12-2013

Modeling Preferences with Availability Constraints

Bingtian DAI
Singapore Management University, btdai@smu.edu.sg

Hady W. LAUW
Singapore Management University, hadywlauw@smu.edu.sg

DOI: https://doi.org/10.1109/ICDM.2013.41

Modeling Preferences with Availability Constraints

Bing Tian Dai Hady W. Lauw
School of Information Systems
Singapore Management University
Email: {btdai, hadywlauw}@smu.edu.sg

***Abstract*—User preferences are commonly learned from historical data whereby users express preferences for items, e.g., through consumption of products or services. Most work assumes that a user is not constrained in their selection of items. This assumption does not take into account the availability constraint, whereby users could only access some items, but not others. For example, in subscription-based systems, we can observe only those historical preferences on subscribed (available) items. However, the objective is to predict preferences on unsubscribed (unavailable) items, which do not appear in the historical observations due to their (lack of) availability. To model preferences in a probabilistic manner and address the issue of availability constraint, we develop a graphical model, called Latent Transition Model (LTM) to discover users' latent interests. LTM is novel in incorporating transitions in interests when certain items are not available to the user. Experiments on a real-life implicit feedback dataset demonstrate that LTM is effective in discovering customers' latent interests, and it achieves significant improvements in prediction accuracy over baselines that do not model transitions.**

***Keywords*—latent interests; topic transition; topic model; graphical model; user preferences**

I. Introduction

By understanding user preferences, commercial companies are able to increase their sales by promoting more products and services. For example, media companies providing cable TV programs are always interested to attract their existing customers to subscribe to more channels at higher subscription fees, thus to make higher profits. In order to do this, these media companies need to recommend unsubscribed channels which users are likely to subscribe, which makes understanding user preferences very critical.

There are various types of user behaviors from which we can learn user preferences. Most of the previous work studies *ratings* behavior [1], i.e., how a user evaluates a product or a service on a scale. In some cases, they study *adoption* behavior [2], i.e., the binary decision of adopting (e.g., purchasing a product, befriending another user).

In this work, we are interested in modeling user preferences from *consumption* behavior, how a user consumes a product or a service. For instance, in the cable TV industry domain, a user chooses what channel to watch from a selection of available channels. In the music industry domain, a user chooses which song to listen to from a set of available songs. Similarly, in other domains such as online radio. On one hand, consumption behavior is useful because we could observe the user consuming the same item (e.g., a channel, a song) again and again. In contrast, most of the time, users will only rate or adopt a specific product once. On the other hand, consumption behavior also introduces a new constraint we need to factor in, which we term the availability constraint.

Availability constraint is the constraint imposed on users to restrict which items are available to each user, i.e., users do not have access to those items which are not specified by the availability constraints. For example, a user can only watch the available cable TV channels, i.e., those that she has subscribed to. Similarly, a user can only listen to songs that she has purchased. The implication of this constraint is that we can only observe consumption behaviors from available channels, but not from unavailable channels.

This gives rise to several challenges in modeling user preferences. Let us illustrate this using an example. Bundling is a common practice in the cable TV industry [3], whereby users are to subscribe to one or more bundles and not allowed to cherry pick channels within a bundle. Suppose there are two bundles: A containing channels $\{A1, A2, A3\}$, and B containing $\{B1, B2\}$. Table I shows the channel watching activities for three users: Kat, Linda, Maggie. The cell values are the time units that each user spends on each channel.

	Bundle A			Bundle B	
	$A1$	$A2$	$A3$	$B1$	$B2$
Kat	100	20	25	N.A.	N.A.
Linda	15	40	90	0	0
Maggie	105	15	25	10	75

TABLE I. Users and their channel activities

The first challenge is the need to factor in availability in interpreting the preferences of users. Kat only subscribes to bundle A, and therefore we cannot observe her activities on bundle B (N.A. or not available). This does not mean that in reality, Kat does not like the channels in B. It could well be that Kat's favorite channels may be $A1$ and $B2$, but $B2$ is simply not available to her. Kat's situation is in contrast to Linda's. The latter subscribes to bundle B, but does not watch the channels there (zero activity). In Linda's case, this is an indication of not liking channels in B.

The second challenge is that the availability constraint restricts the inference of user preferences among similar users. For instance, in inferring whether Kat's preferences is similar to Maggie's, we may want to take into account their activities on bundle A alone, because B is not available to Kat. On the other hand, whether Linda is similar to Maggie would depend on their activities on all available channels.

Tackling these challenges in factoring availability constraint is useful in several respects. For one thing, it leads to more accurate modeling of user preferences. For another thing, it focuses our attempt of prediction on the set of unavailable items, using information from the available items. In contrast, this notion of availability has not been widely considered in previous work. In traditional recommendation systems work, it

is frequently assumed that items are all available to the users, and users are free to choose any item they like.

Problem. Our objective in this work is to build a preference model for each user. We adopt a probabilistic framework for its interpretability. In particular, we would like to model the probability that a user will consume a particular item (either an available item or an unavailable item). However, modeling this probability at the item level directly is not practical, because of the potentially large number of items, most of which are unavailable and therefore not directly observable.

We propose to first put items that similar users tend to like into groups, and model the probability that a user will like each group, as well as the probability of an item within each group. We observe that explicit groupings may not always accurately reflect a grouping of items that a user may like. For instance, a cable TV bundle is unlikely to contain all the channels that a user may like, because the company may spread popular channels over multiple bundles to get customers to subscribe to more bundles. Thus, we would like to learn *latent preference* groups to be inferred from consumption behavior data.

Our approach is to realize the user preference model through generative topic modeling framework. While there are existing such models such as LDA [4], they are not sufficient for the problem because they do not expressly factor in the notion of availability. Hence, we build a new generative model, which we call *Latent Transition Model* or *LTM*, with the following intuition. When a user would like to consume an item that is not available, she will substitute it with another available item. The substitution is modeled by a transition from a first-choice latent group (or "topic") to a second-choice group. This gives rise to different consumption behaviors such as picking an available item directly, or picking an unavailable item followed by transitioning to an available item.

In this paper, we mainly discuss the domain of cable TV in the examples and experiments, partially due to the presence of a suitable dataset in this domain. However, as will be evident in Section III, our model is general enough to cover other cases of consumption behaviors, where the notion of availability can be properly defined. This includes predicting which other music tracks a user will like based on her listening behaviors on music she already has access to. Another example in product recommendation is when some items are unavailable to a user due to her budget constraint (assuming the budget is known).

Contributions. In this paper, we make the following contributions to tackle the above problem.

- *First*, we identify the availability constraint as an important factor in modeling user preferences based on consumption behaviors.
- *Second*, we propose a generative model, called **Latent Transition Model** (**LTM**), which incorporates the notion of transition among latent preference groups based on availability of items to individual users.
- *Third*, we design a randomized algorithm for inferring LTM based on Gibbs sampling. Importantly, the algorithm has to be able to handle "triplet" latent variables that arise because of the transition from a first-choice to a second-choice group before picking an available item. We further propose an optimization that improves efficiency by two orders of magnitude.
- *Fourth*, we conduct a comprehensive evaluation of the proposed model on a real-life proprietary dataset from the cable TV industry. The application task is to predict the next bundle that a customer is likely to subscribe to, given the consumption behavior on existing bundles of the customer.

Organization. Our paper is organized as follows. We review previous work in Section II. We describe our proposed Latent Transition Model in Section III, and its inference algorithm in Section IV. This is then followed by the experiments with a real-life dataset in Section V. We then conclude in Section VI.

II. Related Work

In terms of problem. The study of modeling user preferences is an area of interest in personalization or recommendation systems [1]. We are different from the majority of previous work in this area in two ways. First, in terms of output, we focus on predicting *consumptions*, not ratings. Second, in terms of input, instead of explicit ratings or meta-data [5], we work with implicit feedback dataset. Implicit feedback is of interest in several domains such as cable TV [6], search [7], [8], music [9], Internet radio [10] and so on. What is common across all these cases is that the users do not explicitly express their preferences, but rather indicate them indirectly through their behaviors, e.g., which TV shows they watched and for how long, which music tracks they listened to.

