improvement of our model (H) versus one without censoring (D) shows that incorporating censorship knowledge also significantly enhances the modeling accuracy.

5.3 Use Case: Predicting the number of episodes in a session

We consider predicting the number of episodes a user will watch given information available at the beginning of a session. Noting that we observe the covariates \mathbf{x}_i and the censorship threshold h_i when session starts, we predict the number of episodes a user will watch \hat{v}_i as,

$$\hat{v}_i = \mathbb{E}(v; \boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{k=1}^K \pi_k \mathbb{E}(v; \boldsymbol{\beta}_k) \quad (11)$$

where $\mathbb{E}(v; \boldsymbol{\beta}_k) = \sum_{j=0}^{h_i-1} j \Pr(j; \lambda_{i,k}) + h_i \Pr_k(v \geq h_i; \lambda_{i,k})$ is the censored expectation of a single Poisson distribution, with $\lambda_{i,k} = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_k)$ as the session dependent Poisson rates and the parameters $\boldsymbol{\beta}_k$ and π_k are learned from the training set.

We predict the number of episodes viewed in a session as the expectation based on the session dependent covariates and censorship threshold. It is noted that Eq. (11) guarantees that $\hat{v}_i \leq h_i$ in prediction using our censored model.

Table 4: The predictive root-mean-square-error (Pred.RMSE) and the predictive mean absolute error (Pred.MAE) of the proposed model against various statistical and regression baselines.

Description	ID	Pred.RMSE	Pred.MAE
Single Poisson	A	1.573 \pm 0.048	1.027 \pm 0.011
Poisson Regression	B	1.513 \pm 0.047	0.995 \pm 0.009
Mixture Poisson	C	1.573 \pm 0.048	1.027 \pm 0.011
Poisson Regression with Latent Factors	D	1.513 \pm 0.049	0.993 \pm 0.009
Censored Poisson	E	1.517 \pm 0.049	0.920 \pm 0.016
Censored Poisson Regression	F	1.456 \pm 0.051	0.895 \pm 0.015
Censored Mixture Poisson	G	1.514 \pm 0.050	0.926 \pm 0.017
Proposed Model	H	1.452 \pm 0.052	0.895 \pm 0.016
Linear Regression	L2	1.517 \pm 0.047	0.993 \pm 0.008
Thresholded Linear Regression	TL2	1.512 \pm 0.046	0.988 \pm 0.008
ℓ_1 -Regularized Linear Regression	Lasso	1.624 \pm 0.048	0.924 \pm 0.015

The problem of predicting v_i when a session starts can be viewed as a regression problem where the response variable is v_i and the regression features are \mathbf{x}_i and h_i . Therefore, we compare our algorithm against standard linear regression, ℓ_1 regularized linear regression (Lasso), and Poisson regression (log-linear regression, the setting ‘B’ as in Table 3). Additionally, we perform a version of linear regression that considers the censorship threshold h_i , such that for linear coefficients $\boldsymbol{\beta}$, the inferred value is $\hat{v}_i = \min(h_i, \mathbf{x}_i^\top \boldsymbol{\beta})$, we refer to this as *Thresholded Linear Regression*.

As evaluation metrics, we compute the predictive root mean square error (RMSE) and mean absolute error (MAE) on the validation set using 5-fold cross-validation. We compare our model against the statistical baselines in Section 5.2 and the regression baselines, using the same setting id as in Table 3 for consistency. The mean and standard deviation across 5-fold cross-validation for Predictive RMSE and MAE are summarized in Table 4.

As it is observed in the table, the proposed model (‘H’) outperforms the regression baselines (linear regression, Lasso, Poisson regression) in both RMSE and MAE metrics. We note that all linear regression methods directly minimize the square error as training objective, hence they are given the advantage in term of RMSE metric. Therefore the improvement is not expected to be significant. Our model, however, outperforms these baselines in prediction while not explicitly optimizing RMSE or MAE. This suggests

that our model can better predict the user behavior given the external factors. We note that the censored Poisson Regression ('F') can predict as accurate as our model, due to the fact that we use the expectation in Eq. (11) and the latent mixture assignment is marginalized.

