Identifying and Characterizing Subpopulations in Massive Open Online Courses

by

Cody A. Coleman

B.S. Electrical Engineering and Computer Science, Massachusetts Institute of Technology (2013)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2015

Copyright © 2015 Cody A. Coleman, All rights reserved.
The author hereby grants to M.I.T. permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole and in part in any medium now known or hereafter created.

Author .................................................................
Department of Electrical Engineering and Computer Science
January 30, 2015

Certified by ..............................................................
Isaac Chuang
Professor of Electrical Engineering and Computer Science; Professor of Physics; Senior Associate Dean of Digital Learning
Thesis Supervisor

Accepted by .............................................................
Professor Albert R. Meyer
Chairman, Master of Engineering Thesis Committee
Identifying and Characterizing Subpopulations in Massive
Open Online Courses

by

Cody A. Coleman

Submitted to the Department of Electrical Engineering and Computer Science
on January 30, 2015, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

The large and diverse student populations in Massive Open Online Courses (MOOCs) present an unprecedented opportunity to understand student behavior and learn about learning. A tremendous amount of information on students is collected by logging their behaviors. However, despite this wealth of data, little has been done to identify important subpopulations and understand their strengths and weaknesses. This thesis focuses on the potential of various learner subpopulations to succeed and contribute to the course.

First, I investigate teacher enrollment in 11 MITx MOOCs showing that teachers represent a potentially large and untapped resource. Depending on their expertise, teachers could provide additional instruction or guidance to struggling students or a way to extend the reach of MOOCs into traditional classrooms. They could also provide MOOCs with another source of revenue through accreditation opportunities.

Second, inspired by the phenomenon widely known as the “spacing effect,” I look at how students choose to spend their time in 20 HarvardX MOOCs in order to identify observational evidence for the benefits of spaced practice in educational settings. While controlling for the effect of total time on-site, it is shown that the number of sessions students initiate is an important predictor of certification rate, particularly for students who only spend a few hours in a course.

Finally, by adapting Latent Dirichlet Allocation, I discover probabilistic use cases that capture the most salient behavioral trends in a course. Not only do these use cases provide insights into student behavior, they also serve as an effective method of dimensionality reduction for additional analysis and prediction. Together, the studies in this thesis represent a step forward in digital learning that illuminates subpopulations that are important to the future success of MOOCs.

Thesis Supervisor: Isaac Chuang
Title: Professor of Electrical Engineering and Computer Science; Professor of Physics; Senior Associate Dean of Digital Learning
Acknowledgments

This thesis is the culmination of the determination and passion of many people. The support I received from people inside and outside of MIT was incredible. Without their help, I would not have been able to craft the following work.

My research advisor, Professor Isaac Chuang, is one of the most “explosively productive” people I know. He inspired me to push myself and revive the explorer I had been as a kid. He gave me the opportunity to ponder what I wanted to do and how I was going to do it. This freedom was a mind-altering shift from the rigid requirements and relentless work from my undergraduate degree. Thanks to Isaac, I am now free to embrace my curiosity and have optimism for what is possible.

However, my MIT experience would not have been complete without a number of other critical people and organizations. Ken Zolot, senior lecturer in Electrical Engineering and Computer Science, demystified entrepreneurship, making me realize that my pipe dream of starting a company was not impossible. Anne Hunter and the Undergraduate Office for the Electrical Engineering and Computer Science department created a home away from home, where I always felt welcomed. Finally, Professor Sanjay Sarma started my journey in the Office of Digital Learning by offering a research assistantship position for my Master of Engineering degree. Working in the Office of Digital Learning has been an amazing opportunity, where I met many amazing people and got to work closely with HarvardX. Together, these experiences and more have made my time at MIT unforgettable, and I would like to thank everyone that made these experiences possible.

Outside of MIT, I have enjoyed a wealth of support and encouragement. My teachers at Winslow Township High School did whatever they could to support and help me. In particular, Norm Ingram, Nancy Young, and Chantel Smith have been like family to me. Chantel even welcomed me into her home whenever I needed a place to stay. She was the closest thing I had to a loving and caring mother. I have been extremely fortunate to meet such kindhearted people. Thank you for all of your support!
## Contents

Cover page 1

Abstract 2

Acknowledgments 3

Contents 5

List of Figures 9

List of Tables 17

1 Introduction 19

1.1 The Massive Open Online Course .......................... 19

1.2 The problem: One size doesn’t fit all ....................... 20

1.3 Contributions of this thesis .............................. 21

2 Teacher Enrollment in MIT MOOCs: Are We Educating Educators? 23

2.1 Introduction ............................................. 24

2.1.1 MOOCs past and present .............................. 24

2.1.2 Enrollment Surprises in MITx courses .................. 25

2.2 Survey Methodology and Forum Analysis .................... 26

2.2.1 Entrance Survey: Teachers “are” enrolling in MITx open online courses ......................... 27
3 Beyond Time-on-Task: The Relationship Between Spaced Study and Certification in MOOCs

3.1 Introduction ................................................. 38
   3.1.1 Early MOOC Research on Participation and Time-on-Task . 38
   3.1.2 What is the Spacing Effect? ............................. 39
   3.1.3 From Experiments to Education ............................ 39

3.2 Data .............................................................. 41
   3.2.1 Data Collection .......................................... 41
   3.2.2 Measures ................................................ 43

3.3 Results .......................................................... 44
   3.3.1 Exploratory Analysis of Key Features in MOOC Data .... 44
   3.3.2 Identifying a Benefit of Spacing via Student-to-Student Comparisons ................................................. 48
   3.3.3 Identifying a Benefit of Spacing via Within-Student Comparisons 51

3.4 Discussion ...................................................... 54
   3.4.1 Differences Between the Current Results and Spacing Effects
       Studied in Psychology .......................................... 55
   3.4.2 Possibilities of Other Confounding Factors that Drive Certification ......................................................... 56
   3.4.3 Interventions for Applying Spaced Practice to MOOCs .... 57
4 Probabilistic Use Cases

4.1 Introduction ................................................. 60
4.2 Related Work ................................................ 62
4.3 Course Studied and Dataset ............................... 63
4.4 Methods ...................................................... 64
  4.4.1 LDA for Traditional Topic Modeling .................. 66
  4.4.2 LDA for Probabilistic Use Cases ..................... 68
  4.4.3 Interaction Representations ............................ 69
  4.4.4 Evaluating Interaction Representations .............. 69
  4.4.5 Evaluating Probabilistic Use Cases .................. 70
4.5 Results ...................................................... 71
  4.5.1 Selecting an Interaction Representation ............ 71
  4.5.2 Identifying the Number of Use Cases Through Log Perplexity 73
  4.5.3 Probabilistic Use Cases ............................... 74
  4.5.4 Predicting Certification ............................... 76
4.6 Discussion .................................................. 83
4.7 Conclusions ................................................ 84

5 Contributions ................................................. 85

A Supplementary Materials for Teacher Enrollment .............. 87
  A.1 Highlights ................................................. 87
  A.2 Discussion Forums ....................................... 97

B 50-Use Model .................................................. 99
  B.1 Probability Distributions .................................. 99

Bibliography .................................................... 105
List of Figures

2-1 Percentage of self-identifying teachers, whether they have or currently
 teach the course topic, and the percentage of currently practicing teach-
 ers. Percentages are relative to the number of survey respondents. . . 27

2-2 Instructional context of self-identifying teachers. Survey question al-
 lowed for multiple responses. All percentages are relative to total num-
 ber of responses within each course. . . . . . . . . . . . . . . . . . . . . 29

2-3 (Left) Average percentages of enrollees that were teacher respondents,
 non-teacher respondents, and non-respondents across MITx courses on
edX in the Spring of 2014, and (right) average percentage of comments
 across courses that the aforementioned groups made. Although survey
 respondents identifying as teachers represent only 4.5% of the overall
 population (left), they generate 22.4% of all comments (right). Cate-
gories are based on the entrance survey. . . . . . . . . . . . . . . . . . 32
3-1 Illustrations of session and time on-site. (3-1a) Schematic of the definition of session and time on-site: A session was defined as a collection of click events separated by periods of inactivity that lasted more than 30 minutes, in accordance with the Google Analytics standard for defining sessions of website usage\(^1\). Total time spent on-site was calculated as the sum of the lengths of all sessions by a student. (3-1b) Example sessions taken from the data, which demonstrate some variety and combinations of event types initiated by students. The top example demonstrates a session focused on video usage, interleaved with page navigation events. The bottom example demonstrates a session that focuses on quiz problems, with occasional page navigation. Note that the set of event types shown in these examples (9 event types) does not include all possible event types. (A comprehensive list and description of event types can be found at the edX research guide\(^2\)).

3-2 Exploratory analysis of time on-site, spacing, and certification rates in MOOCs. (3-2a) Histograms of students’ hours on-site levels are shown, with each colored line representing students from a different course, for a total of 20 colors representing 20 courses, as detailed in the legend to the right. Note the log-scale x-axis. (3-2b) Histograms of students’ spacing levels are shown for each of the 20 courses. Note that the spikes at the low spacing levels are due to the fact that the number of sessions can only take integer values.

---

\(^1\)https://support.google.com/analytics/answer/2731565?hl=en
\(^2\)http://edx.readthedocs.org/projects/devdata/en/latest/internal_data_formats/event_list.html
Exploratory analysis of time on-site, spacing, and certification rates in MOOCs. (3-3a) Certification rates are plotted as a function of session count (10 levels). Students from each course were divided equally into deciles of session counts, and the average certification rates for each decile is plotted as a function of session count for that decile, to form the colored lines, one color for each course. The mean certification rates and session counts were then averaged across all 20 courses for each decile to obtain the bold black line, representing the mean relationship between certification rate and session count across courses. (3-3b) Certification rates are plotted as a function of time on-site, after dividing students into deciles of time on-site in an analogous fashion to session count in (3-3a). (3-3c) Time on-site is plotted as a function of session count after dividing students into deciles of session counts, as in (3-3a).
Across-student analysis of the spacing effect in MOOCs.  

(3-4a) Certification rates are plotted as a function of time on-site (10 levels) and spacing (3 levels) to form $10 \times 3 = 30$ total groups of students taken from the 20 HarvardX courses in the analysis. First, students within each course were divided equally into deciles of time on-site, and then students of each decile were divided further into terciles of low, mid, and high spacing groups (light gray, gray, and dark gray dots). The certification rates and time on-site levels of the 30 groups were then averaged across the 20 courses to form what is shown here.  

(3-4b) The top bar graphs in light red show the average time on-site, session count, and certification rate aggregated across high levels of time on-site corresponding to the light red region in (3-4a), for low (light gray), mid (gray) and high (dark gray) spacing groups. The bottom bar graphs in light blue are analogous plots representing the time on-site, session count, and certification rates aggregated across the low levels of time on-site corresponding to the light blue region of (3-4a).  

Note that the lowest (leftmost) and highest (rightmost) deciles of time on-site were not included in the calculation of these bar graphs for a fairer comparison since the levels of time on-site among these spacing groups were not well matched; however, the trend remains within these excluded spacing groups where higher certification is generally associated with higher spacing. Error bars indicate standard errors of the mean, across the 20 courses. The large error bars reflect the large variability across courses noted in Figure 3-2, but the statistical comparisons are highly significant because comparisons are evaluated in a paired fashion within-course. **: $p < 0.01$, ***: $p < 0.001$ . . . . . . . . 50
3-5 Within-student analysis of the spacing effect in MOOCs. (3-5a) Session counts are shown for 5 different variations of changes from one course to the other within a course-pair. Dark gray represents students who greatly increased their session count from low to high terciles between courses. Terciles were computed within each course, to account for course-to-course differences in session counts. Light gray represents those who mildly increased their session count from low to mid, or mid to high terciles between the pair of courses. Red and dark orange represent students who greatly and mildly decreased their session count, respectively, in an analogous manner. Light orange represents those who remained within the same tercile for both courses. Students session counts were averaged within course for each course-pair, and furthermore these course-pair session counts were then averaged across all course-pairs, to form what is shown here. (3-5b) The mean time on-site levels are shown corresponding to the 5 different variations of session count changes in (3-5a). By construction, these levels of time on-site were largely maintained between the courses within each course-pair, in contrast to the session counts of (3-5a). Only students who maintained their time on-site within the same decile for both courses within each course-pair were selected for this analysis, in order to control for the effect of time on-site. (3-5c) Certification rates are shown for the 5 different variations of session count changes corresponding to the colors of (3-5a) and (3-5b). Error bars indicate the standard error of the mean computed across all course-pairs. ***: p < 0.001 . . . . . . 53

4-1 Course structure visualization. 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. . . . . . . . . . . . . . . . . . . . . . 65
4-2 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, and tutorial sequences.

4-3 The graphical model for Latent Dirichlet Allocation [BNJ03] when applied to topic modeling. Each node represents a random variable. The hidden variables for the topic distributions ($\theta^t$), topic assignments ($z_i$), and topics ($\beta_z$) comprise the unshaded nodes, while the observed words ($w_i$) are shaded. The plates represent replicates for the $n_t$ words, $T$ documents, and $K$ topics.

4-4 The graphical model for Latent Dirichlet Allocation when applied to use case modeling. Every node represents a random variable. The hidden variables for the use case distributions ($\phi^t$), use case assignments ($u_i$), and use cases ($\beta_u$) are represented by the unshaded nodes, while observed course interactions ($m_i$) are shaded. The plates represent replicates for the $n_t$ interactions, $T$ students, and $K$ use cases.

4-5 5-fold log perplexity for a varying number of use cases.

4-6 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 – similar to that found in Figure 4-1 – is below the lowermost probability distribution. All probability distributions are coded according to this visual aid: color – the type of resource, width – density of resources, and height – weight toward final grade (height of bars).

4-7 The evolution of the shopping use case from the 3-Use Case model (Figure 4-6a) over the 18 weeks of 8.02x. Figure 4-7g contains the course structure visual aid.

