Modeling User Consumption Sequences

Austin R. Benson*
Stanford University
Stanford, CA
arbenson@stanford.edu

Ravi Kumar
Google Inc.
Mountain View, CA
ravi.k53@gmail.com

Andrew Tomkins
Google Inc.
Mountain View, CA
atomkins@gmail.com

ABSTRACT

We study sequences of consumption in which the same item may be consumed multiple times. We identify two macroscopic behavior patterns of repeated consumptions. First, in a given user's lifetime, very few items live for a long time. Second, the last consumptions of an item exhibit growing inter-arrival gaps consistent with the notion of increasing boredom leading up to eventual abandonment.

We then present what is to our knowledge the first holistic model of sequential repeated consumption, covering all observed aspects of this behavior. Our simple and purely combinatorial model includes *no* planted notion of lifetime distributions or user boredom; nonetheless, the model correctly predicts both of these phenomena. Further, we provide theoretical analysis of the behavior of the model confirming these phenomena. Additionally, the model quantitatively matches a number of microscopic phenomena across a broad range of datasets.

Intriguingly, these findings suggest that the observation in a variety of domains of increasing user boredom leading to abandonment may be explained simply by probabilistic conditioning on an extinction event in a simple model, without resort to explanations based on complex human dynamics.

Keywords. sequence mining; repeat consumption; boredom

1. INTRODUCTION

Under the rubric of recommender systems, researchers over the last many decades have developed a rich body of work predicting a user's likelihood to respond well to a particular item. Typically the item is unfamiliar to the user, and the system is an aid to discovery.

In the datasets we consider, between 15% and 59% of consumptions are in fact items already consumed by the user. However, our understanding of repeat consumption remains relatively poor. We have little sense of when a user is prone to re-consume versus seeking variety, and even knowing the user will re-consume, we have limited understanding of which already-seen item would be preferred. Finally, we lack detailed models of macroscopic elements of repeat consumption: the processes by which users encounter new material, become enamored, consume frequently, and then slowly transfer attention to other alternatives.

Here we present a first attempt at a holistic model of repeat consumption behavior that explains a number of microscopic and macroscopic properties we observe in our data. This contrasts with earlier work on repeat consumption, which studies either time independent consumption sequences [4], or only individual user-item pairs [7, 18]. Our model captures consumption sequences in three

*This work was done while the author was at Google, Mountain View, CA.

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883025.

dependent of the total lifetime, until the very end when boredom sets in. We therefore argue that more complex models formalizing these notions explicitly are not required, at least not to reproduce the statistics we consider.

Finally, we study personalization versions of our model. Earlier work finds that the likelihood to re-consume an item that was consumed i steps ago falls off as a power law in i, attenuated by an exponential cutoff. We consider fitting such a function to each user individually. While the fit of the earlier model to global distributions over many users is satisfying both visually and likelihood-wise, the fit with respect to a particular user is less accurate. Instead, we find that a double Pareto distribution can be fit to each user with a significant increase in overall likelihood. We hypothesize that the double Pareto naturally captures a regime of recency in which a user recalls consuming the item, and decides whether to re-consume it, versus a second regime in which the user simply does not bring the item to mind in considering what to consume next; these two behaviors are fundamentally different, and emerge as a transition point in the function controlling likelihood to re-consume.

For the importance of time in repeat consumption, we show that the situation is complex. Different datasets show wildly varying time dependencies with respect to repeat consumptions within a day or two. Hence we do not hypothesize a single functional form to incorporate time across our datasets; we allow the importance of time to remain non-parametric, and simply characterize the improvement in likelihood that is possible by extending the model to incorporate different levels of granularity of temporal features.

2. DATA

We now describe our datasets and some macroscopic observations about the way users consume items. We have made an effort to use several public datasets so that our results can be reproduced.

2.1 Datasets

We collected a variety of datasets in order to study user consumption patterns. The datasets fall under three broad categories: (i) music and video, where songs, videos, or artists are the consumed items; (ii) clicks data from internet browsing history, where we interpret clicks on web pages as a form of consumption intent; (iii) "check-in" data, where consumptions are physical locations. Each dataset consists of a sequence of consumptions for several users. Each user sequence is a list of consumption activities for that user, and each consumption activity consists of an identifier of the consumed item and a timestamp. The datasets are described below and the consumption statistics are summarized in Table 1.

LASTFM and LASTFMARTISTS. These datasets are derived from the complete listening habits of users on the music streaming service last.fm [6]. Users can select individual songs or listen to "stations" based on a genre or artist, where the sequence of songs comes from a recommendation system. We consider two consumption sequences: one where the consumptions are songs and one where the consumptions are artists. The data is publicly available.[2]

YOUTUBE and YOUTUBEMUSIC. These datasets contains sequences of videos watched by anonymized users on YouTube. We consider up to the last 10,000 videos watched by anonymized users with at least 100 video watches. We only consider a video "watched" if it was played for at least half of the video length. For privacy reasons, the videos are anonymized and we only consider videos watched by at least 50 distinct users. YOUTUBEMUSIC is a subset of YOUTUBE, consisting of only the music videos.

[2] http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/lastfm-1K.html

Table 1: Summary statistics of data sets. Many consumptions in the datasets are repeats. In other words, users tend to consume items that they have consumed in the past.

Dataset	# users	# unique items	# repeat consumptions	fraction repeat
LASTFM	992	1.50M	14.5M	0.69
LASTFMARTISTS	992	174K	18.2M	0.95
YOUTUBE	696K	1.44M	182M	0.26
YOUTUBEMUSIC	694K	497K	83.1M	0.44
BRIGHTKITE	51.4K	773K	3.63M	0.51
GPLUS	18.4K	1.81M	2.36M	0.31
MAPCLICKS	432K	216K	43.5M	0.38
WIKICLICKS	852K	528K	54.2M	0.15

MAPCLICKS. This dataset consists of clicks on business entities, e.g., restaurants and movie theaters, on Google Maps. We consider the sequence of all entity clicks issued by anonymized users, so consumed items are businesses. For privacy reasons, we only consider businesses clicked on by at least 50 distinct users, and only consider users with at least 100 clicks.

WIKICLICKS. This dataset consists of clicks on English Wikipedia pages by Google users. We consider the sequence of all pages clicked on by anonymized users, so consumed items are web pages. For privacy reasons, we only consider pages clicked on by at least 50 distinct users, and only consider users with at least 100 clicks.

BRIGHTKITE. BrightKite was a location-based social networking website where users could check in to physical locations. Here we consider the consumed items to be all latitude-longitude pairs of anonymized user check-ins. The data is publicly available.[3]

GPLUS. On Google+, users can share physical location and choose to make this check-in public. The dataset consists of all public check-ins made by several thousand users. The data is public.[4]

2.2 Macroscopic observations

We now make two empirical macroscopic observations about the datasets. First, items have finite lifetimes and tend to follow a heavy-tailed Pareto distribution. Second, gaps in consumption sequences of an item tend to be larger at the end of the item's lifetime. The growing gaps are evidence for user boredom with the items. In Sections 4 and 5, we present a generative model for user consumption sequences that captures these macroscopic properties of the data both empirically and theoretically.