5.4 Use Case: Predicting if the user will watch the next episode

Another application of our viewership model is to predict the continuation of a viewing session. Specifically, given the number of episodes a user has already watched in the middle of a session v_c , we would like to predict if they will continue to watch the next episode. Given our model, we use $\Pr(v > v_c | v_c; \beta, \pi)$ as the probability of continuation, where β, π are learned from the training set. This conditional probability can be calculated as,

$$\begin{aligned} \Pr(v > v_c | v_c; \beta, \pi) &= \sum_{z=1}^K p_z(v > v_c | v_c; \beta, \pi) p(z | v_c; \beta, \pi) \\ &= \sum_{z=1}^K \frac{p_z(v > v_c; \beta, \pi)}{p_z(v \geq v_c; \beta, \pi)} p(z | v_c; \beta, \pi) \end{aligned} \quad (12)$$

where $p(z | v_c; \beta, \pi)$ is the posterior likelihood of component membership given the current watched episodes as in Eq. (7).

We then predict the event of continuation by thresholding this conditional probability. Note that this prediction problem is a standard binary classification problem where class labels are 1 if the user continued to watch or 0 if the user stopped. Therefore, we compare our model against the standard classification algorithms using logistic regression and support vector machines (SVMs).

Table 5: The Area Under Curve (AUC) and Cross Entropy of the proposed model compared against various statistical and classification baselines.

Description	ID	AUC	Cross Entropy
Single Poisson	A	0.438 ± 0.003	-0.803 ± 0.006
Poisson Regression	B	0.539 ± 0.004	-0.754 ± 0.006
Mixture Poisson	C	0.530 ± 0.009	-0.727 ± 0.004
Poisson Regression with Latent Factors	D	0.618 ± 0.003	-0.703 ± 0.004
Censored Poisson	E	0.572 ± 0.006	-0.711 ± 0.009
Censored Poisson Regression	F	0.642 ± 0.008	-0.670 ± 0.009
Censored Mixture Poisson	G	0.631 ± 0.007	-0.652 ± 0.005
Proposed model	H	0.687 ± 0.005	-0.636 ± 0.006
ℓ_2 -Regularized Logistic Regression	LR	0.631 ± 0.005	-0.664 ± 0.002
Linear SVM	SVM	0.617 ± 0.005	NA

For each session i with v_i television episodes, we can construct v_i training or testing samples at the end of each episode. We first create 5-fold cross-validation on the sessions and further convert each session in training and validation sets into continuation data. We consider a collection of performance metrics: the Area Under Curve (AUC), the predictive cross entropy, and the correct classification rate evaluated on validation sets. The cross entropy can be computed as $\sum_i \sum_{j=1}^{v_i} y_j \log(p_j^+) + (1 - y_j) \log(1 - p_j^+)$ averaged over all test samples, where y_j is the label of continuation and p_j^+ is the probability of continuation for j -th test sample.

Table 5 summarizes the prediction result for our model against the statistical baselines summarized in Section 5.2 and the classification baselines. The mean and standard deviation across the 5-folds are reported. We observe that in Table 5, the proposed model outperforms the compared approaches in all performance metrics.

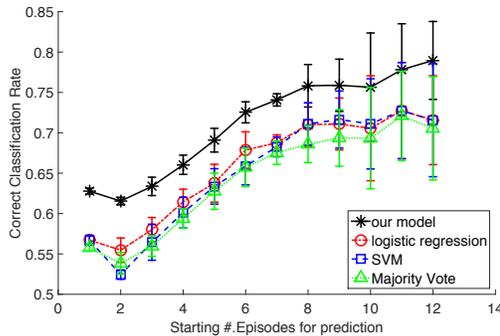


Figure 5: Classification rates on validation set when considering the current episode for the user.

We also investigate how the prediction accuracy changes as the user watches more episodes. Specifically, we design the experiment to only predict the continuation when v_c is above a certain threshold. Figure 5 summarizes the results for our model and standard classification algorithms. Note that the number of training/validation samples are smaller and more biased when the threshold for v_c increases. Therefore, we also include the simple majority votes rules as baselines. For simple majority vote, we predict a user would continue after watching v episodes if most training samples at the end of v episodes continued. As we can see in Figure 5, the prediction accuracy increases as the threshold on v_c increases. This matches the intuition that the event of continuation is easier to predict if we have more observation history in a single session. However, our algorithm consistently outperforms the baselines. This shows that our model can accurately infer the behavior type of a session and make predictions as we gather more observations in a session.

