

# Detecting Changes in Student Behavior from Clickstream Data

Jihyun Park  
Department of Computer Science  
University of California, Irvine  
Irvine, CA 92697  
jihyunp@ics.uci.edu

Kameryn Denaro  
Teaching and Learning  
Research Center  
University of California, Irvine  
Irvine, CA 92697  
kdenaro@uci.edu

Fernando Rodriguez  
School of Education  
University of California, Irvine  
Irvine, CA 92697  
fernandr1@uci.edu

Padhraic Smyth  
Department of Computer Science  
University of California, Irvine  
Irvine, CA 92697  
smyth@ics.uci.edu

Mark Warschauer  
School of Education  
University of California, Irvine  
Irvine, CA 92697  
markw@uci.edu

## ABSTRACT

Student clickstream data can provide valuable insights about student activities in an online learning environment and how these activities inform their learning outcomes. However, given the noisy and complex nature of this data, an ongoing challenge involves devising statistical techniques that capture clear and meaningful aspects of students' click patterns. In this paper, we utilize statistical change detection techniques to investigate students' online behaviors. Using clickstream data from two large university courses, one face-to-face and one online, we illustrate how this methodology can be used to detect when students change their previewing and reviewing behavior, and how these changes can be related to other aspects of students' activity and performance.

## CCS Concepts

•Information systems → Data mining; Web log analysis; •Computing methodologies → Machine learning approaches; •Applied computing → Learning management systems;

## Keywords

Student clickstream data; Change detection; Regression; Poisson models

## 1. INTRODUCTION

One of the major goals in educational data mining (EDM) is to use student clickstream data to describe and understand students' behavioral patterns. While past findings

have advanced our understanding what we can learn from clickstream data, one of the remaining challenges involves devising statistical techniques that help us identify students who are changing behavior in the middle of a term. There are a number of reasons motivating this problem; one is to identify students who are in need of assistance during the course, another is to identify reasons that students are changing their behavior so that a course could be improved overall. The analysis of clickstream data within a course can also provide invaluable information to course instructors and to education researchers, and there is a need to be able to both summarize and visualize the results in a straightforward manner.

In this paper we will focus on clickstream data from two courses at a large university: one face-to-face course and one online course, both from the 2015-2016 academic year. For each course, clickstream data is obtained through a course management system in the form of {student ID, time stamp, activity}. The types of activities recorded correspond to broad categories of student behavior, such as previewing lecture notes, submitting assignments, or posting and responding to discussion board questions. For instance, one of the courses we examine in this paper had 377 registered students who generated approximately 380,000 click events over a 10-week period. Figure 1 displays each of the individual student clickstreams over the 85 days of the course, with each row corresponding to a student. While the plot shows some general increases in click activities around quiz and exam dates, it is not easy to see much else, nor to understand how individual student behaviors are related to the overall population due to significant variability in students' click patterns. Furthermore, we are unable to determine whether students change their click behaviors in any significant way, or whether or not these behaviors are correlated with course performance.

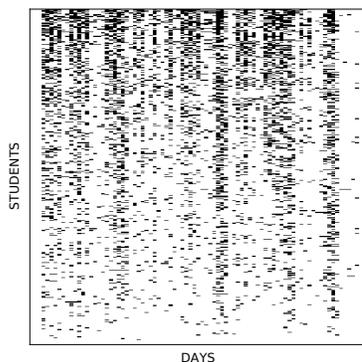
As discussed in more detail in the next section, student clickstream data has been the subject of a number of prior studies, such as the investigation of potential predictive relationships between online student activity and student outcomes (such as course grades). Here we focus instead on detecting changes in individual student activity over time, relative to the activity of the class as a whole. In particular

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

LAK '17, March 13-17, 2017, Vancouver, BC, Canada

© 2017 ACM. ISBN 978-1-4503-4870-6/17/03...\$15.00

DOI: <http://dx.doi.org/10.1145/3027385.3027430>



**Figure 1: A plot of student clickstream activity in the 10-week face-to-face course over time, where each row represents an individual student and each column represents a day. A black marker in cell  $i, t$  indicates clickstream activity for student  $i$  on day  $t$ .**

we investigate the use of statistical change detection techniques (e.g., [9]) to automatically detect changes in activity over time for each student. We model the activity of each student relative to the aggregate activity of all students in the class and compare two models on a per student basis; a model where there is no change in student activity versus a model where there is a significant change in activity at some unknown point during the period of the course. Likelihood-based techniques are used to fit both models on a per student basis and model selection criteria is implemented in order to determine whether each student is best modeled under the “change” or “no-change” model.

The paper proceeds as follows. In Section 2 we discuss related work. Section 3 outlines the change-detection methodology that we propose, and Section 4 provides illustrative results on simulated data sets. Section 5 discusses the course data sets that provide an illustration of the methods discussed in Section 3 and Section 6 describes the results of applying our change-detection methodology to these data sets. The paper concludes with discussion and conclusions in Section 7. The primary novel contribution of this work is the development of a systematic quantitative approach for detecting significant changes in a student’s clickstream over time.

## 2. RELATED WORK

Clickstream data analysis in an educational setting has focused on what the clickstream can say about the students in terms of learning behavior through a variety of features derived from the clickstream. Much of the prior work on clickstream data analysis for understanding student behavior has occurred in the context of Massive Open Online Courses (MOOC) setting. Many of these analyses have focused on using the clickstream data to predict MOOC completion (for example in [5]) and to predict learning outcomes within a MOOC. For example, the relationship between the number of posts and the learning gains of the students [16] has been investigated, as well as how discussion forum views are potentially related to learning outcomes [1]. There has also been research focused on improving predictions of learning outcomes by incorporating clickstream events as well as sum-

maries of the clickstream [3].

A secondary research topic has focused on describing students with similar clickstreams (e.g., [15]), the activities that the students are engaging in, and in understanding the student’s typical online interaction within a class. As an example, clickstream data analysis was used to better understand whether or not students were following a defined learning path [6]. In other work, students’ clickstreams were grouped into similar plans of action to better understand learning pathways [14]; how discussion forums and other activities in the MOOC were related to country and culture [12]; and examined whether engagement on discussion forums increased based on the type of video a student watched [2]. All of these clickstream analyses have an underlying goal of describing student behaviors through the clickstream and to draw meaningful conclusions about those students.