Finite lifetimes. Table 2 quantifies finite item lifetimes by measuring the fraction of items appearing in the first 20% of a user's consumption lifetime that do not appear in the last 20% of the user's lifetime. We measure the user's lifetime in terms of total number of items consumed (*index lifetime*) and the real elapsed time between the first and last items consumed (*temporal lifetime*). Indeed, the lifetimes are finite in all of our datasets. In most datasets, fewer than 10% of the items consumed in the first part of the lifetime are still consumed in the last part of the lifetime.

Figure 1 shows the distribution of the number of times that an item is consumed by a user. We only considered items whose first and last consumption were in the middle 60% of the user's index lifetime. This filtering excludes censored sequences that began before or finished after the timeframe in which the data was collected. The count lifetime distributions are all heavy-tailed, roughly following a Pareto distribution. In the YOUTUBE, YOUTUBEMUSIC, MAPCLICKS, and WIKICLICKS data, the slope parameter is at

[3] snap.stanford.edu/data/loc-brightkite.html

[4] The data is available through the Google+ API.

Table 2: Finite item lifetimes: of items consumed in the first 20% of the sequence (in terms of index or absolute time), the table lists the fraction that do not appear in the last 20% of the sequence in terms of the number of item consumed (index), or the user lifetime (temporal). This fraction is large for all data.

Dataset	Fraction finite lifetimes Index	Temporal
LASTFM	0.78	0.86
LASTFMARTISTS	0.63	0.75
YOUTUBE	0.98	0.99
YOUTUBEMUSIC	0.94	0.96
BRIGHTKITE	0.89	0.93
GPLUS	0.96	0.98
MAPCLICKS	0.99	0.99
WIKICLICKS	0.98	0.99

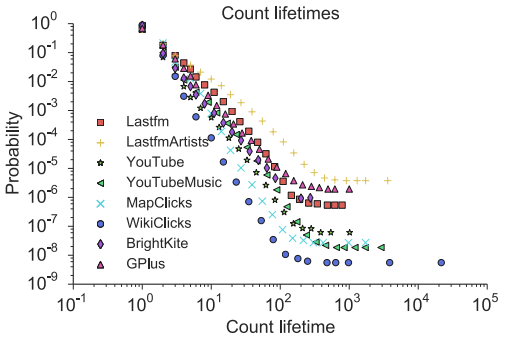

Figure 1: Probability distribution of the number of times an item is consumed by a user over all datasets. In all cases, the lifetimes tend to follow a heavy-tailed distribution.

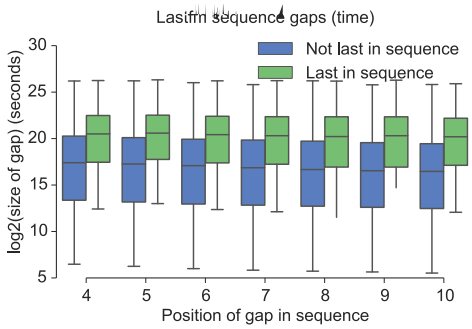

Figure 2: Sizes of temporal gaps between consumptions of the same item in LASTFM, conditioned on whether or not the gap is the last in the sequence. The last gaps are much larger.

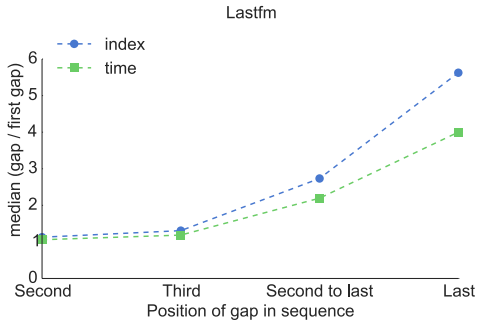

Figure 3: Median normalized index and temporal gap sizes in the LASTFM dataset. Gaps at the end of an item's lifetime are larger than gaps at the beginning of the item's lifetime. This is consistent with user's experiencing boredom.

least 3 and hence they are characterized by a distribution with finite mean and variance [24]. The other datasets all have slope greater than 1.5 and have a finite mean. Consequently, it is reasonable to expect finite lifetimes in at least some of these datasets.

Boredom and growing gaps. The notion of *boredom* in user consumption has been studied in a variety of domains, including psychology [28], consumer brand marketing [17, 22], and the web [13, 18]. For example, Kapoor et al. explicitly model user-item interactions in "sensitization" or "boredom" states [18]. Here we provide some evidence for boredom by looking at the gaps between a user's consumption of the same item. In Section 5, we will show how our model also captures boredom. An important distinction of our model is that it is generative, i.e., it provides a way to generate the entire user consumption sequence out of which boredom is a *consequence* of the model (Theorem 5.4). In contrast, prior work has focused solely on modeling a given user-item pair.

We define the *temporal gaps* as the difference between the times at which a user consumes a particular item. Formally, if the k consumption timestamps are $t_1, \ldots, t_k$, the temporal gaps are $\delta_{i+1} = t_{i+1} - t_i$, $i < k$. Figure 2 shows the gaps δ_i for various gap positions (i) in the LASTFM dataset, conditioned on whether or not the gap is the last gap in the consumption sequence. We see that there is a clear trend for the last gap to be longer, regardless of the total sequence length. This is consistent with users experiencing boredom with an item before eventually ceasing to consume the item.

We can also look at boredom in terms of the number of items consumed between consumption of the same item. If the user's consumptions are numbered $1, \ldots, N$ and a particular item is consumed at indices $j_1, \ldots, j_k$, we define the *index gaps* by $g_{i+1} = j_{i+1} - j_i$, $i < k$. Figure 3 shows the median behavior of the temporal and index gaps relative to the first gap in the LASTFM dataset. Consistent with our observations from Figure 2, we see that the final temporal gap δ_{final} tends to be 3.5 times the size of the first temporal gap δ_1, and the final index gap g_{final} is roughly 5 times the size of the first index gap g_1. Furthermore, the gaps tend to grow over time. The second-to-last gap is also larger than the first gap, but not as large as the final gap, and the second gap is still roughly the same size as the first gap.

3. RELATED WORK

Our holistic modeling of consumption sequences follows a breadth of research across computer science, marketing, and psychology.

Repeat consumption. A key component of our consumption sequence modeling is the notion of *repeat consumption* of items already consumed in the past. Anderson et al. model repeat consumption as a combination of recency (recently consumed items are likely to be consumed again) and quality (popular or high-quality items are more likely to be consumed) [4]. However, their models are time-independent. Here, we explicitly model time, in order to capture consumption inter-arrival times (Section 4.1) and to improve the repeat consumption selection model (Section 4.3). In fact, we find absolute elapsed time to be more predictive than item quality. Repeat consumption has also been studied in several domains such as visitation of web pages [1, 2], purchasing behavior in online commerce [9], and web search queries [31, 32]. Finally, Chen et al. studied short-term repeat consumption behaviors on location-based social network data [7]. While this work has focused on modeling the interaction between a user-item pair, we model the full consumption sequence of an individual, capturing interactions with many items.

Boredom in consumption sequences. The notion of boredom has long been of interest to advertisers and marketers seeking to keep users interested in a product. Here, the tradeoff between trying new products or sticking with the same brand is called *variety-seeking behavior* [17, 22]. This notion is also rooted in the classical exploration-exploitation tradeoff [12, 21, 30]. There are also studies of boredom in consumption on the web. Das Sarma et al. explicitly model boredom from a utility maximization perspective in order to understand cyclic trends [13]. Kapoor et al. model user-item pairs in states of sensitization (active consumption) or boredom [18], and find that items consumed at a lower rate tend to spend more time in the boredom state.