6. BINGE WATCHING CHARACTERIZATION AND DISCUSSION

The experimental study in Section 5 validates that our proposed model both fits to the data and also has applications in inferring user viewing behavior. We can also use this model to segment particular behaviors automatically from our model inference algorithm. In this section, we discuss how our model can be used to categorize and interpret distinct types of viewing behavior. In particular, we focus on the over-consumption of content, or “binge watching” behavior.

6.1 Identifying Binge Watching Behavior

We begin by considering the real-world implications of our learned mixture model. Given $K = 3$ mixture components, Figure 6 depicts the distribution of each component of the learned model for the three most popular television series in the dataset, “Walking Dead”, “Homeland”, and “The Big Bang Theory”.

We find that the learned mixture components reveal three distinct behaviors that are consistent across the dataset. (1) The component with the smallest λ captures the sessions where the user watches a few (*i.e.*, one or two) episodes in a session. (2) The component with the λ value in between those of the other two corresponds to the sessions in which user consumes an above average number of episodes (*i.e.*, 3 to 7) as indicated by the green curves in Figure 6. (3) The third component with the largest consumption rate λ (the red curves in Figure 6) represents users who watch an extreme number of episodes in a session.

We find these different types of behavior to be consistent not only across the examples in Figure 6 but also hold for all data sam-

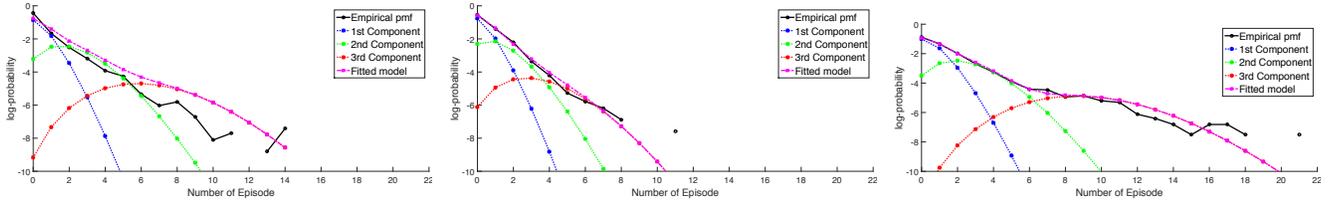


Figure 6: Empirical pmf and the fitted distribution of our model of television watching sessions for “The Walking Dead” (left), “Homeland” (middle), and “The Big Bang Theory” (right). The $K = 3$ mixture components are shown.

ples. This is validated by Figure 7 in which we plot the distributions for the learned Poisson parameters $\lambda_{i,k}$ for each component ($k = 1, 2, 3$) and across all viewing sessions³. Overall, the consumption rates for the three components are distinct. The average consumption rate for the second component is 2.9 episodes per session and is 6.8 for the third component. These two components are in accordance with the “binge watching” in prior studies [3, 17]. This motivates us to define these two component as the binge-watching components.

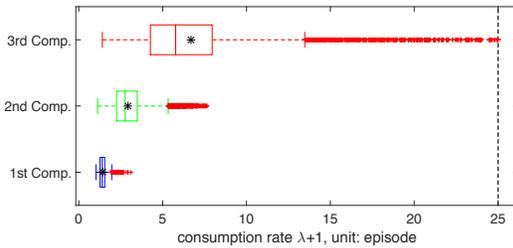


Figure 7: Box-plot of the distribution of $\lambda_{i,k} + 1$ across the three components. The central mark is the median, the edges of the box are the 25th and 75th percentiles. * indicates the mean. The color scheme is consistent with Figure 6.

A Data-Driven Binge Watching Definition - Given the two components associated with binge watching behavior, we then identify a binge-watching session by calculating $\Pr(z_i = k | v_i; \beta, \pi)$ according to (7), the posterior mixture assignment of session i to component k . Formally, we define a session to be a binge-watching session if $k = 2$ or $k = 3$ is the most likely mixture assignment of the session. This provides a fully data-driven approach to define and infer binge-watching behavior.