MOOCs are typically used by people as a way to learn new skills or keep up-to-date with current ones. Because most MOOCs do not offer formal degrees, there are no serious consequences for doing poorly or dropping out. In contrast, college course grades determine whether students succeed or fail (whether they advance to the next course, remain in their intended major, or graduate). Thus, findings from MOOC clickstream studies cannot offer broad explanations about student learning experiences in higher education settings. So while MOOCs and college courses share some similarities, in terms of course management systems and clickstream data, studying college courses may require a different set of goals and statistical techniques.

For instance, one important area of higher education research focuses on student engagement. Studies find that students who are not engaged with the learning process—that is, students who do not put in the time and energy into purposeful learning—are at greater risk for failing courses and dropping out of college [10]. While this finding is not new, understanding how to quickly identify these students, especially at the course-level, remains a significant challenge.

Clickstream data has the potential to address this since the data is obtained in real time. Researchers can provide instructors with immediate insights how students are engaging with the course management system. This is especially important in courses with large enrollments, where problems with student engagement can often go unnoticed [13]. Some recent work has found that student engagement with the course management system, as indexed by number of days students visited the site relative to their peers, was positively related to course outcomes [11]. Our work adds to this area of research by using statistical change detection techniques to further understand course engagement.

More broadly, changepoint detection techniques for event time-series is a widely studied topic and a variety of statistical methodologies have been developed (e.g., [7, 9]), with much of this work focused on single (univariate) time-series. Web user behavior has been analyzed to detect changes in an individual’s behavior, to report “interesting” sessions, and to detect changes in user activity [8]. There has not been any prior work (to our knowledge) on change detection applied to multiple clickstreams of students in an educational setting.

Thus far, previous work in the analysis of clickstream data in an educational setting has focused on grouping students into similar groups, understanding possible dropout, predicting student success in a course, and defining learning

pathways. Our goal is to add to the current body of research in a meaningful way by using changepoint detection techniques as a proxy for understanding student engagement. By detecting whether student behavior changes in a significant manner over the time-period of a particular term, we hope to identify students who increase, decrease, or show no change in their clickstream activities, and whether these changes relate to course performance.

### 3. METHODOLOGY

We discuss below our approach for modeling and change detection of student activity. We begin by defining some general notation and then introduce two different models: a Bernoulli model for binary data and a Poisson model for count data. The section concludes with a description of changepoint detection for both of these models.

#### 3.1 Notation

Let  $N$  be the number of individual students in a course where  $i$  is an index that refers to an individual student in the class,  $i = 1, \dots, N$ . We will assume below that time is discrete<sup>1</sup> with  $T$  discrete time-points and  $t = 1, \dots, T$  being an index running from the first to the last time-period of clickstream logging for the course. Below we will refer to  $t$  on a daily time-scale for convenience but in general other time-periods—such as days or weeks—could be used.

Let  $X$  be the observed data for a course, represented as an  $N \times T$  array whose entries are counts  $x_{it} \in \{0, 1, 2, \dots\}$ . Note that  $x_{it}$  represents the number of click events for student  $i$  on day  $t$ , where  $1 \leq i \leq N$  and  $1 \leq t \leq T$ . We will also consider a binarized version of the data  $x'_{it} = I(x_{it} > 0)$ , where  $I(\cdot)$  is an indicator function (as in Figure 1 for example). The number of clicks  $x_{it}$  (counts) by student  $i$  on a given day  $t$  in principle contains more information than the binarized version  $x'_{it}$ , but could also be quite noisy in the sense that more clicks might not necessarily correlate well with relevant student activity. We explore both options since the choice of looking at a count versus the binarized version in practice will depend on the context of a particular analysis.

#### 3.2 Bernoulli Models for Binary Data

For the binary data,  $x'_{it}$ , let  $\pi_{it}$  be the probability that each student  $i$  is active on day  $t$  (i.e., the probability that student  $i$  generates one or more clicks on day  $t$ ). The log-odds of  $\pi_{it}$  is modeled as:

$$\log \frac{\pi_{it}}{1 - \pi_{it}} = \mu_t + \alpha_i \quad (1)$$

where  $\mu_t, t = 1, \dots, T$  can be viewed as a time-varying population mean for the log-odds and  $\alpha_i, 1 \leq i \leq N$  is a student-dependent offset to account for individual-level variation in student behavior.

The role of  $\alpha_i$  in this model is to modulate the time-varying population mean  $\mu_t$  in a student-specific manner. A positive value of  $\alpha_i$  for student  $i$  will increase the log-odds above the population mean  $\mu_t$ , which in turn means that student  $i$  tends to click more than the mean student as represented by  $\mu_t$ . A negative value of  $\alpha_i$  has the opposite effect; student  $i$  has a lower probability of clicking compared

<sup>1</sup>A changepoint methodology using a continuous-time model could in principle also be developed in a manner similar to the discrete-time methodology we describe in this paper.

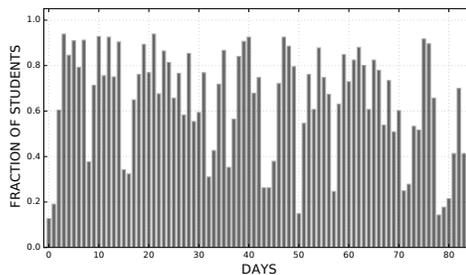


Figure 2: Proportion of students who click each day during a 10-week course.

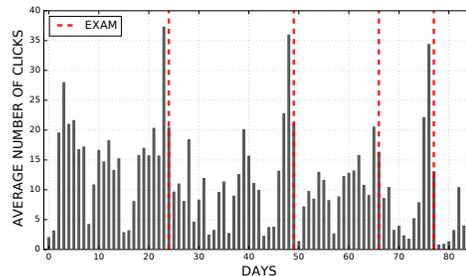


Figure 3: Average number of click events per student each day during a 10-week course.

to the average student.  $\mu_t$  represents time-varying population behavior on a log-odds scale.

Our approach to change detection relies on modeling each student’s activity relative to that of the overall student population in the class. This population (or background) rate  $\mu_t$  will typically vary significantly as a function of time  $t$  since student behavior is strongly affected by temporal effects such as days of lectures, weekday versus weekend effects, assignment deadlines, exams, and so on. As an example, Figure 2 shows the proportion of students who clicked on a file each day, summarizing the data shown earlier in Figure 1.