Related to boredom is the study of user engagement, and determining when users are likely to cancel services is known as *churn prediction* [15]. Most relevant to our work is the hazard-based modeling of Kapoor et al. [19] on LASTFM. Other related work includes user engagement patterns in social networks [20, 35].

Navigation on the web. Finally, we note that a special case of user consumption sequences is the traversal of web pages. One of the first models in this domain is the famous random surfer [26]. Detailed user studies have been performed on, for example, navigation behavior of Wikipedia [33, 34] and the relationship between queries and subsequent web navigation [8, 14].

4. A HOLISTIC CONSUMPTION MODEL

User consumption sequences are formally represented as lists of tuples (x_i, t_i), $i = 1, 2, \ldots$, where x_i is the item and t_i is the time. We model these sequences with the following high-level procedure:
(i) **Temporal model (to capture inter-arrival times).** The inter-arrival times $\Delta_{i+1} = t_{i+1} - t_i$ are generated from a stochastic process.
(ii) **Novelty model (to capture repeat versus novel consumption).** Given $\Delta_1, \ldots, \Delta_i$, determine whether or not x_i is a *novel* consumption or a *repeat* consumption.
(iii) **Choice model (to capture item identities).** If x_i is novel, draw from some distribution; otherwise, draw from the history following some distribution. The distributions depend on inter-arrival times.

In the remainder of this section, we detail these three model components. While it is possible to model these components in various orders, i.e., changing the conditioning, we choose the above ordering for several reasons. First, we treat the second step as a supervised learning problem, and it is useful to have Δ_is as features. Second, we build upon work from Anderson et al. [4] on repeat consumption in order to effectively model which items are selected assuming we condition on the item being a repeat consumption. Third, we find that modifying these repeat consumption models to account for time can significantly improve performance. Consequently, we prefer to condition on Δ_{i+1}.

In this section, our model will be *global*, i.e., all of the processes and distributions will be the same for each user. However, the models may depend on individual user behavior. For example, we use a logistic regression for determining novelty vs. re-consumption, which depends on the user's past history. However, we train this model over all users. In Section 6, we consider personalization.

4.1 Temporal model

We begin by modeling the inter-arrival times, i.e., the times between a user's consumptions. The simplest model of inter-arrival times allows IID choices from an inter-arrival distribution. However, Figure 4 shows that for some of our datasets, inter-arrivals are not independent: users exhibit different behavior when they are using the service in a "session" of active consumption as opposed to time between sessions. This behavior makes sense in, for example, music consumption, where users actively listen to several songs in a single session before taking a break from the service.

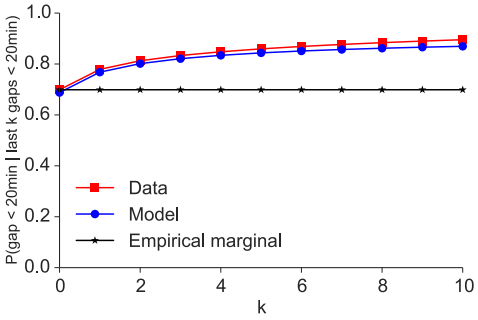

Figure 4: Evidence for active states: the probability of the next time gap in a consumption sequence being less than 20 minutes given that the last k gaps were less than 20 minutes as a function of k for the YOUTUBE data. The data and model show "active states" of continuous consumption. The marginal probability is the probability that the gap is less than 20 minutes, independent of k (equivalent to the $k = 0$ point for the data).

Behavior within a session is simple, capturing continuous consumption. Therefore, in addition to considering natural choices for IID inter-arrival distributions, we also consider a semi-Markov model of inter-arrival times, in which a user moves between two states: "within" and "between" sessions, each of which has its own inter-arrival distribution. This is in a sense the simplest possible form of dependence between arrivals, in the form of a single bit determining the state. We show that the likelihood improvement of moving from a one-state IID model to a two-state session model is large enough to justify the mild increase in complexity.[5] Also, the inter-arrival distributions within each state is simpler and natural.

A semi-Markov model. Our semi-Markov model captures active and inactive user consumption states. The following generative model captures this behavior:
(i) Draw $K \sim D_1$ consumptions for the current session.
(ii) Draw K intra-session gaps from D_2.
(iii) Draw an inter-session gap $s \sim D_3$.
We note that for unknown D_1, D_2, and D_3, this is an explicit duration hidden Markov model [23].

After analyzing the relevant data, we fix upon the following constituent distributions.

Session length distribution D_1: The number of items K in a session drawn from D_1 follows a power law with exponential cutoff,

$$\Pr(K = k) \propto k^{-\alpha} e^{-\beta k}.$$

Intra-session gap distribution D_2: An intra-session gap G_2 drawn from D_2 follows a double Pareto [25, 29] distribution,

$$\Pr(G_2 = g_2) \propto \begin{cases} g_2^{\eta-1} & g_2 \leq \gamma \\ g_2^{-\nu} & g_2 > \gamma \end{cases}. \quad (1)$$

Inter-session gap distribution D_3: Finally an inter-session gap G_3 draw from D_3 follows a simple power law,

$$\Pr(G_3 = g_3) \propto \lambda^{-g_3}.$$

Figure 5 shows the fit of each of these 3 distributions for the LASTFM dataset. Overall, these model families describe the data quite well. There is some discrepancy in the tail of the distributions (and the head of D_2 for LASTFM), which is due to a lack of data at those points. Furthermore, the inter-session gap distribution (D_3) does not capture the increased probability near 24 hours.

[5]For all model training in Section 4, we use likelihood maximization on the complete dataset.

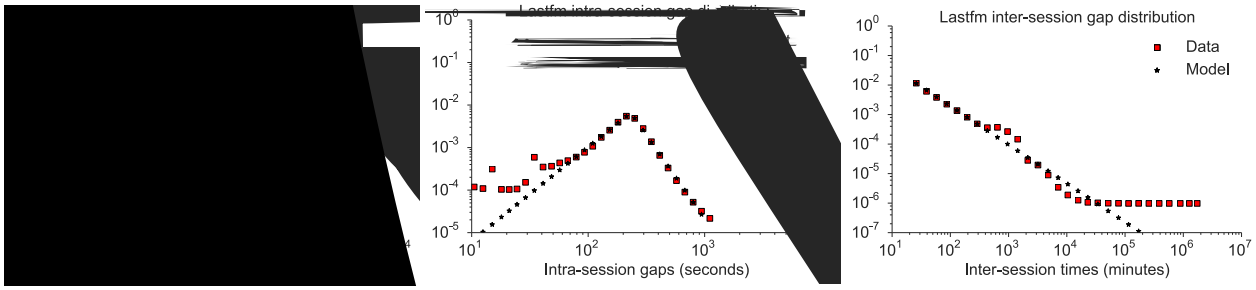

Figure 5: Distribution of session length (D_1, left), intra session gaps (D_2, middle), and inter-session gaps (D_3, right) for the LASTFM dataset and for the semi-Markov model. Figure 6 shows the full inter-event distribution for data simulated according to the model.