4-8 The evolution of the disengaging use case from the 3-Use Case model (Figure 4-6b) over the 18 weeks of 8.02x. Figure 4-8g contains the course structure visual aid.
4-9 The evolution of the completing use case from the 3-Use Case model (Figure 4-6c) over the 18 weeks of 8.02x. Figure 4-9g contains the course structure visual aid.

4-10 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 4-10i and Figure 4-10j 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.

A-1 Percentage of self-identifying teachers, whether they have or currently teach the course topic, and the percentage of currently practicing teachers. Percentages are relative to the number of survey respondents (see Table A.4).

A-2 Instructional context of self-identifying teachers. Survey question allowed for multiple responses (see Question 3 in Table A.1). All percentages are relative to total number of responses within each course.

A-3 Current teachers' responses to the exit survey questions from Table A.2.

A-4 The average number of posts (A-4a), comments (A-4b), and forum-click interactions (A-4c) for teachers (orange) versus non-teacher respondents (navy). Activity is significantly higher in 4 out of 11 courses in each category. The distributions of forum activity are highly skewed toward zero, often with a few outliers leading to significant tails toward high activity.
B-1 Probability distributions for use cases 1 through 10 in a 50-Use Case Model trained on all 18 weeks of logs from 8.02x. In contrast to the 3-Use Case Model, the 50-Use Case model provides higher granularity into disengaged and engaged behavior trends. Figure B-1i and Figure B-1j 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.

B-2 Probability distributions for use cases 11 through 20 in a 50-Use Case Model trained on all 18 weeks of logs from 8.02x. In contrast to the 3-Use Case Model, the 50-Use Case model provides higher granularity into disengaged and engaged behavior trends. Figure B-2i and Figure B-2j contain the course structure visual aid.

B-3 Probability distributions for use cases 21 through 30 in a 50-Use Case Model trained on all 18 weeks of logs from 8.02x. In contrast to the 3-Use Case Model, the 50-Use Case model provides higher granularity into disengaged and engaged behavior trends. Figure B-3i and Figure B-3j contain the course structure visual aid.

B-4 Probability distributions for use cases 31 through 40 in a 50-Use Case Model trained on all 18 weeks of logs from 8.02x. In contrast to the 3-Use Case Model, the 50-Use Case model provides higher granularity into disengaged and engaged behavior trends. Figure B-4i and Figure B-4j contain the course structure visual aid.

B-5 Probability distributions for use cases 41 through 50 in a 50-Use Case Model trained on all 18 weeks of logs from 8.02x. In contrast to the 3-Use Case Model, the 50-Use Case model provides higher granularity into disengaged and engaged behavior trends. Figure B-5i and Figure B-5j contain the course structure visual aid.
List of Tables

2.1 Surveys of teacher enrollment were distributed in 11 MITx courses on edX in the Spring of 2014. The questions identified three categories: past or present teachers, current teachers, and teachers with experience teaching the courses topic. These categories are not mutually exclusive. Totals and average percentages across courses are provided relative to survey respondents. ................................. 28

3.1 Course Information for Twenty 2012-2014 HarvardX Courses ........ 42

4.1 True positive rates (TPR) and true negative rates (TNR) for identifying certificate earners with different interaction representations. .... 72

4.2 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. ................................................. 82

A.1 Entrance survey teacher questions. ................................. 89

A.2 Cross course average percentages of current teachers' responses to exit survey questions. ..................................................... 90

A.3 The 11 MITx courses used in studying teacher enrollment. ........ 92
A.4 Teacher enrollment in 11 MITx Courses. Students who answered Question 1 in Table A.1 were included in “Surveyed.” Students who answered “Yes” to Question 1 were included in “Teachers.” Students who answered “Yes” to Question 4 were included in “Current Teachers”. Students who answered “Yes” to Question 2 were included in “Teach / Taught Topic”.
Chapter 1

Introduction

1.1 The Massive Open Online Course

Massive Open Online Courses (MOOCs) have advanced digital learning by leveraging the power of the Internet to expand the reach of educators to students from around the world. They allow professors from universities, including MIT, Harvard, and Stanford, to develop and deploy courses on topics ranging from Ancient Greek Heroes to Electricity and Magnetism. These courses adopt digital technologies to enhance the learning experience. Lectures are captured on video, so students can play, pause, and even speed up the recordings. This allows them to consume material at their own pace. Through automatic and peer grading, students can submit assignments asynchronously and receive almost immediate feedback. Combined with course discussion forums, these raw components and modules can provide an engaging experience for students free of charge. In the first year alone, HarvardX and MITx courses drew over 600,000 students [HRN+14]. Even a single course can attract a massive audience. For example, 8.02x, an introductory physics MOOC on Electricity and Magnetism, drew over 41,307 registrants from over 150 countries [SRN+14].

Such large numbers, combined with diverse backgrounds and enrollment motivations, presents an unprecedented opportunity to understand student behavior and
learn about learning. MOOC platforms, such as edX\(^1\) and Coursera\(^2\), capture click-stream data – stored records of user interactions with course content. This includes playing or pausing a video, submitting a problem, and posting on the discussion forums are logged. Together, these interactions form what is almost equivalent to a body language for users of MOOCs. Every action can be transformed into a piece of a student’s story: X did Y on module Z at time T. Aggregated over a course, these actions form a story containing a great deal of insight that can used to inform course development and provide more learning experiences adapted to individuals.

1.2 The problem: One size doesn’t fit all

Despite such a wealth of data, MOOCs have yet to provide an experience that matches the diversity of their students. Currently, MOOCs have largely presented material in a one-size-fits-all manner. No matter the student’s age, level of education, or country of origin the same material is presented in a prescribed structure devised by the course team. This in turn burdens both instructors and students. Instructors struggle to grasp who their students are and what motivates them. Dealing with such uncertainty becomes a daunting task as the course team must balance the accessibility of the class to a wide audience with the academic rigor of the material. Consequently, students are left to define their experience by filling any knowledge gaps and identifying the key components they care about. For example, a Ph.D. student taking an introductory physics class and a high school student taking the same class are given the same level of guidance and are evaluated on the same scale. This may embody the idea of meritocracy, but it fails to realize the true potential of digital learning to democratize education and close the achievement gap. Instead, it threatens to restrict success to the more educated and well versed [CSA\(^+\)13], and misses an opportunity to foster a robust and sustainable network of learners.

\(^1\)edX “offers free online courses and classes from the world’s best universities including MIT, Harvard, Berkeley, UT and others.”
\(^2\)Coursera “provides universal access to the world’s best education, partnering with top universities and organizations to offer courses for anyone to take, for free.”
This lack of adaptation and personalization might have led to the low retention rates that have plagued the advancement of MOOCs and been the subject of many studies \cite{KNDC13, RGH13, YSAR13, WJ09, KPS13}. Even though tens of thousands of people sign up for a course, only roughly 5% of registrants actually earn a certificate \cite{KNDC13}. Although retention is only part of the story, this discrepancy between the number of registrants and those that earn a certificate hints at an opportunity to more effectively identify and engage students. Rather than creating courses that toss students into the wild, where only the strong will survive, MOOCs could engage far more people by partitioning the massive population they serve. By identifying important subpopulations and understanding their strengths and weaknesses, MOOCs could construct a virtual ecosystem of learners, educators, and professionals. Different subpopulations could be targeted for specialized roles that would not only benefit the individuals but the larger community. For example, content experts could help novices master the material. Such relationships could make a course more engaging and accessible. If MOOCs are to truly democratize education and serve a varied population of learners, we must continue to identify and understand the most important subpopulations, so we can provide a more tailored experience that goes beyond one-size-fits-all.

\section*{1.3 Contributions of this thesis}

This thesis aims to break down the large populations of learners in MOOCs by (1) identifying important subpopulations and (2) characterizing their behaviors, in order to uncover valuable use cases for MOOCs to target. In the following chapters, we present an analysis of data from several MOOCs across MITx\textsuperscript{3} and HarvardX\textsuperscript{4}.

Chapter 2 investigates teacher enrollment in 11 MITx MOOCs showing that teachers represent a potentially large and untapped resource. Depending on their expertise, teachers could provide additional instruction or guidance to struggling students or a

\textsuperscript{3}MIT’s institution for creating MOOCs

\textsuperscript{4}Harvard University’s institution for creating MOOCs
way to extend the reach of MOOCs into traditional classrooms. They could also provide MOOCs with another source of revenue through accreditation opportunities.

Chapter 3 explores how distributed engagement within a course affects performance and certification rates. Inspired by the phenomenon widely known as the “spacing effect,” this chapter looks at how students choose to spend their time in 20 HarvardX MOOCs, in order to identify observational evidence for the benefits of spaced practice in educational settings. While controlling for the effect of total time on-site, it is shown that the number of sessions students initiate is an important predictor of certification rate, across students in all courses. In particular, for the subpopulation of students who only spend a few hours in a course, spaced practice can dramatically improve their performance.

Chapter 4 defines a machine learning approach for identifying and characterizing subpopulations in MOOCs. By adapting Latent Dirichlet Allocation (LDA) from natural language processing to model user behavior in MOOCs, we discover probabilistic use cases that capture the most salient behavioral trends in a course. Not only do these use cases provide insights into student behavior, they also serve as an effective method of dimensionality reduction for additional analysis and prediction.

Each of these chapters resulted from a collaborative effort between HarvardX and MITx. As part of this collaboration, I concentrated on characterizing student behavior. In the investigation of teacher enrollment, I focused on understanding the impact of teachers in forum discussions, while the other contributors handled conducting and analyzing the surveys. In order to explore the potential impact of the “spacing effect” in MOOCs, I extracted session information from the raw clickstream data and provided feedback throughout the study. Finally, as a step towards a more automated approach to characterizing student behavior, I led the initiative to adapt LDA to model student behavior. I performed the majority of the programming, data analysis, visualization, and writing in that work. Together, the studies in this thesis represent a step forward in digital learning that strives to illuminate subpopulations that are important to the future success of MOOCs.
Chapter 2

Teacher Enrollment in MIT MOOCs: Are We Educating Educators?

Participants in Massive Open Online Courses (MOOCs) come from an incredibly diverse set of backgrounds and act with a wide range of intentions [CSA+13, HRN+14]. Recent surveys of 11 MITx courses on edX in the spring of 2014 show that teachers (versus traditional college students) are a significant fraction of MITx MOOC participants. This suggests many ways to improve and harness MOOCs, including the potential arising from the collective professional experience of participants, opportunities for facilitating educator networks, MOOCs as a venue for expert-novice interactions, and possible added value from enhancing teacher experience through accreditation models and enabling individual teacher re-use of MOOC content. In this chapter, we present data in detail from these teacher enrollment surveys, illuminate teacher participation in discussion forums, and draw lessons for improving the utility of MOOCs for teachers.

The work done in this chapter was part of a collaborative effort with Daniel T.
Seation\textsuperscript{1}, Jon P. Daries\textsuperscript{2}, and Isaac Chuang\textsuperscript{3}. I supported Daniel T. Seaton throughout this study by analyzing the forum activity of teachers, visualizing the results, and helping craft the studies narrative. The original paper, “Teacher Enrollment in MITx MOOCs: Are We Educating Educators?”, can be found on the Social Science Research Network [SCDC14].

\section{Introduction}

\subsection{MOOCs past and present}

One of the earliest precursors to modern MOOCs was intended for high school teachers in the United States. In 1958, a post-war interpretation of introductory physics called “Atomic-Age Physics” debuted at 6:30AM on the National Broadcasting Company’s (NBC) “Continental Classroom”. Daily viewership was estimated at roughly 250,000 people [Ran60, Lac59, Kel62, Car74] and over 300 institutions partnered to offer varying levels of accreditation for the course. Roughly 5,000 participants were certified in the first year [GM66, Car74] and teachers were estimated to be 1 in 8 of all certificate earners [Kel62], indicating reach beyond the target demographic of high school teachers. Through its expansion of courses between 1958 and 1963, the Continental Classroom represented a bold approach in utilizing technology to address national needs in education reform. In contrast, the current MOOC era has largely focused on student-centric issues like democratizing access [Aga13] and reducing costs in higher education.

Nevertheless, a few scholars have reaffirmed the tremendous potential for engaging teachers within the current MOOC movement. Douglas Fisher highlights the sense of community emerging from the “nascent and exploding online education movement” [Fis12], while summarizing his perspective on using MOOC resources

\footnote{\textsuperscript{1}Institutional Research Group, Office of the Provost, Massachusetts Institute of Technology, Cambridge, MA 02139} \footnote{\textsuperscript{2}Institutional Research Group, Office of the Provost, Massachusetts Institute of Technology, Cambridge, MA 02139} \footnote{\textsuperscript{3}Office of Digital Learning, Massachusetts Institute of Technology, Cambridge, MA 02139}
from another university in his classes. Others have begun to recognize the substantial role that MOOCs could play in reforming teacher professional development [KWF13, JÖS14]. More pragmatically, Samuel Joseph has described the importance of community teaching assistants (TAs) within edX, along with his own efforts to support and organize contributions of over 250 volunteer TAs in a single MOOC [Jos13].

MOOCs are generating unprecedented dialogue about the current state and future of digital education [Pap12], and it is clear that interest goes beyond typical scholars studying education. One may even argue that MOOCs are an ideal hub for hosting enthusiastic educators to discuss, refine, and share pedagogy. Paramount to such an ideal is a simple question: are a substantial number of educators already enrolling?

2.1.2 Enrollment Surprises in MITx courses

At MITx (MIT’s MOOC organization), early MOOC experiments indicated that teachers are indeed enrolling. The MITx course known as 8.MReV: Mechanics Review, which derived from an introductory on-campus physics course at MIT, was initially advertised as a challenging course for high school students. Upon completion, however, course staff recognized that high school teachers were an active contingent [FRT+13]. In response, the 8.MReV team went so far as to partner with the American Association of Physics Teachers to offer Continuing Education Units (CUEs) in subsequent offerings.

