Probabilistic Use Cases: Discovering Behavioral Patterns for Predicting Certification

Cody A. Coleman
Massachusetts Institute of Technology
Cambridge, MA 02139
colemanc@mit.edu

Daniel T. Seaton
Massachusetts Institute of Technology
Cambridge, MA 02139
dseaton@mit.edu

Isaac Chuang
Massachusetts Institute of Technology
Cambridge, MA 02139
ichuang@mit.edu

ABSTRACT
Advances in open-online education have led to a dramatic increase in the size, diversity, and traceability of learner populations, offering tremendous opportunities to study detailed learning behavior of users around the world. This paper adapts the topic modeling approach of Latent Dirichlet Allocation (LDA) to uncover behavioral structure from student logs in a MITx Massive Open Online Course, 8.02x: Electricity and Magnetism. LDA is typically found in the field of natural language processing, where it identifies the latent topic structure within a collection of documents. However, this framework can be adapted for analysis of user-behavioral patterns by considering user interactions with courseware as a "bag of interactions" equivalent to the "bag of words" model found in topic modeling. By employing this representation, LDA forms probabilistic use cases that clusters students based on their behavior. Through the probability distributions associated with each use case, this approach provides an interpretable representation of user access patterns, while reducing the dimensionality of the data and improving accuracy. Using only the first week of logs, we can predict whether or not a student will earn a certificate with 0.81 ± 0.01 cross-validation accuracy. Thus, the method presented in this paper is a powerful tool in understanding user behavior and predicting outcomes.

Author Keywords
Latent Dirichlet Allocation; Student Behavior; Use Case Modeling; Massive Open Online Courses

ACM Classification Keywords
I.5.2 Design Methodology: Feature evaluation and selection

INTRODUCTION
Massive Open Online Courses (MOOCs) create a tremendous opportunity to study learning from the perspective of large and diverse populations of students. In the first year alone, HarvardX[1] and MITx[2] courses enrolled roughly 600,000 unique users from around the world [10]. Such large numbers, combined with diverse backgrounds and enrollment motivations, implies variation in how users choose to interact with material. Clickstream data – stored records of user interactions with course content – provide the opportunity to understand such variation. Previous studies have aggregated clickstream data to inform broad metrics such as the unique number of resources accessed within a course [5], while others offered more detailed activity such as pause and play clicks within a single lecture video [12]. These data provide a great deal of insight into student behavior, but enumerating all possible student-interaction patterns is nearly impossible. Furthermore, interpreting such patterns remains a daunting task for researchers and course teams alike.

In this paper, we make the problem of modeling student behavior more tractable by adapting the approach of Latent Dirichlet Allocation (LDA) [4]. LDA is an unsupervised probabilistic model, which has had great success illuminating shared topics in large collections of texts [2, 3, 4]. Along with natural language processing, LDA has been adapted in areas such as genetics [15] and web page recommendation [22]. In the latter, LDA discovered latent topics associated with the semantics of user webpage access patterns, while delivering better performance compared to conventional clustering techniques [22]. Inspired by these adaptations, we use LDA to distill user interactions in an educational context by considering user interactions with resources making up a course.

Our adaptation of LDA results in a finite set of use cases representing the probability distributions of a participant interacting with each resource in the courseware. Behavioral patterns can be deduced from the most probable resources within each use case. Within any digital course containing unique resource identifiers, these probabilities offer a natural interpretation of behavioral patterns in a course. An additional feature of LDA is the mixed-membership model, where student behavior is represented as different proportions of a shared set of use cases, rather than hard cluster assignments. This enabled us to compare students by their relative proportions, define behavioral patterns, and reduce the dimensionality of the data for further analysis and prediction. Detecting such patterns is important to handle the openness of MOOCs, which has been tied to a variety of behavioral patterns, as evidenced

L@S 2015, March 14–18, 2015, Vancouver, BC, Canada

http://dx.doi.org/10.1145/2724660.2724662

[1] Harvard University's institution for creating MOOCs
[2] MIT's institution for creating MOOCs

by large initial enrollments, low percentages of completions, and widely varying resource use [14, 10, 18].

In this paper, we adapt LDA to edX clickstream data in order to address the following questions:

- Can LDA serve as an unsupervised approach for discovering the behavioral trends of MOOC participants?

- Can the mixed-membership model from LDA predict certification?

Our application involves one MITx MOOC, an introductory physics course called 8.02x: Electricity and Magnetism from the spring of 2013.

The paper continues as follows. The Related Work section summarizes work related to modeling learner behavior as context for our work. The Course Studied and Dataset section overviews the data examined in this paper. The Methods section describes the theory behind LDA and how it is adapted to and evaluated in an educational context. The Results section provides the outcome from applying LDA to 8.02x. The Discussion section outlines strengths and weaknesses of this approach. The Conclusion section summarizes the key contributions of this paper.