To the best of our knowledge, ours is the first work to deal with the *availability constraint* directly and systematically. A related but different concept is competition [11]. Among the items presented to a user, which one will she pick? This is a different problem, because it focuses on relative preference among available items, whereas we focus on extending preference to unavailable items by factoring availability explicitly.

Among the previous work on implicit feedback, the work in [6] also used dataset from the cable TV domain. However, there are two crucial differences from our work. First, [6] attempted to predict which channels (among those that a user had watched before) she would watch again. In contrast, we attempt to predict which channels (other than those the user has subscribed to) she is likely to want to subscribe to in the future. In the former, there is a "direct" signal for the channels to be predicted (previous watching sessions). In the latter, there is not. The reason for the absence of direct signal in the latter is the unavailability of some items (e.g., some channels are unsubscribed, and therefore cannot generate watching data).

In terms of approach. The second difference is in terms of approach. [6]'s solution was based on matrix factorization (MF) [12], [13], [14], [15], [16], which are popular in recommender systems because of its wide applicability to ratings. In addition to MF, other rating prediction approaches include collaborative filtering (CF) [1]. The approach of rating prediction (MF or CF) is not appropriate in our scenario. For one thing, there is no "rating" in our case. It is possible, but inappropriate, to model the length of time a user watches a show as "rating" for several reasons. First, the model will be optimized to predict the length of time a user is likely

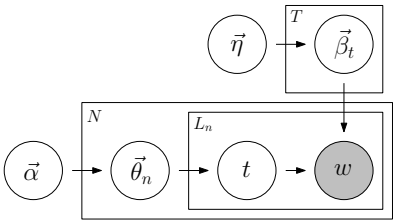


Fig. 1. Latent Dirichlet Allocation (LDA)

to watch an item, which is not directly relevant to whether the user is likely to subscribe to an unavailable item. Second, by predicting absolute lengths of time on unavailable channels, it results in the unrealistic scenario of predicting the user will spend much longer absolute amount of time in aggregate over the collection of all items. It is not necessarily the case that customers will spend more time watching TV if they subscribe to more channels. It is more useful to learn how their preferences are distributed among all the channels. Therefore, we adopt a generative modeling approach that expresses preferences in terms of probability distributions.

While we model preferences over individual items, in Section V, we validate the proposed approach on a real-life cable TV dataset by predicting bundles. This is because at the point of subscription, customers choose bundles, rather than individual channels. Several previous works concern the recommendation of bundles. [17] looked into how to configure items into a bundle in for viral marketing. [18] looked into how to personalize bundles for individual customers. Their main issue is *bundle configuration*. This is a complementary problem, and is not applicable to our setting because in our case the cable TV bundles were already specified in the dataset.

In terms of modeling topic-based preferences. Our topic modeling approach is related to *Latent Dirichlet Allocation* or *LDA* [4], which is widely used to model topics in documents. The graphical model of LDA is shown in Figure 1. Each document n has a distribution over topics $\vec{\theta}_n$. To generate the document, we repeatedly pick a topic t from this distribution, and generate a word w from the topic's word distribution $\vec{\beta}_t$. $\vec{\alpha}$ and $\vec{\eta}$ are Dirichlet priors for $\vec{\theta}_n$ and $\vec{\beta}_t$ respectively. Compared to LDA, our model is significantly novel in a few respects. First, in terms of modeling, we model availability-based transition between topics (vs. no transition in LDA). Second, in terms of inference, this transition gives rise to "triplet" latent variables (vs. singleton latent variables in LDA). Third, in terms of optimization, we propose collapsing multiple "related" triplet latent variables. To validate these differences, we will use LDA as a baseline in experiments.

Transition between topics based on availability captures a specific type of dependency between two topics. There are other topic modeling approaches that focus on "dependencies", but none captures the concept of availability. Unlike correlated topic models [19], [20], [21] with symmetric correlations between topics, our work models directed transitions. Other topic models may define transitions based on time [22], [23] or distributional similarity [24], but not availability.

III. Latent Transition Model

Our objective in this section is to develop a model for user preferences that factors in the fact that some items are never observed in a user's historical data because they are unavailable to the user, and not because the user does not like them. As input, we have a set of observations of users' consumptions of various items, as well as which subset of items are available to every user. As output, we would like to learn a model (for every user) for how these consumptions could have been generated, so as to help in the prediction of future consumptions. To help with the description of the model, we maintain a list of notations in Table II.

Notation	Description
$\vec{\theta}_n$	user n's probability distribution over topics
$\vec{\beta}_t$	topic t's probability distribution over items
$\vec{\tau}_t$	topic t's distribution of transition probabilities to other topics
$\theta_{n,t}$	probability of user n choosing topic t
$\beta_{t,c}$	probability of topic t generating item c
$\tau_{t,t'}$	probability of transitioning from topic t to topic t'
$\vec{\alpha}$	Dirichlet prior for $\vec{\theta}_n$ for all users
$\vec{\eta}$	Dirichlet prior for $\vec{\beta}_t$ for all topics
$\vec{\psi}_t$	Dirichlet prior for $\vec{\tau}_t$ for topic t
λ	parameter controlling within-topic vs. across-topic transition
A_n	subset of items available to user n
$\bar{A}_n$	subset of items unavailable to user n
T	total number of topics
N	total number of users
C	total number of items (e.g., channels)
L_n	total consumption instances (e.g., watching sessions) for user n
$c_{n,l}$	an instance of consumption (e.g., a watching session) by user n

TABLE II. Notations

A. Modeling Preference

We begin with the consideration of how to model *preference* itself. While the end outcome is to estimate a user's preference for individual items, it is not feasible nor desirable to model this directly. It is not feasible because the observation is not complete, i.e., we can observe the user's consumption behavior for only the items that are available to her. It is not necessarily desirable because such item-specific estimation may overfit the data. A common assumption in previous work is items share some form of "similarity" in the latent space, and it is thus sufficient to model preferences in this latent space.

We thus associate each user n with a vector $\vec{\theta}_n$ of T latent factors, where the value corresponding to each latent factor t reflects the degree of preference of user n for that factor. In contrast to matrix factorization-based framework, where this $\vec{\theta}_n$ is simply a vector of real values with no other interpretation, in this work we attach a semantic interpretation to these values as a probability distribution over the latent factors. Each value is thus the probability that a user n prefers a latent factor t. To relate user preferences to the items, we also associate each item with these latent factors. Each latent factor t is associated with a probability distribution $\vec{\beta}_t$ over the items. To borrow the terminology in [4], we refer to each latent factor as a *topic*.

Applied to the cable TV scenario, which we will experiment with later, an instance of consumption refers to a session of watching a channel. When a user wants to watch TV, she first thinks of some "topic" to watch. A topic captures the association of several channels (items) that a significant number of users tend to watch. For example, a topic may be a group of channels with similar broadcasting patterns (TV series at certain time period), with similar genre (e.g., non-fiction such as documentaries and news), or with similar language (a topic on "Chinese shows" may have high probabilities for a variety show, a news channel, as well as a movie channel).

One naive way to learn $\vec{\theta}_n$ and $\vec{\beta}_t$ for various users and items is by using *LDA* [4]. In this case, to generate a user's watching data, we would repeatedly sample a topic t from the user's topic distribution $\vec{\theta}_n$, and then sample a channel c from the sampled topic's distribution $\vec{\beta}_t$. This naive way suffers from a shortcoming, which we will explain shortly.

B. Modeling Transition

One crucial issue with the naive modeling by LDA is the assumption that any item (channel) is available for consumption, and thus could be generated from a topic's distribution $\vec{\beta}_t$. This assumption does not hold in scenarios where only some subset of items are available to the users. For instance, in the cable TV domain, a user could only watch those channels that she has subscribed to, and therefore no watching data could be generated for the unsubscribed channels. This is a serious issue because although LDA's model parameters allow the generation of all items, many of those "possible items" are never actually observed.