Different Types of Binge Watching - While existing studies characterize binge-watching as a single type of viewing behavior [3, 17], our model reveals that there are in fact two sub-classes of binge-watching. Specifically, the component $k = 3$ captures the sessions in which a significantly large number of episodes is consumed. This is a rather extreme behavior than the component $k = 2$. Therefore, we define the component $k = 3$ as “hyper-binge” watching and corresponding sessions as “hyper-binge” watching sessions. We believe this gives a finer-grain analysis of binge-watching behavior. Similarly, we refer to sessions corresponding to component $k = 2$ as the “binge” watching sessions, and the ones corresponding to component $k = 1$ as the “regular” sessions.

Using this definition, we find that 22.2% of all watching sessions fall in one category of binge watching, *i.e.*, binge watching sessions or hyper-binge watching sessions. To be more explicit, 20.1% of sessions are binge sessions and 2.1% of sessions are hyper-binge sessions. This indicates that “hyper-binge” is a rare yet extreme viewing behavior.

³Noting that we subtract 1 from v_i to simplify the model derivation, we plot $\lambda + 1$.

Content and Context-Aware Binge Watching Definition - Our model learns differences between the viewing behavior for different types of shows, days of the week, and devices used for watching. These are summarized in Figures 8, 9, and 10. In these figures, the distributions of $\lambda_{i,k}$ across the three components are shown.

Recall that in most binge-watching studies so far, a one-size-fits-all definition of bingeing is used [17]. However, Figure 8 shows that binge-watching in our model is adapted to the content of specific watching sessions. For example, in Figure 8, a binge watching session ($k = 2$) for the 20-minutes-long comedy “How I Met Your Mother” has on average more than five episodes while the 1-hour long action drama “Homeland” has on average only two episodes when bingeing. Similarly, a hyper-binge watching session ($k = 3$) for “How I Met Your Mother” has on average more than 10 episodes per session while that of “Homeland” reveals only four episodes. We can also observe that the less story-focused comedy, “The Big Bang Theory”, is binge-watched and hyper-binge watched with a smaller consumption rate than “How I Met Your Mother”, which is narrative-driven (*i.e.*, the episodes are connected by a common storyline).

Figures 9 and 10 demonstrate that our binge watching definition is context-aware. We observe that binge and hyper-binge watching sessions on Friday, Saturday, and Sunday have more episodes viewed compared to that of weekdays. Similarly, the dominance of using TV devices (*i.e.*, tv and blu-ray) over mobile (*i.e.*, iPad, and iPhone) appears to be more significance for binge watching and hyper-binge watching sessions compared to regular sessions.

Finally, we consistently observe the distinction between the hyper-binge component $k = 3$ and the binge component $k = 2$ in terms of consumption rate across different contents and context in Figures 8, 9, and 10. In addition, we note that the consumption rates for the first component are between one and two episodes per session. This suggests that the regular component $k = 1$ captures a uniform behavior across all sessions.

6.2 Binge Watching Behavior Characterization

A characterization of binge watching behavior is presented here in terms of (1) binge watching behavior across individual users, (2) difference in viewing patterns between bingeing and non-bingeing sessions, and (3) the transition patterns among different viewing behavior types.

User Characterization - Given the most likely component assignment for all sessions of a particular user, we calculate the fraction of sessions that correspond to one of the three components. For each component, we sort the fraction of users in ascending order and summarize the results in Figure 11. We note that the points on different curves with the same x-value may not correspond to the same user. Figure 11 indicates that binge-watching is not a uniform behavior across the user-population. We find that 64% of users in our dataset binge-watched at least once while 11% of users hyper-

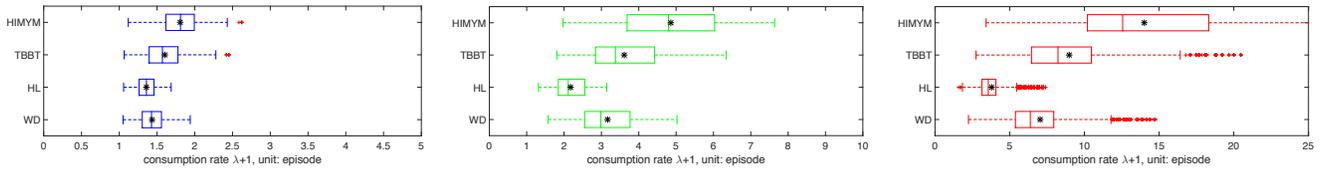


Figure 8: Box-plot of the distribution of $\lambda_{i,k}$ across the three components for four TV series: “The Walking Dead” (WD), “Homeland” (HL), “The Big Bang Theory” (TBBT), and “How I Met Your Mother” (HIMYM). * indicates the mean. **Left:** Regular, **Middle:** Binge, **Right:** Hyper-binge.