RELATED WORK

Understanding student behavior has been a consistent theme in MOOC research. Most studies aim to group students by their behavior, and then better understand how discovered behavior leads to educational advancement. A central challenge to any study includes significant aggregation of raw data sets, often requiring advanced methods that scale to large data sets.

Many researchers have employed machine learning and pattern recognition techniques to distill raw clickstream data into more interpretable models of student behavior. Kizilcec et al. [13] applied clustering techniques to gain insight into student disengagement and course completion. They represented students by their weekly activity, capturing whether or not students were "on track," "behind," "auditing," or "out" each week. Using these features, they performed k-means clustering and constructed four learner subpopulations: "completing", "auditing", "disengaging", and "sampling". These student subpopulations were then compared in terms of their demographics, surveyed experience, forum activity, and video streaming index to analyze retention. Rameh et al. [16] used the graphical model of Probabilistic Soft Logic (PSL) to distinguish forms of engagement in MOOCs. In contrast to Kizilcec et al., Rameh et al. viewed engagement/disengagement as latent variables and focused on social behaviors such as posting and subscribing in addition to more traditional behaviors such as following course material and completing assessments. Their study illustrated the informative role peer-to-peer interactions can play in user modeling. With a similar methodology Yang et al. [23] used social behavior for a survival analysis of students in MOOCs, finding that social engagement within the course was correlated with retention.

In this paper, we provide another perspective. Rather than rigidly defined feature sets, we use LDA to uncover behavioral patterns directly from the data in a unsupervised manner. This preserves much of the statistical information from the original dataset, while still forming an interpretable representation. Unlike the studies above, we don't sacrifice much granularity for interpretability.

COURSE STUDIED AND DATASET

8.02x: Electricity and Magnetism is an MITx MOOC offered by edX in the spring of 2013, based on an introductory physics course at MIT. Between January, 17 and September 8, enrollment reached 43,758 people (MIT has since removed this course), from around the world with a wide range of educational backgrounds [17]. Courseware interactions from these enrollees led to 37,394,406 events being recorded in the edX tracking logs [19]. Courseware in this context refers to the individual learning components and software features available to 8.02x participants.

The resources in 8.02x included a variety of videos, problems, textual content, and simulation activities. The major assessments consisted of weekly problem sets (18%), interactive simulations with concept questions (2%), and examinations – three midterms (15% each) and a final (30%). To promote engagement and self-assessment, weekly lectures were typically broken into roughly 5-10 video segments, each interspersed with graded multiple choice questions. These resources are organized hierarchically. Chapters, sequences, and verticals are container objects that form organizational units in the course. Within these containers are the course resources [7]. In order to better understand the course structure of 8.02x, a screenshot is provided in Figure 1. Unique resources are navigated in two ways: the left navigation bar provides a link to sequences of content that are organized into chapters (represented by weeks of material), while the top navigation provides access to individual resources and verticals. More information about 8.02x can be found in the MITx Course report [19].

METHODS

This section explains how LDA is adapted to model user behavior in MOOCs and the processes used to predict certification. Beginning with an overview of the theoretical background of LDA, we cover its original use for topic modeling in natural language processing [4] and draw a parallel between topic modeling and user modeling, which forms the basis for probabilistic use cases. Then, we explain how a students interactions are represented and quantified. This representation is evaluated according to their ability to predict certification and serves as a baseline for LDA. The section concludes with an explanation of the two-part evaluation process for LDA.

LDA for Traditional Topic Modeling

Traditionally, LDA has been used to understand the latent topic structure of textual documents. Topics are thought of as probability distributions over a fixed and shared vocabulary. LDA is an unsupervised technique, meaning initially there are no keywords or tags that can be used for categorization

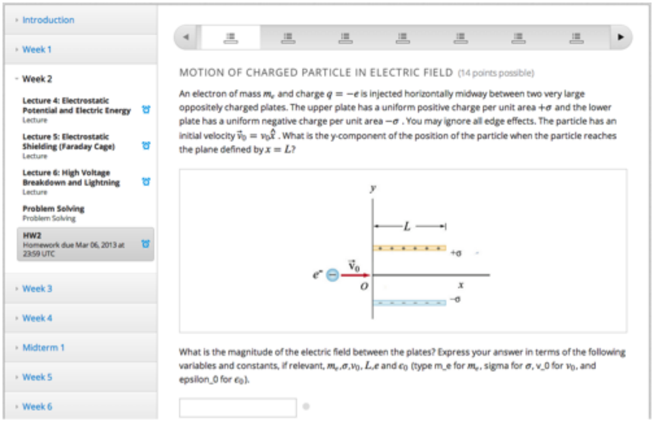


Figure 1: Screenshot of student display for 8.02x courseware. The left navigation bar provides access to weekly chapters, while the main display, to its right, offers videos, problems, and html pages packaged in lecture, problem set, tutorial, and exam sequences.

by topic. Hence, the topics, their distributions, and the topic assignments of each word are hidden variables that need to be estimated. These hidden variables are combined with the observable variables – document word counts for each word of the fixed vocabulary – to form a generative process that defines a joint distribution over the hidden and observable variables. This distribution is used to form a posterior distribution for the hidden variables that is optimized through an approximation to the Expectation-Maximization (EM) Algorithm [3, 4].