This has two implications. First, because generative models such as LDA are learned from the observations, these lack of observations that are expected by the model will affect the learning of the model parameters. Second, it implies that there needs to be a mechanism that allows us to learn a user's preference of unsubscribed channels even when no historical data for those channels have been observed.

One way to get around this issue is to assume that whenever the model generates an unavailable item, it simply fails. This is not a realistic scenario. For instance, when a user wishes to watch TV, she does not stop watching just because the channel that she likes is not available. More likely, she will pick a different channel to watch. In this scenario, we say that the user transitions from the former to the latter.

We thus propose the notion of *transition based on availability*. When a user n picks a topic t_1 from $\vec{\theta}_n$, and then picks a channel c_1 from $\vec{\beta}_{t_1}$, there are two possible outcomes. First, c_1 is available to the user, and the model simply generates a watching session. Second, c_1 is unavailable to the user, and she picks an alternative channel c_2 to watch.

In the second scenario, in theory, it is possible that even c_2 may again be unavailable, resulting in another transition, which again may be to an unavailable channel. Carried to the extreme, this may lead to infinite transitions, which realistically would not really occur in real life. In practically all cases, a user will eventually decide on an available channel to watch. To avoid the degenerate cases, and for simplicity of modeling, in this work we focus on the case where at most one transition will occur, i.e., either the first or the second channel picked (c_1 or c_2) will be observed. This should cover most cases, and we will keep the extension to modeling a single instance of observation due to multiple transitions (c_1 to c_2 to c_3 and so on) as future work.

When a transition occurs, how does the user pick the alternative channel c_2? One possibility is that the user has stayed on the same topic t_1, in which case we simply pick c_2 from $\vec{\beta}_{t_1}$. Another possibility is that the user now "transitions" to another topic t_2, and then picks c_2 from $\vec{\beta}_{t_2}$. This transition from one topic to another is modeled by a vector $\vec{\tau}_t$. For each topic t, $\vec{\tau}_t$ is a probability distribution of transitioning to various topics (including t itself).

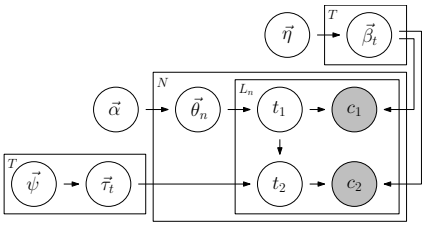


Fig. 2. Latent Transition Model (LTM)

In correspondence to the two possibilities of choosing channel c_2 when channel c_1 is not available, we name them *Within-Topic Transition* and *Across-Topic Transition* respectively. Within-topic transitions reckon the topic for c_1 and c_2 are the same, whereas across-topic transitions consider c_1 and c_2 are generated by different topics. Intuitively, for certain topics, e.g., "kids", within-topic transitions dominate across-topic transitions as kids are not interested in anything else. Other less "addictive" topics would demonstrate more across-topic transitions than within-topic transitions. Therefore, we distinguish within-topic transitions from across-topic transitions for each different topic.

C. Generative Process

We therefore build a generative model that incorporates modeling such preferences and transitions, which we refer to as ***Latent Transition Model*** or **LTM**, as shown in Figure 2. We now describe the generative process of LTM.

$\mathcal{M} = \{\vec{\alpha}, \vec{\eta}, \Psi\}$ where $\Psi = \{\vec{\psi}_1, \vec{\psi}_2, \ldots\}$. $\vec{\alpha}$, $\vec{\eta}$ and Ψ are the three parameters of LTM, serving as the Dirichlet priors for user's topic distribution, topic's channel distribution and topic's transition distribution respectively. As we discussed above, some topics may be more prone to transitions than other topics. Therefore, the topic transition prior for each topic ought to be different, which is why there exists one topic transition prior $\vec{\psi}_t$ for each topic t.

We assume that A_n or the subset of channels available to user n is known and is given as input. The subset of unavailable channels $\bar{A}_n$ is the complement of A_n. For N users, T topics and C channels in total, the observation is generated as follows:

1) For each topic t ($1 \leq t \leq T$):
   a) generate the topic-channel probability distribution $\vec{\beta}_t$, where $\beta_{t,c}$ is the probability of watching channel c with topic t

$$\vec{\beta}_t \sim \text{Dirichlet}(\vec{\eta})$$

   b) generate the topic-transition probability distribution $\vec{\tau}_t$, where $\tau_{t,t'}$ is the probability of transitioning from topic t to topic t'

$$\vec{\tau}_t \sim \text{Dirichlet}(\vec{\psi}_t)$$

2) For each user n ($1 \leq n \leq N$):
   a) generate the user-topic probability distribution $\vec{\theta}_n$, where $\theta_{n,t}$ is the probability of user n choosing t

$$\vec{\theta}_n \sim \text{Dirichlet}(\vec{\alpha})$$

b) generate the list of user n's observed channel-watching sessions $c_{n,l}$ for $1 \leq l \leq L_n$ where L_n is the total number of watching sessions observed for user n, as follows:

i) generate a topic $t_1 : 1 \leq t_1 \leq T$ based on the user n's topic distribution $\vec{\theta}_n$

$$t_1 \sim \text{Multinomial}(\vec{\theta}_n)$$

ii) generate a channel $c_1 : 1 \leq c_1 \leq C$ based on topic t_1's channel distribution $\vec{\beta}_{t_1}$

$$c_1 \sim \text{Multinomial}(\vec{\beta}_{t_1})$$

A) If c_1 is available to n, i.e., $c_1 \in A_n$, then we observe:

$$c_{n,l} = c_1$$

B) Otherwise, i.e., $c_1 \in \bar{A}_n$, then we have a transition:

- generate a topic $t_2 : 1 \leq t_2 \leq T$ based on topic t_1's transition distribution $\vec{\tau}_{t_1}$

$$t_2 \sim \text{Multinomial}(\vec{\tau}_{t_1})$$

- generate a watching session $c_2 : 1 \leq c_2 \leq C$ based on t_2's channel distribution $\vec{\beta}_{t_2}$

$$c_2 \sim \text{Multinomial}(\vec{\beta}_{t_2})$$

- implicitly $c_2 \in A_n$, therefore we observe:

$$c_{n,l} = c_2$$

To summarize, as shown in Figure 2, the three Dirichlet priors determine all other variables in this model. $\vec{\eta}$ and Ψ each generate T instances of $\vec{\beta}$ and T instances of $\vec{\tau}$. The topic-channel probability distribution and the topic-transition probability distribution for topic t are denoted by $\vec{\beta}_t$ and $\vec{\tau}_t$ respectively. $\vec{\alpha}$ generates N instances of $\vec{\theta}$, i.e., $\vec{\theta}_n$ for user n. For each of the L_n watching sessions of user n, we assume the observation is either c_1 or c_2 where c_1 is directly chosen by the first-choice topic t_1 and c_2 is chosen by the second-choice topic t_2 according to $\vec{\beta}_{t_1}$ and $\vec{\beta}_{t_2}$ respectively. The transition from t_1 to t_2 in the latter case involves the topic-transition probability distribution $\vec{\tau}_{t_1}$.

For the priors, as is common in topic models, we may set $\vec{\alpha}$ and $\vec{\eta}$ to be uniform, i.e., all $\alpha_t = \alpha$ and all $\eta_c = \eta$ for a pair of scalars α and η. We will experiment with different values of α and η in the experiments. However, it is not adequate to set $\vec{\psi}_t$ to be uniform, as within-topic transitions are expected to be different from across-topic transitions. We therefore introduce two scalars ψ and λ to model the topic transition prior and the difference between the two kinds of transitions. The across-topic transitions are parameterized by $\psi_{t,t'} = \psi$ given $t \neq t'$, while the within-topic transitions involve the additional λ, making $\psi_{t,t} = \lambda\psi$. A larger value of λ makes within-topic transitions more likely. Therefore, the topic transition probability $\vec{\tau}_t$ for topic t is generated by $\text{Dirichlet}(\vec{\psi}_t) = \text{Dirichlet}(\psi, \ldots, \psi, \lambda\psi, \psi, \ldots, \psi)$.