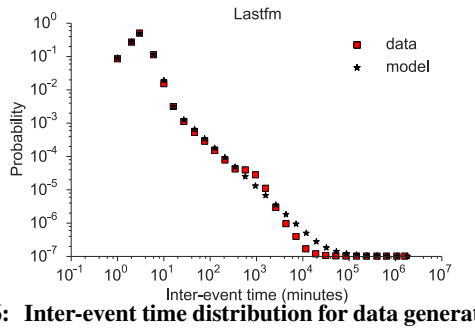

Figure 6: Inter-event time distribution for data generated with our semi-Markov model compared to the true distribution of the LASTFM dataset. For this dataset, the inter-event time is the elapsed time between the start of two consecutive song plays.

Thus, the overall model of inter-arrival times may be captured by six parameters: (α, β) for the number of items in a session, (η, ν, γ) for the inter-arrival time within a session, and λ for inter-arrival time across sessions. The values of the best fit parameters for our datasets are shown in Table 3. Figure 6 shows the overall inter-arrival distribution for the inter-arrival time from a simulation with the model and the empirical data from LASTFM. The model simulation matches the data quite well.

We also compared our semi-Markov model to several common inter-arrival models, where the samples are IID. For each IID inter-arrival distribution we consider, Table 4 shows the likelihood of this model relative to our semi-Markov model. This ratio is computed by first computing the maximum likelihood estimator for each of the IID distributions and then taking the geometric mean of the ratio of the likelihoods over every inter-arrival time in the data. It is clear from the table that our semi-Markov model out-performs these common IID inter-arrival models.

Table 3: Time arrival parameters. Music and video use 20 minute gaps, others 60. (LASTFMARTISTS parameters are the same as those for LASTFM by the construction of the dataset).

Dataset	α	β	η	ν	γ	λ
LASTFM	0.54	0.04	2.77	3.23	230.49	1.30
YOUTUBE	1.36	0.08	0.98	1.34	275.62	1.29
YOUTUBEMUSIC	1.61	0.12	1.13	4.23	241.99	1.20
BRIGHTKITE	2.47	0.00	0.00	0.50	215.94	1.16
GPLUS	2.94	0.00	0.00	0.71	55.60	1.10
MAPCLICKS	0.78	0.32	0.14	1.05	3.63	1.00
WIKICLICKS	2.46	0.00	0.00	0.69	137.94	1.06

Optimizing model parameters. To simplify parameter estimation, we fix the sessions to be all items consumed without taking a break of at least B minutes. From analyzing the data, we found that $B = 20$ is appropriate for the music and video data sets, and $B = 60$ is appropriate the clicks and check-in data. After making this assumption, the number of items per session follows a power law with exponential cutoff with maximum value B, the between-session times follow a power law with minimum value B, and we know the empirical distribution for the within-session arrival times. After performing the decomposition, we bin the data logarithmically and use a non-linear solver to best fit the probability density function. While there are more sophisticated maximum likelihood maximization procedures for estimating these parameters [11, 25], we find that simple curve-fitting works well in practice.

Table 4: Relative likelihoods of common IID distributions for inter-arrival times to the likelihood of our semi-Markov model. In all cases, our model is much more likely. This is unsurprising since the data exhibit dependence (Figure 4).

Dataset	exponential	Pareto	log normal
LASTFM	0.05	0.13	0.43
YOUTUBE	0.02	0.12	0.22
YOUTUBEMUSIC	0.01	0.07	0.12
BRIGHTKITE	0.02	0.04	0.07
GPLUS	0.02	0.03	0.06
MAPCLICKS	0.00	0.21	0.20
WIKICLICKS	0.02	0.04	0.08

4.2 Novelty model

Next, we discuss the issue of predicting when a user will re-consume. We treat this prediction as a supervised learning problem, and use features such as the user's proclivity for re-consuming, the number of items consumed by the user so far, whether or not the last few consumptions were repeat consumptions, and the time since the last consumption. The response variable is whether or not the user re-consumed (or a probability for repeat consumption).

We tested several learning algorithms in the `scikit-learn` library [27] and found that logistic regression performed the best in general. Table 5 lists the error rates for logistic regression on each dataset. We find that we get the best results on the LASTFM and LASTFMARTISTS datasets, both of which contain a large number of repeat consumptions. We note that our performance is similar to a related sequence prediction problem studied by Chen et al. on the LASTFM data [7].

4.3 Choice model

The last part of our model determines which items to consume. There are two cases: when the consumption is novel and when the consumption is a repeat.

Novel consumptions. Here, we assign identities to items being consumed for the first time by this user. Accurately determining

Table 5: Performance of logistic regression on each data set for predicting whether a consumption is a repeat consumption (repeat consumption corresponds to a positive response variable).

Dataset	Accuracy	Precision	Recall	RMSE
LASTFM	0.859	0.885	0.934	0.321
LASTFMARTISTS	0.957	0.961	0.957	0.189
YOUTUBE	0.834	0.765	0.515	0.352
YOUTUBEMUSIC	0.761	0.757	0.677	0.404
BRIGHTKITE	0.817	0.842	0.942	0.353
GPLUS	0.763	0.779	0.769	0.401
MAPCLICKS	0.677	0.623	0.388	0.447
WIKICLICKS	0.848	0.557	0.040	0.347

which item a user is likely to consume draws upon the wealth of research in recommendations and response prediction [3, 5]. We therefore assume that this section is a black box, assigning identities to items as they are first consumed by each user.

However, in order to proceed with a reasonable model, we instantiate this black box with a simple placeholder. We assign identities so as to match the popularity of items across the entire dataset. Formally, let n_e be the number of users that consume item e at least once. Let $I(\cdot)$ denote the binary indicator function. Let $x_1, \ldots, x_{i-1}$ be the consumptions of a single user and suppose that we have determined that the ith consumption is novel. Then we choose an as-yet unconsumed item from a multinomial distribution whose parameters are n_e,

$$\Pr(x_i = e \mid x_i \text{ is novel}) = \frac{I(e \notin \{x_1, \ldots, x_{i-1}\})n_e}{\sum_{e'} I(e' \notin \{x_1, \ldots, x_{i-1}\})n_{e'}}.$$

This ensures that the expected number of users that consume an item is the same as in the datasets.

Repeat consumptions. We first consider a fixed time at which a user will re-consume an item, and study in detail which item will be re-consumed. From earlier work [4], we know that a key factor in determining which item will be re-consumed at a particular time is the sequence of items consumed during prior time steps. In particular, recent work by Anderson et al. [4] used the following repeat consumption model for the ith consumption:

$$\Pr(x_i = e) = \frac{\sum_{j<i} I(x_j = e) w(i-j) s(x_j)}{\sum_{j<i} w(i-j) s(x_j)}, \tag{2}$$

where w are the recency weights and s are the item quality scores.

In this model, an item is "copied" from the past[6] with probability dependent on the number of intervening items, and the quality of the item. The model factors these two contributions by assuming they are combined as a product, and then normalizes the result. We introduce a third factor, time, which turns out to be more important than the popularity/quality of each item (see Table 6). We incorporate time with the following selection model:

$$\Pr(x_i = e) = \frac{\sum_{j<i} I(x_j = e) w(i-j) s(x_j) T(t_i - t_j)}{\sum_{j<i} w(i-j) s(x_j) T(t_i - t_j)}, \tag{3}$$

for some nonnegative function T. In this new model, an item is copied with probability that is a product of a time factor, a distance factor (number of intervening items) and a quality factor. As these factors are optimized jointly, one may view the time factor as being the change in likelihood of copying a particular item from i steps back, depending on how long ago in absolute time that past consumption occurred.

[6]Following Equation 2, the item could be copied from one of several locations if it has been consumed more than once in the past.