A widely covered anecdotal example of teacher enrollment involved the inaugural MITx course, 6.002x: Circuits and Electronics. An MIT alumnus teaching electrical engineering to high school students in Mongolia enrolled in 6.002x alongside his students. He used the online content to flip his classroom, asking his students to complete all assignments related to the 16-week course. This experiment even led to one exceptional student from this class being admitted to MIT in 2013 [Pap13]. Such an example raises questions regarding how many other teachers, not just alumni, may be practicing similar strategies without the knowledge of MOOC providers.

The examples above represent only fractions of courses and participants, but pro-
vide signals that an important demographic may be hidden to MOOC providers and course developers. If a substantial number of teachers are indeed enrolling in MOOCs, the educational possibilities are considerable: expert-novice pairings in courses, networking educators around pedagogy or reusable content, and generally tailoring courses to satisfy the needs of teachers. In response to these initial, anecdotal findings, we used a systematic survey protocol to address specific questions related to teacher enrollment, their backgrounds, and their desire for accreditation and access to materials to use in their own courses.

2.2 Survey Methodology and Forum Analysis

In the spring of 2014, entrance and exit surveys addressing the motivations and backgrounds of participants were given in 11 MITx courses. Although these surveys addressed multiple issues concerning participants, teacher enrollment was a significant concern with questions addressing the following broad themes:

1. Are a significant number of teachers enrolling in MITx open online courses?

2. If so, do these teachers come from traditional instructional backgrounds?

3. Do teachers completing a course desire accreditation opportunities and broader usability of MITx resources?

Over 33,000 participants responded to the entrance surveys, which focused largely on themes 1 and 2. The exit surveys had over 7,000 respondents, with questions addressing theme 3 for current teachers. The survey questions used to address these broad themes can be found the supplementary material in Appendix A.

In addition to survey data, a tremendous amount of participant interaction data is available. Discussion forums entries and click-stream data allow one to check if participating teachers are actively pursuing an important aspect of their profession, namely, instructing other participants within a course. Using the responses to the entrance surveys, we can compare the behavior of teachers versus non-teachers.
2.2. Survey Methodology and Forum Analysis

2.2.1 Entrance Survey: Teachers “are” enrolling in MITx open online courses

Recent entrance surveys of over 33,000 participants in 11 MITx courses from the Spring of 2014 reveal a substantial number of enrolling teachers (Table 2.1). Cross-course averages indicate 28.1% of respondents identify as past or present teachers, while 8.8% identify as current teachers, and 5.9% have taught or currently teach the course subject. Note, these percentages are even more striking when enumerated. Across all survey respondents, there are 9,628 self-identifying teachers, 2,901 practicing teachers, and 1,909 participants that have or currently teach the topic. The average survey response rate was 16.9%, meaning that if respondents were a random sample of registrants, the actual numbers of teachers would be approximately 6 times larger. Although teachers are likely to respond at greater rates, we argue that the baseline numbers are themselves numerically significant.

![Figure 2-1: Percentage of self-identifying teachers, whether they have or currently teach the course topic, and the percentage of currently practicing teachers. Percentages are relative to the number of survey respondents.](image)

The entrance surveys also included questions contextualizing respondents’ instructional backgrounds. Figure 2-2 contains distributions of response options, where 73.0% of responses indicate what we consider traditional teaching backgrounds: Primary/Secondary School, College/University, or Support Staff, e.g., Teaching Assis-
Table 2.1: Surveys of teacher enrollment were distributed in 11 MITx courses on edX in the Spring of 2014. The questions identified three categories: past or present teachers, current teachers, and teachers with experience teaching the courses topic. These categories are not mutually exclusive. Totals and average percentages across courses are provided relative to survey respondents.

A notable exception is Entrepreneurship (15.390x), which has only 60.2% of respondents with traditional teaching backgrounds (also see supplementary material in Appendix A). The course topic may offer a signal that explains the variation of teacher enrollment across courses. One can make alternative arguments that the large enrollment in Global Warming (12.340x) and Street-Fighting Math (6.SFMx) may be partially explained by interest of teachers in seeing materials and alternative views on instruction.

Taking a closer look at the breakdown of instructional backgrounds (see Figure 2-2), “College or University” teachers account for the largest populations, but may range from faculty to teaching assistants. The size of the “College or University” population is likely due to all surveyed courses being college level. K-12 teachers make up 25.0% of instructional context found in Figure 2-2 (found by summing Pri-
mary School and Secondary School categories). The categories “Outside the scope of traditional schools” and “Other” indicate respondents that consider themselves teachers outside the available choices within the survey (see supplementary material in Appendix A). Free response submissions were allowed within the “Other” category, and a number of respondents identified as tutors or working in corporate training. Each category provides a number of hypotheses for “who” enrolls in these courses, but future must dig deeper into “why”.

**Figure 2-2:** Instructional context of self-identifying teachers. Survey question allowed for multiple responses. All percentages are relative to total number of responses within each course.
2.2.2 Exit Survey: Course completing teachers desire accreditation and use of resources

Within the last two weeks of each course, questions related to teachers were added to exit surveys assessing the course impact on teachers, and whether course developers were missing opportunities to provide professional development and accreditation for participating teachers. Out of 7,149 survey respondents, 1,002 (15.6%) again identified as current teachers and answered questions pertaining to accreditation, course influence, and use of MITx MOOC resources. Across all 11 MITx courses on edX, an average of 53.9% (560) of current teachers answered “yes” to having interest in accreditation opportunities, while 15.1% answered “no” and 26.4% answered “unsure” (see supplementary material in Appendix A). Greater than 70% of responding teachers slightly-agree or strongly-agree they will use MITx material in their current teaching, and greater than 70% would be interested in using material from other courses. Even so, many MOOC providers have yet to adopt Open Educational Resource models [Par13] aimed at facilitating content sharing.

2.2.3 Discussion Forum Analysis: Teachers are actively discussing

Each MITx MOOC employed a threaded discussion forum to support enrollee interactions. Any participant can contribute written content, and it is of central interest whether teachers are actively engaging with other participants. Forum data collected by the edX platform allow us to monitor such activity in the form of three allowed interactions: “posts” initialize a discussion thread, “comments” are replies within those threads, and “upvotes” allow users to rate a post or comment. Comments are often generated as a response to help seeking, representing an interaction that a teacher would be well suited to undertake. We focus on comments below – noting that similar analyses applied to posts were found to be nearly identical.

One way of framing teacher behavior in the forums is simply counting their total number of textual contributions relative to all participants, i.e., what is the likelihood
of a participant receiving a comment from a teacher. For all participants – teacher respondents, non-teacher respondents, and non-respondents – a total of 57,621 comments were generated across all 11 MITx courses. Figure 2-3 highlights the percentage of comments relative to population sizes. Despite representing only 4.5% of the total population, teacher respondents generated 22.4% of all comments. Non-teacher respondents generated 33.8% of total comments, but were twice as large a population (10.3%). All other participants (non-respondent) generated 43.8% of total comments, but made up 85.2% of the population.

More notably, 1 in 5 comments were written by survey responding teachers, 1 in 12 were by a current teacher, and 1 in 16 were by teachers who teach or have taught the subject. It must be noted that MITx course staff did not advertise to attract teachers, nor did they make special considerations for their contributions to a course.

In addition, we have also analyzed distributions of posts, comments, and overall discussion activity to search for statistically significant trends between teacher and non-teacher respondents. The mean number of discussion comments by teacher respondents is statistically higher than non-teacher respondents in 4 out of 11 courses (see supplementary material in Appendix A), but not statistically distinguishable in the others. We note that distributions of forum activity are often highly skewed, typically due to only a few users contributing the majority of text [HDG+14]. For example, course staff members designate community-teaching assistants to moderate discussion forums and generate a tremendous amount of textual contributions [Jos13]. Of the 15 participants assigned this duty within the 11 MITx courses studied here, 10 took the entrance survey, 5 of which identified as teachers.

Finally, we note that trends found through analysis of posts – initializing a discussion thread – are nearly identical to those found above for comments. This implies that teachers do not reserve their forum behavior to only replies in threads, but are also actively initializing discussion (see supplementary material in Appendix A).
2.3 Discussion

2.3.1 Implications for Courses: Networked Instruction

Teacher enrollment and participation has implications for platform design in terms of how educators are networked with other participants. Forums could be optimized to promote expert-novice dialogue, while novel assessment types such as peer grading [PHC+13] could make use of participant profile information in assigning graders. In addition, platforms could provide tools that allow teachers to discuss specific content or pedagogy, while simultaneously contributing feedback to course staff. The collection and maintenance of profile information will be crucial, with questions remaining for platform designers on whether such information should be collected publicly (e.g., LinkedIn for MOOCs) or privately (surveys). Public and private data collection issues are particularly relevant considering the recent discussion of student privacy in MOOCs [DRW+14].

Teacher enrollment also has implications for course design, where participating teachers could begin taking on aspects of group discussion or tutoring sessions within...
a course. A recent experiment in the HarvardX course on Copyright [Fis14] has begun experimenting with such ideas using cohorting tools to divide participants into small-enrollment sections, each led by a Harvard Law School student. When considering worldwide enrollment in MOOCs, it makes sense that teachers from specific cultural background could lead students from their own regions. Within any model of networked instruction, consideration of the impact on both teachers and students should be taken into account.

### 2.3.2 Implications for Teachers: Professional Development

Teacher Professional Development [VG04] is an ongoing focus of federal educational policy [SHC11, BP12]. A recent report from the Center for Public Education emphasized that educational reform movements like the Common Core Standards require equal reform for teacher professional development [Gul13]. Key issues include moving away from one-day workshops, delivering professional development in the context of a teacher’s subject area, and developing peer (or coaching) networks to facilitate implementation of new classroom techniques.

Already, MOOCs directly related to professional development [KWF13] are emerging; Coursera has launched a “teacher professional development” series serving pedagogical needs of a variety of educators, and edX has just announced a professional development initiative focusing on AP high school courses. Nonetheless, the huge catalogue of available courses raises the question of whether professional development should be provided within the context of a specific topic. For example, a 2009 survey indicated that only 25% of high school physics teachers were physics majors [WT10]. Because of the tremendous amount of content being produced for MOOCs, it seems quite possible that teacher training could leverage this technology to be performed within a Pedagogical Content Knowledge framework [Shu86].

Providers will face challenges in addressing the broad meaning of accreditation and professional development in regard to worldwide access and the diversity of teacher backgrounds [BP12]. In the United States, MOOC participation will also need to be defined in the context of current accreditation models. Costs of Continuing Education
Units (CEUs) are often charged by contact hours in workshop formats taking place over one to two days; how does one translate a 16 week MITx course where mean total-time spent by certificate earners is 100 hours [SBC+14]? However, exploring this issue may help MOOC providers identify potential revenue models; spending estimates for professional development in the United States range between $1000 and $3000 per teacher per year [JÖS14].

2.3.3 Teacher Utilization of MOOC Resources

One of the central themes of the teacher survey respondents is a strong desire to be able to use the MITx course materials in their own teaching. This suggests that ideally, teachers would be able to employ a personalized sub-selection of assessment problems, text, and video content, of a given MOOC, and provide this with their own schedule of material release and due dates, synchronized with their own classrooms schedules, and in harmony with local curricula. Moreover, perhaps ideally, teachers would also be able to enroll their own personal cohort of students, and be able to see their student’s progress and scores. And in such an environment, students would likely benefit from being able to discuss the content in the personalized cohort of individuals defined by the teacher.

Unsurprisingly, however, MOOC platforms like edX have not been designed to work this way. On the other hand, the Khan Academy does offer “coach’s” functionality, which provides essentially all these ideal capabilities to successfully engage teachers. This presents an excellent model that MOOC providers might emulate.

In many ways, what is needed is mechanisms for MOOCs to easily transform into “Personal Online Courses,” dropping the “massive” and “open,” to enable teachers to become a strong point of contact between students and the rich content of MOOCs such as that provided by MITx courses on edX. Such personal online courses could also be natural stepping stones to transforming the digital learning assets of MOOCs into open educational resources, opening doors not just to re-use, but also, collaborative authoring and “social coding” of course content by teachers building off each others’ work.
2.4 Conclusions

Teachers are heavily enrolled and engaged in MITx courses on edX, and evidence indicates that they are playing a substantial role in discussion forums. Measuring the current impact of teachers on other participants is an important area for future research, and one that might help develop learning frameworks that better partner teachers with course staff and other participants. The motivations of teachers will play a key role, whether engaging in life-long learning, life-long instruction, or searching for new pedagogy and peer support. Regarding pedagogy and peer support, adoption of new teaching practices is a major challenge facing teachers and school districts in the United States [Gre14]. MOOCs targeting the needs of teachers and providing mechanisms for MOOCS to become Personal Online Courses, can potentially provide a space for educators to overcome adoption barriers, and a sustainable foundation for the continued existence of MOOCs.

Teacher enrollment clearly represents an unrecognized, meaningful audience for MOOC providers. Recent reports have largely focused on demographics within MOOCs [CSA+13, HRN+14], even leading to criticism that the typical participant is older and in possession of an advanced degree [Ema13]. Recognition of the significance of large teacher enrollments in MOOCs may shift perspectives toward course design and MOOC platform capabilities more attuned to expert participants. Teacher participants in MOOCS are a resource to be respected and valued.
Chapter 3

Beyond Time-on-Task: The Relationship Between Spaced Study and Certification in MOOCs

A long history of laboratory and field experiments has demonstrated that dividing study time into many sessions is often superior to massing study time into few sessions, a phenomenon widely known as the “spacing effect.” Massive open online courses (MOOCs) collect abundant data about student activity over time, but little of its early research has used learning theory to interrogate these data. Taking inspiration from this psychology literature, in this chapter we use data collected from MOOCs to identify observational evidence for the benefits of spaced practice in educational settings. We investigated tracking logs from 20 HarvardX courses to examine whether there was any relationship between how students allocated their participation and what performance they achieved. While controlling for the effect of total time on-site, we show that the number of sessions students initiate is an important predictor of certification rate, across students in all courses. Furthermore, we demonstrate that when students spend similar amounts of time in multiple courses, they perform better in courses where that time is distributed among more sessions, suggesting the benefit of spaced practice independently of student characteristics. We conclude by proposing interventions to guide students’ study schedules and leverage such an effect.
The work done in this chapter was part of a collaborative effort with Yohsuke R. Miyamoto, Joseph Jay Williams, Jacob Whitehill, Sergiy Nesterenko, and Justin Reich from Harvard University. I supported Yohsuke R. Miyamoto throughout this study by extracting session information from raw click events and providing feedback on the development of this study. The original paper, “Beyond Time-on-Task: The Relationship between Spaced Study and Certification in MOOCs”, can be found on the Social Science Research Network [MCW+15].