Modeling the log-odds as a linear function (Equation 1) is a standard technique in generalized linear modeling and ensures that the resulting probability  $\pi_{it}$  above lies between 0 and 1, i.e., Equation 1 above can be rewritten as

$$\pi_{it} = \frac{1}{1 + e^{-(\mu_t + \alpha_i)}}. \quad (2)$$

#### 3.3 Estimation of Model Parameters

The parameters  $\mu = \{\mu_1, \dots, \mu_T\}$  and  $\alpha = \{\alpha_1, \dots, \alpha_N\}$  are estimated from the  $N \times T$  data array  $X'$  with entries  $x'_{it} \in \{0, 1\}, 1 \leq i \leq N, 1 \leq t \leq T$ . Since the  $x'_{it}$ ’s are binary the likelihood for each individual data point  $x'_{it}$  can be written as:

$$L(\mu, \sigma | x'_{it}) = \pi_{it}^{x'_{it}} (1 - \pi_{it})^{(1-x'_{it})}, \quad (3)$$

where  $\pi_{it}$  is defined in Equation 2. The likelihood of the full data set  $X'$  is then defined as:

$$\begin{aligned} L(\mu, \alpha | X') &= P(X' | \mu, \alpha) \\ &= \prod_{i=1}^N \prod_{t=1}^T \pi_{it}^{x'_{it}} (1 - \pi_{it})^{(1-x'_{it})}. \end{aligned} \quad (4)$$

Here we make the assumption that the observed data for

each student on each day is conditionally independent of all other observations (for students and for days) given the parameters  $\mu$  and  $\alpha$ . This is a simplification since it ignores (for example) possible time-varying trends in student behavior. Nonetheless, as we will see later in the experimental results it provides a useful basis for change detection.

We use a two-stage procedure for parameter estimation as follows<sup>2</sup>. We first generate an estimate  $\hat{\mu}_t$  for the population mean as follows:

$$\hat{\mu}_t = \log \frac{\hat{q}_t}{1 - \hat{q}_t}, \quad 1 \leq t \leq T \quad (5)$$

where  $\hat{q}_t = \frac{1}{N} \sum_{i=1}^N x'_{it}$ , which is the proportion of students (across all students) that generated a click on day  $t$ .

In the second step, we fit a regression model for each student  $i$  in Equation 1 with the population mean  $\hat{\mu}_t$  set as an offset.  $\alpha_i$  can be thought of as a student-specific intercept term for each student  $i$ .

### 3.4 Poisson Models for Count Data

We can also model the counts  $x_{it}$  directly, where  $x_{it}$  can have values  $\{0, 1, 2, \dots\}$ . A natural model in this context is the Poisson model.

We develop the count model in a manner similar to that for binary case earlier. In particular, we model the logarithm of the mean of the Poisson distribution,  $\log \lambda_{it}$  as a linear function of a time-varying population rate  $\mu_t$  and an individual student effect  $\alpha_i$ :

$$\log \lambda_{it} = \mu_t + \alpha_i. \quad (6)$$

Note that although for convenience we use the same notation,  $\mu$  and  $\alpha$ , for our two sets of parameters, and they play an analogous role as their “namesake” parameters in the binary model, these parameters are different from those in the binary model described earlier.

Figure 3 shows the average number of click events for each student per day, reflecting the type of time-varying population behavior that  $\mu_t$  is intended to capture. The red dashed lines are the dates for the three midterms and the final, and we can see much more click activities right before the exam dates.

We can write the likelihood function for a single count  $x_{it}$  as

$$P(x_{it}|\mu_t, \alpha_i) = \frac{\lambda_{it}^{x_{it}} e^{-\lambda_{it}}}{x_{it}!}, \quad (7)$$

where  $\lambda_{it}$  is defined in Equation 6. As with the binary case, assuming that the observations  $x_{it}$  are conditionally independent given the parameters, the full likelihood can be written as:

$$\begin{aligned} L(\mu, \alpha|X) &= P(X|\mu, \alpha) \\ &= \prod_{i=1}^N \prod_{t=1}^T \frac{\lambda_{it}^{x_{it}} e^{-\lambda_{it}}}{x_{it}!}. \end{aligned} \quad (8)$$

We again make a conditional independence assumption for the Poisson model. A two-stage parameter estimation process is carried out as before. In the first step we estimate  $\hat{\mu}_t$  as follows:

$$\hat{\mu}_t = \log \hat{m}_t, \quad 1 \leq t \leq T \quad (9)$$

<sup>2</sup>The estimation could be done in a single-step; we would expect similar results to what we obtain in the two-step approach.

where  $\hat{m}_t = \frac{1}{N} \sum_{i=1}^N x_{it}$ , representing the average number of click events across the population that were generated on day  $t$ . In the second step we fit a Poisson regression model for each student  $i$  as in Equation 6 with an offset  $\hat{\mu}_t$  to get an estimate for each  $\alpha_i$ .

### 3.5 Detecting Changes in Activity

To detect changes in activity we allow for the possibility that each student’s activity rate changes at some unknown time point during the course. The proposed approach that we describe below works in the same manner for both the Bernoulli binary model and the Poisson count model, the only difference being in how the likelihood is defined and the parameters are estimated for each (as described earlier). For simplicity, the reader can assume below that we are using either the Bernoulli or Poisson model, and the issue is whether to fit a model with a change or with no change.

We fit two different models for each student  $i$ . The first model is the one where we assume that the student’s rate of activity  $\alpha_i$ , defined relative to the background activity  $\mu_t$ , does not change over time. In the second model, the changepoint model, we assume that a student’s activity rate switches at some unknown changepoint. We fit both models to the data for each student and use a data-driven model selection technique to select which model is justified given the observed data.

In the changepoint model we assume that there is one activity rate  $\alpha_{i1}$  for student  $i$  before changepoint  $\tau_i$  and a different activity rate  $\alpha_{i2}$  after the changepoint  $\tau_i$ . The changepoint model for binary data (for example) can be written as follows, where  $\mathbf{I}$  is an indicator function:

$$\log \frac{\pi_{it}}{1 - \pi_{it}} = \mu_t + \alpha_{i1} \mathbf{I}(t < \tau_i) + \alpha_{i2} \mathbf{I}(t > \tau_i) \quad (10)$$

with a similar definition for the Poisson model. We can interpret this model as fitting two regression models with different means on either side of the changepoint.

The value of the changepoint  $\tau_i$  for each student is unknown. Since time  $t$  is discrete the values of  $\tau_i$  can take one of  $T - 1$  possible values, corresponding to the  $T - 1$  boundaries between the  $T$  observation times.