Table 6: Relative likelihoods of models: recency weights capture most of the likelihood, but the relative consumption times of items is more important than the quality or popularity.

	Learned scores		
Dataset	w	w and s	w and T
BRIGHTKITE	0.91	0.92	0.98
GPLUS	0.87	0.92	0.94
LASTFM	0.99	0.99	1.00
LASTFMARTISTS	0.96	0.96	1.00
YOUTUBE	0.91	0.94	0.96
YOUTUBEMUSIC	0.92	0.93	0.97
MAPCLICKS	0.81	0.82	0.99
WIKICLICKS	0.78	0.81	0.91

Learning model parameters. Let R_u be the set of all repeat consumptions for user u, and let $x_i^{(u)}$ and $t_i^{(u)}$ be the ith consumption of user u and the corresponding timestamp. Following the model in Equation 3, the negative log-likelihood of the repeat consumptions in the dataset is

$$-\log\left[\prod_u \prod_{i \in R_u} \frac{\sum_{j<i} I(x_j^{(u)} = x_i^{(u)}) w(i-j) s(x_j^{(u)}) T(t_i^{(u)} - t_j^{(u)})}{\sum_{j<i} w(i-j) s(x_j^{(u)}) T(t_i^{(u)} - t_j^{(u)})}\right].$$

This equation is not jointly convex in w, s, and T, but it is convex in each function with the other two fixed. Thus, we employ a block coordinate descent method, using a standard gradient descent procedure to maximize the likelihood with respect to w or s or T. The log-likelihood function splits with respect to any consumption of any user, so there is ample room for parallelizing these procedures.

We now compute the gradients with respect to T. The gradients with respect to w and s are similar. Since the arguments to T are continuous, we model the function as a piecewise constant with a logarithmic binning scheme. Let there be m bins, so that we can represent the time scores T as a vector in $\mathbb{R}_+^m$. Furthermore, let $b(t) \in \{1, \ldots, m\}$ be the bin index for any time difference t. Finally, for a fixed index i and a binary predicate I that depends on j, denote

$$A_{u,i}(I) = \sum_{j<i} I(j) w(i-j) s(x_j^{(u)}) T(t_i^{(u)} - t_j^{(u)}),$$

$$B_{u,i}(I) = \sum_{j<i} I(j) w(i-j) s(x_j^{(u)}).$$

For a given user and a given repeat consumption, the gradient with respect to T for a given repeat consumption is

$$\frac{\partial LL}{\partial T_k} = \sum_{u, i \in R_u} \frac{B_{i,u}(b(t_j^{(u)} - t_i^{(u)}) = k)}{A_{i,u}(1)} - \frac{B_{i,u}(x_i^{(u)} = x_j^{(u)},\ b(t_j^{(u)} - t_i^{(u)}) = k)}{A_{i,u}(x_i^{(u)} = x_j^{(u)})}.$$

Empirical analysis of time model. We conclude by analyzing the learned time scores T that are used in the repeat consumption selection process (Equation 3). The time scores measure the relative importance of elapsed time *after* accounting for the proclivity to consume recently-consumed items (w) and the relative quality of each item (s). We learned the function T as a piecewise-constant function with exponentially-spaced bins of $30 \cdot 1.1^k$ seconds, $k = 0, 1, \ldots$. We chose k large enough so that the inter-arrival times are covered by the binning.

The top two rows of Figure 7 show the learned time scores T. Interestingly, the behavior of the time scores are quite different for the music and video data. In LASTFM, time scores tend to be large around the duration of a song (3-4 minutes)[7] and then flatten out after an hour. However, for YOUTUBE, the time scores actually decrease around the time of a song/video (3-5 minutes) and peak

[7]If our data included duration information, we would instead model the time from completed consumption of item i to start of consumption of $i+1$.

in the time range of one hour to one day. This is possibly due to album effects—the last.fm service lets users play albums of the same artist, whereas these are often organized into single tracks when uploaded to YouTube.[8] For the clicks data (WIKICLICKS and MAPCLICKS) and check-in data (BRIGHTKITE and GPLUS), there is a consistent trend towards larger time scores for small time intervals (30 seconds to 2 minutes). This is partly due to double-clicks, e.g., from slow page loads, in the clicks data or from power users in the check-in data that tend to check in several times. Finally, we observe that the time scores capture cyclic behavior in the check-in data around daily and weekly marks. This behavior is particularly strong for the BRIGHTKITE dataset, where cyclic behavior has been observed [10].

To eliminate short-term effects, we also learned the time score function T with bins of $12 \cdot 2^k$ hours. The bottom two rows of Figure 7 show the results. In this case, the scores for LASTFM and LASTFMARTISTS are near-uniform, i.e., absolute time does not play a role aside from recency (which is captured by w). However, repeat consumption in YOUTUBE and YOUTUBEMUSIC still shows a preference for videos seen in the last day.

Finally, we evaluated the importance of time compared to recency (w) and global item quality (s). Table 6 lists the relative likelihood of the model when only optimizing a subset of the parameters compared to learning the full model (here we used the time score binning of $30 \cdot 1.1^k$ seconds). We see that the recency weights (w) account for most of the likelihood in all datasets. Interestingly, likelihood is improved more with time scores (T) than with item scores (s), suggesting that *when* users consume is more predictive than *what* they consume. This makes sense with, e.g., the check-in data:es 380(o)5echeck-iT101Tfe5.97761dte.97761dte.9776d2353). How-

were consumed in the middle 60% of the user's lifetime, i.e., the sequence begins after the first 20% of the user's lifetime and ends before the final 20% of the user's lifetime. Again, this latter requirement avoids data censoring at the beginning and end of the user lifetime and ensures that we examine finite lifetimes.

The top row of Figure 10 plots the median of the first, second, second-to-last, and last normalized gaps over all sequences in the YOUTUBE, LASTFM, and GPLUS datasets. We observe that for YOUTUBE and LASTFM, the normalized gaps are monotonically increasing in the sequence. This holds for both the data and the model and for both the normalized index gaps and normalized time gaps. This means that users tend to become bored with items over time until eventually the item is never consumed again.

In the LASTFM dataset, the model simulation produces normalized index and time gaps that are slightly smaller than the true data. However, the shapes of the curves for the model simulation and the data are close. We scaled each curve by the mean ratio of δ_i/δ_1 and g_i/g_1 for the model to the data. These are plotted in the bottom row of Figure 10. Here, we see a very close match between model simulation and data for all datasets. We conclude that our model slightly over-estimates the first gap in the sequence.

Analysis. Although we do not explicitly model boredom, we do observe it both in the data and in our model. We now show that this boredom is actually just a consequence on conditioning on the finiteness of an item's lifetime. Again, we will assume the recency model (Equation 4) for simplicity.

We begin with two lemmas that will help us prove our main result in Theorem 5.4. We consider a user sequence with consumptions of item e at indices $j_1, j_2, \ldots$ and index gaps $g_{i+1} = j_{i+1} - j_i$.