More formally, LDA assumes there are K topics in the collection of T documents that have a fixed vocabulary (V). These topics are indexed by $\hat{z} = 1, ..., K$ and represent a probability distribution over V called $\beta_{\hat{z}}$, where each word ($\hat{w}$) has a probability $\beta_{(\hat{w}|\hat{z})}$. Each document ($d^t$) in the collection can be represented as a bag of n_t words, i.e. $d^t = ({w_1}^t, ..., {w_{n_t}}^t)$. Although all of the documents share the same set of topics, each document has its own topic proportions (θ^t). θ^t is a categorical distribution sampled from a Dirichlet distribution with parameters α, where $\theta^t_{\hat{z}}$ is the probability of topic $\hat{z}$ in document d^t. This categorical distribution in turn is the basis for a multinomial distribution used to assign each word in d^t to a topic, i.e. $z^t_1, ..., z^t_{n_t}$. Using this formulation gives rise to an expansion of the joint distribution, $\prod_{t=1}^{T} P(d^t, z^t_1, ..., z^t_{n_t}, \phi^t, \beta; \alpha)$, as shown in Equation 1.

$$\prod_{z=1}^{K} P(\beta_z) \prod_{t=1}^{T} P(\theta^t|\alpha)\{\prod_{i=1}^{n_t} \theta^t_z \beta_{w_i|z}\} \tag{1}$$

Unfortunately, the posterior distribution for hidden variables defined by LDA is normally intractable because of the marginal distribution for the documents [4]. To approximate a solution, we use the python package gensim [8], which implements an online variational Bayes algorithm as proposed by Hoffman et al. [11].

LDA for Probabilistic Use Cases

To model behavior, we represent students as a bag of interactions with the courseware. Each of the static resources in the backbone of the course, as defined in the Course Studied and Dataset section, has a unique module identifier. These module identifiers ($\hat{m}$) form a fixed course vocabulary or structure ($\hat{m} \in C$) for LDA. In 8.02x, there were 1,725 unique module identifiers. With this information, a student in a course with T registrants can be modeled as $s^t = (m^t_1, ..., m^t_{n_t})$, where m^t_i represents an interaction with a course module. By substituting the students in a course for the collection of documents, topics describe behavioral patterns rather than words. For clarity, we refer to these interaction-based topics as *probabilistic use cases*. As such, use cases are similarly indexed by $\hat{u} = 1, ..., K$ and define a probability distribution over C called $\beta_{\hat{u}}$, where each module has an interaction probability of $\beta_{\hat{m}|\hat{u}}$. Students, like documents, share the same set of use cases, but in different proportions defined by ϕ^t. Equation 2 shows the expansion of the the joint distribution, $\prod_{t=1}^{T} P(d^t, u^t_1, ..., u^t_{n_t}, \phi^t, \beta; \alpha)$, that is parallel to the topic modeling application. This model builds on the existing infrastructure for topic modeling and allows us to investigate the hidden behavioral structure within a course.

$$\prod_{u=1}^{K} P(\beta_u) \prod_{t=1}^{T} P(\phi^t|\alpha)\{\prod_{i=1}^{n_t} \phi^t_u \beta_{m_i|u}\} \tag{2}$$

Quantifying Interactions

In applying LDA to model the behavior of MOOC participants, each student is represented as a bag of interactions with the courseware, where we only consider browser issued events. To quantify these interactions, we used time spent in seconds on each course module. Time was calculated by taking the difference between browser event timestamps. Breaks over 30 minutes long were discarded. This use of time spent is unique to the context of modeling user behavior. The traditional topic modeling application are limited to binary indicators or word counts.

While using time spent to quantify user interactions, the bag of interactions model was tested based on its ability to accurately predict whether or not a student would earn a certificate. For each week in 8.02x's 18 week runtime, we separately generated each of the interaction representations using the logs from the beginning of the course to the end of the given week. The performance of each representation was quantified by 5-fold cross validation of a Support Vector Machine (SVM) classifier for certification, where Different Error Costs (DEC) compensated for the class imbalance [20]. This provided a baseline to compare the predictive performance of a student's use case proportions (ϕ^t).

