Mining MOOC Clickstreams: Video-Watching Behavior vs. In-Video Quiz Performance

Christopher G. Brinton, *Student Member, IEEE*, Swapna Buccapatnam, Mung Chiang, *Fellow, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*—Student video-watching behavior and quiz performance are studied in two Massive Open Online Courses (MOOCs). In doing so, two frameworks are presented by which video-watching clickstreams can be represented: one based on the sequence of events created, and another on the sequence of positions visited. With the event-based framework, recurring subsequences of student behavior are extracted, which contain fundamental characteristics such as reflecting (i.e., repeatedly playing and pausing) and revising (i.e., plays and skip backs). It is found that some of these behaviors are significantly correlated with changes in the likelihood that a student will be Correct on First Attempt (CFA) or not in answering quiz questions, and in ways that are not necessarily intuitive. Then, with the position-based framework, models of quiz performance are devised based on positions visited in a video. In evaluating these models through CFA prediction, it is found that three of them can substantially improve prediction quality, which underlines the ability to relate this type of behavior to quiz scores. Since this prediction considers videos individually, these benefits also suggest that these models are useful in situations where there is limited training data, e.g., for early detection or in short courses.**

***Index Terms*—Clickstream data, data mining, performance prediction, MOOC, learning analytics, social learning networks.**

I. Introduction

OVER the past decade, technology advances have been influencing the ways we can learn. One of the prominent innovations has been the Massive Open Online Course (MOOC). MOOC providers such as Coursera, edX, and Udacity have offered courses reaching out to tens and even hundreds of thousands of students within single sessions [1].

One salient feature of MOOCs is high dropoff rates, with typically less than 10% of students that initially enroll in a course receiving a certificate of completion at the end [2]. This has motivated several research studies in recent years on student engagement, and on identifying factors associated with participation levels (*e.g.,* [3]–[7]). Various factors have been cited as affecting engagement in turn; for one, the open nature of MOOCs attracts a diverse set of students with a range of motivations for why they enrolled in a course in the first place [2], [8], [9], with only a fraction identifying completion as one of their goals. Among those targeting completion, other factors that have been claimed to lower retention are the high student-to-teacher ratios exhibited in these courses, and the asynchronous nature of interaction between students and instructors [2].

Despite the challenges that MOOCs present, the data that these platforms collect brings substantial opportunities to study the process of student learning. The backend infrastructures driving them capture detailed measurements on students as they interact with the different forms of learning integrated into the courses. To see this, consider the standard three learning modes that MOOC platforms offer to students: video lectures, assessments (*e.g.,* in-video quizzes, homework assignments, and exams), and social networking (usually through discussion forums) [2]. For video content, individual clickstream events are captured, with a click event generated and stored each time a student interacts with a video, specifying the particular action (*e.g.,* pause, ratechange, skip), position, and time at which it occurred. For assessments, the specific responses to individual questions are recorded, and for the discussion forums, all posts and comments are stored.

After a few years and thousands of course offerings through these platforms, researchers have begun to take data mining approaches to studying MOOC student behavior (*e.g.,* [5], [10]–[13]). This data is also motivating the design of mechanisms to help improve student outcomes in MOOCs. Notable examples include algorithms (i) for early detections of students with high likelihoods of dropping out of the course before its completion [4], [11] and of performing poorly on quizzes/exams [12], [14], (ii) for recommendations of discussion participation [5] and of certain peer grading allocations [15], and (iii) for individualization of the content delivered to each student through machine-learning-based user modeling [2].

What remain understudied are the *relationships* between how students interact on the different learning modes. In particular, we are interested to know: *How is a student's behavior related to his/her performance in a MOOC?* Developing such an understanding would have implications not only for theories about how humans process information, but also for systems aiming to improve student learning experiences. For one, early detection performance prediction systems that are usually driven by past performance history—which tends to be a sparse source of information in MOOCs—could be augmented with behavioral signals that were identified as being correlated with low or high student performance [12], [14]. Additionally, algorithms for updating user models in individualization could be expanded to include behavioral signals in making determinations as to the most suitable path of learning for each student to take, as is suggested in [2]. Furthermore, these relationships could be

Manuscript received September 30, 2015; revised January 29, 2016; accepted March 2, 2016. Date of publication March 24, 2016; date of current version May 25, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tsung-Hui Chang. This work was in part supported by the Army Research Office under MURI Grants W911NF-11-1-0036 and W911NF-14-1-0190.

C. G. Brinton, M. Chiang, and H, V. Poor are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: cbrinton@princeton.edu; chiangm@princeton.edu; poor@princeton.edu).

S. Buccapatnam is with IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: sbuccap@us.ibm.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2016.2546228

provided to course instructors directly, in the form of extended learning and content analytics [16], [17]. The behavioral signals could give instructors insight into which parts and/or types of their content are causing confusion.

A. Measuring Behavior and Performance

Our work is motivated by the question of how behavior is related to performance in MOOCs. To do this, we will employ two datasets coming from two different MOOCs we have instructed on Coursera, described in Section II-A. We measure behavior and performance from our datasets as follows:

Behavior. We focus on the behavior students exhibit while watching lecture videos. This is the dominant mode of instruction that we provided in our courses, and is where users spend the majority of their time on MOOC platforms [10], [13]. These behaviors are captured through clickstream logs, which we detail in Section II-B.