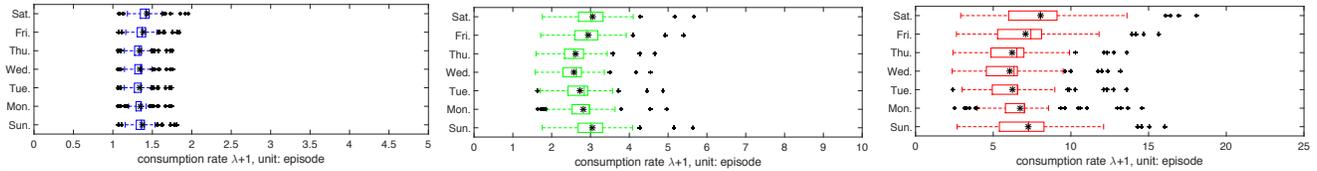


Figure 9: Box-plot of the distribution of $\lambda_{i,k}$ across the three components for different days in the week. * indicates the mean. **Left:** Regular, **Middle:** Binge, **Right:** Hyper-binge.

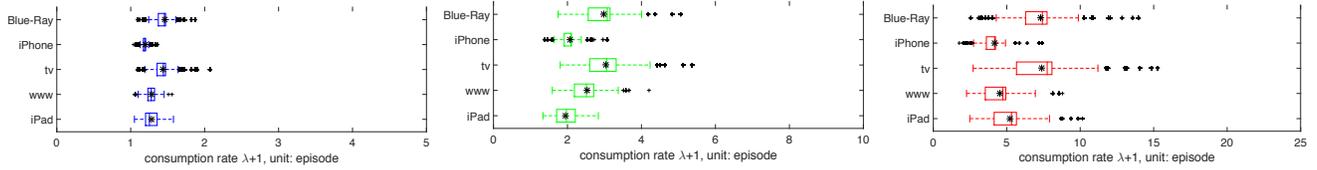


Figure 10: Box-plot of the distribution of $\lambda_{i,k}$ across the three components for a group of devices. * indicates the mean. ‘www’ denotes watching sessions on web-browsers from non-mobile devices. **Left:** Regular, **Middle:** Binge, **Right:** Hyper-binge.

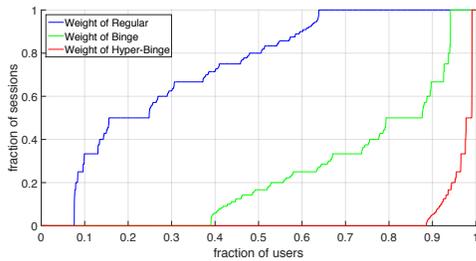


Figure 11: Distribution of different types of viewing sessions across users.

binged at least once. These numbers are consistent with the ones reported in prior user studies [3, 17], conveying that binge watching is popular among viewers. We also observe that for 7.6% of users, all their watching sessions are binge-watching. Similarly, 20% of users binge-watch in more than half of their sessions.

Viewing Patterns - While it is common that episodes are watched in a sequential order in a session to catch-up content, we find that this is not necessarily true for binge-watching. For instance, user might watch episode 10 and then go back to episode 7 or jump to episode 15 in a session. Table 6 summarizes the percentage of sessions that are watched in a sequential order across different components. While regular sessions are almost entirely sequential (97%), we observe 84% of binge watching sessions and only 76% of the hyper-binge watching sessions are sequential, indicating that binge viewers are more likely to watch content out-of-order.

A detailed investigation shows that this distinction in viewing patterns also depends on specific content. For example, story-driven series “Walking Dead”, “Homeland”, and “How I Met Your Mother” are more likely to be watched in sequential order for both binge and non-binge watching sessions. In contrast, binge-watching implies a different viewing pattern for series like “The Big Bang Theory”, “Modern Family”, and “NCIS” where each

Table 6: Fraction of sessions with sequential viewing across components.