Evaluating Probabilistic Use Cases

Using the best interaction representation from the above process, LDA was evaluated on how well it modeled the data in addition to its predictive performance. Traditionally, model selection, i.e. selecting the optimal number of use cases, is based upon log perplexity [4] per interaction. This method is equivalent to the negative log likelihood of a corpus (approximated by the Evidence Lower Bound) divided by the number of interactions within that corpus, as in Equation 3. This is commonly used in natural language processing to evaluate language models. We use log perplexity per interaction here to reduce perplexity to a reasonable numerical scale.

$$log\{perplexity(corupus)\} = \frac{-log\{P(corpus|\alpha,\beta)\}}{\sum_{corpus} n_t}. \tag{3}$$

Using the models that fit well without an excessive number of use cases, we evaluated how well LDA predicted certification. LDA was trained on a weekly basis, where only the logs from the beginning of the course to the end of the given week were used. Students were then represented by their use case proportions (ϕ^t) in the 5-fold cross validation of a SVM classifier for certification, where DEC compensated for the class imbalance [20]. This approach demonstrated the predictive power of a student's use case proportions (ϕ^t) and quantified the effect that varying the number of use cases has on performance.

RESULTS

The results from applying LDA to 8.02x are broken into 4 subsections. The Interaction Representation subsection evaluates the time spent bag of interactions representation. Employing this representation, the Identifying the Number of Use Cases Through Log Perplexity subsection explores how well LDA fits the data for a varying number of use cases. The Probabilistic Use Cases subsection visualizes and explains the resulting use cases through their probability distribution over the course structure. In the final subsection, Predicting Certification, students' use case proportions are used in order to predict certification.

Predicting Certification with a Bag of Interactions Model

A student's activity within the edX platform can be quantified in a number of ways. In this section, we evaluate the ability of the bag of interactions model to predict certification of students. User interactions are quantified by time spent in seconds on each course module. We use 5-fold cross-validation over the 8.02x's 18 week duration as our primary heuristic. Table 1 shows the average scores broken down into true positive rates (TPR) for identifying certificate earners and true negative rates (TNR) for identifying non-certificate earners. Comparing both metrics illuminates any asymmetries in performance due to class imbalance [1].

Based on Table 1, using time spent for the bag of interactions model achieved high performance on non-certificate earners very early in the course, while performance on certificate earners lagged behind. This asymmetry is likely due to the overwhelming number of non-certificate earners. Only 4.2% of registrants earned a certificate in 8.02x [19], making the classes of non-certificate earners and certificate earners extremely imbalanced. Thus, the early performance of the bag of interactions model must be taken with skepticism because

	Time Spent					
Week	1	2	3	4	5	6
TPR	0.38	0.39	0.39	0.45	0.48	0.54
TNR	0.92	0.95	0.97	0.98	0.98	0.98
Week	7	8	9	10	11	12
TPR	0.76	0.81	0.91	0.91	0.93	0.94
TNR	0.97	0.97	0.97	0.97	0.97	0.97
Week	13	14	15	16	17	18
TPR	0.95	0.95	0.96	0.96	0.96	0.97
TNR	0.97	0.97	0.98	0.98	0.97	0.98

Table 1: True positive rates (TPR) and true negative rates (TNR) for identifying certificate earners with different interaction representations.

a trivial majority classifier could achieve 95% overall accuracy by labeling the entire feature space as non-certificate earners. Balanced prediction between class is much more difficult.

Identifying the Number of Use Cases Through Log Perplexity

Using the time spent on modules as the underlying representation, 5-fold cross-validation of log-perplexity per interaction is displayed in Figure 2. The optimal number of use cases appeared to be around 50, however, it is unclear from cross-validation alone how much of effect such a large number of use cases has on our ability to interpret the model. Determining the right balance between predictive performance and interpretability is currently an open issue in probabilistic topic modeling [2]. Although there has been some work that tries to quantify interpretability [6], our vocabulary of course modules is only understood by a small set of individuals, making it difficult for us to apply those strategies here. Hence, in the next section we chose to visually explore how use cases vary and separate as the number of use cases increases.

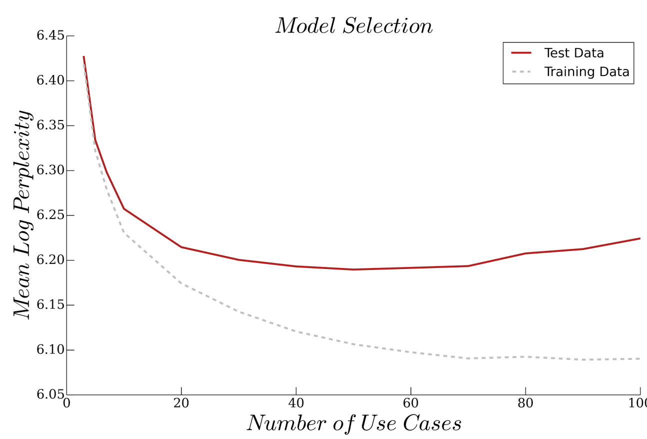


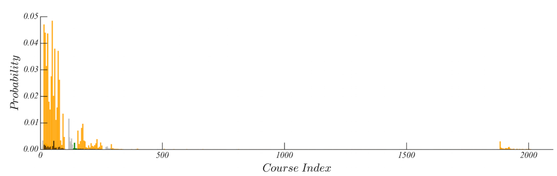
Figure 2: 5-fold log perplexity for a varying number of use cases.