Note that both c_1 and c_2 are partially observed, i.e., for each watching session, we do not know the observed channel is c_1 or c_2. It is common that there exists multiple kinds of observations, and it is usually controlled by a switch determined by a model parameter. However, in our case, c_2 depends on the availability of c_1, so c_2 is observed only if c_1 is not available. This characteristic of our model is called *observations with dependencies*. In general, observations with dependencies cannot be modeled by a switch. In the next section, we will elaborate how we deal with observations with dependencies and infer the variables $\vec{\theta}$, $\vec{\beta}$ and Ψ.

IV. Inference

There are several approaches to infer the parameters of a generative models. One of the well-known approaches that are used for statistical inference is *Gibbs Sampling* [25].

A. Gibbs Sampling

We first look at the basic scenario without transition. Let $\boldsymbol{c}$ denote the set of observed watching sessions, and $\boldsymbol{z}$ denote the set of latent variables. For each observed watching session c, Gibbs sampling samples a value for z from $\{1, 2, \ldots, T\}$ according to a calculated probability distribution, as the topic assignment for c. As the probability distribution used for sampling a particular z depends on the value of other z's, Gibbs sampling usually takes many iterations to converge to a local optimum that maximizes the posterior probability $p(\boldsymbol{z}|\boldsymbol{c})$.

To introduce transition into this sampling process, we need to figure out whether c_1 or c_2 is being observed for each watching session. The approach of incorporating a switch into LTM is not feasible, as discussed in Section III. Note that c_2 is only observed when c_1 is not available, therefore, when c_2 is observed, c_1 becomes latent. Our proposal is thus to use "special" latent variables to represent the two different kinds of observations. Specifically, for a user n:

- If we observe c_1, we only have t_1 being latent, the latent variable is thus one of the T topics
- Otherwise, we observe c_2, and all t_1, c_1 and t_2 are latent. We therefore use a triplet (t_1, c_1, t_2) to represent a latent variable here. Note that c_1 is not available to n, i.e., $c_1 \in \bar{A}_n$.

Let $z_{n,l}$ be the latent variable determining the observed watching channel $c_{n,l}$, we have $z_{n,l} \in \{1, 2, \ldots, T\}$ or $z_{n,l} \in \{1, 2, \ldots, T\} \times \bar{A}_n \times \{1, 2, \ldots, T\}$. In the former case, $z_{n,l} = t_1$ when c_1 is observation. In the latter case, $z_{n,l} = (t_1, c_1, t_2)$ when c_2 is the observation.

The probability of observing all $\boldsymbol{c}$ with $\boldsymbol{z}$ is therefore:

$$\begin{aligned} p(\boldsymbol{z}, \boldsymbol{c}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\tau}|\mathcal{M}) &= \prod_{n=1}^{N} p(\vec{\theta}_n|\vec{\alpha}) \prod_{t=1}^{T} p(\vec{\beta}_t|\vec{\eta}) \prod_{t=1}^{T} p(\vec{\tau}_t|\vec{\psi}_t) \\ &\quad \cdot \prod_{n=1}^{N} \prod_{l=1}^{L_n} p(z_{n,l}, c_{n,l}|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathcal{M}) \end{aligned}$$

With $z_{n,l} \in \{1, 2, \ldots, T\}$, we have:

$$p(z_{n,l}, c_{n,l}|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathcal{M}) = \theta_{n,z_{n,l}} \beta_{z_{n,l},c_{n,l}}$$

.

With $z_{n,l} \in \{1, 2, \ldots, T\} \times \bar{A}_n \times \{1, 2, \ldots, T\}$, we have:

$$p(z_{n,l}, c_{n,l}|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathcal{M}) = \theta_{n,z_{n,l}^{(1)}} \beta_{z_{n,l}^{(1)},z_{n,l}^{(2)}} \tau_{z_{n,l}^{(1)},z_{n,l}^{(3)}} \beta_{z_{n,l}^{(3)},c_{n,l}}$$

where $z^{(1)}_{n,l} = t_1$, $z^{(2)}_{n,l} = c_1$ and $z^{(3)}_{n,l} = t_2$ as in Figure 2.

Integrating $p(\boldsymbol{z}, \boldsymbol{c}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\tau}|\mathcal{M})$ over $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$, we have:

$$
\begin{aligned}
p(\boldsymbol{z}, \boldsymbol{c}|\mathcal{M}) \quad = \quad & \prod_{n=1}^{N} \frac{\Gamma(\sum_{t=1}^{T} \alpha_t)}{\prod_{t=1}^{T} \Gamma(\alpha_t)} \frac{\prod_{t=1}^{T} \Gamma(m^{(1)}_{n,t} + \alpha_t)}{\Gamma(\sum_{t=1}^{T} m^{(1)}_{n,t} + \alpha_t)} \\
& \cdot \prod_{t=1}^{T} \frac{\Gamma(\sum_{c=1}^{C} \eta_c)}{\prod_{c=1}^{C} \Gamma(\eta_c)} \frac{\prod_{c=1}^{C} \Gamma(m^{(2)}_{t,c} + \eta_c)}{\Gamma(\sum_{c=1}^{C} m^{(2)}_{t,c} + \eta_c)} \\
& \cdot \prod_{t=1}^{T} \frac{\Gamma(\sum_{t'=1}^{T} \psi_{t,t'})}{\prod_{t'=1}^{T} \Gamma(\psi_{t,t'})} \frac{\prod_{t'=1}^{T} \Gamma(m^{(3)}_{t,t'} + \psi_{t,t'})}{\Gamma(\sum_{t'=1}^{T} m^{(3)}_{t,t'} + \psi_{t,t'})}
\end{aligned}
$$

where $m^{(1)}_{n,t}$, $m^{(2)}_{t,c}$ and $m^{(3)}_{t,t'}$ are the number of times user n chooses topic t as her first-choice topic, the number of times channel c is assigned to topic t and the number of times topic t transitions to topic t' over all users, respectively.

For a particular user n_0, and a watching channel c_{n_0,l_0}, by considering the three possible outcomes $z_{n_0,l_0} = t_1$, $z_{n_0,l_0} = t_1c_1t_1$ and $z_{n_0,l_0} = t_1c_1t_2$, we have:

$$
\begin{aligned}
& p(z_{n_0,l_0} = t_1, \boldsymbol{z}^{-}_{n_0,l_0}, \boldsymbol{c}|\mathcal{M}) \propto (m^{(1)-}_{n_0,t_1} + \alpha_{t_1}) \cdot \frac{m^{(2)-}_{t_1,c_1} + \eta_{c_1}}{sum^{(2)-}_{t_1}} \\
& p(z_{n_0,l_0} = t_1c_1t_1, \boldsymbol{z}^{-}_{n_0,l_0}, \boldsymbol{c}|\mathcal{M}) \qquad (1) \\
& \propto (m^{(1)-}_{n_0,t_1} + \alpha_{t_1}) \cdot \frac{(m^{(2)-}_{t_1,c_1} + \eta_{c_1})(m^{(2)-}_{t_1,c_2} + \eta_{c_2})}{sum^{(2)-}_{t_1}(sum^{(2)-}_{t_1} + 1)} \cdot \frac{m^{(3)-}_{t_1,t_1} + \psi_{t_1,t_1}}{sum^{(3)-}_{t_1}} \\
& p(z_{n_0,l_0} = t_1c_1t_2, \boldsymbol{z}^{-}_{n_0,l_0}, \boldsymbol{c}|\mathcal{M}) \\
& \propto (m^{(1)-}_{n_0,t_1} + \alpha_{t_1}) \cdot \frac{m^{(2)-}_{t_1,c_1} + \eta_{c_1}}{sum^{(2)-}_{t_1}} \cdot \frac{m^{(2)-}_{t_2,c_2} + \eta_{c_2}}{sum^{(2)-}_{t_2}} \cdot \frac{m^{(3)-}_{t_1,t_2} + \psi_{t_1,t_2}}{sum^{(3)-}_{t_1}}
\end{aligned}
$$

where $sum^{(2)-}_{t} = \sum_{c=1}^{C} m^{(2)-}_{t,c} + \eta_c$ and $sum^{(3)-}_{t} = \sum_{t'=1}^{T} m^{(3)-}_{t,t'} + \psi_{t,t'}$.