Title	Regular	Binge	Hyper-binge
All Series	97%	84%	76%
Walking Dead	98%	87%	85%
Homeland	98%	94%	96%
HIMYM	97%	88%	77%
Big Bang Theory	87%	56%	52%
Modern Family	83%	59%	78%
NCIS	98%	71%	67%

Table 7: Transition probability of next session being type j (j -th column) given current session being type i (i -th row).

	Regular	Binge	Hyper-binge
Regular	0.82	0.16	0.02
Binge	0.66	0.30	0.04
Hyper-binge	0.59	0.32	0.09

episode is self-contained. We find that only 56% of binge watching and 52% of hyper-binge watching sessions of “The Big Bang Theory” are viewed in a sequential order. In other words, potentially half of the binge-watching sessions of “The Big Bang Theory” are not primarily for “catching-up” on the series.

Transition between Behavior Types - Given the component assignment of each session, we can calculate the probability of the next session being type j given the previous session being type i of the same user. This standard transition probability matrix is illustrated in Table 7. We find that a user currently bingeing is twice as likely to binge in their next session compared with someone currently in a regular session. We also observe that a majority of binge watchers will not binge on content in their next session, indicating that binge watching is not a consistent behavior for users.

7. CONCLUSION

The combination of video on-demand services and access to en-

tire collections of television episodes has led to the rise of binge watching behavior. In this paper, we offered a “first of its kind” study of viewer binge watching habits from real-world video on-demand records. We characterized the real-world observations to determine relevant context covariates and limitations – such as data censorship. Using these insights, we constructed a statistical viewership model and validated that it fits better to the data than all other tested models. We then showed how this viewership model can be used to infer features of the user viewing session, such as the number of episodes viewed and whether the user will continue watching. Finally, we exploit the properties of this model to introduce a methodology that automatically extracts attributes of binge watching behavior. Our characterization reveals different types of binge watching behavior, the prevalence of binge watchers viewing content out-of-order, and that binge watching is not a consistent behavior among our users. Next steps include modeling user-level behavior, examination of mechanisms to influence binge watching behavior, analysis of binge consumption of other types of products (e.g., video games, gambling, etc.) and how it differs from television binge watching.

8. REFERENCES

- [1] Alcohol Research Group Datasets. <http://arg.org/resources-tools/databases/>.
- [2] The Transparency Project. <http://www.thetransparencyproject.org/>.
- [3] Netflix declares binge watching is the new normal. <http://www.prnewswire.com/news-releases/netflix-declares-binge-watching-is-the-new-normal-235713431.html>, 2013.
- [4] H. Abrahamsson and M. Nordmark. Program popularity and viewer behavior in a large TV-on-demand system. In *ACM Internet Measurement Conference*, pages 199–210, 2012.
- [5] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.
- [6] L. Carlinet, T. Huynh, B. Kauffmann, F. Mathieu, L. Noirie, and S. Tixeuil. Four months in daily motion: Dissecting user video requests. In *IEEE Wireless Communications and Mobile Computing Conference*, pages 613–618, 2012.
- [7] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Networking*, 17(5):1357–1370, 2009.
- [8] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, pages 1–38, 1977.
- [9] G. Dupret and M. Lalmas. Absence time and user engagement: evaluating ranking functions. In *ACM international conference on Web search and Data Mining*, pages 173–182, 2013.
- [10] L. A. Hannah, D. M. Blei, and W. B. Powell. Dirichlet process mixtures of generalized linear models. *The Journal of Machine Learning Research*, 12:1923–1953, 2011.
- [11] Y. He and K. Wang. Inferring search behaviors using partially observable markov model with duration. In *Proceedings of the fourth ACM International Conference on Web search and Data Mining*, pages 415–424, 2011.
- [12] K. Kapoor, M. Sun, J. Srivastava, and T. Ye. A hazard based approach to user return time prediction. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1719–1728, 2014.
- [13] D. Karlis, P. Papatla, and S. Roy. Finite mixtures of censored Poisson regression models. *Statistica Neerlandica*, 2015.
- [14] R. Kay. Proportional hazard regression models and the analysis of censored survival data. *Applied Statistics*, pages 227–237, 1977.
- [15] C. McCulloch and S. Searle. *Generalized linear and mixed Models*. John Wiley and Sons, 2004.
- [16] S. Nagano, Y. Ichikawa, N. Takaya, T. Uchiyama, and M. Abe. Nonparametric hierarchical Bayesian modeling in non-contractual heterogeneous survival data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 668–676, 2013.
- [17] TiVO. Viewers’ New TV Reality: A “Season” = One Day. <http://pr.tivo.com/press-releases>, 2014.
- [18] R. A. Volberg. The prevalence and demographics of pathological gamblers: implications for public health. *American Journal of Public Health*, 84(2):237–241, 1994.
- [19] G. Willmot. Mixed compound Poisson distributions. *Astin Bulletin*, 16(S1):S59–S79, 1986.
- [20] W. C.-H. Wu, M.-Y. Yeh, and M.-S. Chen. Predicting winning price in real time bidding with censored data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1305–1314, 2015.
- [21] H. Yeh, B. Gajewski, P. Mukhopadhyay, and F. Behbod. The zero-truncated Poisson with right censoring: an application to translational breast cancer research. *Statistics in Biopharmaceutical Research*, 4(3):252–263, 2012.
- [22] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng. Understanding user behavior in large-scale video-on-demand systems. In *ACM SIGOPS Operating Systems Review*, volume 40, pages 333–344, 2006.