3.1 Introduction

3.1.1 Early MOOC Research on Participation and Time-on-Task

Much of the early research in Massive Open Online Courses (MOOCs) has focused on measures of student participation, such as time spent on-site, number of click events produced, minutes of video watched, or number of assignments completed. These studies have repeatedly observed two commonplace findings: 1) that the level of participation along one dimension of a course is a predictor of participation along other dimensions, and 2) that the level of participation is a predictor of better grades and course completion [Col13, MGK+14, REN+14, WDR14].

These results align with the total-time law in psychology that the amount of learning is a direct function of study time [CP67, Und70]. Psychologists, however, have long known that not all uses of study time produce equal benefits. One striking exception to the total-time law is known as the spacing effect: for most learning outcomes, shorter, more spaced out study sessions are preferable to massed study sessions. Here we investigate the contribution of the spacing effect to student performance in MOOCs, going beyond existing MOOC research by focusing not only on levels of participation, but also on the allocation of student study time into multiple sessions and its potential benefits for student performance. The decades of research in experimental psychology, education, and cognitive science should help in guiding
practical research [Wil13]; we demonstrate that well-established learning theories can produce important hypotheses for learning analytics, and that learning theory can guide the discovery of effective uses of student learning time.

3.1.2 What is the Spacing Effect?

The spacing effect, initially documented by Herman Ebbinghaus in 1885, is the remarkable phenomenon where distributed presentations of material result in better long-term retention than that resulting from massed presentations of the same material, for a given amount of study time [CPV+06, Dem89, Ebb13, Gre89, Mel70]. For example, the long-term retention of some element A is much stronger following a distributed practice session such as [X, A, Y, Z, A, W, A], rather than a massed practice session such as [X, Y, A, A, A, Z, W], even though the total time spent on studying A is the same. The dependability and robustness of this effect is demonstrated by its replication in a wide variety of experimental tasks over numerous studies [Mel70, MDM+06, PWKS63, SLBP00, Und70, Wau70, You67]. Moreover, the quantitative advantage earned from a spaced schedule is remarkable, with reports that two spaced presentations are about twice as effective as two massed presentations [Hin74, Mel70], and the difference between them increasing further as a function of the number of presentations [Und70].

3.1.3 From Experiments to Education

Despite such potential for bolstering learning, the spacing effect has suffered a history of slow translation into standard educational practices, despite knowledge of the effect in its community for more than a century [Dem88]. In an effort to mitigate this issue, studies have appeared in recent years of real-world classroom demonstrations of the spacing effect in many contexts, such as learning vocabulary [Bir10, CPC09, GO09, RT06, SCK11], and surgical motor tasks [MDM+06]. Furthermore, technology has enabled what is perhaps the most visible application of the spacing effect today – spaced repetition software, sophisticated systems of digital flashcards that help
sequence the presentation of study material in a spaced manner [GJ10].

Our study extends this research on the applications of spaced practice into the realm of MOOCs, and demonstrates new opportunities for research made possible by the massive yet granular data sets of student activity collected by online learning systems. Much of the research on spaced practice takes advantage of experimental designs that can demonstrate causal relationships and internal validity in very specific circumstances. Complementary to this work, our data allows us to observe large numbers of students who study diverse course content and spend their time across multiple courses in diverse ways. While we cannot draw causal conclusions from our data, we can investigate the effects of spaced practice “in the wild,” beyond settings where experimental psychologists control critical elements of course design or student time. Moreover, studies of spaced practice often demonstrate their effects on specific measures of memory and retention. We demonstrate the effects of spaced practice on more generally applicable measures of course completion.

We begin our research with the hypothesis that, broadly speaking, among students who spend similar amounts of time on a MOOC, those who distribute their time into more sessions will perform better than those with fewer sessions. We test this hypothesis with two research designs. First, we examine all students in all courses and show that among students with similar total-time within a course, those with higher session counts perform better.

This correlational study, however, leaves open the possibility that students who spaced their practice differ from students who clustered their practice in other important but unobservable ways. Therefore, in the second level of our investigation, we control for student-to-student differences by examining changes in performance within-individuals, focusing on students who spend similar amounts of time across multiple different courses. We find evidence even within individual student behavior that spaced practice is associated with better performance.

These findings suggest a simple way of improving student learning, agnostic to the nature of specific activities and course content, by providing interventions to encourage students to complete the course with more sessions for the amount of total
study time they have available. In the sections that follow we present our data, methods, and findings, and suggest a set of possible experimental interventions to encourage spaced practice.

3.2 Data

3.2.1 Data Collection

In order to explore the relationship between student performance and how students distribute their time in MOOCs, we examined the timing of click events recorded in the tracking logs of 101,913 unique students (with a combined 127,868 course registrations) with non-zero grades in 20 HarvardX courses (Table 3.1), and how they relate to students’ certification rates (rate of achieving a passing grade) in these courses.

HarvardX is an online learning initiative of Harvard University, and provides the 20 courses examined here on the edX MOOC platform. The different types of online click events triggered by students range from page navigation, to lecture video plays, to problem submissions, as shown in Figure 3-1b, and the diverse set of topics of these courses ranged from biostatistics (Health in Numbers), to philosophy (Justice). Some courses were offered on multiple occasions, such as The Ancient Greek Hero, and Justice, while many others were offered on a single occasion. ChinaX is a series of separate courses with an ongoing theme that continued for 6 modules, and thus tends to have overlapping, devoted students, resulting in high certification rates. Certification rates from its later modules also tend to be high because of attrition from earlier modules. Since we use certification as the outcome of interest in subsequent analysis, we restrict our investigation to students who answer at least one problem correctly, to exclude casual browsers and auditors [HRN+14].
### Table 3.1: Course Information for Twenty 2012-2014 HarvardX Courses

<table>
<thead>
<tr>
<th>Course title</th>
<th>Course code</th>
<th>Start date</th>
<th>End date</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Ancient Greek Hero</td>
<td>CB22x</td>
<td>3/13/13</td>
<td>8/26/13</td>
</tr>
<tr>
<td>The Ancient Greek Hero</td>
<td>CB22.1x</td>
<td>9/3/13</td>
<td>12/31/13</td>
</tr>
<tr>
<td>Justice</td>
<td>ER22x</td>
<td>3/2/13</td>
<td>7/26/13</td>
</tr>
<tr>
<td>Justice</td>
<td>ER22.1x</td>
<td>4/8/14</td>
<td>7/17/14</td>
</tr>
<tr>
<td>Unlocking the Immunity to Change</td>
<td>GSE1x</td>
<td>3/11/14</td>
<td>6/30/14</td>
</tr>
<tr>
<td>Leaders of Learning</td>
<td>GSE2x</td>
<td>7/8/14</td>
<td>8/25/14</td>
</tr>
<tr>
<td>Fundamentals of Clinical Trials</td>
<td>HSPH</td>
<td>10/14/13</td>
<td>2/14/14</td>
</tr>
<tr>
<td>Health in Numbers: Quantitative Methods in Clinical &amp; Public Health Research</td>
<td>PH207x</td>
<td>10/15/12</td>
<td>1/30/13</td>
</tr>
<tr>
<td>United States Health Policy</td>
<td>PH210x</td>
<td>4/7/14</td>
<td>6/30/14</td>
</tr>
<tr>
<td>Human Health and Global Environmental Change</td>
<td>PH278x</td>
<td>5/15/13</td>
<td>7/25/13</td>
</tr>
<tr>
<td>Data Analysis for Genomics</td>
<td>PH525x</td>
<td>4/7/14</td>
<td>6/30/14</td>
</tr>
<tr>
<td>The Political and Intellectual Foundations of China (ChinaX)</td>
<td>SW12x</td>
<td>10/31/13</td>
<td>12/23/13</td>
</tr>
<tr>
<td>The Creation and End of a Centralized Empire (ChinaX)</td>
<td>SW12.2x</td>
<td>1/2/14</td>
<td>1/30/14</td>
</tr>
<tr>
<td>Cosmopolitan Tang: Aristocratic Culture (ChinaX)</td>
<td>SW12.3x</td>
<td>2/13/14</td>
<td>3/6/14</td>
</tr>
<tr>
<td>A New National Culture (ChinaX)</td>
<td>SW12.4x</td>
<td>3/20/14</td>
<td>4/10/14</td>
</tr>
<tr>
<td>From Global Empire to Global Economy (ChinaX)</td>
<td>SW12.5x</td>
<td>4/24/14</td>
<td>5/8/14</td>
</tr>
<tr>
<td>The Last Empire (ChinaX)</td>
<td>SW12.6x</td>
<td>5/22/14</td>
<td>6/19/14</td>
</tr>
<tr>
<td>Global Health: Case Studies from a Biosocial Perspective</td>
<td>SW25x</td>
<td>2/25/14</td>
<td>5/31/14</td>
</tr>
<tr>
<td>Tangible Things</td>
<td>USW30x</td>
<td>6/2/14</td>
<td>8/2/14</td>
</tr>
</tbody>
</table>
3.2.2 Measures

We use certification in a course as our outcome of interest. Certification was awarded to students who achieved a minimum level of performance on a combination of quiz scores, homework assignments, project results, and so on, with content and a threshold that varied from course to course. We use certification as a proxy for learning, but since courses differ in the type and rigor of assessments, the specific meaning of certification varies. It is important to note that most studies of spacing effects focus on measures of memory and retention, rather than this overall measure of course performance.

We extract two measures from course tracking logs to predict certification: (1) total time on-site, and (2) the number of sessions among which their time was distributed, inspired by the spacing effect documented in the field of psychology. As shown in Figure 3-1a, a session was defined as a collection of click events separated by periods of inactivity that lasted more than 30 minutes, in accordance with the Google Analytics standard for defining sessions for website usage\(^1\). Total time spent on-site was thus calculated by summing the lengths of all sessions by a student. A few example sessions shown in Figure 3-1b, taken from CB22x and SW12x, demonstrate the patterns of click events that might occur during a session.

Our metric of student participation, total time on-site, includes the time spent on all activities on-site regardless of whether it was spent watching lecture video clips, solving problem sets, readings chapters, taking quizzes, and so on. Noting that the patterns and frequencies of these event types can vary widely from course-to-course, we took this simplified approach of treating all types of activity equally in the form of total time on-site to keep our analyses agnostic to the wide course-to-course differences in content, structure, and certification requirements. This may allow potential interventions inspired by our findings to be flexibly applied to a wide variety of courses.

\(^1\)https://support.google.com/analytics/answer/2731565?hl=en
3.3 Results

3.3.1 Exploratory Analysis of Key Features in MOOC Data

With these data, we embarked on an exploratory analysis of the relationships between certification rate, time on-site, and number of sessions, inspired by the spacing effect from psychology, where spaced presentation of material benefits long-term recall. An initial look at the data revealed widely varying levels of certification rate, time on-site, session count from course to course. The certification rates of students with nonzero
grades in each course vary widely, from 11% to 70% (Table 3.1). In Figure 3-2, we plot histograms of the number of students by log-scale session count and total course hours, respectively. These histograms reveal that time on-site and session count of courses vary widely, with the median time on-site level of courses varying from as little as 2 hours (USW30x) to about 17 hours (PH207x), an over 8-fold difference, and median sessions from 5 sessions (SW12.5x) to 26 sessions (PH207x), an over 5-fold difference. Since these courses differed greatly in these descriptive statistics and their substantive structural features, we chose to analyze each course individually so as to prevent misinterpreting course-to-course differences in time on-site, session count, and certification as effects of individual students.

A further look at the histograms of Figure 3-2 also reveals widely varying levels of engagement and number of sessions across students within each course – for example, total time on-site of individual students across all courses varied from as little as a few minutes to as much as over 100 hours, and session counts from 1 to over 100. We wondered if this rich variability of behavior across individual students could explain differences in their course performance. Are users with high session count or time on-site more likely to achieve certification?

Thus, we next plotted total time on-site and session count against certification rate. In Figure 3-3a, we divided students within each course into deciles based on their session counts from low to high, and computed each decile’s certification rate to obtain each of the colored lines in Figure 3-3a, with colors representing separate courses. The bold black line corresponds to the mean relationship between session count and certification rate across courses, which was obtained by averaging the session count and certification rate of each the deciles across the 20 courses. These relationships appear sigmoidal, with small effects on certification at small session counts less than 10 sessions, followed by a sharp increase and later a gradual plateauing as the certification rates approach its maximum of 1 (note the log-scale of the x-axis).

When we analyzed time on-site versus certification in the same manner as with session count (Figure 3-3b), here too we found a strong positive correlation with certification rate, consistent with expectations from the MOOC literature [Col13,
CHAPTER 3. BEYOND TIME-ON-TASK: THE RELATIONSHIP BETWEEN SPACED STUDY AND CERTIFICATION IN MOOCs

MGK$^{+14}$, REN$^{+14}$, WDR14]. This relationship appeared to be very similar to that of session count from Figure 3-3a, where certification rates increased in a sigmoidal fashion also with respect to time on-site. It is also perhaps interesting to note the dip that occurs in certification rate in some courses from the lowest decile to the next, for both session count and time on-site; this may reflect the behavior of some
3.3. RESULTS

(a) Certification rates are plotted as a function of session count (10 levels). Students from each course were divided equally into deciles of session counts, and the average certification rates for each decile is plotted as a function of session count for that decile, to form the colored lines, one color for each course. The mean certification rates and session counts were then averaged across all 20 courses for each decile to obtain the bold black line, representing the mean relationship between certification rate and session count across courses.