LEMMA 5.2. *If the recency weights w are monotonically decreasing, then $\Pr(g_j = k)$ monotonically decreasing in k.*

PROOF. We prove this pointwise. For any gaps $g_1, \ldots, g_{j-1}$,

$$\begin{aligned}
&\Pr(g_j = k \mid g_1, \ldots, g_{j-1}) \\
&= \Pr(g_j \geq k-1 \mid g_1, \ldots, g_{j-1}) \cdot \left(w_k + w_{k+g_{k-1}} + \cdots + w_{k+g_{k-1}+\ldots+g_1}\right) \\
&\geq \Pr(g_j \geq k \mid g_1, \ldots, g_{j-1}) \cdot \left(w_{k+1} + w_{k+1+g_{k-1}} + \cdots + w_{k+1+g_{k-1}+\ldots+g_1}\right) \\
&= \Pr(g_j = k+1 \mid g_1, \ldots, g_{j-1}).
\end{aligned}$$

The inequality follows from the facts that the w's are monotonically decreasing and that $g_j \geq k+1$ implies $g_j \geq k$. □

LEMMA 5.3. *Let p_k be a discrete probability distribution monotonically decreasing in k. Let a_k be a monotonically increasing sequence such that $q_k = p_k a_k$ is a discrete probability distribution. Then for any random variable X, $\mathsf{E}_{\sim p}(X) \leq \mathsf{E}_{\sim q}(X)$.*

PROOF. Since a_k is monotonically increasing and q_k is a probability distribution, there must exist a K such that $0 \leq a_k \leq 1$ for $k < K$ and $a_k > 1$ for $k \geq K$. Thus, q_k just shifts mass in p_k from $k < K$ to $k \geq K$. Since p_k is monotonically decreasing, the expectation must increase. □

The following theorem says that conditioning on the fact that an item will not be consumed again, its last gap will be larger.

THEOREM 5.4. *Suppose that the recency weights w are monotonically decreasing. Let $\mathcal{E}$ be the event that an item is consumed exactly j times. Then $\mathsf{E}(g_j \mid \mathcal{E}) \geq \mathsf{E}(g_j)$.*

PROOF. First, by applying Bayes' theorem,

$$\mathsf{E}(g_j \mid \mathcal{E}) = \sum_{k=1}^{\infty} k \Pr(\mathcal{E} \mid g_j = k) \Pr(g_j = k) / \Pr(\mathcal{E}).$$

By Lemma 5.2, the sequence $p_k = \Pr(g_j = k)$ is monotonically non-increasing. Let $a_k = \Pr(\mathcal{E} | g_j = k) / \Pr(\mathcal{E})$ Then a_k monotonically non-decreases with k since w is monotonically non-increasing. Finally, since $q_k = p_k a_k = \Pr(g_j = k \mid \mathcal{E})$ is a probability distribution, the result then follows by Lemma 5.3. □

6. PARSIMONIOUS PERSONALIZATION

Finally, we consider personalizing our model for each user. We restrict ourselves to just personalizing the recency weights w because they are the largest component of the model (both in terms of number of parameters and effect on likelihood). Our goal is to find *parsimonious* personalization, i.e., we want to model each user with just a few parameters. This allows us to avoid over-fitting and gain interpretability by examining particular parameterized families.

6.1 Model

After analyzing the data and experimenting with different parameter families, we found the double Pareto distribution (Equation 1) to be a good fit for modeling the recency weights w. Formally,

$$w(\delta) \propto \begin{cases} (\delta/\gamma)^{\beta-1} & \delta \leq \gamma \\ (\delta/\gamma)^{-\alpha-1} & \delta > \gamma \end{cases},$$

where $\delta = 1, 2, \ldots$, $\alpha > 1$, $\beta > 0$ and $\gamma \geq 1$. Thus, our personalized double Pareto (PDP) model finds triples (α, β, γ) for each user. The result is a parsimonious personalized model, where each user's behavior is captured by just three parameters. We compare this distribution to the power-law with exponential cutoff (PLECO) model, which was previously proposed for global recency weights [4].

Learning model parameters. Note that w is differentiable (although not continuously differentiable, specifically at $\delta = \gamma$). Therefore, we can run gradient descent on the negative log-likelihood for each user. The gradients of w with respect to each parameter are

$$\frac{\partial w(\delta)}{\partial \alpha} = \begin{cases} 0 & \delta \leq \gamma \\ -\ln(\delta/\gamma) w(\delta) & \delta > \gamma \end{cases}, \quad \frac{\partial w(\delta)}{\partial \beta} = \begin{cases} \ln(\delta/\gamma) w(\delta) & \delta \leq \gamma \\ 0 & \delta > \gamma \end{cases},$$

$$\frac{\partial w(\delta)}{\partial \gamma} = \begin{cases} -(\beta-1)(\delta/\gamma)^{\beta}/\delta & \delta \leq \gamma \\ (\alpha+1)(\delta/\gamma)^{-\alpha}/\delta & \delta > \gamma \end{cases}.$$

With an easy application of the chain rule, we optimize the parameters to minimize the negative log-likelihood (Section 4) for each user, given that we have already learned the rest of the model.

In practice, we observed that the gradient descent algorithm is sensitive to the initial value of γ. Therefore, we perform the optimization over equally-spaced points in some interval $[\gamma_{\min}, \gamma_{\max}]$.

6.2 Experiments

We evaluated our personalization model on "heavy users" in the LASTFM, YOUTUBE, and BRIGHTKITE datasets. Specifically, we consider 25 users from LASTFM with at least 80,000 consumptions and 40,000 repeat consumptions each, 250 users from YOUTUBE with at least 8,000 consumptions and 2,000 repeats each, and 500 users from BRIGHTKITE with at least 500 consumptions and 250 repeats each. We restrict our personalization experiments to such heavy users because they actually have sufficient data for training the models described in Section 6.1. However, we note that these users only represent a small portion of the user population.

To form an upper bound on the improvement from personalization, we first optimize a non-parametric weight vector w for each user and compute the likelihood. Note that this tends to overfit the weights w—for example, if one user happens to not repeat any items from 17 positions prior, then $w(17)$ will be set to zero for that user in order to maximize the likelihood. After, we compare the likelihood from the overfit weights to the model with the global weight vector w, i.e., learning the full model w for just a single

Table 7: Relative likelihoods of parsimonious personalized models to the (overfitted) model of learning w for each user.

Base dataset	# heavy users	form of w Global	PDP	PPLECO
LASTFM	25	0.80	0.89	0.88
YOUTUBE	250	0.68	0.77	0.73
BRIGHTKITE	500	0.45	0.99	0.90

user in the optimization procedure discussed in Section 4.3. The likelihood ratios provide an upper bound on the total possible improvement in the model from personalization.

Table 7 lists the relative likelihoods. For LASTFM, the global model has already captured 80% of the possible likelihood. The table also lists the relative likelihoods of the PDP model to the overfitted, non-parametric model. In all cases, personalization captures over 75% of the available likelihood. For BRIGHTKITE, PDP captures essentially all of the likelihood.

Figure 8 shows the PDP parameters for a heavy LASTFM user. We see that this family effectively captures inflection points in the true (overfit) weights. In other examples, wee see the double Pareto distribution learns a slightly positive slope, which cannot be captured by simple power laws or related distributions.

Compared to the non-personalized situation, in which earlier work (confirmed in our analysis) has shown a good fit for the PLECO distribution of weights, we see in the personalized setting a significant likelihood improvement using PDP, which requires the same number of fit parameters as PLECO. There is nothing inconsistent about this, as the appropriate mixture of personalized PDP distributions may be well-approximated by a PLECO in the global setting.