Probabilistic Use Cases

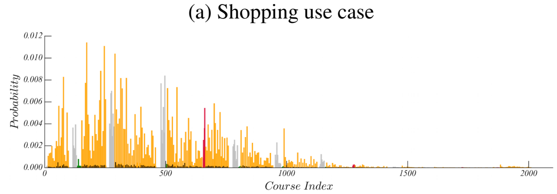
This section illustrates the descriptive power of probabilistic use cases by plotting their probability distributions according to the course structure. With the 3-use case model as a baseline, we describe the resulting behavioral patterns. Subsequently, we investigate how these use cases evolved over the course's duration and subdivide as the number of use cases is increased.

Figure 3 shows the probability distributions of the 3-use case model after all 18 weeks of 8.02x. Each probability distribution is color-coded according to the course structure visual aid at the lower most x-axis. Color indicates the type of resource, and the length of each vertical bar is the weight towards the final grade. In order to ensure consistency, all figures in this section use the same visual aid in conveying course structure within each probability distribution.

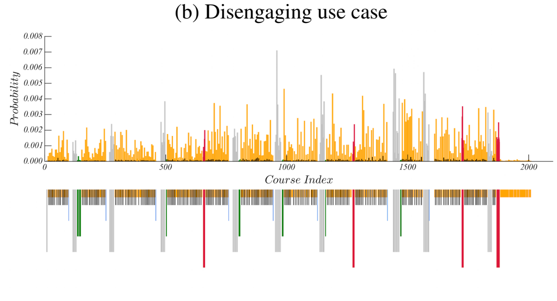
Each of the presented use cases in Figure 3 illuminates a distinct behavioral pattern in 8.02x. The use case in Figure 3a concentrated the majority of its probability on videos from the first week of the courses. This skewed distribution resulted from the large population of registrants that only watched the videos at the beginning of the course before stopping activity.



(a) Shopping use case



(b) Disengaging use case



(c) Completing use case

Figure 3: Probability distributions from a 3-Use Case Model of 8.02x over all released content during the 18 week running. A course structure visual aid is below the lowermost probability distribution. Each bar is a set of resources, where color and length represents the type of resource and its weight toward final grade, respectively. Orange - lecture videos, black - lecture questions, gray - homework, green - simulations, red - exams, and blue - problem solving videos.

Based on our observations, we hypothesize that these users were simply "shopping", although many other possibilities exist. Figure 3b captures users who actively participated in the course yet disengaged midway through. Finally, the distribution in Figure 3c remains roughly uniform throughout the course, signifying significant engagement with the majority of the course material. Together these 3 use cases represent intuitive groups (shopping, disengaging, and completing) for students based on their interactions with the courseware.

These three use cases were evident from the very beginning of the course. The shopping use case remained virtually unchanged after the first two weeks of the course, while the disengaging and completing use cases slowly spread their distributions out, as new material was released. Figures for each use-case per week will be made available online. Students in the completing cohort engaged with material as soon as it was released, following the instructor's intentions. The disengaging use case had a similar, although delayed, progression to the completing use cases, where students increasingly lagged behind as new material was released. Overall the behavioral patterns captured in Figure 3 remained well-defined throughout the course, defining consistent archetypes for students.

Increasing the number of use cases breaks these archetypes into additional behavioral patterns based on the types of materials accessed and the percentage of the course utilized. Figure 4 depicts the 10-use case model trained on all 18 weeks of 8.02x. Users that failed to make it to the end of the course are represented by use cases in Figures 4a, 4b, 4c, 4d, and 4e. The shopping use case (see Figure 3a) reemerges most clearly in Figure 4a. Figure 4b potentially illuminates a shopping variant, where users are only attempting the first problem set. Figures 4c, 4d, and 4e resemble the disengaging use case from Figure 3b, highlighting potential inflection points in the course. The remaining 6 use cases embody the completing use case, as they spread their probability distributions across the course. Going from Figure 4f to Figure 4j there is a clear shift in probability from videos to assessments. Such separation indicates the degree to which students depended on videos, ranging from users that primarily audited the class to potential experts that attempted the problem sets and exams with little instruction. Therefore, we get higher granularity into the behavioral trends with the course by varying the number of use cases.