With Equation 1, we can inference the three sets of probability distributions: the user-topic probability distributions $\vec{\theta}_n$, the topic-channel probability distributions $\vec{\beta}_t$ and the topic-transition probability distributions $\vec{\tau}_t$, by Gibbs sampling. Algorithm 1 outlines the LTM inference on sampling latent variables $z_{n,l}$ for each observed watching channel $c_{n,l}$. As explained earlier, there are two forms[1] of $z_{n,l}$, which are $z_{n,l} \in \{1, 2, \ldots, T\}$ or $z_{n,l} \in \{1, 2, \ldots, T\} \times \bar{A}_n \times \{1, 2, \ldots, T\}$. We update the count according to the two forms of $z_{n,l}$. The latent variables $z_{n,l}$ of the first form update the user-topic count $\boldsymbol{m}^{(1)}$ and the topic-channel count $\boldsymbol{m}^{(2)}$ once, but the latent variables of $z_{n,l}$ of the second form update the topic-channel count $\boldsymbol{m}^{(2)}$ with an additional count, as well as update the topic-transition count $\boldsymbol{m}^{(3)}$. This is shown in Function UpdateCnt, which either increases or decreases the count of $\boldsymbol{m}^{(1)}$, $\boldsymbol{m}^{(2)}$ and $\boldsymbol{m}^{(3)}$ by $\delta = +1$ or -1.

B. Optimization

Algorithm 1 suffers from the inefficiency due to the large search space for latent variables $z_{n,l}$ (particularly the triplets). Because the number of unavailable channels $|\bar{A}_n|$ can be close to the total number of channels C, the number of possible values of $z_{n,l}$ is $\mathcal{O}(C \cdot T^2)$. In order to optimize the Gibbs sampling process, we reduce the space for latent variables $z_{n,l}$ to $\mathcal{O}(T^2)$ by combining unsubscribed channels for each user. This is achieved by collapsing the set of latent variables $z_{n,l}$ of the second form (triplets) with the same pairs of topics t_1 and t_2, i.e., we use $z'_{n,l} = (t_1, t_2)$ to represent $z_{n,l} \in \{t_1\} \times \bar{A}_n \times \{t_2\}$. Thus the number of possible values of $z'_{n,l}$ is $T + T^2$. This optimization effectively reduces the running time of Gibbs sampling by a factor proportional to C (in our case it cuts down the time to $\frac{1}{50}$ to $\frac{1}{100}$ as compared to the original Gibbs sampling on the larger space of latent variables).

In the original Gibbs sampling (without optimization), for each watching channel, we build one probability distribution over all possible values of the latent variable, and sample from it once for the value of the latent variable, shown from line 10 to 13 in Algorithm 1. It thus requires multiple calculations of

[1] $\{1, 2, \ldots, T\}$ is short-formed to $[1, T]$ in the Algorithm 1 and Function UpdateCnt

Algorithm 1: Algorithm for LTM inference

input : A list of observed watching channels $\{c_{n,l}|l = 1, \ldots, L_n\}$ for each user n
output: latent variable assignment $z_{n,l}$ for each $c_{n,l}$

1 **foreach** *user* n **do**
2 **foreach** *watching channel* $c_{n,l}$ **do**
3 Randomly choose z_0 from $[1, T] \cup ([1, T] \times \bar{A}_n \times [1, T])$;
4 $z_{n,l} \leftarrow z_0$;
5 `UpdateCnt` $(z_{n,l}, \boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}, \boldsymbol{m}^{(3)}, +1)$;
6 **foreach** *iteration till convergence* **do**
7 **foreach** *user* n **do**
8 **foreach** *watching channel* $c_{n,l}$ **do**
9 `UpdateCnt` $(z_{n,l}, \boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}, \boldsymbol{m}^{(3)}, -1)$;
10 **foreach** $z \in [1, T] \cup ([1, T] \times \bar{A}_n \times [1, T])$ **do**
11 calculate $p(z_{n,l} = z, \boldsymbol{z}^{-}_{n,l}, \boldsymbol{c}|\mathcal{M})$ by formulae in Equation 1;
12 Randomly choose z_0 from $[1, T] \cup ([1, T] \times \bar{A}_n \times [1, T])$ with probabilities $p(z_{n,l} = z, \boldsymbol{z}^{-}_{n,l}, \boldsymbol{c}|\mathcal{M})$;
13 $z_{n,l} \leftarrow z_0$;
14 `UpdateCnt` $(z_{n,l}, \boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}, \boldsymbol{m}^{(3)}, +1)$;

Function `UpdateCnt` $(z_{n,l}, \boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}, \boldsymbol{m}^{(3)}, \delta)$

if $z_{n,l} \in [1, T]$ **then**
  $m^{(1)-}_{n,z_{n,l}} \leftarrow m^{(1)-}_{n,z_{n,l}} + \delta$;
  $m^{(2)-}_{z_{n,l},c_{n,l}} \leftarrow m^{(2)-}_{z_{n,l},c_{n,l}} + \delta$;
else
  $m^{(1)-}_{n,z^{(1)}_{n,l}} \leftarrow m^{(1)-}_{n,z^{(1)}_{n,l}} + \delta$;
  $m^{(2)-}_{z^{(1)}_{n,l},z^{(2)}_{n,l}} \leftarrow m^{(2)-}_{z^{(1)}_{n,l},z^{(2)}_{n,l}} + \delta$;
  $m^{(3)-}_{z^{(1)}_{n,l},z^{(3)}_{n,l}} \leftarrow m^{(3)-}_{z^{(1)}_{n,l},z^{(3)}_{n,l}} + \delta$;
  $m^{(2)-}_{z^{(3)}_{n,l},c_{n,l}} \leftarrow m^{(2)-}_{z^{(3)}_{n,l},c_{n,l}} + \delta$;

the probability distributions for a user, one for each watching channel. In the optimized Gibbs sampling, we can consolidate watching channels from the same user, and build just one probability distribution at the user level. For each user at each iteration, we first subtract all latent variables for her watching channels from the topic-channel count $\boldsymbol{m}^{(2)}$ and the topic-transition count $\boldsymbol{m}^{(3)}$, but retain the user-topic count $\boldsymbol{m}^{(1)}$, and then calculate the probabilities according to Equation 2, where $sum_{n,t}^{(4)-} = \sum_{c \in \bar{A}_n} m_{t,c}^{(2)-} + \eta$.

$$
\begin{aligned}
& p(z'_{n_0,l_0} = t_1, \boldsymbol{z}'^{-}_{n_0,l_0}, \boldsymbol{c}|\mathcal{M}) \propto (m_{n_0,t_1}^{(1)-} + \alpha) \cdot \frac{m_{t_1,c_1}^{(2)-} + \eta}{sum_{t_1}^{(2)-}} \\
& p(z'_{n_0,l_0} = t_1 t_1, \boldsymbol{z}'^{-}_{n_0,l_0}, \boldsymbol{c}|\mathcal{M}) \qquad (2) \\
& \propto (m_{n_0,t_1}^{(1)-} + \alpha) \cdot \frac{sum_{n,t_1}^{(4)-}(m_{t_1,c_2}^{(2)-} + \eta)}{sum_{t_1}^{(2)-}(sum_{t_1}^{(2)-} + 1)} \cdot \frac{m_{t_1,t_1}^{(3)-} + \lambda\psi}{sum_{t_1}^{(3)-}} \\
& p(z'_{n_0,l_0} = t_1 t_2, \boldsymbol{z}'^{-}_{n_0,l_0}, \boldsymbol{c}|\mathcal{M}) \\
& \propto (m_{n_0,t_1}^{(1)-} + \alpha) \cdot \frac{sum_{n,t_1}^{(4)-}}{sum_{t_1}^{(2)-}} \cdot \frac{m_{t_2,c_2}^{(2)-} + \eta}{sum_{t_2}^{(2)-}} \cdot \frac{m_{t_1,t_2}^{(3)-} + \psi}{sum_{t_1}^{(3)-}}
\end{aligned}
$$

This however is done at the cost of losing one count on topic-channel count $\boldsymbol{m}^{(2)}$ since this set of latent variables do not record which unsubscribed channel was considered as the first topic, i.e., we use the observed channels to estimate the topic-channel distribution, not the set of unsubscribed channels that users considered before changing to an available channel.