7. CONCLUSIONS

We developed a general model for user consumption sequences. We showed empirically and theoretically that this model captures: (1) the heavy-tailed distribution of item lifetimes and (2) a notion of boredom in terms of growing gaps between consumptions before an item is abandoned by a user. Importantly, our model does not need to explicitly incorporate these properties; rather, they arise from the simple generative process. In future work, it would be useful to evaluate these ideas on a concrete task such as predicting the number of unique items consumed in a specified time interval. We also studied personalization in the recency weights for repeat consumption item selection. Interestingly, for some datasets, the likelihood gains were only modest, suggesting that user consumption patterns are quite similar (even if they consume different items). Determining when personalization will significantly improve the model is an interesting avenue for future research.

Finally, our datasets cover entertainment, social check-ins, and web browsing behavior, but there are a number of additional domains where we can apply our models in future work. For example, sequences in consumer goods shopping (groceries, clothes, etc.) have many repeat consumptions due to brand loyalty [16] and also contain temporal dynamics from seasonal purchasing.

Acknowledgments. We thank the reviewers for their suggestions.

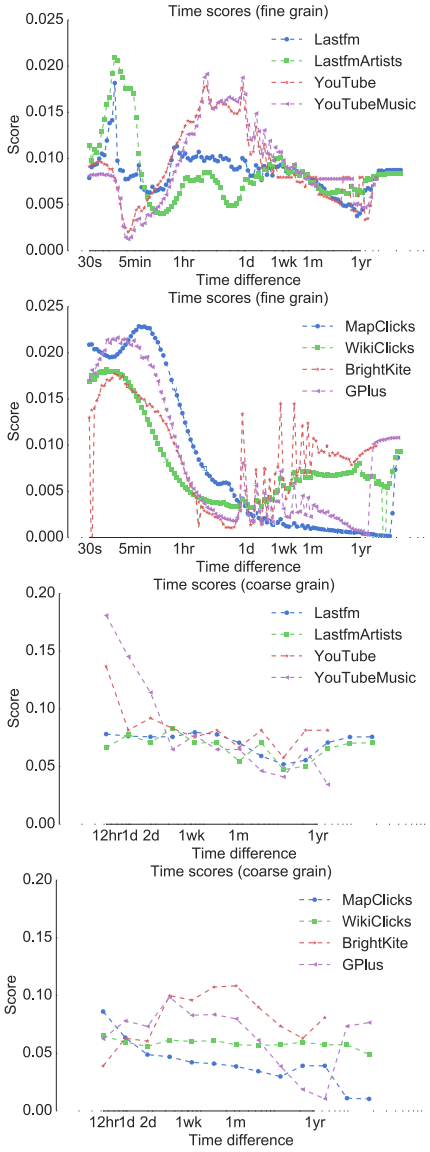

Figure 7: Learned time scores $T(\cdot)$ (Equation 3). Top two rows: Fine-grain time scores with binning of $30 \cdot 1.1^k$ seconds. The scores capture information such as album effects (LASTFMARTISTS, first row) and periodic behavior (BRIGHTKITE, second row). Bottom two rows: Coarse-grain time scores with binning of $12 \cdot 2^k$ hours. YouTube users have a preference for videos watched in the past 24 hours (third row).

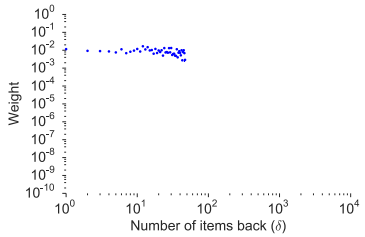

Figure 8: Personalized recency weights w for a heavy user in LASTFM. The double Pareto model is able to captures inflection points in the weights near 200 (top) and near 10 (bottom).

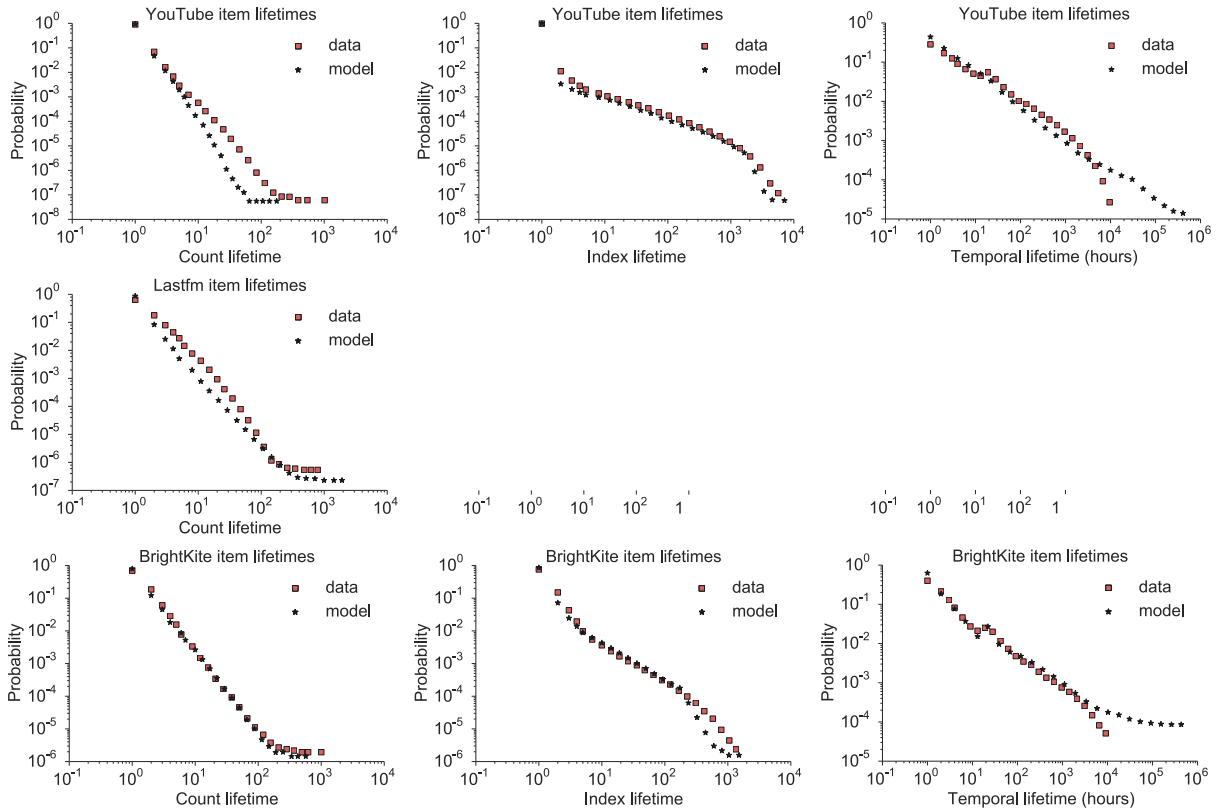

Figure 9: Distributions of count (left), index (middle), and temporal (right) lifetimes for three datasets. In all cases, the lifetime distributions in the model-simulated data tends to match the lifetimes in the data.