APPENDIX

We present the details of derivation for Algorithm 1 for parameter estimation in M-step. Recall that the overall objective is,

$$\Theta^{(t)} = \arg \max \sum_{i=1}^N \sum_{k=1}^K \tau_{i,k}^{(t-1)} [\log(\pi_k) + (1 - c_i)(v_i \log(\lambda_{i,k}) - \log(v_i!) - \lambda_{i,k}) + c_i \log(\Pr_k(v_i \geq h_i))] \quad (13)$$

subject to $\sum_k \pi_k = 1$, $\pi_k \geq 0$. For mixture weights it is straightforward to show $\pi_k^{(t)} \propto \sum_i \tau_{i,k}^{(t-1)}$. For β_k ’s, we optimize,

$$\max_{\beta_k} \sum_{i=1}^N \tau_{i,k}^{(t-1)} [(1 - c_i)(v_i \log(\lambda_{i,k}) - \lambda_{i,k}) + c_i \log(\Pr_k(v_i \geq h_i))] \quad (14)$$

where $\lambda_{i,k} = \exp(\mathbf{x}_i^\top \beta_k)$ and \Pr_k ’s also depend on β_k . We solve this optimization numerically using gradient descent. We first calculate $\nabla_{\beta_k} \log(\Pr_k(v_i \geq h_i))$. For $h_i > 0$, we have,

$$\begin{aligned} \nabla_{\beta_k} \log(\Pr_k(v_i \geq h_i)) &= \frac{1}{\Pr_k(v_i \geq h_i)} \nabla_{\beta_k} \left\{ \sum_{j=h_i}^{\infty} \frac{\lambda_{i,k}^j}{e^{\lambda_{i,k}} j!} \right\} \\ &= \frac{1}{\Pr_k(v_i \geq h_i)} \sum_{j=h_i}^{\infty} \frac{\lambda_{i,k}^j \nabla e^{-\lambda_{i,k}} + e^{-\lambda_{i,k}} \nabla \lambda_{i,k}^j}{j!} \quad (14) \end{aligned}$$

$$= \frac{\lambda_{i,k} \mathbf{x}_i}{\Pr_k(v_i \geq h_i)} \sum_{j=h_i}^{\infty} \frac{\lambda_{i,k}^{j-1} e^{-\lambda_{i,k}}}{j!} + \frac{\lambda_{i,k}^{j-1} e^{-\lambda_{i,k}}}{(j-1)!} \quad (15)$$

$$= \frac{\lambda_{i,k} \Pr_k(v_i = h_i - 1)}{\Pr_k(v_i \geq h_i)} \mathbf{x}_i \quad (16)$$

On the other hand, $\nabla_{\beta_k} \log(\Pr_k(v_i \geq h_i)) = 0$ if $h_i = 0$. Overall, we have,

$$\nabla_{\beta_k} l(\Theta) = \sum_{i=1}^N \tau_{i,k}^{t-1} [(1 - c_i)(v_i - \lambda_{i,k}) + c_i \lambda_{i,k} \frac{\Pr_k(v_i = h_i - 1)}{\Pr_k(v_i \geq h_i)}] \mathbf{x}_i$$