Multiple $z'_{n,l}$ are then sampled from the same probability distribution, one for each watching channel. This is much more efficient since one probability distribution is utilized by multiple watching channels. Since count $\boldsymbol{m}^{(1)}$ is retained from the previous iteration and the counts $\boldsymbol{m}^{(2)}$ and $\boldsymbol{m}^{(3)}$ are hardly affected by just one user, the modified probability distribution is very close to the probability distributions built for sampling the latent variables one by one. With all latent variables sampled for all her watching channels, we finally update the counts $\boldsymbol{m}^{(1)}$ (after a reset), $\boldsymbol{m}^{(2)}$ and $\boldsymbol{m}^{(3)}$ together.

V. Experiments

A. Dataset Description

Availability constraint is a novel concept, and current public datasets have not included this information. Similarly to previous works on TV datasets [6], we need to rely on a proprietary dataset, because there is a lack of suitable public dataset. An Asian media company provided us a dataset on customers' TV subscriptions and their watching histories over a month. Due to the restrictions from the company, we cannot disclose too much details about the dataset.

Our dataset includes approximately 100 channels which are grouped into a handful of bundles. In the actual setting, customers select bundles to subscribe. As mentioned in Section II, the appropriate evaluation task is thus to predict bundles, rather than individual channels, because a customer may effectively subscribe to some channels not due to preferences, but simply due to their being in the same bundle as other preferred channels. For each customer with at least 4 bundle subscriptions, we randomly "hide" one bundle, and predict this hidden bundle based on only the watching history from the remaining bundles. For example, if a customer subscribes to bundles 1, 2, 3 and 5, and 1 is randomly chosen as the bundle to be "hidden", we then remove all channels in bundle 1 from her watching history. As customers must subscribe at least 3 bundles, it makes no sense to predict the third bundle with two bundles, thus we consider customers with at least 4 bundles.

We further define a watching session as a continuous interval on a channel at least 15 minutes long, because some short intervals may be due to channel surfing. To more effectively learn their preferences, we select those customers who have 100 watching sessions or more. Finally, we obtain a set of approximately 7000 customers, whose average number of watching sessions is about 142.

B. Prediction Measures

For each customer, LTM outputs topic distributions $\vec{\theta}_n$ for each customer n, and topic-channel distribution $\vec{\beta}_t$ for each topic t. The preference $p_{n,c}$ of customer n on channel c, is thus computed by $p_{n,c} = \sum_{t=1}^{T} \theta_{n,t}\beta_{t,c}$.

To predict the next bundle that she may subscribe, we need to compute the preferences over bundles from the preferences over channels. As mentioned in Section II, bundle configuration [17], [18] is not the focus of our problem, because the bundles are specified in the dataset. The summation aggregation used in previous work [17], [18] is inappropriate because it favors larger bundles, and in our case the bundles vary in sizes. Thus, to compute the preferences over bundles more equitably, we adopt two aggregate measures: *Average* and *Maximum*.

1) A customer may subscribe to a bundle when she generally likes all channels in the bundle. *Average* measure takes the average of $p_{n,c}$ over all the channels in a bundle. For a bundle $b \in B - S_n$, where B is the set of all bundles and S_n is the set of bundles customer n has subscribed to, the average preference of customer n on bundle b is computed by $\text{avg}(n,b) = \frac{1}{|b|}\sum_{c \in b} p_{n,c}$.

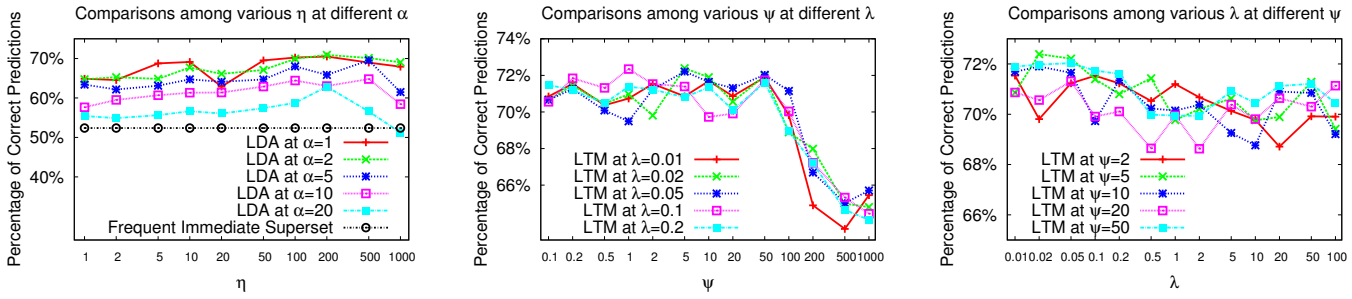
2) A customer may also subscribe to a bundle because she particularly likes one channel in the bundle, and she does not have the option to subscribe to that channel only. *Maximum* measure takes the maximum value of $p_{n,c}$ over all channels in a bundle, i.e., for $b \in B - S_n$, the maximum preference of customer n on bundle b is computed by $\max(n,b) = \max_{c \in b} p_{n,c}$.

The predicted bundle b for user n is the bundle with the highest measure, i.e., either $\text{argmax}_{b \in B - S_n}\ \text{avg}(n,b)$ or $\text{argmax}_{b \in B - S_n}\ \max(n,b)$. Accuracy is the percentage of users for which the predicted bundle b is the correct "hidden" bundle.

C. Baselines

As explained in Section II, we would focus on comparisons with other approaches that model preferences as probability distribution over channels/bundles, rather than absolute rating prediction (MF or CF). The first baseline, Frequent Immediate Superset, models probabilities based on the frequencies of bundles alone, which showcases a comparison to a non-topic modeling approach. The second baseline, LDA, models probabilities based on topics but not transition, which showcases a comparison to a *non-transition* topic modeling approach.

Frequent Immediate Superset (FIS). This approach models the conditional probability that a customer will pick a



(a) Comparisons of LDA on the *Average* measure among various values of η at different α

(b) Comparisons of LTM on the *Average* measure among various ψ at different λ

(c) Comparisons of LTM on the *Average* measure among various λ at different ψ

Fig. 3. Effect of model parameters on LDA and LTM

bundle b, given that she has already adopted a subset of bundles $S \subset B$, where B is the universal set of bundles. Let $\|S\|$ be the number of customers who subscribe to at least the bundles in S, and $\|S \cup \{b\}\|$ be the number who subscribe to S as well as b. The probability $p(S \cup \{b\}|S)$ is estimated by $\frac{\|S \cup \{b\}\|}{\|S\|}$. This is equivalent to finding the most frequent immediate superset as the prediction for customers with subscription S, i.e., $\arg\max_{b \in B-S} \|S \cup \{b\}\|$. We call this *Frequent Immediate Superset* (FIS) method, and it predicts the "hidden" bundles correctly for 52.4% of the customers.

LDA. As introduced in Section II, LDA is a randomized algorithm, we therefore ran LDA at $T = 5$, thirty times with different random number generator seeds, and took the average. Considering the number of channels (~ 100) is not large compared to the number of words when discovering topics from documents, we first took a 5% sample, ran the Gibbs sampling for 100 iterations, and then ran for the whole set of customers for 10 iterations. Figure 3(a) presents LDA's prediction accuracy on the *Average* measure at various values of $\eta \in \{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000\}$ when $\alpha \in \{1, 2, 5, 10, 20\}$. When α is too large (> 20), i.e., the topic preference is more homogeneous for each user, LDA does not outperform the FIS method since difference on topic preferences among different users is less pronounced. However, when $\alpha \leq 10$, there is a significant improvement over the FIS baseline, shown in Figure 3(a).