In effect this changepoint model has 3 parameters (assuming  $\mu_t$  is known): the two activity rates and the changepoint. We generate maximum likelihood estimates of these parameters by maximizing the log-likelihood defined as follows (for each student  $i$ )

$$\begin{aligned} l_i(\alpha_{i1}, \alpha_{i2}, \tau_i, \mu) \\ = \sum_{t < \tau_i} \log P(x'_{it}|\alpha_{i1}, \mu_t) + \sum_{t > \tau_i} \log P(x'_{it}|\alpha_{i2}, \mu_t) \end{aligned} \quad (11)$$

(with a similar equation for counts  $x_{it}$  and the Poisson model).

To fit this model, we use a similar two-stage approach as for the model with no-change described earlier. In the first stage we fit the background rate  $\mu_t$  using the data across all students, in the same manner as for the no-change model. In the second stage we find the values  $\alpha_{i1}, \alpha_{i2}, \tau_i$ , for each student  $i$ , that maximize the log-likelihood defined above. Since  $\tau_i$  is discrete we can reduce the optimization problem to finding the values of  $\alpha_{i1}$  and  $\alpha_{i2}$  for a fixed  $\tau_i$  and then iterate over the  $T - 1$  possible values of  $\tau_i$ . For each fixed value of  $\tau_i$ , the log-likelihood splits into the two parts on the right-hand side of Equation 11 above, a log-likelihood term containing  $\alpha_{i1}$  and a second log-likelihood term con-

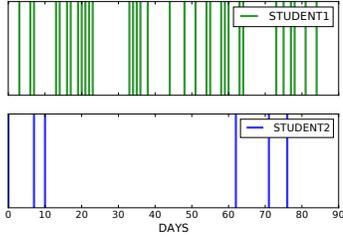


Figure 4: Simulated activity data for two students.

taining  $\alpha_{i2}$ . Each can be optimized independently using the same procedure described earlier for estimating  $\alpha_i$  for the no-change model.

For each student  $i$ , once the parameters of both the no-change and the changepoint models have been estimated, we select the best model from the two candidate models. The likelihood (or log-likelihood), evaluated at the maximum likelihood values of the parameters, is not useful for model selection since the changepoint model will always have a likelihood value that is at least as high as the no-change model (this is because the changepoint model contains the no-change model as a special case).

There are a variety of model selection techniques in the statistical literature to handle the issue of how to fairly compare models (in the case where models have different numbers of parameters) including techniques such as penalized likelihood, Bayesian criteria, and cross-validation [4]. In the results in this paper we use the Bayesian Information Criterion (BIC) which is a well-established and easily interpretable method for model selection. The BIC score is defined for each student as

$$BIC_{iM} = -2l_{iM} + p_M \log T \quad (12)$$

where  $M$  indicates a particular model ( $M = 1$  corresponds to the no-change model, and  $M = 2$  corresponds to the changepoint model),  $l_{iM}$  is the log-likelihood for model  $M$  for student  $i$ 's data evaluated at the maximum likelihood values of the parameters,  $p_M$  is the number of parameters in each model ( $p_1 = 1$ ,  $p_2 = 3$ , for the no-change and changepoint models respectively)<sup>3</sup>, and  $T$  is the number of observations per student. The second term in the BIC,  $p_M \log T$ , can be interpreted as a penalty for having additional parameters in a model.

The BIC method selects the model with the lowest BIC score for each student. In particular, in the context of our changepoint application, we can use BIC to detect if there is evidence that a student's rate of activity changed, i.e., if  $BIC_{i2} < BIC_{i1}$  then the evidence supports the changepoint model over the no-change model for student  $i$ .

## 4. RESULTS FOR SIMULATED DATA

To illustrate how the change-detection methods work, we simulated daily binary time-series of student click activity for 400 students over 85 days (numbers that are roughly similar to the larger of the two classes we analyze later in the paper). The true population rate  $\mu_t$  switched between two different values over time, one with a high rate and one with

<sup>3</sup>Technically we should also count the background model parameters  $\mu$  here, but since this is the same for both models we can omit it.

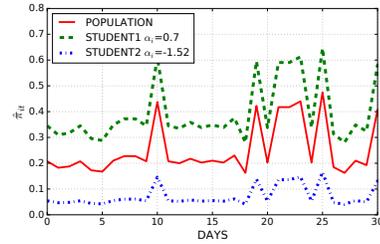


Figure 5: Estimated activity probabilities ( $\hat{\pi}_{it}$ ) for the two simulated students and for the population.

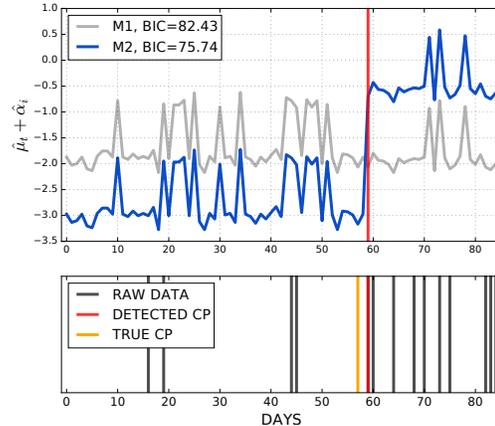
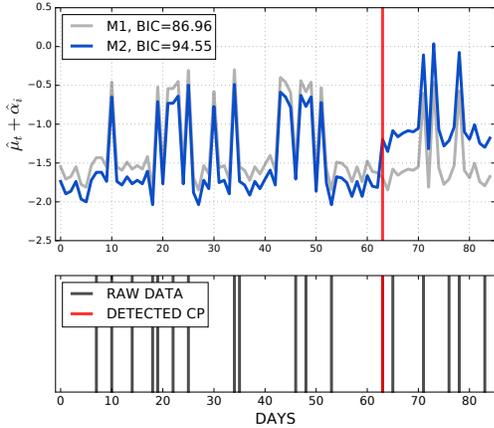


Figure 6: Log-odds of  $\hat{\pi}_{it}$ , for M1 and M2, and simulated data of a student with a changepoint at  $t = 57$ .

a low rate. The variability in the simulation roughly corresponds to what we observed in the real student data. The offsets,  $\alpha_{ij}$ , for each student were sampled independently from a normal distribution;  $\alpha_{ij} \sim Normal(0, \sigma = 1.5)$ . Half of the students were simulated with one  $\alpha_{i1}$ , i.e. no change in behavior over time. The other half of the students had two different offsets sampled,  $\alpha_{i1}$  and  $\alpha_{i2}$ , on either side of a changepoint  $\tau_i$  which was sampled independently from a uniform distribution;  $\tau_i \sim U(15, 70)$ .