Predicting Certification Through Probabilistic Use Cases

By substituting in students' use case proportions, we effectively reduce the dimensionality of the data from thousands of resources to a small number of use cases. This allows for more accurate predictions of user outcomes. Through 5-fold cross validation, we test this hypothesis on a weekly basis in 8.02x, using certification as our outcome of choice. Table 2 presents the overall accuracy rates (ACC), true positive rates (TPR), and true negative rates (TNR) for 3, 5, 10, and 50-use case models. Despite the initial drop in TNR in comparison to the base representation of time spent in Table 1, the use case formulations yield much higher TPR, providing balanced prediction performance between certificate and non-certificate earners. Moreover, as the number of use cases in-

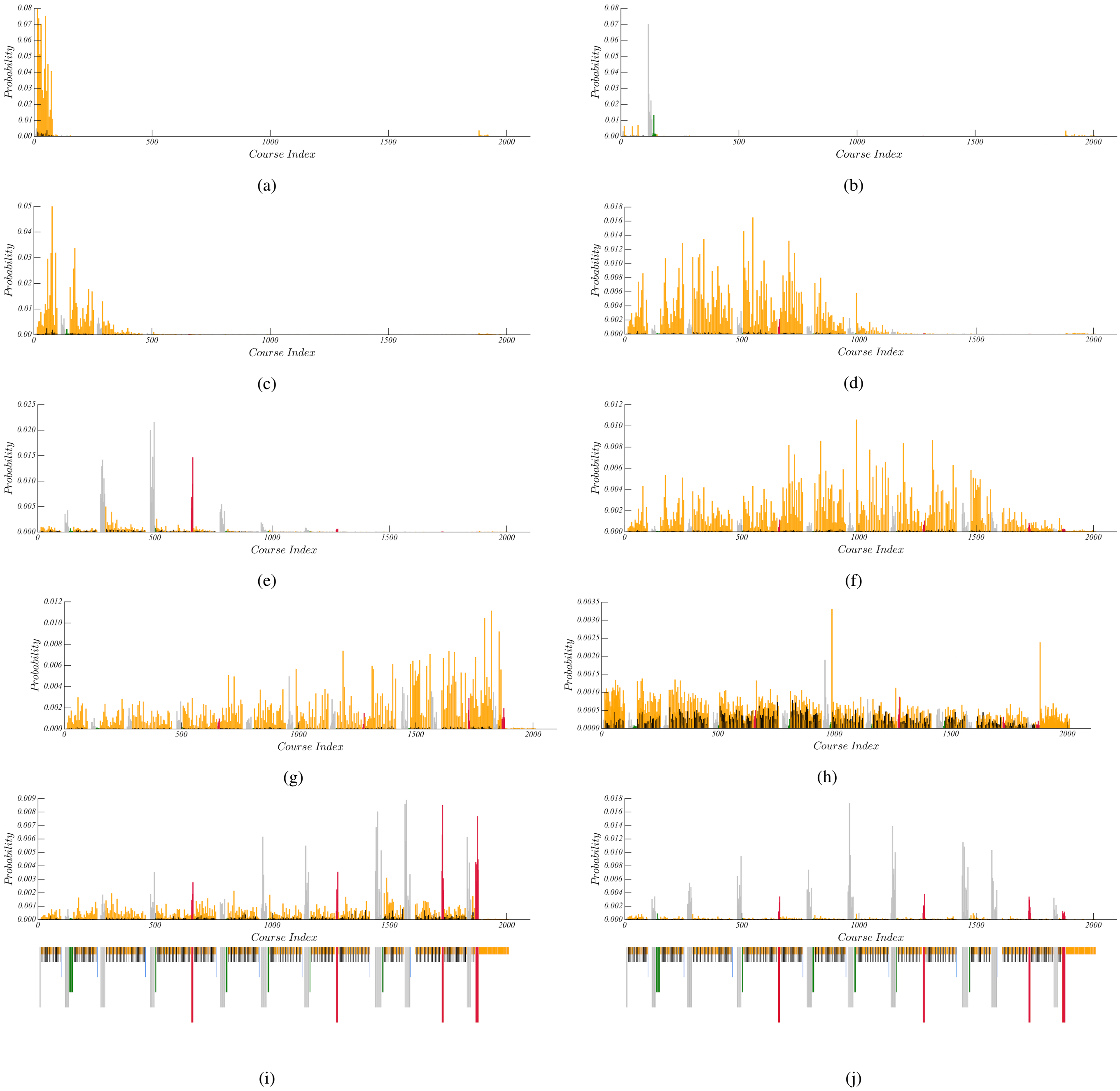


Figure 4: Probability distributions for each use case in a 10-Use Case Model trained on all 18 weeks of logs from 8.02x. In contrast to the 3-Use Case Model, the 10-Use Case model provides higher granularity into disengaged and engaged behavior trends. Figure 4i and Figure 4j contain the course structure visual aid. Each bar is a set of resources, where color and length represents the type of resource and its weight toward final grade, respectively. Orange - lecture videos, black - lecture questions, gray - homework, green - simulations, red - exams, and blue - problem solving videos.