Because of LDA's better performance than FIS, we will subsequently compare against only LDA, adopting the same setting which does best for LDA, namely $\alpha = 2$ and $\eta = 200$.

D. Effect of Transition Priors on LTM

We consider the effect of LTM's transition priors ψ and λ. Figure 3(b) and Figure 3(c) show the prediction accuracies on the *Average* measure at various pairs of ψ and λ for LTM at $T = 5$. Each line in Figure 3(b) is plotted with a fixed λ, while each line in Figure 3(c) is plotted with a fixed ψ.

Figure 3(b) shows the prediction accuracies on the *Average* measure with $\lambda \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$. All curves decline when ψ goes large. The larger ψ is, the less dominant is the count $m^{(3)}_{t1,t2}$ in determining the transition probability from t_1 to t_2. Thus, larger ψ makes the difference between across-topic transitions (transitions from t_1 to t_2, compared with transitions from t_3 to t_2) less obvious (note that the difference between within-topic transitions and across-topic transitions are controlled by λ). If we had assumed that the second-choice topic is independent of the first-choice topic in Section III, larger ψ should not have made an impact to the prediction accuracy since transition probabilities do not matter much in determining the second-choice topic. However, the prediction accuracies go down, which means the transition patterns indeed depend on the topic t_1 a user n takes as her first-choice topic. That shows it is important to consider topic transitions rather than to take the her second-choice topic t_2 as another sample from her topic preference θ_n. On the other hand, the performance goes down as λ increases in Figure 3(c). As within-topic transitions are parameterized by $\lambda\psi$, this shows the transitions happen less likely within the same topics, but more often, from one topic to another.

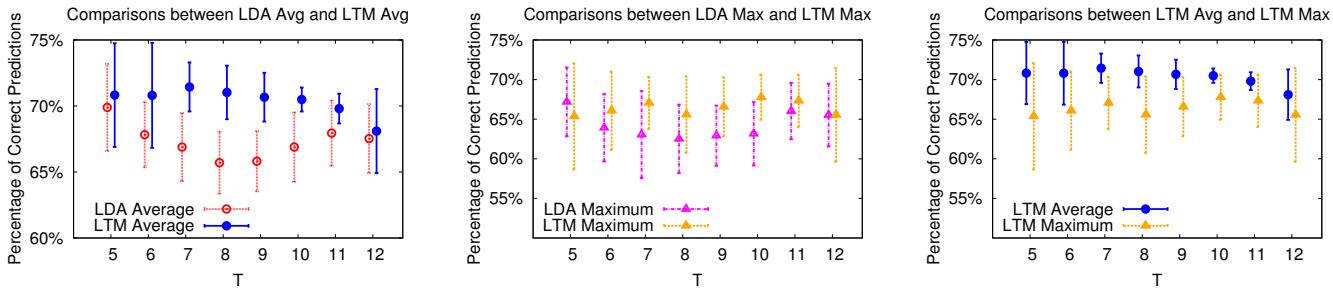
The best setting for ψ and λ at $\psi = 5$ and $\lambda = 0.02$ gives the prediction accuracy of 72.4%, i.e., LTM correctly predicts for three quarters of the customers on their "hidden" bundles. We will use this setting to compare against LDA subsequently.

E. Comparisons between LTM and Baselines

We now compare LTM against LDA for different T. In experiments, we find that both LTM and LDA perform best for relatively small T. We hypothesize that this is due to the relatively low number of bundles. High T, such as ≥ 30, cause overfitting, with all methods dropping below 60% accuracy.

LTM vs. LDA Figures 4(a) and 4(b) show the prediction accuracies on both the *Average* and *Maximum* measures respectively. For each T, we run both LDA and LTM thirty times each (with 30 seeds), and calculate the mean and standard errors of the prediction accuracy. As shown in Figures 4(a) and 4(b), with error bars presenting the standard errors, LTM generally performs better than LDA on both measures, as shown by LTM's generally higher mean in prediction accuracy.

To test the statistical significance of the outperformance by LTM, we conduct one-tailed paired-sample t-test [26] for each T. The null hypothesis H_0 is there is no difference between the two samples. Table III summarizes the p-values with different alternative hypotheses H_1 (first column). p-value is the probability of erroneously rejecting H_0 in favor of H_1. We can reject H_0 and accept H_1 if $p \leq \gamma$, where γ is the significance level (usually 0.01 or 0.05). The lower the p-value, the more significant is the result. In the first row, H_1 is LTM performs better than LDA on the *Average* measure. Except for



(a) LDA vs. LTM: Average Measure (b) LDA vs. LTM: Maximum Measure (c) LTM: Average vs. Maximum Measures

Fig. 4. Comparisons between LDA and LTM on both *Average* and *Maximum* measure with different number of topics

H_1	$T=5$	$T=6$	$T=7$	$T=8$	$T=9$	$T=10$	$T=11$	$T=12$
LTM Avg better than LDA Avg	2e-01*	3e-04	7e-10	2e-10	9e-10	6e-09	4e-04	2e-01*
LTM Max better than LDA Max	9e-01*	3e-02	6e-05	6e-03	1e-04	7e-06	9e-02*	5e-01*
LTM Avg better than LTM Max	2e-08	5e-11	7e-13	3e-10	1e-09	6e-08	1e-05	3e-04
LDA Avg better than LDA Max	2e-06	4e-07	1e-05	2e-06	7e-06	4e-08	3e-04	6e-04

TABLE III. P-VALUE ON PAIRED-SAMPLE T-TEST (ENTRIES WITHOUT ASTERISKS ARE STATISTICALLY SIGNIFICANT)

the few asteriated entries, the p-values are very low, and we can accept H_1 at 0.01 significance level. In the second row, H_1 is that LTM performs better than LDA on the *Maximum* measure. Again, except for the asteriated entries, we can accept H_1 at 0.05 significance level for T = 6, and at 0.01 significance level for the rest. We therefore conclude LTM is significantly better than LDA for many of the T settings (6 to 11). For $T > 12$, there is insufficient statistical evidence to reject H_0. This is still acceptable as both LDA and LTM overfit and perform worse than for $T \leq 12$. Therefore, we do not show them here due to the overfitting issue.

Average v.s. Maximum Measures. In most of the figures above, we only plotted the prediction accuracies on the *Average* measure, since the *Average* measure reports a higher accuracy than the *Maximum* measure in all settings of T for LTM, which is shown in Figure 4(c). *Average* is also better than *Maximum* for LDA, though the corresponding figure for LDA is not shown here due to space constraint. This outperformance by *Average* measure is statistically significant, both for LTM and LDA, as shown by the last two rows of Table III, where the low p-values show that we can accept H_1 at 0.01 significance level. This means most people would subscribe to a bundle because they generally like the channels in the bundle, rather than a particular channel in the bundle. This is intuitive since a consumer spends more on items she likes packaged together, while she may be less willing to spend if her preferences towards the items in the package are very different.

Running Time. As explained in Section IV, the optimization by collapsing latent variables results in a two-order-of-magnitude improvement in the runtime of our algorithm. Due to the extremely large improvement, we do not show the runtime for the unoptimized algorithm, because then it will obscure the much smaller difference between the optimized LTM and the baseline LDA. Figure 5 shows that there is not much difference between the running times of LDA and LTM. Based on per iteration, the difference increases as T goes larger. In most cases, LTM's running time is only marginally higher, which gives a difference of less than 2 seconds. Therefore, our optimized algorithm is considered efficient.

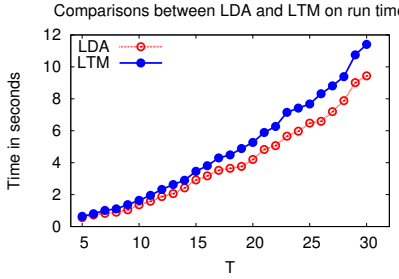


Fig. 5. Comparisons between LDA and LTM on running time per iteration with different number of topics

F. A Case Study

We give a case study here with $T = 10$, $\alpha = 2$, $\eta = 200$, $\psi = 5$ and $\lambda = 0.02$, to illustrate the discovered topics and the transitions among topics. The topics in terms of channels are presented in Table IV. The second column are the names that we manually give to each topic for ease of identification. The upper row of the third column are the leading channels that are present in the topic t, sorted by ascending order of $\beta_{t,c}$. The lower row of the third column lists the destinations that the topic transitions to. The percentages are the transition probabilities. Those with probability $< 5\%$ are not shown.