Figure 4 is a plot of binary data for two simulated students who did not have changepoints. Student 1 is much more active than Student 2, and therefore Student 1 is going to have a larger estimated value for  $\alpha_i$ . The estimated  $\pi_{it}$ 's for these students over the first 30 days (time  $t = 1, \dots, 30$ ) is shown in Figure 5. The plot illustrates how the estimated activity varies relative to the population probability  $\mu_t$  (the red solid curve). The more active student (green dashed line) has higher probabilities of clicking over time, while the less active student (blue dotted line) has lower probabilities, and both probabilities rise and fall relative to the behavior of the population. For example, when student activity on average rises on a particular day such as day 10 (e.g., due to an assignment), the click probability for both students rises.

Next we show the results of two different simulated students, one with a changepoint and the other without a changepoint. Figure 6 is a plot from a student with a changepoint, with the raw data in the lower plot and the fitted model (plotted on a log-odds scale) in the upper plot. There is a clear change in student behavior around day 59, and this is visible both in the raw data (lower plot) and the fitted



**Figure 7: Log-odds of  $\hat{\pi}_{it}$ , for M1 and M2, and simulated without a changepoint.**

changepoint model (top plot). The BIC for the model with the changepoint ( $M = 2$ ) is significantly smaller than that of a model without the changepoint ( $M = 1$ ) for this simulated student’s data, i.e., the BIC method was able to successfully detect that a model with a changepoint is preferred over a model with no changepoint for this data. In contrast, Figure 7 displays the results for a simulated student with no changepoint. Both plots show the results of fitting the changepoint model M2. The changepoint model puts a change at day 63, but the BIC method selects the no-changepoint model since the BIC for no change  $M = 1$  is much smaller than that of the changepoint model  $M = 2$ .

The BIC method for binary simulated data reliably detected changepoints when the magnitude of the change in  $\alpha_{i1}$  and  $\alpha_{i2}$  (before and after the changepoint  $\tau_i$ ) was relatively large, but as the change became smaller it became more conservative. Out of the 400 simulated cases, BIC detected a change in 100 cases, with a precision of 91% (91 out of the 100 detected were true changes) and a recall of 46% (91 out of the 200 true changes were detected). The remaining 54% of true changes had much lower magnitude changes (0.96 on average) compared to the detected cases (magnitude 2.61 change on average).

## 5. CLICKSTREAM DATA SETS

The clickstream data that we used in our study was recorded via the Canvas learning management system (LMS). Canvas is an open-source LMS that serves as a supplemental instructional technology for students. It has been adopted as the campus-wide LMS system by a number of US universities, including UC Irvine. Students use Canvas to download course content, take online quizzes, watch videos, and submit assignments. The most common data available are clickstream data; every time a student clicks on a URL within the Canvas LMS, the click is recorded and logged with student ID, URL, and time-stamp.

Canvas provides an application programming interface (API) which we used to extract a log of all Canvas clickstream data for a course in addition to other relevant course data such as a list of lecture files, pages, and etc. Our Canvas API crawler for non-clickstream portion of the data is publicly available on github at <https://github.com/dkloz/canvas-api-python>.

**Table 1: Number of students who showed increase, decrease, or no change in their activities for each activity data type for the 10-week face-to-face course.**

| Event Type      | $N_{Increase}$ | $N_{Decrease}$ | $N_{NoChange}$ |
|-----------------|----------------|----------------|----------------|
| Preview, binary | 7              | 9              | 361            |
| Preview, count  | 112            | 96             | 169            |
| Review, binary  | 39             | 23             | 315            |
| Review, count   | 121            | 159            | 97             |

Two types of student behavior are considered for the students described in this paper. The first is a student’s previewing behavior. An event is defined as a Preview Event when a student views or downloads a file prior to the event start date. This could indicate how well a student is performing in terms of being prepared for the course. The second type of behavior is related to the students’ reviewing activities. A Review Event is defined as an event when a student views or downloads a file after the event end date, e.g., a student downloading a lecture file after the class in which the material was covered. We found that focusing on these two types of events allowed us to screen out less relevant information in the clickstream data and extract more meaningful information about students’ activities. While in this paper we will focus on events related to our definitions of previewing and reviewing activity, our methodology for change detection is applicable to arbitrary sets of clickstream events.

Extracting each of the previewing and reviewing events results in an activity matrix of size of  $N \times T$  matrix, where the cell  $i, t$  indicates that the number of previewing or reviewing events by student  $i$  on day  $t$ . The data can be binarized to create a binary representation for the Bernoulli model described in section 3.2.

We used data sets from two courses at UC Irvine in our study, both offered during the 2015-2016 academic year. The first is a face-to-face 10-week course with 377 enrolled students. Lectures were held three times a week, and there were 3 midterms and one final exam. Figure 3 shows the average number of click events on each day per student. There is significant variation in students’ clicking activity over time. For example, students tended to be much more active during days close to the exams (shown as red dashed lines).

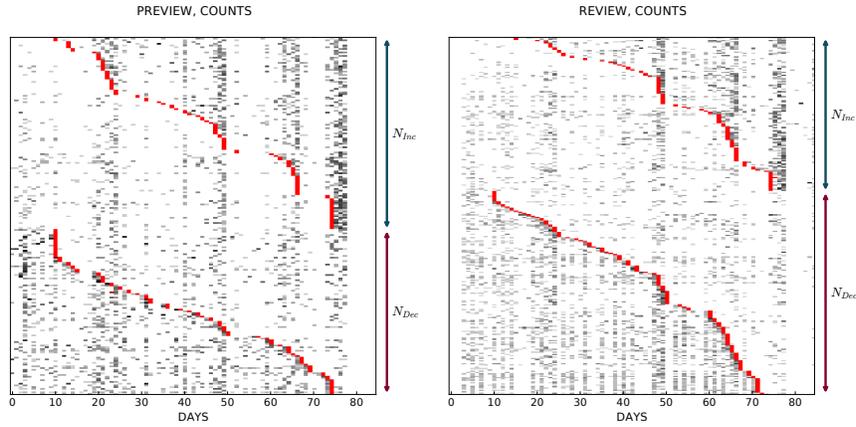
The second data set is somewhat different from the first in that it was an online course offered for 5 weeks. There were 176 enrolled students in this course. This data set is significantly smaller than that for the first course both in terms of the number of students  $N$  and the number of days  $T$ . There were 25 video lectures in total and students were supposed to watch one lecture per day from Monday through Friday. The final exam was held on campus after the 5 lecture weeks.