(b) Certification rates are plotted as a function of time on-site, after dividing students into deciles of time on-site in an analogous fashion to session count in (3-3a).

(c) Time on-site is plotted as a function of session count after dividing students into deciles of session counts, as in (3-3a).

Figure 3-3: Exploratory analysis of time on-site, spacing, and certification rates in MOOCs. (3-3a) Certification rates are plotted as a function of session count (10 levels). Students from each course were divided equally into deciles of session counts, and the average certification rates for each decile is plotted as a function of session count for that decile, to form the colored lines, one color for each course. The mean certification rates and session counts were then averaged across all 20 courses for each decile to obtain the bold black line, representing the mean relationship between certification rate and session count across courses. (3-3b) Certification rates are plotted as a function of time on-site, after dividing students into deciles of time on-site in an analogous fashion to session count in (3-3a). (3-3c) Time on-site is plotted as a function of session count after dividing students into deciles of session counts, as in (3-3a).
students using a secondary account to earn a certificate in a single session, after initial exposure to test questions on a primary account (a behavior that in certain contexts might be described as cheating).

Our analyses thus far of Figures 3-3a and 3-3b suggest a positive relationship of certification, our proxy for learning, with session count and with time on-site. However, Figure 3-3c, which plots total time versus session count, depicts one challenge of further interrogating these relationships. We found that time on-site and session count themselves were highly correlated with each other (Figure 3-3c), with a correlation coefficient of $0.87 \pm 0.04$ (mean $\pm$ s.d. across courses), indicating, as one might expect, that students who work for many sessions also tend to spend a lot of time doing so overall. This multicollinearity of our features makes it difficult to determine whether students certification rates are driven by time on-site, by session count, or both. It reflects a challenge of analyzing real-life, messy data sets of learning “in the wild,” in that we lack experimental control over variables of interest, making it difficult to ascertain the effect of one variable versus another.

3.3.2 Identifying a Benefit of Spacing via Student-to-Student Comparisons

Our data set enables us to reduce ambiguity in interpreting these correlations despite issues of multicollinearity, through its scale of over 100,000 students distributed over 20 courses. Because there are so many students in our data set, every student can be matched with other students who exhibit similar levels of time on-site but different session count, allowing us to compare the effect of session count on students while controlling for time on-site. By looking at the changes in session count while controlling for time on-site, we are able to measure the extent to which the same amount of time on-site is spaced across distinct sessions, corresponding closely to the idea of spaced practice in the psychology literature. Accordingly, we divided students at each of the deciles of time on-site corresponding to the location of the black dots of Figure 3-3b, into low, mid, and high-spacing (session count) terciles, and compared their
certification rates (Figure 3-4a). Figure 3-4a provides compelling visual evidence for the benefit of the spacing in MOOCs, showing that the high-spacing subgroup (black) at each level of time on-site consistently exhibited higher certification rates than their corresponding low and mid-spacing subgroup counterparts at each decile of time on-site (light gray and dark gray dots). The strength of this effect seems to be largest at low levels of time on-site (light blue region) rather than for high level of time on-site (light red region).

The bar graphs in Figure 3-4b help to clarify the differences in the relationship between spacing and certification at different levels of time on site. The top three bar graphs summarize the total time, session number, and certification rate for high levels of time on-site, corresponding to the red area in Figure 3-4a. The light, mid and dark gray bars correspond to the average of the low, mid, and high-spacing terciles in this red region. By design, these low, mid, and high-spacing terciles have similar levels of time on-site (Figure 3-4b, top-left) and highly contrasting session counts (Figure 3-4b, top-middle). The top-rightmost bar graph shows that certification rates were significantly higher for the high-spacing than for the mid-spacing group (Figure 3-4b, top right, paired t-test, p < 0.001), and both of these groups had rates significantly higher than that of the low spacing group (Figure 3-4b, top right, paired t-test, p < 0.001).

We show analogous results for the low levels of time on-site, shown in light blue (Figure 3-3b, bottom). At these low levels of time on-site, the low, mid, and high spacing groups for the low levels of time on-site (light blue region, Figure 3-4a) had similar average time on-site (Figure 3-4b, bottom left) and highly contrasting average session counts, by construction (Figure 3-4b, bottom middle). As in high time on-site levels, these terciles exhibited average certification rates that were significantly higher in the high-spacing groups (Figure 3-4b, bottom right p < 0.001) and significantly lower in the low-spacing groups (p < 0.01) as compared to mid-spacing groups. Note that the lowest (leftmost) and highest (rightmost) deciles of time on-site were not included in the calculation of these bar graphs to allow for a fairer comparison since the levels of time on-site among these spacing groups were not well matched, as one
Figure 3-4: Across-student analysis of the spacing effect in MOOCs. (3-4a) Certification rates are plotted as a function of time on-site (10 levels) and spacing (3 levels) to form $10 \times 3 = 30$ total groups of students taken from the 20 HarvardX courses in the analysis. First, students within each course were divided equally into deciles of time on-site, and then students of each decile were divided further into terciles of low, mid, and high spacing groups (light gray, gray, and dark gray dots). The certification rates and time on-site levels of the 30 groups were then averaged across the 20 courses to form what is shown here. (3-4b) The top bar graphs in light red show the average time on-site, session count, and certification rate aggregated across high levels of time on-site corresponding to the light red region in (3-4a), for low (light gray), mid (gray) and high (dark gray) spacing groups. The bottom bar graphs in light blue are analogous plots representing the time on-site, session count, and certification rates aggregated across the low levels of time on-site corresponding to the light blue region of (3-4a). Note that the lowest (leftmost) and highest (rightmost) deciles of time on-site were not included in the calculation of these bar graphs for a fairer comparison since the levels of time on-site among these spacing groups were not well matched; however, the trend remains within these excluded spacing groups where higher certification is generally associated with higher spacing. Error bars indicate standard errors of the mean, across the 20 courses. The large error bars reflect the large variability across courses noted in Figure 3-2, but the statistical comparisons are highly significant because comparisons are evaluated in a paired fashion within-course. **: $p < 0.01$, ***: $p < 0.001$
can observe in Figure 3-4a; however, the trend remains within these excluded deciles where higher certification is generally associated with higher spacing.

Thus, by comparing the behavior of students with one another within each course, we found that students who distribute their time into a larger number of sessions tend to have higher certification rates than those who do not, while accounting for time on-site. Interestingly, we found that this effect was larger for students at low levels of time on-site, perhaps due to a saturation effect at higher levels of time on-site and certification, suggesting that the potential benefit of spacing may be most prominent at lower levels of participation time.

3.3.3 Identifying a Benefit of Spacing via Within-Student Comparisons

Thus far, we found that students who distribute their time into a larger number of sessions had higher levels of certification than students who spent similar total time on-site divided into fewer sessions. This effect was particularly prominent within low levels of time on-site. However, these correlational observations could be confounded by unobserved variables: it could simply be that high-performing students also tend to have large numbers of sessions, but the act of distributing time into a larger number of sessions may not be a causal factor in better performance. To investigate this question, we analyzed our data more finely to examine whether the effect of sessions can be observed within individuals. We investigated whether students with different levels of session counts across different courses show higher certification rates in courses where they space out their time more. Here we take advantage of our data set of over 100,000 total students to examine the subset of students enrolled in multiple courses. Looking within the overlapping students of a particular pair of courses, some may have distributed their time into more sessions in one than the other, while maintaining their level of time on-site. Although this may happen rarely, the massive scale of the data set gives us the ability to observe a sizeable number of such events. Thus we can examine if a student’s change in session count is associated
with a corresponding change in performance that resembles the spacing effect.

In this analysis, we examined the change in students’ behavior across different pairs of courses; we thus confined our analysis only to students enrolled in multiple courses. Furthermore, we focused on students who maintained similar levels of total time on-site in relation to their peers within each pair of courses, so that we could look at the effect of spacing while controlling for levels of time on-site. We only included students who remained within the same decile of time on-site in both courses, to minimize changes in levels of time on-site across pairs of courses. Because this method substantially reduces the number of students we can include in our analysis, we only considered pairs of courses with at least 250 overlapping students before filtering students based on their levels of time on-site to maintain some level of reliability for examining each course-pair. This threshold allowed us to consider 45 out of the 190 possible pair-wise comparisons that could be made across courses. We further eliminated redundant comparisons that might occur when students were enrolled in more than 3 courses, to avoid double-counting students’ comparisons; such comparisons were eliminated in reverse chronological order such that the comparisons from the earlier courses were retained. As a result, fewer than 3500 out of the over 100,000 students were included in this within-student analysis.

We divided students of each course into 3 groups – low, mid, and high-session groups, corresponding to the low, mid, and high terciles of session counts relative to their peers within each course. Terciles were computed within each course separately, to account for course-to-course differences in session counts. We then compared the changes in certification rate between courses for different paths traversed among these terciles, resulting in 5 different paths: (1) those who greatly decreased from high session counts to low, shown as red in Figure 3-5a, (2), those who mildly decreased from high to mid, or mid to low session counts, shown as dark orange in Figure 3-5a, (3) those who largely maintained their session counts within the same tercile, shown as pale orange in Figure 3-5a, (4) those who mildly increased their session counts from low to mid, or mid to high, shown as light gray in Figure 3-5a, and finally (5) those who greatly increased their session counts from low to high, shown as dark
Figure 3-5: Within-student analysis of the spacing effect in MOOCs. (3-5a) Session counts are shown for 5 different variations of changes from one course to the other within a course-pair. Dark gray represents students who greatly increased their session count from low to high terciles between courses. Terciles were computed within each course, to account for course-to-course differences in session counts. Light gray represents those who mildly increased their session count from low to mid, or mid to high terciles between the pair of courses. Red and dark orange represent students who greatly and mildly decreased their session count, respectively, in an analogous manner. Light orange represents those who remained within the same tercile for both courses. Students session counts were averaged within course for each course-pair, and furthermore these course-pair session counts were then averaged across all course-pairs, to form what is shown here. (3-5b) The mean time on-site levels are shown corresponding to the 5 different variations of session count changes in (3-5a). By construction, these levels of time on-site were largely maintained between the courses within each course-pair, in contrast to the session counts of (3-5a). Only students who maintained their time on-site within the same decile for both courses within each course-pair were selected for this analysis, in order to control for the effect of time on-site. (3-5c) Certification rates are shown for the 5 different variations of session count changes corresponding to the colors of (3-5a) and (3-5b). Error bars indicate the standard error of the mean computed across all course-pairs. ***: p < 0.001
gray in Figure 3-5a. All groups were computed separately for each course-pair, and then averaged together across course-pairs to obtain the time on-site levels, session counts, and certification rates in Figure 3-5. Although these students on average could have big changes in session count (Figure 3-5a), they largely maintained their level of time on-site between the pair of courses (Figure 3-5b), as we intended through the construction of these groups.

Figure 3-5c shows the changes in certification rate for these five groups, providing further evidence for the benefits for spaced practice in MOOCs. The progression of these five paths from greatly decreased (red) to greatly increased (dark gray) session counts corresponds to a monotonically increasing progression in certification rate; furthermore, this positive correlation is preserved across courses ($p < 0.001$). Perhaps most striking is the difference in change of certification rate (Figure 3-5c, $p < 0.001$) between the greatly decreasing (red) and greatly increasing (dark gray) session count paths, with a striking difference between the changes of almost 0.5. A similar comparison can be made for the mildly decreasing and mildly increasing paths in light orange and light gray ($p < 0.001$), with a difference in certification rate of over 0.2. Taken together, these results suggest a benefit in spacing study time in MOOCs, measured within the behavioral changes of individual students.

### 3.4 Discussion

Motivated by the psychology literature on the spacing effect, we examined student behavior in HarvardX data with the hypothesis that for a given level of study time, greater levels of spacing, i.e. larger number of sessions, would be associated with higher certification rates. In agreement with this hypothesis, our analysis comparing different students within-course revealed that those with higher levels of spacing had a higher chance of achieving certification than those who did not, while controlling for their time spent on-site. Interestingly, we found that this benefit may be greater for students at low time on-site levels. Furthermore, by examining students enrolled in multiple courses, we discovered that students who increased their spacing levels from
one course to another while maintaining their levels of time on-site demonstrated a corresponding benefit in their tendency to achieve certification. These findings arose from analyses of a variety of courses, and our metrics, total time on-site and number of sessions, were agnostic to the specific course activities, content, and structure. Taken together, these results strongly suggest benefits to distributing study time into a larger number of sessions, which may have flexible applications for improving student performance in MOOCs.

3.4.1 Differences Between the Current Results and Spacing Effects Studied in Psychology

Our findings cohere with previous research on spaced practice, but there are important differences in our study and previous literature. While our observational study allows us to examine for evidence of spaced practice “in the wild” in diverse settings without parameters controlled or affected by experimentalists, our research designs cannot prove that the effect of spacing of the current findings is directly due to the conventional spacing effect that is recognized by the psychology literature. The spacing effect, as typically recognized, is the benefit in long-term retention following the spaced presentations of particular items of knowledge. However, it is unclear to what extent certification in a course reflects long-term retention as opposed to short-term retention, since student activity is largely self-scheduled; in fact, massed presentation has sometimes been shown to be more beneficial for short-term retention [PSHL62, PHS62]. Our outcome measure, certification, is less precisely defined than long-term retention, but may represent more holistic dimensions of achievement and learning.

The effect of spacing identified in the current study, therefore, could perhaps reflect a separate mechanism for learning. For instance, the spacing effect identified here might be more related to motivation, rather than retrieval of memories per se. It may be that students can stay more excited about learning when their interactions with MOOCs are spaced out. Regardless of the underlying nature of this spacing
effect, however, the findings from this study cohere with previous findings on spaced practice, indicate practical advantages to spacing out time on-site in MOOCs, and motivate practical steps toward leveraging these benefits.