The topics found in Table IV are generally intuitive. For instance, most of the Chinese language channels are grouped together into t_9. In our dataset, most of the people who watch such channels are not likely to watch shows in other languages (e.g., English shows). However, kids channels are split into two topics, t_3 and t_8, where channels in t_3 are more for toddlers and younger kids, but channels in t_8 are for elder kids, even teenagers. This can be even observed from the transitions. As elder kids are more exposed to other channels, we shall expect to see some transitions from unavailable channels (which they might have watched before at friends' house) to t_8. This is certainly true as topics t_1, t_2, t_5 and t_7 all transition to t_8 with probabilities more than 5%, but not to t_3.

Similarly, there are also two topics for Education channels, t_2 and t_5. We think the difference between t_2 and t_5 is that,

	Manual Label	Top Channels and Top Transitions
t_1	Mixed	One HD, Asian Food Channel, Discovery Channel, National Geographic Channel
		t_5 (21.9%), t_1 (20.5%), t_{10} (20.2%), t_8 (18.2%), t_2 (14.0%)
t_2	Education-A	BBC Knowledge, Discovery Channel, National Geographic Channel, History
		t_8 (68.9%), t_{10} (10.1%), t_5 (15.3%)
t_3	Kids-Younger	Disney Junior, Nick Jr, Baby TV, CBeebies, JimJam, Boomerang
		t_3 (88.0%)
t_4	Entertainment	FOX, AXN, Universal Channel, FOXCRIME, Animax, WarnerTV
		t_{10} (33.0%), t_5 (31.8%), t_2 (16.1%), t_4 (7.9%), t_6 (6.1%)
t_5	Education-B	History, Crime & Investigation Network, National Geographic Ch, Discovery Ch
		t_2 (50.8%), t_8 (24.6%), t_{10} (12.7%), t_4 (6.1%)
t_6	News	Sky News, CNBC, BBC World News, CNN, FOX News Channel
		t_{10} (39.1%), t_2 (27.8%), t_5 (21.3%), t_4 (7.5%)
t_7	Entertainment-HD	Star World HD, FOX HD, AXN HD, Universal Channel HD, FOXCRIME HD
		t_8 (53.2%), t_2 (17.0%), t_5 (13.6%), t_{10} (7.8%)
t_8	Kids-Elder	Disney Channel, Nickelodeon, Cartoon Network, Boomerang, Disney Junior
		t_8 (90.6%)
t_9	Chinese	TVBS Asia, One, CTI TV, E City, Star Chinese Channel, TVBS News
		t_9 (97.5%)
t_{10}	Lifestyle	Star World, E! Entertainment, Food Network Asia, BBC Lifestyle
		t_{10} (95.8%)

TABLE IV. CASE STUDY: TOPICS BY CHANNELS AND TRANSITIONS

t_5 contains more serious educational channels with narrower range of audiences, e.g., Crime & Investigation Network. So transitions happen from harder to understand channels to easier ones, for example, there are significantly more transitions from t_5 to t_2 than from t_2 to t_5, there are also more transitions from other topics to t_{10} (Lifestyles) than from t_{10}. Users who wish to find substitutes for channels in t_2 may probably turn to t_8 (easier) than to t_5. Less transitions are observed from t_5 to t_8 than from t_2 to t_8, which confirms that the transitions reduce with the difference between the levels of target audiences.

Another interesting finding is that, HD channels (e.g., t_7) do not transition to channels in the same category but with normal resolution (e.g., t_4). This goes against our initial intuition that, when people find HD channels are not available, they would replace them by the corresponding normal channels. But the transition pattern of topic t_7 suggests otherwise. When we consider it further, this actually makes sense. The same channels, whether in HD or normal resolution, are perfect substitutes of one another. Most people think of "interests" in terms of the content, and not the resolution of the channels. Thus the transitions from a HD channel to its normal channel is never observed. Meanwhile, we would expect that an HD topic and the corresponding normal topic share similarities between their transition patterns. This is indeed the case, demonstrated by the frequent transitions from t_4 or t_7 to t_{10}, t_5 and t_2.

VI. CONCLUSION

We propose Latent Transition Model (LTM) to model user preferences in the presence of availability constraints. The key novel concept is availability-based topic transition, whereby a user transitions from one topic to another if the first-chosen item is not available for consumption. LTM is validated with a real dataset, and shown to be effective and efficient in predicting unsubscribed bundles. As future work, we plan to investigate the issue of multiple transitions for a single observation, where the observed channel is not necessarily the first choice or the second choice, but rather the n-th choice. It is also interesting to factor in economic considerations, such as when items are available at different prices.

ACKNOWLEDGMENT

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

REFERENCES

[1] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *TKDE*, vol. 17, no. 6, 2005.

[2] F. C. T. Chua, H. W. Lauw, and E.-P. Lim, "Predicting item adoption using social correlation," in *SDM*, 2011.

[3] S. Chae, "Bundling subscription tv channels: A case of natural bundling," *International Journal of Industrial Organization*, vol. 10, no. 2, 1992.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, 2003.

[5] D. Agarwal and B.-C. Chen, "flda: matrix factorization through latent dirichlet allocation," in *WSDM*, 2010.

[6] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *ICDM*, 2008.

[7] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *SIGIR*, 2005.

[8] D. Kelly and J. Teevan, "Implicit feedback for inferring user preference: a bibliography," *SIGIR Forum*, vol. 37, no. 2, 2003.

[9] D. Yang, T. Chen, W. Zhang, Q. Lu, and Y. Yu, "Local implicit feedback mining for music recommendation," in *RecSys*, 2012.

[10] N. Aizenberg, Y. Koren, and O. Somekh, "Build your own music recommender by modeling internet radio streams," in *WWW*, 2012.

[11] S.-H. Yang, B. Long, A. J. Smola, H. Zha, and Z. Zheng, "Collaborative competitive filtering: learning recommender using context of user choice," in *SIGIR*, 2011.

[12] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, 2009.

[13] R. Bell, Y. Koren, and C. Volinsky, "Modeling relationships at multiple scales to improve accuracy of large recommender systems," in *KDD*, 2007.

[14] T. Hofmann, "Latent semantic models for collaborative filtering," *TOIS*, vol. 22, no. 1, 2004.

[15] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *NIPS*, vol. 20, 2008.

[16] ——, "Bayesian probabilistic matrix factorization using markov chain monte carlo," in *ICML*, 2008.

[17] D.-N. Yang, W.-C. Lee, N.-H. Chia, M. Ye, and H.-J. Hung, "On bundle configuration for viral marketing in social networks," in *CIKM*, 2012.

[18] M. Xie, L. V. Lakshmanan, and P. T. Wood, "Breaking out of the box of recommendations: from items to packages," in *RecSys*, 2010.

[19] D. M. Blei and J. D. Lafferty, "Correlated topic models," in *NIPS*, 2006.

[20] ——, "A correlated topic model of science," *AAS*, vol. 1, no. 1, 2007.

[21] K. Salomatin, Y. Yang, and A. Lad, "Multi-field correlated topic modeling," in *SDM*, 2009.

[22] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *ICML*, 2006.

[23] Y. Wang, E. Agichtein, and M. Benzi, "Tm-lda: efficient online modeling of latent topic transitions in social media," in *KDD*, 2012.

[24] Q. Liu, E. Chen, H. Xiong, and C. H. Ding, "Exploiting user interests for collaborative filtering: interests expansion via personalized ranking," in *CIKM*, 2010.

[25] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *PNAS*, vol. 101, no. suppl. 1, 2004.

[26] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability and statistics for engineers and scientists*. Prentice Hall, 1998.