## 6. EXPERIMENTAL RESULTS

In this section we will discuss the application of our change detection methodology to the two clickstream data sets described in the previous section.

### 6.1 Example 1: 10-Week Face-to-Face Course

The clickstream data spanned 85 days, which included 10 weeks of instruction as well as activity before and after the 10 weeks. We applied our change-detection methodology to



**Figure 8:** Student preview and review activity data over time, for the students who increased or decreased their behavior in the 10-week face-to-face course. The gray marker at  $t$ -th column in each row means that there was click activity on day  $t$  for that student, with darker colors reflecting larger counts (more clicks).

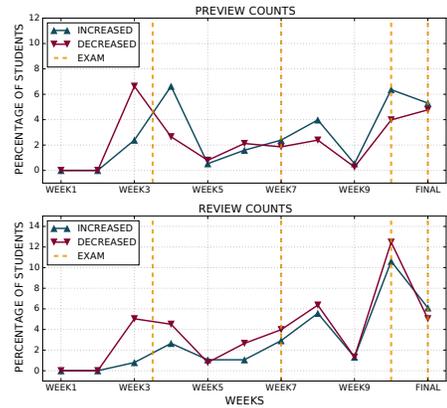
4 different versions of the  $N \times T$  data matrices: for preview and review events, in binary and count form. We restricted changepoints to be in the range from day 10 to day 75, since changepoint detection at the beginning or end of the sequences (i.e., outside of this range) tends to be unreliable due to small sample sizes and not so meaningful in terms of interpreting actual student behavior.

The students that were considered to have changed by the BIC scores were categorized into two groups: students who increased their click activity and students who decreased their click activity. We will refer to these groups as “Increased” and “Decreased” respectively. Note that these terms should be interpreted in a relative sense, since increase and decrease is for the  $\alpha_i$  coefficient for each student *relative to the background rate*  $\mu_t$ . Thus, a detected increase for student  $i$  means in effect that the student is ranked higher in the class in terms of activity relative to other students after the changepoint  $\tau_i$ , compared to their rank before  $\tau_i$  (and conversely for a decrease).

The numbers of students detected as belonging to each group, for each event type, are shown in Table 1. The Poisson count model detects significantly more student changes than the Bernoulli binary model, for both preview and review event types. This is to be expected since the Poisson model has more information to work with (and thus has better sensitivity) compared to the Bernoulli model which only sees a binarized version of the daily counts (and thus has less information per day about student activity). In the discussion below we focus primarily on the Poisson results with counts given its better sensitivity.

Figure 8 shows the click data for each of the students for which a change was detected, with one student per row, and one plot per type of event (Preview and Review). The students are split into two groups within each plot depending on whether their detected changes were increases or decreases, and rows were then ordered within each group based on the chronological location of the changepoint per student. The changepoint locations are marked in red and the plots show a clear distinction between the days with more activity and the days with less activity.

Figure 9 provides a week-by-week summary of the information in Figure 8, showing the number of detected student



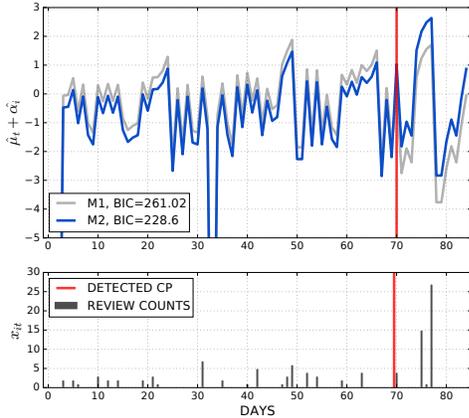
**Figure 9:** Percentage of 10-week face-to-face course students who increased or decreased within each week.

changes per week, for each type of event. The vertical lines that are visible in the two count matrices are the exam dates. There are some obvious temporal patterns in this data. For example, the upper plot (preview events) shows that more than a quarter of the students increased their previewing activities in the third week, which is the week before the first midterm. This agrees with the intuition that prior to the first major exam in a class we would expect to see some significant shifts in student activity. The lower plot shows that the most of the changes in reviewing activity happened towards the end of the quarter, particularly during week 10 before the final exam. Again it makes sense that there are significant changes across students in their relative rates of reviewing activity prior to the final exam. We can also see in both plots that the number of detected changes per week, for increases and for decreases, are strongly correlated. As mentioned earlier this is to be expected with this model since increase and decrease for this model is defined relative to overall mean population behavior.

We also investigated how detected changes in preview and review activities were correlated with student outcome in

**Table 2: Probability of a student getting a passing grade (A, B, C) depending on which group the student is in.**

|                         | $P(\text{Pass} \text{Inc})$ | $P(\text{Pass} \text{Dec})$ | $P(\text{Pass})$ |
|-------------------------|-----------------------------|-----------------------------|------------------|
| Probability             | 0.93                        | 0.76                        | 0.83             |
| $\Delta\text{Pass}$ (%) | 12.1                        | -7.4                        | 0                |
| p-value                 | 0.0025                      | 0.0458                      | -                |

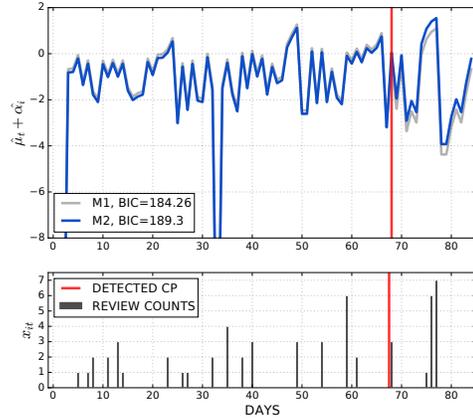


**Figure 10: Log of  $\hat{\lambda}_{it}$  from M1 and M2, and the raw data of a student from the 10-week face-to-face course. The BIC method selected the model with changepoint (M2).**

terms of the students’ final grades in the class. We calculated the probability of a student getting a passing grade given that the student is in the Increased group,  $P(\text{Pass}|\text{Increase})$ , or in the Decreased group,  $P(\text{Pass}|\text{Decrease})$ , and compared these numbers with the marginal (unconditional) probability of a student passing  $P(\text{Pass})$ . For both preview and review count events we used a two-sided binomial test with  $P(\text{Pass})$  as the null hypothesis to compute p-values for  $P(\text{Pass}|\text{Increase})$  and  $P(\text{Pass}|\text{Decrease})$ .