3.4.2 Possibilities of Other Confounding Factors that Drive Certification

It is important to note other factors of student behavior that may also contribute to certification rate that our current analyses have not accounted for. One possibility, for example, is that students with high session counts may also tend to have sessions later in the course, when more course material is accessible and more discussions are available to view on forums. These increased resources could also contribute to increased certification and thus confound the positive effect of session counts on certification. However, we find such an effect unlikely to have a strong influence on the current results; when we quantified students’ session time by calculating the median time of their sessions relative to the course, we found that it was only weakly correlated with session count, with a correlation coefficient of 0.22±0.13 (mean±s.d. across courses). In comparison, recall the correlation of 0.87±0.04 of session count with time on-site that we discovered in our earlier analysis. In addition, the correlation of certification with session time is lower than that with session count in every course, further making session time unlikely to be driving the current observed effects. Exploring further, if we repeat the across-student analyses of Figure 3-4 after omitting all students who did not initiate a session during the last 20% of the course duration, the overall positive effect of spacing remains (p < 0.001), despite omitting more than 70,000 users, over half of the data set. We did not consider this analysis for the within-student analysis of Figure 3-5 as this would severely compromise its already small sample size and hence invalidate any ensuing results.
3.4.3 Interventions for Applying Spaced Practice to MOOCs

Recent years have seen a growing demand for ways to apply the benefits of spacing into practice, in light of criticism of the disconnect between psychological research and educational practice, perhaps best described by the title of an article by Frank Dempster in the American Psychologist [Dem88], “The spacing effect: A case study in the failure to apply the results of psychological research.” Approaches that have recently gained in popularity for applying the spacing effect, however, e.g. through flashcard systems of spaced repetition software, may not be an ideal one to apply to the wide variety of course structures and course topics present in MOOCs. Topics can range from computer science to philosophy, to guitar learning and entrepreneurship, whose content might be neither easily nor appropriately converted to a sequence of flashcards by students. Another recent approach, by a start-up company named SpacedEd [Lam09], has been to call upon instructors to design courses specifically in a spacing-friendly format of a list of questions and answers, to offer online education that is designed around the spacing effect. These approaches both require a costly overhead in which either the students or the instructors must figure out a way to mold the curriculum into a spacing-friendly format. Directly applying this spacing effect to MOOCs could require redesigning of course content and structure, an overbearing cost, considering the rapid growth of the number and variety of MOOC courses.

The present results suggest the potential benefit of designing and applying relatively simple and inexpensive interventions to students taking MOOCs that encourage them to distribute their time into a larger number of shorter sessions, rather than a small number of long sessions. Possible effective interventions might include, for example, incentives for more frequent sessions such as a daily login reward, or perhaps the sending of reminders for logging in, as well as reminders to take breaks within the middle of a session. One useful idea for motivating students to increase the number of their sessions might be the division of assignments into smaller modules that appear more frequently in time. For instance, many courses release their content weekly or bi-weekly, but courses might consider releasing certain elements mid-week, or even daily.
Course developers might divide papers and projects into smaller sub-assignments, release course materials in smaller increments multiple times a week rather than a large release once a week, or assign smaller and more frequent homeworks and quizzes rather than lengthy midterm and final exams. Such experimental interventions could validate the causal relationship between spaced practice and certification. The most useful experimental designs will test competing theories of mechanisms for spacing effects in MOOCs.

No doubt we will see new advances in open online learning and perhaps entirely new generations of learning technologies. The findings here serve as an important reminder that well-established learning theory can be put in the service of new technologies. We have over a century of research on learning theories related to spaced practice. Course and platform developers should attend to findings over the last century of educational research. The field of learning analytics, too, can be greatly enriched when we turn to learning theory to identify what is worth tracking, investigating, and analyzing.
Chapter 4

Probabilistic Use Cases

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 chapter adapts the topic modeling approach of Latent Dirichlet Allocation (LDA) to uncover behavioral structure from student logs in a MITx Massive Open Online Course, namely, 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 cluster 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 chapter is a powerful tool in understanding user behavior and predicting outcomes.

The work done in this chapter was part of a collaborative effort with Daniel
Thomas Seation\textsuperscript{1} and Isaac Chuang\textsuperscript{2}. I served as the project’s lead, performing the majority of the programming, data analysis, visualization, and writing in this chapter. The original paper, “Probabilistic Use Cases: Discovering Behavioral Patterns for Predicting Certification”, can be found on the in the proceedings on the Second ACM conference on Learning@Scale conference [CSC15].

\section{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\textsuperscript{3} and MITx\textsuperscript{4} courses enrolled roughly 600,000 unique users from around the world [HRN\textsuperscript{+}14]. 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 [BPD\textsuperscript{+}13], while others offered more detailed activity such as pause and play clicks within a single lecture video [KGS\textsuperscript{+}14]. 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 chapter, we make the problem of modeling student behavior more tractable by adapting the approach of LDA [BNJ03]. LDA is an unsupervised probabilistic model, which has had great success illuminating shared topics in large collections of texts [Ble12, BL09, BNJ03]. Along with natural language processing, LDA has been adapted in areas such as genetics [PSD00] and web page recommendation [XZY08]. In the latter, LDA discovered latent topics associated with the semantics of user web

\textsuperscript{1}Institutional Research Group, Office of the Provost, Massachusetts Institute of Technology, Cambridge, MA 02139
\textsuperscript{2}Office of Digital Learning, Massachusetts Institute of Technology, Cambridge, MA 02139
\textsuperscript{3}Harvard University’s institution for creating MOOCs
\textsuperscript{4}MIT’s institution for creating MOOCs
4.1. INTRODUCTION

page access patterns, while delivering better performance compared to conventional clustering techniques [XZY08]. 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 directly 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 by large initial enrollments, low percentages of completions, and widely varying resource use [KNDC13, HRN+14, SNM+14].

In this chapter, 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 chapter 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 chapter. 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 chapter.

4.2 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. [KPS13] applied clustering techniques to gain insight into student disengagement or course completion. They represented students by their weekly activity, capturing whether or not students were “on track”, “behind,” “auditing,” or “out” each week. Using this feature representation, 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. [RGH+13] used the graphical model of Probabilistic Soft Logic (PSL) to distinguish between forms of engagement in MOOCs. In contrast to Kizilcec et al., Rameh et al. viewed engagement and 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. [YSAR13] used social behavior for a survival analysis of students in MOOCs, finding that social engagement within the course was correlated with retention.

Preceding MOOC research, the educational data mining community has pursued a broad range of techniques in modeling student behavior and outcomes. Kordan
and Conati [KC11] used the k-means algorithm to group students via their logs in an interactive visualization program for learning Artificial Intelligence. From the resulting groups, association rule mining was used to derive the learning behavior of these clusters and train a classifier to distinguish between high and low learners. In a learning-by-teaching environment with 40 eighth grade students, Kinnebrew and Biswas [KB12] combined sequence and episode mining techniques to categorize students as high or low performance as well as identify periods of productivity and counter-productivity for each student. McCuaig and Baldwin [MB12] also focused on performance and employed decision trees to divide up the space of learners and predict final grades. They used a mix of digital logs, formal observations, and self-assessment to create a holistic understanding of their 122 participants in a C programming class. Finally, building on the initial work of Kardan and Conati [KC11], Davoodi et al. [DKC13] used the combination of k-means clustering and association rule mining to evaluate user behavior in an educational game, finding that students with more prior knowledge were more engaged in the game.

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

4.3 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. The resources in 8.02x included a variety of videos, problems, textual content, and simulation activities. To promote engagement and self-assessment, weekly lectures were typically broken into roughly 5-10 video segments, each interspersed with graded multiple choice questions. A course structure guide indicating resource density and weight toward final grade is provided in Figure 4-1. Major assessment included weekly homework (18%), inter-
active simulations with concept questions (2%), and examinations - three midterms (15% each) and a final (30%). Other supplemental components included problem-solving videos, an eTextbook, and a threaded discussion forum. Certification was granted to students scoring greater than 60% on graded course components. The course ran for 18 weeks, equivalent to the length of the semester long course at MIT.

Between January 17 and September 8, enrollment reached 43,758 people (MIT has since removed this course), with participants coming from around the world and representing a wide range of educational backgrounds [RSB+13]. Courseware interactions from these enrollees led to 37,394,406 events being recorded in the edX tracking logs [SRN+14]. Courseware in this context refers to the individual learning components and software features available to 8.02x participants. In this study, only components making up the backbone of 8.02x were analyzed. This included chapters, sequences, verticals, problems, videos, and html pages. Courses on the edX platform are organized hierarchically, so components are nested inside one another. Chapters, sequences, and verticals are container objects that form organizational units in the course. Generally, chapters contain sequences, which in turn contain verticals. Within these containers are the course resources [edX].

In order to better understand the course structure of 8.02x, a screenshot is provided in Figure 4-2. 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 [SRN+14].

4.4 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 [BNJ03] and draw a parallel between topic modeling and user modeling,
Figure 4-1: Course structure visualization. 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.
which forms the basis for probabilistic use cases. Following this introduction, alternative definitions for quantifying interactions with course modules are explained. These alternatives are evaluated according to their ability to predict certification, which is explained subsequently along side the evaluation process for LDA.

4.4.1 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 could be used for categorization 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
4.4. METHODS

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 [BL09, BNJ03].

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 \( \hat{z} = 1, \ldots, K \) 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. \( d^t = (w_1^t, \ldots, w_{n_t}^t) \). Although all of the documents share the same set of topics, each document has its own topic proportions \((\theta^t)\). \( \theta^t \) is a categorical distribution sampled from a Dirichlet distribution with parameters \( \alpha \), 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. \( z_1^t, \ldots, z_{n_t}^t \). Using this formulation gives rise to the graphical model shown in Figure 4-3 and an expansion of the joint distribution, \( \prod_{t=1}^{T} P(d^t, z_1^t, \ldots, z_{n_t}^t, \phi^t, \beta; \alpha) \), as shown in Equation 4.1.

\[
\prod_{z=1}^{K} P(\beta_{z}) \prod_{t=1}^{T} P(\theta^t || \alpha) \prod_{i=1}^{n_t} \theta^t_{z_1^{i} w_i || z} \tag{4.1}
\]

**Figure 4-3:** The graphical model for Latent Dirichlet Allocation [BNJ03] when applied to topic modeling. Each node represents a random variable. The hidden variables for the topic distributions \((\theta^t)\), topic assignments \((z_i)\), and topics \((\beta_{z})\) comprise the unshaded nodes, while the observed words \((w_i)\) are shaded. The plates represent replicates for the \( n_t \) words, \( T \) documents, and \( K \) topics.
Unfortunately, the posterior distribution for hidden variables defined by LDA is normally intractable because of the marginal distribution for the documents [BNJ03]. To approximate a solution, we use the python package gensim\(^5\), which implements an online variational Bayes algorithm as proposed by Hoffman et al. [HBB10].

### 4.4.2 LDA for Probabilistic Use Cases

In order 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 \(\phi^t\). With this construction, Figure 4-4 shows a graphical model that is parallel to the topic modeling application and yields the expansion of the joint distribution, \(\prod_{t=1}^{T} P(d^t, u^t_1, ..., u^t_{n_t}, \phi^t, \beta; \alpha)\), shown in Equation 4.2.

This model builds on the existing infrastructure for topic modeling and allows us to investigate the hidden behavioral structure within a course.

\[
\prod_{u=1}^{K} P(\beta_u) \prod_{t=1}^{T} P(\phi^t|\alpha) \prod_{i=1}^{n_t} \phi^t_{\hat{u}} \beta_{\hat{m}|\hat{u}} \tag{4.2}
\]

\(^5\)http://radimrehurek.com/gensim/
4.4. METHODS

Figure 4-4: The graphical model for Latent Dirichlet Allocation when applied to use case modeling. Every node represents a random variable. The hidden variables for the use case distributions ($\phi^t$), use case assignments ($u_i$), and use cases ($\beta_u$) are represented by the unshaded nodes, while observed course interactions ($m_i$) are shaded. The plates represent replicates for the $n_t$ interactions, $T$ students, and $K$ use cases.

4.4.3 Interaction Representations

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. In order to quantify these interactions, we considered the following:

- **Binary indicators** that are 1 if the student ever visited the module and zero otherwise.
- **Counts** of the browser events logged by a student for a specific module.
- **Time spent** in seconds on each course module, where time is calculated by taking the difference between browser event timestamps. Breaks over 30 minutes long were discarded.

Binary indicators and word counts are common in the traditional topic modeling applications, while time spent is unique to the context of modeling user behavior.

4.4.4 Evaluating Interaction Representations

Each representation 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 [VCC+99]. The representation that performed the best across these weekly prediction tasks served as the bag of interactions for fitting LDA, saving us from having to consider each interaction representation for a varying number of use cases. This also provided a baseline to compare the predictive performance of a student’s use case proportions ($\phi^t$).

4.4.5 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 it’s predictive performance. Traditionally, model selection, i.e. selecting the optimal number of use cases, is based upon log perplexity [BNJ03] 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 4.3. 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.

$$\log\{\text{perplexity(corpus)}\} = \frac{-\log\{P(\text{corpus}\|\alpha, \beta)\}}{\sum_{\text{corpus}} n_t}. \quad (4.3)$$

Using the models that fit well without an excessive number of use cases, we evaluated how well LDA predicted outcomes like 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 ($\phi^t$) in the 5-fold cross validation of a SVM classifier for certification, where DEC compensated for the class imbalance [VCC+99]. This approach demonstrated the predictive power of a student’s use case proportions ($\phi^t$) and quantified the effect that varying the number of use cases has on performance.
4.5 Results

The results from applying LDA to 8.02x are broken into 4 subsections. The Selecting an Interaction Representation subsection compares the various interaction representations. Employing the optimal 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.

4.5.1 Selecting an Interaction Representation

A student’s activity within the edX platform can be quantified in a number of ways. In this section, we compare the ability of our activity representations – binary indicators, interaction counts, and time spent on each module – to predict certification of students. In this section, binary indicators, interaction counts, and time spent on each modules are compared based on their ability to predict certification of students. We use 5-fold cross-validation over the 8.02x’s 18 week duration as our primary heuristic. Table 4.1 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 [BP13].