	3-use case model			5-use case model			10-use case model			50-use case model		
Week	ACC	TNR	TPR	ACC	TNR	TPR	ACC	TNR	TPR	ACC	TNR	TPR
1	0.71±0.01	0.70	0.79	0.77±0.01	0.77	0.76	0.81±0.01	0.81	0.75	0.81±0.01	0.81	0.74
2	0.79±0.01	0.78	0.93	0.83±0.01	0.82	0.90	0.83±0.02	0.82	0.89	0.85±0.02	0.85	0.90
3	0.87±0.02	0.86	0.96	0.84±0.02	0.83	0.96	0.88±0.02	0.87	0.96	0.90±0.02	0.90	0.94
4	0.90±0.01	0.89	0.97	0.91±0.02	0.90	0.97	0.91±0.02	0.90	0.97	0.93±0.02	0.93	0.95
5	0.87±0.02	0.86	0.98	0.91±0.02	0.91	0.98	0.91±0.02	0.91	0.96	0.93±0.02	0.93	0.96
6	0.90±0.02	0.90	0.99	0.91±0.02	0.90	0.99	0.92±0.02	0.91	0.98	0.94±0.02	0.94	0.98
7	0.92±0.02	0.91	0.99	0.91±0.02	0.90	0.99	0.92±0.02	0.92	0.98	0.95±0.02	0.95	0.97
8	0.92±0.02	0.91	0.99	0.94±0.02	0.94	0.99	0.94±0.01	0.93	0.99	0.96±0.02	0.96	0.97
9	0.94±0.01	0.93	0.99	0.95±0.01	0.95	0.98	0.94±0.01	0.94	0.99	0.96±0.01	0.96	0.97
10	0.93±0.02	0.93	0.99	0.94±0.02	0.93	1.00	0.96±0.01	0.96	0.98	0.97±0.01	0.97	0.97
11	0.93±0.02	0.93	1.00	0.95±0.01	0.95	1.00	0.96±0.01	0.96	0.99	0.97±0.01	0.97	0.98
12	0.93±0.02	0.93	1.00	0.93±0.02	0.93	0.99	0.96±0.01	0.96	0.99	0.98±0.01	0.98	0.97
13	0.92±0.02	0.91	0.99	0.95±0.01	0.95	0.99	0.97±0.01	0.97	0.99	0.98±0.01	0.98	0.98
14	0.96±0.01	0.95	0.97	0.97±0.01	0.97	0.99	0.97±0.01	0.97	0.99	0.98±0.01	0.98	0.98
15	0.92±0.02	0.92	0.99	0.95±0.01	0.95	0.99	0.96±0.01	0.96	0.99	0.99±0.01	0.99	0.98
16	0.96±0.01	0.96	1.00	0.95±0.01	0.94	1.00	0.97±0.01	0.97	0.99	0.99±0.01	0.99	0.98
17	0.96±0.01	0.95	1.00	0.97±0.01	0.97	0.98	0.97±0.01	0.97	0.99	0.98±0.01	0.98	0.98
18	0.96±0.01	0.96	1.00	0.96±0.01	0.96	1.00	0.97±0.01	0.97	0.99	0.99±0.00	0.99	0.98

Table 2: Overall accuracy rates (ACC), true positive rates (TPR), and true negative rates (TNR) for 3, 5, 10, and 50-use case models at predicting certification.

creases, both the TNR and TPR increase. At the peak of 50 use cases, a SVM classifier with DEC achieves 0.81±0.01 accuracy at predicting certification with just one week of data. Even with only 3 use cases the, prediction accuracy is still at 0.71±0.01 with only one week of data.

DISCUSSION

Applying LDA to 8.02x generates probabilistic use-cases that transform massive amounts of statistical information into a set of behavioral trends that are more easily characterized and communicated. Investigating the probability distributions associated with each use case can help researchers distinguish archetypes such as auditors, completers, and even experts. The true descriptive power of LDA, nevertheless, comes from its mixed-membership model. Because students have their own proportions for each use case, important differences between users are preserved, which is critical in prediction.

Despite the preserved statistical information, the implementation of LDA in this paper involves two assumptions regarding the student data. First, LDA assumes that the order of the interactions does not matter when determining the use cases. The joint distribution in Equation 2 indicates this assumption, as permutations of interactions do not affect the overall likelihood of the model. However, the order of student interactions can encode valuable information about behavioral patterns. For example, consider playing a video in a lecture sequence and answering a follow up question. Answering the question before watching the video alludes to a very different behavior than the reverse. Rather than following the natural order of the course, a student might be trying to optimize their behavior to get through the material as quickly as possible. To relax this constraint, the work of Wallach [21] or Griffiths et al. [9] could be adapted for use case modeling.

The second assumption is that the ordering of the students does not matter. Because enrollment took place throughout the running of 8.02x, this is not entirely true. The release and due dates for course content were spread across roughly 16 weeks, meaning students ultimately had different user experiences depending on the date they enrolled. Such course features could potentially have a dramatic effect on behavior, which traditional LDA does not currently capture.