Table 2 shows the results for review count data. At the 0.01 level of significance,  $P(\text{Pass}|\text{Increase})$  is significant and  $P(\text{Pass}|\text{Decrease})$  is significant at the 0.05 level. Students in the Increased group have a higher probability of passing the course, while the students in the Decreased group have a higher probability of failing. This means that students who increased their reviewing behavior (relative to all of the students in the course), at some point during the quarter, ended up getting better grades on average than those that did not. For preview counts, the probabilities were also in the direction of increases in previewing leading to better outcomes on average (and vice versa), but these changes were not statistically significant. This may suggest, for this particular course, that changes in review activities are better predictors of student outcomes than preview activities.

Finally, for the 10-week course, we analyzed in more detail the results for two specific students (using their Review data) to illustrate how the model can be used to interpret clickstream activity at the individual student level. Figure 11 illustrates the results for a student where the lower plot shows the observed daily review clicks, and the upper plot shows the Poisson models for the no-change model and



**Figure 11: Log of  $\hat{\lambda}_{it}$  from M1 and M2, and the raw data of a student from the 10-week face-to-face course. The BIC method selected the no-changepoint model (M1).**

**Table 3: Number of students who showed increase, decrease, or no change in their activities for each activity data type for online course.**

| Data Type       | $N_{\text{Increase}}$ | $N_{\text{Decrease}}$ | $N_{\text{NoChange}}$ |
|-----------------|-----------------------|-----------------------|-----------------------|
| Preview, binary | 6                     | 8                     | 162                   |
| Preview, counts | 41                    | 40                    | 95                    |
| Review, binary  | 11                    | 6                     | 159                   |
| Review, counts  | 47                    | 66                    | 63                    |

the changepoint model (with a detected change at day 70). For this student the BIC method preferred the changepoint model over the no-change model, with  $BIC_2 < BIC_1$  by a large margin. This is reflected in the observed data in the lower plot where the number of counts for this student increase significantly after the changepoint.

Figure 11 shows the same type of plot for a student where the BIC method selected the model without the changepoint. From the raw counts (lower plot) it looks like the student’s activity level could have changed (increased) after day 68. However, relative to the background activity (particularly around days 76 to 78, leading up to the final exam) this student’s activity level is not sufficiently different to the mean population behavior to justify the additional parameters in the changepoint model, as reflected in the BIC scores ( $BIC_1 < BIC_2$ ).

## 6.2 Example 2: Online 5-Week Course

The second course we analyzed was a 5-week online summer course. The event data set for this course we analyzed was smaller than the first in terms of both the number of students ( $N = 176$ ) and number of days with clickstream activity ( $T = 50$ ). The course was offered online and the students were expected to watch a lecture video on every weekday over the 5 weeks, leading to more uniformity and less variability in student clickstream activity over time. In addition, the 10-week class had 3 midterm exams and a final exam, while the 5-week online class only had a single final exam at the end of the course.

The numbers of students detected for each of the Increased

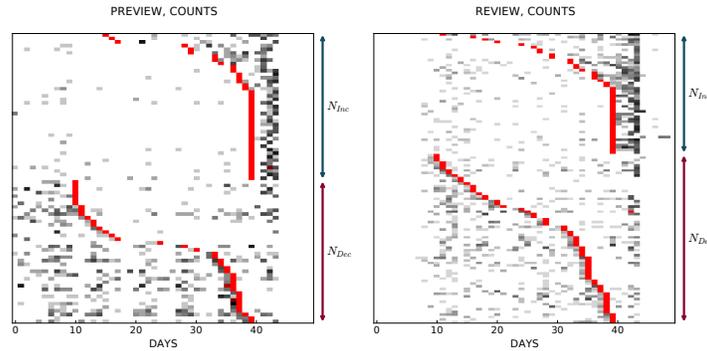


Figure 12: Student preview and review activity data over time, for the students who increased or decreased their behavior in the 5-week online course. The gray marker at  $t$ -th column in each row means that there was click activity on day  $t$  for that student, with darker colors reflecting larger counts (more clicks).

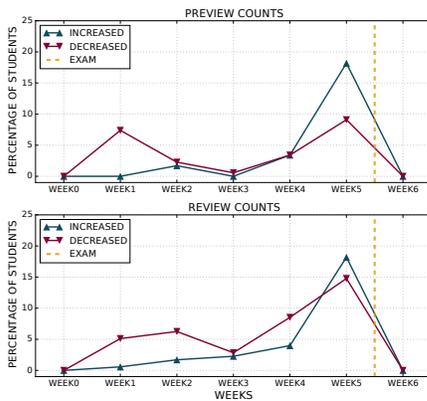


Figure 13: Percentage of students with detected increase or decrease in activity for each week in the online 5-week course.

and Decreased groups, for both preview and review events, are shown in Table 3. We see a similar overall pattern to that for the 10-week class, namely that the Poisson model using counts detects considerably more changes than the Bernoulli method using binary data. The overall proportions of changes detected are roughly similar across both classes, with about 50% of students having increased or decreased count activity relative to the population, for each of the two types of events. One difference we found between the two courses was the proportion of students who exhibited no change at all, for either preview or review events: 13% of students in the 10-week course and 25% in the 5-week courses. This difference might be due to the intermediate exams (3 midterms) in the 10-week course, leading to more variability in student behavior compared to the 5-week course which only had a final exam.

The clickstreams for the students with detected changes are shown in Figure 12. We observe very high activities at the end of the course session for students in the Increased group, for both Preview and Review event types. The majority of the changepoints occur just before the darker area of the plot. Figure 13 shows that, among the students who had an increased change that most of them had a changepoint in the fifth week, which is the last week of the course before the

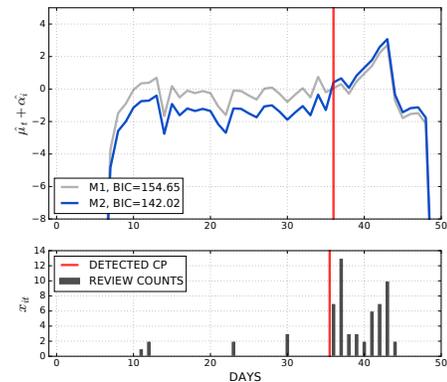


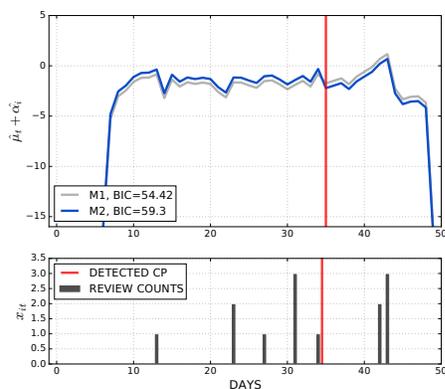
Figure 14: Log of  $\hat{\lambda}_{it}$  from M1 and M2, and the raw data of a student from the 5-week online course. The BIC method selected the changepoint model (M2).

final. We did not analyze the relationship of click activity and course outcomes for this course since fewer than 5% of the students received grades of D or F in the class, resulting in a sample size that is too small for reliable inferences.