Based on Table 4.1, binary indicators failed to accurately identify certificate earners, labeling the entire sample space as non-certificate earners. Even with DEC to compensate for class imbalance, there was not enough of a distinction between non-certificate and certificate earners, which is likely a vestige from some non-certificate students quickly scanning all of the course’s content. Interaction counts and time spent performed better at identifying certificate earners, as these two representations were not as susceptible to such extensive exploration.

Time spent had a slight advantage over interaction counts, making our choice for
<table>
<thead>
<tr>
<th>Week</th>
<th>TNR</th>
<th>TPR</th>
<th>TNR</th>
<th>TPR</th>
<th>TNR</th>
<th>TPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.99</td>
<td>0.00</td>
<td>0.92</td>
<td>0.33</td>
<td>0.92</td>
<td>0.38</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>0.00</td>
<td>0.96</td>
<td>0.36</td>
<td>0.95</td>
<td>0.39</td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
<td>0.00</td>
<td>0.98</td>
<td>0.40</td>
<td>0.97</td>
<td>0.39</td>
</tr>
<tr>
<td>4</td>
<td>1.00</td>
<td>0.00</td>
<td>0.98</td>
<td>0.44</td>
<td>0.98</td>
<td>0.45</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>0.00</td>
<td>0.98</td>
<td>0.52</td>
<td>0.98</td>
<td>0.48</td>
</tr>
<tr>
<td>6</td>
<td>1.00</td>
<td>0.00</td>
<td>0.98</td>
<td>0.57</td>
<td>0.98</td>
<td>0.54</td>
</tr>
<tr>
<td>7</td>
<td>1.00</td>
<td>0.00</td>
<td>0.98</td>
<td>0.65</td>
<td>0.97</td>
<td>0.76</td>
</tr>
<tr>
<td>8</td>
<td>1.00</td>
<td>0.00</td>
<td>0.97</td>
<td>0.79</td>
<td>0.97</td>
<td>0.81</td>
</tr>
<tr>
<td>9</td>
<td>1.00</td>
<td>0.00</td>
<td>0.98</td>
<td>0.82</td>
<td>0.97</td>
<td>0.91</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>0.00</td>
<td>0.98</td>
<td>0.83</td>
<td>0.97</td>
<td>0.91</td>
</tr>
<tr>
<td>11</td>
<td>1.00</td>
<td>0.00</td>
<td>0.97</td>
<td>0.93</td>
<td>0.97</td>
<td>0.93</td>
</tr>
<tr>
<td>12</td>
<td>1.00</td>
<td>0.00</td>
<td>0.97</td>
<td>0.94</td>
<td>0.97</td>
<td>0.94</td>
</tr>
<tr>
<td>13</td>
<td>1.00</td>
<td>0.00</td>
<td>0.97</td>
<td>0.94</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>14</td>
<td>1.00</td>
<td>0.00</td>
<td>0.97</td>
<td>0.95</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>15</td>
<td>1.00</td>
<td>0.00</td>
<td>0.97</td>
<td>0.96</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>16</td>
<td>1.00</td>
<td>0.00</td>
<td>0.97</td>
<td>0.96</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>17</td>
<td>1.00</td>
<td>0.00</td>
<td>0.97</td>
<td>0.96</td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td>18</td>
<td>1.00</td>
<td>0.00</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
<td>0.97</td>
</tr>
</tbody>
</table>

**Table 4.1:** True positive rates (TPR) and true negative rates (TNR) for identifying certificate earners with different interaction representations.

The base representation of students for training LDA. It achieved high performance on non-certificate earners very early in the course, while the 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 [SRN+14], 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 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.
4.5. RESULTS

4.5.2 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 4-5. 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 [Ble12]. Although there has been some work that tries to quantify interpretability [CGW+09], 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.

![Model Selection](image)

**Figure 4-5:** 5-fold log perplexity for a varying number of use cases.
4.5.3 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 4-6 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 appearing at the lower most x-axis. The visual aid is simply a rotation of Fig. 4-1, where color indicates type of resources, 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 4-6 illuminates a distinct behavioral patterns in 8.02x. The use case in Figure 4-6a 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. Based on our observations, we hypothesize that these users were simply “shopping”, although many other possibilities exist. Figure 4-6b captures users who actively participated in the course yet disengaged midway through. Finally, the distribution in Figure 4-6c 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 as shown in Figure 4-7, while the disengaging and completing use cases slowly spread their distributions out, as new material was released. Students in the completing cohort engaged with material as soon as it was released (Figure 4-9), which follows
4.5. RESULTS

(a) Shopping use case

(b) Disengaging use case

(c) Completing use case

**Figure 4-6:** 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 – similar to that found in Figure 4-1 – is below the lowermost probability distribution. All probability distributions are coded according to this visual aid: color – the type of resource, width – density of resources, and height – weight toward final grade (height of bars).

the instructors’ 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 (see Figure 4-8). There was a slight anomaly in week
where the disengaging use case looks closer to a completer that only attempted problems. This artifact is likely the result of the low mean log perplexity score of the 3-use case model as shown in Figure 4-5. There are many more underlying use cases than the 3-use case model can show, so the three most predominate at any given time are the ones selected. Nevertheless, the behavioral patterns captured in Figure 4-6 generally remained well-defined throughout the course.

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 4-10 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 4-10a, 4-10b, 4-10c, 4-10d, and 4-10e. The shopping use case (see Figure 4-6a) reemerges most clearly in Figure 4-10a. Figure 4-10b potentially illuminates a shopping variant, where users are only attempting the first problem set. Figures 4-10c, 4-10d, and 4-10e resemble the disengaging use case in Figure 4-6b, 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 4-10f to Figure 4-10j there is a clear shift in probability from videos to assessments. Such separation indicates the degree to which student 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. The 50-Use Case Model is included in Appendix B for further illustration.

4.5.4 Predicting Certification

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 4.2 presents the overall accuracy rates (ACC), true positive
Figure 4-7: The evolution of the shopping use case from the 3-Use Case model (Figure 4-6a) over the 18 weeks of 8.02x. Figure 4-7g contains the course structure visual aid.
Figure 4-8: The evolution of the disengaging use case from the 3-Use Case model (Figure 4-6b) over the 18 weeks of 8.02x. Figure 4-8g contains the course structure visual aid.
4.5. RESULTS

(a) Week 1
(b) Week 3
(c) Week 6
(d) Week 9
(e) Week 12
(f) Week 15
(g) Week 18

Figure 4-9: The evolution of the completing use case from the 3-Use Case model (Figure 4-6c) over the 18 weeks of 8.02x. Figure 4-9g contains the course structure visual aid.
Figure 4-10: 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 4-10i and Figure 4-10j 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.
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 4.1, 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 increases, both the TNR and TPR increase. At the peak of 50 use cases, a SVM classifier with DEC achieves $0.81 \pm 0.01$ accuracy at predicting certification with just one week of data. Even with only 3 use cases the, prediction accuracy is still at $0.71 \pm 0.01$ with only one week of data.
<table>
<thead>
<tr>
<th>Week</th>
<th>3-use case model</th>
<th>5-use case model</th>
<th>10-use case model</th>
<th>50-use case model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACC</td>
<td>TNR</td>
<td>TPR</td>
<td>ACC</td>
</tr>
<tr>
<td>1</td>
<td>0.71±0.01</td>
<td>0.70</td>
<td>0.79</td>
<td>0.77±0.01</td>
</tr>
<tr>
<td>2</td>
<td>0.79±0.01</td>
<td>0.78</td>
<td>0.93</td>
<td>0.83±0.01</td>
</tr>
<tr>
<td>3</td>
<td>0.87±0.02</td>
<td>0.86</td>
<td>0.96</td>
<td>0.84±0.02</td>
</tr>
<tr>
<td>4</td>
<td>0.90±0.01</td>
<td>0.89</td>
<td>0.97</td>
<td>0.91±0.02</td>
</tr>
<tr>
<td>5</td>
<td>0.87±0.02</td>
<td>0.86</td>
<td>0.98</td>
<td>0.91±0.02</td>
</tr>
<tr>
<td>6</td>
<td>0.90±0.02</td>
<td>0.90</td>
<td>0.99</td>
<td>0.91±0.02</td>
</tr>
<tr>
<td>7</td>
<td>0.92±0.02</td>
<td>0.91</td>
<td>0.99</td>
<td>0.91±0.02</td>
</tr>
<tr>
<td>8</td>
<td>0.92±0.02</td>
<td>0.91</td>
<td>0.99</td>
<td>0.94±0.02</td>
</tr>
<tr>
<td>9</td>
<td>0.94±0.01</td>
<td>0.93</td>
<td>0.99</td>
<td>0.95±0.01</td>
</tr>
<tr>
<td>10</td>
<td>0.93±0.02</td>
<td>0.93</td>
<td>0.99</td>
<td>0.94±0.02</td>
</tr>
<tr>
<td>11</td>
<td>0.93±0.02</td>
<td>0.93</td>
<td>1.00</td>
<td>0.95±0.01</td>
</tr>
<tr>
<td>12</td>
<td>0.93±0.02</td>
<td>0.93</td>
<td>1.00</td>
<td>0.93±0.02</td>
</tr>
<tr>
<td>13</td>
<td>0.92±0.02</td>
<td>0.91</td>
<td>0.99</td>
<td>0.95±0.01</td>
</tr>
<tr>
<td>14</td>
<td>0.96±0.01</td>
<td>0.95</td>
<td>0.97</td>
<td>0.97±0.01</td>
</tr>
<tr>
<td>15</td>
<td>0.92±0.02</td>
<td>0.92</td>
<td>0.99</td>
<td>0.95±0.01</td>
</tr>
<tr>
<td>16</td>
<td>0.96±0.01</td>
<td>0.96</td>
<td>1.00</td>
<td>0.95±0.01</td>
</tr>
<tr>
<td>17</td>
<td>0.96±0.01</td>
<td>0.95</td>
<td>1.00</td>
<td>0.97±0.01</td>
</tr>
<tr>
<td>18</td>
<td>0.96±0.01</td>
<td>0.96</td>
<td>1.00</td>
<td>0.96±0.01</td>
</tr>
</tbody>
</table>

Table 4.2: 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.
4.6 Discussion

Applying LDA to 8.02x provides a data-driven approach to understanding student behavior that transform massive amounts of statistical information into a set of use cases 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 chapter 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 graphical model in Figure 4-4 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 the question that follows it. 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 [Wal06] or Griffiths et al. [GSBT04] 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 date they enrolled. Such course features could potentially have a dramatic effect on behavior, which LDA does not currently capture.

Nevertheless, the application of LDA in this chapter 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 that vary on factors such as population size, course structure, or material.

4.7 Conclusions

Our results show that LDA can be adapted in 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 provide simple intuition about which course components are utilized, and potentially more complex modes related to 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.
Chapter 5

Contributions

The primary aim of this thesis was to illuminate valuable subpopulations in MOOCs. The large and diverse student populations in MOOCs present an unprecedented opportunity to understand student behavior and learn about learning. MOOCs platforms collect a tremendous amount of information on students by logging their behavior, capturing what is essentially the body language of the web. However, despite this wealth of data, little has been done to identify important subpopulations and understand their strengths and weaknesses. The results in this thesis focused on the latent potential of various learner subpopulations.

Surveying students in MITx MOOCs found a substantial and valuable subpopulation of teachers. Approximately 1 in 4 (28.0%) respondents identify as past or present teachers, while nearly one in ten (8.7%) identify as current teachers. Despite representing only 4.5% of the nearly 250 thousand enrollees, survey responding teachers generated 22.4% of all discussion forum comments. More notably, 1 in 12 comments are from current teachers, and 1 in 16 comments are from teachers with experience teaching the subject. With the correct infrastructure in place, teachers could become a strong point of contact between students and the rich content of MOOCs such as the MITx courses on edX. Moreover, teachers could be a valuable source of revenue for MOOCs because an average of 53.9% (560) of current teachers answered yes to having interest in accreditation opportunities. Accommodating these teachers could be invaluable to the success of MOOCs.
Another important contribution of this thesis was the investigation of the “spacing effect” in MOOCs. Looking at how students choose to spend their time in 20 HarvardX MOOCs identified observational evidence for the benefits of spaced practice in educational settings. It was shown that the number of sessions a student initiates is an important predictor of performance. In particular for the subpopulation of students who only spend a few hours in a course, spaced practice can have a dramatic effect on their performance. Even when students spend similar amounts of time in multiple courses, they perform better in courses where their time is distributed among more sessions.

Finally, this thesis contributed a novel approach to identifying and characterizing student behavior in MOOCs. Inspired by the success of Latent Dirichlet Allocation (LDA) in topic-modeling, LDA was adapted to user-modeling by considering users as a “bag-of-interactions”. Applying this method to 8.02x, an introductory physics MOOC from MITx, discovered probabilistic use cases that capture the most salient behavioral trends in the course, such as “shoppers,” “completers,” and “experts”. These probabilistic use cases also served as an effective form of dimensionality reduction, allowing us to predict whether or not a student will earn a certificate with 0.81±0.01 accuracy from the first week of logs.