Nevertheless, the application of LDA in this paper serves as a solid proof of concept. To truly validate the effectiveness of this approach, the methods need to be applied to a broad range of courses. As next steps, we are excited to explore how factors, such as population size, course structure, or material, effect the resulting use cases.

CONCLUSIONS

Our results show that LDA can be adapted to the context of user modeling in MOOCs. The descriptive power of this approach reveals a number of latent use-cases learned from data in the MITx on edX MOOC, 8.02x: Electricity and Magnetism. These use cases have shown distinct patterns of behavior, while preserving important statistical information for additional analysis. Perhaps most important, using only the first week of logs, probabilistic use cases can predict whether or not a student will earn a certificate with 0.81±0.01 accuracy.

Beyond research, it is our hope that this may impact course content teams and platform developers. The probabilistic representation of use cases provides intuition about which course components are utilized and potentially more complex modes of student behavior. The mixed-membership representation of students offered by LDA also has the potential to facilitate similarity queries between students on the basis of their behavior. From a platform perspective, these queries could in turn serve as the basis for intervention studies of specific cohorts. LDA adapted for user modeling provides key insights into behavior via a data-driven approach that could potentially form a foundation for adaptive design in large-scale applications.

ACKNOWLEDGMENTS
We thank the MITx and the HarvardX Research Committee for guidance on this work and computational resources.

REFERENCES
1. Batuwita, R., and Palade, V. Class imbalance learning methods for support vector machines. *Imbalanced Learning: Foundations, Algorithms, and Applications* (2013), 83.
2. Blei, D. M. Probabilistic topic models. *Communications of the ACM 55*, 4 (2012), 77–84.
3. Blei, D. M., and Lafferty, J. D. Topic models. *Text mining: classification, clustering, and applications 10* (2009), 71.
4. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3* (1 2003), 993–1022.
5. Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., and Seaton, D. Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment 8* (2013), 13–25.
6. Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems* (2009), 288–296.
7. Course XML Tutorial – edX Data Documentation documentation. `http://edx.readthedocs.org/en/latest/course_data_formats/course_xml.html`.
8. gensim: Topic modelling for humans. `http://radimrehurek.com/gensim/`.
9. Griffiths, T. L., Steyvers, M., Blei, D. M., and Tenenbaum, J. B. Integrating topics and syntax. In *Advances in Neural Information Processing Systems* (12 2004), 537–544.
10. Ho, A. D., Reich, B. J. F., Nesterko, S. O., Seaton, D. T., Mullaney, T. P., Waldo, J. H., and Chuang, I. HarvardX and MITx: The first year of open online courses, fall 2012-summer 2013. *Social Science Research Network* (1 2014). `http://ssrn.com/abstract=2381263`.
11. Hoffman, M., Bach, F. R., and Blei, D. M. Online learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems* (12 2010), 856–864.
12. Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., and Miller, R. C. Understanding in-video dropouts and interaction peaks inonline lecture videos. In *Proceedings of the first ACM conference on Learning@Scale conference*, ACM (3 2014), 31–40.
13. Kizilcec, R. F., Piech, C., and Schneider, E. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge*, ACM (2013), 170–179.
14. Koller, D., Ng, A., Do, C., and Chen, Z. Retention and intention in massive open online courses: In depth. *Educause Review 48*, 3 (2013).
15. Pritchard, J. K., Stephens, M., and Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics 155*, 2 (2000), 945–959.
16. Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., and Getoor, L. Modeling learner engagement in MOOCs using probabilistic soft logic. In *NIPS Workshop on Data Driven Education* (2013).
17. Rayyan, S., Seaton, D. T., Belcher, J., Pritchard, D. E., and Chuang, I. Participation and performance in 8.02x electricity and magnetism: The first physics MOOC from MITx. *arXiv preprint arXiv:1310.3173* (2013).
18. Seaton, D., Nesterko, S., Mullaney, T., Reich, J., and Ho, A. Characterizing video use in the catalogue of MITx MOOCs. *European MOOC Stakeholders Summit, Lausanne* (2014), 140–146.
19. Seaton, D. T., Reich, J., Nesterko, S. O., Mullaney, T., Waldo, J., Ho, A. D., and Chuang, I. 8.02x electricity and magnetism – spring 2013 MITx course report (MITx working paper# 10). *Social Science Rsearch Network* (1 2014). `http://ssrn.com/abstract=2382328`.
20. Veropoulos, K., Campbell, C., Cristianini, N., et al. Controlling the sensitivity of support vector machines. In *Proceedings of the international joint conference on artificial intelligence*, vol. 1999 (1999), 55–60.
21. Wallach, H. M. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, ACM (2006), 977–984.
22. Xu, G., Zhang, Y., and Yi, X. Modelling user behaviour for web recommendation using LDA model. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, vol. 3, IEEE (2008), 529–532.
23. Yang, D., Sinha, T., Adamson, D., and Rosé, C. P. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-Driven Education Workshop* (2013).