As with the 10-week class, we examine the results for review events for 2 specific students, to illustrate the methodology at the level of individual students. Figure 14 shows the results for a student where the method detected a change in activity at day 35. Figure 15 shows the results for a student where the no-change model was preferred by BIC. Both students exhibited increases in their review activities after day 40, but the magnitude of change for the first student is significantly greater than that for the second student (as can be seen in the lower panels of both plots)—relative to the student population as a whole, the second student did not exhibit a significant change in activity.

## 7. CONCLUSIONS AND FUTURE WORK

Student clickstream data is inherently difficult to work with given its complex and noisy nature. This paper described a statistical methodology for detecting changepoints in such data and illustrated the potential of the approach by applying the methodology to two large university courses. The proposed approach is relatively simple and allows for



**Figure 15: Log of  $\hat{\lambda}_{it}$  from M1 and M2, and the raw data of a student from the 5-week online course. The BIC method selected the no-change-point model (M1).**

a number of possible extensions; the development of more flexible changepoint models (such as systematic drifts in student activity levels), allowing for more than a single changepoint, post hoc adjustments for multiple testing, and using robust estimation techniques for parameters and their respective standard errors. Bayesian methods could also be potentially useful in this context for both parameter estimation and model selection to more fully reflect uncertainty in inferences at the individual student level. A useful extension for educators would be to develop an online detection variant of the offline approach proposed here, potentially allowing for identification of at-risk students, instructor feedback, or interventions while a course is in session.

While the results in this paper are promising and there are interesting methodological avenues to pursue, the most important future direction from an education research perspective will involve more in-depth investigation of the utility of these types of methods in terms of providing actionable insights that are relevant to the practice of education.

## Acknowledgments

This paper is based upon work supported by the National Science Foundation under Grants Number 1535300 (for all authors) and 1320527 (for PS). The authors would like to thank Sarah Eichhorn, Wenliang He, and Dimitris Kotzias for their assistance in acquiring and preprocessing of the clickstream data used in this paper.

## 8. REFERENCES

- [1] Y. Bergner, D. Kerr, and D. E. Pritchard. Methodological challenges in the analysis of MOOC data for exploring the relationship between discussion forum views and learning outcomes. In *Proceedings of the EDM Conference*, pages 234–241. International Educational Data Mining Society (IEDMS), 2015.
- [2] S. Bhat, P. Chinprutthiwong, and M. Perry. Seeing the instructor in two video styles: Preferences and patterns. In *Proceedings of the EDM Conference*, pages 305–312. International Educational Data Mining Society (IEDMS), 2015.
- [3] C. G. Brinton and M. Chiang. MOOC performance prediction via clickstream data and social learning networks. In *Proceedings of the INFOCOM Conference*, pages 2299–2307. IEEE, 2015.
- [4] G. Claeskens. Statistical model choice. *Annual Review of Statistics and its Application*, 3:233–256, 2016.
- [5] S. Crossley, L. Paquette, M. Dascalu, D. S. McNamara, and R. S. Baker. Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, LAK '16*, pages 6–14. ACM, 2016.
- [6] D. Davis, G. Chen, C. Hauff, and G.-J. Houben. Gauging MOOC learners’ adherence to the designed learning path. In *Proceedings of EDM Conference*, pages 54–61. International Educational Data Mining Society (IEDMS), 2016.
- [7] I. A. Eckley, P. Fearnhead, and R. Killick. *Bayesian Time Series Models*, chapter 10 Analysis of changepoint models, pages 205–224. Cambridge University Press, Cambridge, 2011.
- [8] P. Hofgesang and J. P. Patist. Online change detection in individual web user behaviour. In *Proceedings of WWW Conference*, pages 1157–1158. ACM, 2008.
- [9] C. Kirch and J. Tajdudje Kamgaing. Detection of change points in discrete valued time series. In R. Davis, S. Holan, R. Lund, and N. Ravishanker, editors, *Handbook of Discrete Valued Time Series*, chapter 11, pages 219–244. Chapman and Hall, 2014.
- [10] G. D. Kuh, T. M. Cruce, R. Shoup, J. Kinzie, and R. M. Gonyea. Unmasking the effects of student engagement on first-year college grades and persistence. *The Journal of Higher Education*, 79(5):540–563, 2008.
- [11] C. Learning. Community insights: Emerging benchmarks and student success trends from across the civitas. Technical report, December 2016.
- [12] Z. Liu, R. Brown, C. Lynch, T. Barnes, R. S. Baker, Y. Bergner, and D. S. McNamara. MOOC learner behaviors by country and culture; an exploratory analysis. In *Proceedings of the EDM Conference*, pages 127–134. International Educational Data Mining Society (IEDMS), 2016.
- [13] C. Mulryan-Kyne. Teaching large classes at college and university level: Challenges and opportunities. *Teaching in Higher Education*, 15(2):175–185, 2010.
- [14] K. H. R. Ng, K. Hartman, K. Liu, and A. W. H. Khong. Modelling the way: Using action sequence archetypes to differentiate learning pathways from learning outcomes. In *Proceedings of the EDM Conference*, pages 167–174. International Educational Data Mining Society (IEDMS), 2016.
- [15] G. Wang, X. Zhang, S. Tang, H. Zheng, and B. Y. Zhao. Unsupervised clickstream clustering for user behavior analysis. In *CHI Proceedings*, pages 225–236. ACM, 2016.
- [16] X. Wang, D. Yang, M. Wen, K. R. Koedinger, and C. P. Rose. Investigating how student’s cognitive behavior in MOOC discussion forum affect learning gains. In *Proceedings of the EDM Conference*, pages 226–233. International Educational Data Mining Society (IEDMS), 2015.