In conclusion for this thesis I:

- Analyzed the forum activity of teachers, visualized the results, and helped explore potential possibilities for MOOCs given significant teacher enrollment

- Extracted and manipulated the raw click events to calculate session information and helped investigate the psychology phenomenon know as the “spacing effect” in MOOCs

- Adapted Latent Dirichlet Allocation to model user behavior, creating an unsupervised method to identify and characterize the most salient behavioral trends in MOOCs
Appendix A

Supplementary Materials for Teacher Enrollment

A.1 Highlights

In the spring of 2014, 11 MITx courses gave entrance and exit surveys addressing the motivations and backgrounds of participants. Over 33,000 participants responded to the entrance surveys, a 16.9% response rate. Amongst the survey's questions, several were targeted at teacher enrollment, asking whether participants self-identify as an instructor or teacher, with affirmative responses followed by contextual questions addressing the aspects of their instruction (see Table A.1). These questions can be broadly distilled into the following themes:

1. Are a significant number of teachers enrolling in MITx open online courses?

2. If so, do these teachers come from traditional instructional backgrounds?

3. Do teachers completing a course desire accreditation opportunities and broader usability of MITx resources?

The exit survey aimed at course outcomes, and questions were added in order to gauge the influence of a course on current teachers (see questions in Table A.1 and Table A.2). Note, courses in this study did not advertise directly to teachers,
nor has MITx as an organization attempted to reach out to this demographic. We view the results of this analysis as a baseline for teacher enrollment. All surveys were distributed using the Qualtrics survey platform\(^1\) and appeared along side or within the first unit of material in each course. In some courses, reminders were sent to those students having not yet completed the entrance survey.

\(^{1}\text{www.qualtrics.com}\)
<table>
<thead>
<tr>
<th>Number</th>
<th>Question</th>
<th>Response Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Are you currently, or have you ever identified yourself as, an instructor?</td>
<td>Yes; No</td>
</tr>
<tr>
<td></td>
<td><strong>If “Yes” to Question 1</strong></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Are you, or have you, taught material related to this course?</td>
<td>Yes; No; Unsure</td>
</tr>
<tr>
<td>3</td>
<td>In what settings did your instruction take place?</td>
<td>Elementary / Primary School; Secondary / High School;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>College or University; Outside the scope of traditional schools; Support Staff, e.g., Teaching Assistant;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other (open text field);</td>
</tr>
<tr>
<td>4</td>
<td>Are you currently employed as a teacher or instructor?</td>
<td>Yes; No</td>
</tr>
</tbody>
</table>

**Table A.1:** Entrance survey teacher questions.
<table>
<thead>
<tr>
<th>Questions</th>
<th>No</th>
<th>Unsure</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-course Averages</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Would you be interested in teacher accreditation opportunities? e.g.,</td>
<td>15.1%</td>
<td>27.3%</td>
<td>55.9%</td>
</tr>
<tr>
<td>Continuing education units or graduate credit.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I will use material and/or ideas from this course in my future teaching</td>
<td>5%</td>
<td>4.3%</td>
<td>44%</td>
</tr>
<tr>
<td>efforts.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perspective gained in this course will adjust how I teach future courses.</td>
<td>3.8%</td>
<td>3.1%</td>
<td>42.3%</td>
</tr>
<tr>
<td>I would be interested utilizing resources from other MITx courses in my</td>
<td>3.4%</td>
<td>3.4%</td>
<td>45.7%</td>
</tr>
<tr>
<td>own teaching efforts.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table A.2:** Cross course average percentages of current teachers' responses to exit survey questions.
Highlights from the questions on teacher enrollment are shown below:

- The 11 MITx courses capture a wide range of topics from Street-Fighting Math to Entrepreneurship 101 (see Table A.3).

- Across the 11 MITx courses, entrance surveys indicate 1 in 4 (28.0%) survey respondents identify as teachers, while nearly one in ten (8.7%) identify as current teachers. In addition, one in twenty (5.9%) identify as having taught the course topic (see Table A.4).

- The percentages of teachers varied between courses with Global Warming Science (12.340x) attracting the most teachers and Building Mobile Experiences (21W.789x) attracting the least. (see Figure A-1).

- The largest percentage of teachers came from colleges and universities (post-secondary education) (see Figure A-2).

- Teachers were generally interested in taking what they learned from the MITx courses back to their classes (see Figure A-3).
<table>
<thead>
<tr>
<th>Course</th>
<th>Description</th>
<th>Launch Date</th>
<th>Course Length</th>
<th>Registrants by Week 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>21W.789x</td>
<td>Building Mobile Experiences</td>
<td>2014-02-04</td>
<td>12 Weeks</td>
<td>31072</td>
</tr>
<tr>
<td>6.041x</td>
<td>Introduction to Probability - The Science of Uncertainty</td>
<td>2014-02-04</td>
<td>15 Weeks</td>
<td>26569</td>
</tr>
<tr>
<td>6.00.2x</td>
<td>Introduction to Computational Thinking and Data Science</td>
<td>2014-03-05</td>
<td>9 Weeks</td>
<td>15065</td>
</tr>
<tr>
<td>12.340x</td>
<td>Global Warming Science</td>
<td>2014-02-19</td>
<td>12 Weeks</td>
<td>13047</td>
</tr>
<tr>
<td>6.00.1x</td>
<td>Introduction to Computer Science and Programming Using Python</td>
<td>2014-02-19</td>
<td>9 Weeks</td>
<td>22797</td>
</tr>
<tr>
<td>15.071x</td>
<td>The Analytics Edge</td>
<td>2014-03-04</td>
<td>11 Weeks</td>
<td>26530</td>
</tr>
<tr>
<td>16.110x</td>
<td>Flight Vehicle Aerodynamics</td>
<td>2014-03-05</td>
<td>14 Weeks</td>
<td>28653</td>
</tr>
<tr>
<td>15.390x</td>
<td>Entrepreneurship 101: Who is your customer?</td>
<td>2014-03-18</td>
<td>6 Weeks</td>
<td>44867</td>
</tr>
<tr>
<td>6.SFMx</td>
<td>Street-Fighting Math</td>
<td>2014-04-08</td>
<td>7 Weeks</td>
<td>23640</td>
</tr>
<tr>
<td>3.091x,2</td>
<td>Solid-State Chemistry</td>
<td>2014-05-12</td>
<td>15 Weeks</td>
<td>6954</td>
</tr>
<tr>
<td>2.01x</td>
<td>Elements of Structures</td>
<td>2014-06-03</td>
<td>12 Weeks</td>
<td>7705</td>
</tr>
</tbody>
</table>

**Table A.3:** The 11 MITx courses used in studying teacher enrollment.
### Table A.4: Teacher enrollment in 11 MITx Courses

Students who answered Question 1 in Table A.1 were included in “Surveyed.” Students who answered “Yes” to Question 1 were included in “Teachers.” Students who answered “Yes” to Question 4 were included in “Current Teachers.” Students who answered “Yes” to Question 2 were included in “Teach / Taught Topic.”

<table>
<thead>
<tr>
<th>Course</th>
<th>Surveyed</th>
<th>Teachers</th>
<th>Current Teachers</th>
<th>Teach / Taught Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>21W.789x</td>
<td>4217</td>
<td>933</td>
<td>242</td>
<td>115</td>
</tr>
<tr>
<td>6.041x</td>
<td>2400</td>
<td>553</td>
<td>197</td>
<td>116</td>
</tr>
<tr>
<td>6.00.2x</td>
<td>2997</td>
<td>739</td>
<td>216</td>
<td>123</td>
</tr>
<tr>
<td>12.340x</td>
<td>2458</td>
<td>956</td>
<td>318</td>
<td>277</td>
</tr>
<tr>
<td>6.00.1x</td>
<td>3997</td>
<td>956</td>
<td>280</td>
<td>143</td>
</tr>
<tr>
<td>15.071x</td>
<td>3010</td>
<td>838</td>
<td>183</td>
<td>122</td>
</tr>
<tr>
<td>16.110x</td>
<td>1709</td>
<td>441</td>
<td>139</td>
<td>86</td>
</tr>
<tr>
<td>15.390x</td>
<td>4843</td>
<td>1682</td>
<td>405</td>
<td>268</td>
</tr>
<tr>
<td>6.SFMx</td>
<td>4162</td>
<td>1364</td>
<td>499</td>
<td>333</td>
</tr>
<tr>
<td>3.091x_2</td>
<td>1639</td>
<td>506</td>
<td>195</td>
<td>144</td>
</tr>
<tr>
<td>2.01x</td>
<td>2058</td>
<td>483</td>
<td>173</td>
<td>144</td>
</tr>
<tr>
<td>Total</td>
<td>33490</td>
<td>9451</td>
<td>2847</td>
<td>1871</td>
</tr>
<tr>
<td>Average Percent</td>
<td>16.8%</td>
<td>28.0%</td>
<td>8.7%</td>
<td>5.9%</td>
</tr>
</tbody>
</table>

*Table A.4: Teacher enrollment in 11 MITx Courses. Students who answered Question 1 in Table A.1 were included in “Surveyed.” Students who answered “Yes” to Question 1 were included in “Teachers.” Students who answered “Yes” to Question 4 were included in “Current Teachers.” Students who answered “Yes” to Question 2 were included in “Teach / Taught Topic.”*
Figure A-1: Percentage of self-identifying teachers, whether they have or currently teach the course topic, and the percentage of currently practicing teachers. Percentages are relative to the number of survey respondents (see Table A.4).
Figure A-2: Instructional context of self-identifying teachers. Survey question allowed for multiple responses (see Question 3 in Table A.1). All percentages are relative to total number of responses within each course.
Figure A-3: Current teachers' responses to the exit survey questions from Table A.2.
A.2 Discussion Forums

An intriguing aspect of teacher enrollment involves how this group chooses to participate in courses. In addition to survey data, a tremendous amount of participant interaction data is available. Discussion forums entries and click-stream data allow one to check if participating teachers are actively pursuing an important aspect of their profession, namely, instructing other participants within a course.

Discussion forums are the only source for social interaction data. These highly unstructured conversational hubs include help-seeking questions on problem solving, as well as course-wide introductions where participants state their country of origin and background. Only a fraction of participants are active posters (initiating a forum thread) and commenters (responding to posts); the majority of forum interactions simply involve searching through and reading posts [HDG+14].

Participants can post (text initiating a discussion thread), comment (responds within an existing thread), and up-vote (assign one point to a thread, where total score leads to increased visibility). In addition, one can count the total number of forum-click interactions using click-stream data. Figure A-4 shows the average number of posts (Figures A-4a), comments (Figures A-4b), and forum-click interactions (Figures A-4c) for teachers versus non-teachers.
Figure A-4: The average number of posts (A-4a), comments (A-4b), and forum-click interactions (A-4c) for teachers (orange) versus non-teacher respondents (navy). Activity is significantly higher in 4 out of 11 courses in each category. The distributions of forum activity are highly skewed toward zero, often with a few outliers leading to significant tails toward high activity.
Appendix B

50-Use Model

B.1 Probability Distributions

Figures B-1, B-2, B-3, B-4, and B-5 contain the 50 partial probability distributions associated with a 50-use case model trained on all 18 weeks of logs from 8.02x. Each distribution captures the probability of interacting with 8.02x’s main resources, which included a variety of videos, problems, textual content, and simulation activities. Navigational interactions, such as clicking on a chapter or a sequence in the courseware navigation, are not included. As a result, figures like Figure B-1d and Figure B-1h appear to have an empty distribution because they are navigation only use cases. Together, all 50 probability distributions illuminate the diversity of student behavior in 8.02x.
Figure B-1: Probability distributions for use cases 1 through 10 in a 50-Use Case Model trained on all 18 weeks of logs from 8.02x. In contrast to the 3-Use Case Model, the 50-Use Case model provides higher granularity into disengaged and engaged behavior trends. Figure B-1i and Figure B-1j 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.
Figure B-2: Probability distributions for use cases 11 through 20 in a 50-Use Case Model trained on all 18 weeks of logs from 8.02x. In contrast to the 3-Use Case Model, the 50-Use Case model provides higher granularity into disengaged and engaged behavior trends. Figure B-2i and Figure B-2j contain the course structure visual aid.
**Figure B-3:** Probability distributions for use cases 21 through 30 in a 50-Use Case Model trained on all 18 weeks of logs from 8.02x. In contrast to the 3-Use Case Model, the 50-Use Case model provides higher granularity into disengaged and engaged behavior trends. Figure B-3i and Figure B-3j contain the course structure visual aid.
Figure B-4: Probability distributions for use cases 31 through 40 in a 50-Use Case Model trained on all 18 weeks of logs from 8.02x. In contrast to the 3-Use Case Model, the 50-Use Case model provides higher granularity into disengaged and engaged behavior trends. Figure B-4i and Figure B-4j contain the course structure visual aid.
Figure B-5: Probability distributions for use cases 41 through 50 in a 50-Use Case Model trained on all 18 weeks of logs from 8.02x. In contrast to the 3-Use Case Model, the 50-Use Case model provides higher granularity into disengaged and engaged behavior trends. Figure B-5i and Figure B-5j contain the course structure visual aid.
Bibliography


[BPD+13] Lori Breslow, David E Pritchard, Jennifer DeBoer, Glenda S Stump, Andrew D Ho, and DT Seaton. Studying learning in the worldwide


[FRT13] Colin Fredericks, Saif Rayyan, Raluca Teodorescu, Trevor Balint, Daniel Seaton, and David E Pritchard. From flipped to open instruction: The
mechanics online course. In *Sixth Conference of MITs Learning International Network Consortium*, 2013. 25


the first ACM conference on Learning@Scale conference, pages 117–126. ACM, 3 2014. 31, 97


[KGS+14] Juho Kim, Philip J Guo, Daniel T Seaton, Piotr Mitros, Krzysztof Z Gajos, and Robert C Miller. Understanding in-video dropouts and in-

[KNDC13] Daphne Koller, Andrew Ng, Chuong Do, and Zhenghao Chen. Retention and intention in massive open online courses: In depth. Educause Review, 48(3), 2013. 21, 61


[MB12] Judi McCuaig and Julia Baldwin. Identifying successful learners from interaction behaviour. 6 2012. 63


[MDM+06] Carol-Anne E Moulton, Adam Dubrowski, Helen MacRae, Brent Graham, Ethan Grober, and Richard Reznick. Teaching surgical skills: what
kind of practice makes perfect?: a randomized, controlled trial. *Annals of surgery*, 244(3):400, 2006. 39


the international joint conference on artificial intelligence, volume 1999, pages 55–60, 1999. 70


[XZY08] Guandong Xu, Yanchun Zhang, and Xun Yi. Modelling user behaviour for web recommendation using LDA model. In Web Intelligence and