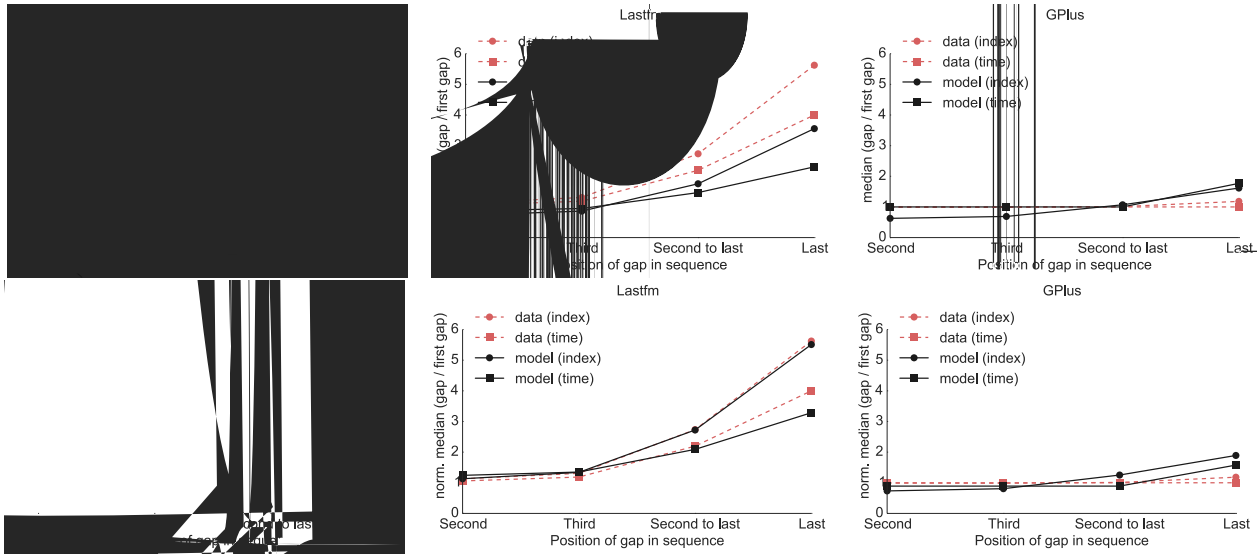

Figure 10: (Top) Median behavior of the normalized gaps in the time between consumptions for a given user-item pair. Gaps are measured in terms of the number of other items consumed (index) and absolute time between repeat consumptions. The last two normalized gaps tend to be larger than the first two gaps (relative to the size of the first gap) for the YOUTUBE and LASTFM data in both the data and model simulation, indicating boredom. (Bottom) Same data as top, except the curves are scaled by the mean ratio of model to data in the top set of curves. The gaps in the LASTFM data now take nearly the same values in the model and in the data.

8. REFERENCES

[1] E. Adar, J. Teevan, and S. T. Dumais. Large scale analysis of Web revisitation patterns. In *CHI*, pages 1197–1206, 2008.

[2] E. Adar, J. Teevan, and S. T. Dumais. Resonance on the Web: Web dynamics and revisitation patterns. In *CHI*, pages 1381–1390, 2009.

[3] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *TKDE*, 17(6):734–749, 2005.

[4] A. Anderson, R. Kumar, A. Tomkins, and S. Vassilvitskii. The dynamics of repeat consumption. In *WWW*, pages 419–430, 2014.

[5] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46:109–132, 2013.

[6] Ò. Celma Herrada. *Music Recommendation and Discovery in the Long Tail*. PhD thesis, Universitat Pompeu Fabra, 2009.

[7] J. Chen, C. Wang, and J. Wang. Will you “reconsume" the near past? Fast prediction on short-term reconsumption behaviors. In *AAAI*, pages 23–29, 2015.

[8] E. H. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions and the Web. In *CHI*, pages 490–497, 2001.

[9] C.-M. Chiu, M.-H. Hsu, H. Lai, and C.-M. Chang. Re-examining the influence of trust on online repeat purchase intention: The moderating role of habit and its antecedents. *Decision Support Systems*, 53(4):835–845, 2012.

[10] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *KDD*, pages 1082–1090, 2011.

[11] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[12] J. D. Cohen, S. M. McClure, and J. Y. Angela. Should I stay or should I go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):933–942, 2007.

[13] A. Das Sarma, S. Gollapudi, R. Panigrahy, and L. Zhang. Understanding cyclic trends in social choices. In *WSDM*, pages 593–602, 2012.

[14] D. Downey, S. Dumais, D. Liebling, and E. Horvitz. Understanding the relationship between searchers’ queries and information goals. In *CIKM*, pages 449–458, 2008.

[15] G. Dror, D. Pelleg, O. Rokhlenko, and I. Szpektor. Churn prediction in new users of Yahoo! answers. In *WWW*, pages 829–834, 2012.

[16] J. Jacoby and D. B. Kyner. Brand loyalty vs. repeat purchasing behavior. *Journal of Marketing Research*, pages 1–9, 1973.

[17] B. E. Kahn, M. U. Kalwani, and D. G. Morrison. Measuring variety-seeking and reinforcement behaviors using panel data. *Journal of Marketing Research*, pages 89–100, 1986.

[18] K. Kapoor, K. Subbian, J. Srivastava, and P. Schrater. Just in time recommendations: Modeling the dynamics of boredom in activity streams. In *WSDM*, pages 233–242, 2015.

[19] K. Kapoor, M. Sun, J. Srivastava, and T. Ye. A hazard based approach to user return time prediction. In *KDD*, pages 1719–1728, 2014.

[20] M. Karnstedt, T. Hennessy, J. Chan, and C. Hayes. Churn in social networks: A discussion boards case study. In *SocialCom*, pages 233–240, 2010.

[21] J. G. March. Exploration and exploitation in organizational learning. *Organization Science*, 2(1):71–87, 1991.

[22] L. McAlister. A dynamic attribute satiation model of variety-seeking behavior. *Journal of Consumer Research*, pages 141–150, 1982.

[23] C. Mitchell, M. Harper, and L. Jamieson. On the complexity of explicit duration HMM’s. *IEEE Transactions on Speech and Audio Processing*, 3(3):213–217, 1995.

[24] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[25] M. Mitzenmacher. Dynamic models for file sizes and double Pareto distributions. *Internet Mathematics*, 1(3):305–333, 2004.

[26] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, InfoLab, Stanford University, 1999.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830, 2011.

[28] R. K. Ratner, B. E. Kahn, and D. Kahneman. Choosing less-preferred experiences for the sake of variety. *Journal of Consumer Research*, 26(1):1–15, 1999.

[29] W. J. Reed and M. Jorgensen. The double Pareto-lognormal distribution—a new parametric model for size distributions. *Communications in Statistics-Theory and Methods*, 33(8):1733–1753, 2004.

[30] K. H. Schlag. Why imitate, and if so, how?: A boundedly rational approach to multi-armed bandits. *JET*, 78(1):130–156, 1998.

[31] J. Teevan, E. Adar, R. Jones, and M. Potts. History repeats itself: Repeat queries in Yahoo’s logs. In *SIGIR*, pages 703–704, 2006.

[32] J. Teevan, E. Adar, R. Jones, and M. A. Potts. Information re-retrieval: Repeat queries in Yahoo’s logs. In *SIGIR*, pages 151–158, 2007.

[33] R. West, A. Paranjape, and J. Leskovec. Mining missing hyperlinks from human navigation traces: A case study of Wikipedia. In *WWW*, pages 1242–1252, 2015.

[34] R. West, J. Pineau, and D. Precup. Wikispeedia: An online game for inferring semantic distances between concepts. In *IJCAI*, pages 1598–1603, 2009.

[35] J. Yang, X. Wei, M. S. Ackerman, and L. A. Adamic. Activity lifespan: An analysis of user survival patterns in online knowledge sharing communities. In *ICWSM*, 2010.