Performance. Our videos are equipped with in-video quiz questions, which are short multiple-choice exercises that we designed to test a student's knowledge recall of the content in the video before he/she proceeds. Our measures of performance are the scores that students obtain on their first attempts at these quizzes, *i.e.,* whether they are Correct on First Attempt (CFA) or not (non-CFA). We use the in-video questions because they serve as immediate feedback of the knowledge a student gained from the behavior they exhibited in the video, thereby reducing the effect of confounding factors (*e.g.,* reviewing other materials, varying information retention based on a student's innate cognitive ability).[1] The first attempt has also been selected in other works as an objective measure of performance (in *e.g.,* [12], [18]).[2]

Our goal, then, is to relate video-watching behavior to in-video quiz performance. After filtering (see Section II-A), our datasets contain 315 K and 416 K clickstream events corresponding to 26 K and 36 K first-attempt quiz submissions.

B. Research Objectives and Contributions

In this paper, we formalize different ways that video-watching clickstreams can be represented as sequences, and apply our frameworks to meet two specific objectives (O1&2):

1) *O1*: To identify recurring video-watching behaviors of students, such as reviewing content or skipping forward repeatedly.
2) *O2*: To assess the impact of video-watching behavior on in-video quiz performance, *i.e.,* how patterns identified in O1 and specific positions visited in each video are signals of knowledge gained.

A few methods for representing student video-watching clickstreams have been proposed previously. Some have taken a higher-level approach and computed aggregate, summary quantities of the behaviors (*e.g.,* fraction of video completed, duration of pause) [12], [13], others have looked at the most frequently visited video positions [10], and others have searched for sequences of events in the clickstreams (*e.g.,* play, then skip, then pause) [11]. Motivated by this, in studying O1&2, we develop frameworks for representing clickstreams as sequences that account for (i) the types of events, (ii) the positions in the video that a student visited, and (iii) the duration or length of time between the events and positions.

More specifically, our investigation is broken down into two components: behavioral motifs and behavior-based prediction.

(1) Behavioral motifs. We first develop an event-based framework to represent clickstreams (Section II), which captures event types and their lengths. Leveraging this framework, we are able to identify video-watching *motifs*, *i.e.,* sub-sequences of student behavior that occur significantly often, in our two datasets. These motifs by themselves are informative of recurring behaviors for O1 (Section III), and for O2 we are able to correlate the occurrence of certain motifs in a dataset with a change in the likelihood of CFA through mixed-effects modeling. For example, we find that a series of behaviors are indicative of students reflecting on material, and tend to be associated with an increase in the chance of CFA in one of our courses and of non-CFA in the other. As another example, we identify motifs that are consistent with rapid-paced skimming through the material, and reveal that these are associated with a decrease in the chance of CFA in both of our courses.

For these motifs, the identified positive and negative correlations with CFA are particularly helpful, because for many of them, either case is conceivable. For one, skimming could intuitively be a sign of a student either correctly or incorrectly perceiving familiarity with the material; our results indicate the latter tends to occur more often. Also, we find that incorporating the lengths in addition to the events is important to these findings, because extracting motifs from sequences of events alone does not reveal these insights.

(2) Behavior-based prediction. In investigating O2, we also develop models for knowledge gained based on the clicks that a student makes in a video. The quality of such a model can be evaluated by considering its ability to generalize to incoming samples through prediction. The higher the quality, the stronger the association between video-watching behavior and quiz performance. To this end, we study student performance prediction (specifically, CFA prediction) for MOOCs. Enhancing CFA prediction is an important area of research in its own right, because such methods can improve systems for early detection *e.g.,* of struggling/advanced students and of easy/difficult material [12], [14].

In seeking appropriate models for behavior-based prediction, we find that while some behavioral patterns of the motifs are significantly associated with quiz performance, their supports across sequences are not sufficient to make large improvements in online CFA prediction. As a result, we propose a second behavioral representation, which is based on the sequence of positions visited in a video (Section IV). In contrast to training over a long course duration as in [12] and [17], we consider CFA prediction on a *per-video* basis, in order to quantify the benefit obtained by the positions in each individual video and to investigate the application of earliest detection. In evaluating four different algorithms using our framework (Section V), we find that

[1] Besides in-video quizzes, our courses also had machine-graded exams. However, the final exams in our MOOCs are much less suitable for quality evaluation than are the in-video quizzes, because (i) only a small fraction of students actually took them (less than 5% in each case), and (ii) there were only a small proportion of correct submissions since the exam questions were designed to be much more difficult than the in-video quizzes.

[2] In general, slip and guess probabilities [19] could be inferred from subsequent quiz attempts. In our datasets, however, less than 9% of submissions have more than one valid attempt registered.

Dataset	Lectures	Lecture Videos	Video Length (min) avg. (s.d.)	Quizzes	Users	Clickstream Events	User-Video Pairs	CFA Score avg. (s.d.)
'FMB'	20	92	16.9 (5.96)	92	3770	314,632	26,250	0.663 (0.473)
'NI'	6	115	5.44 (2.17)	69	2680	416,214	36,464	0.750 (0.433)

Fig. 1. Basic information on the two datasets. The values in the right group of columns are the final numbers after data filtering.

likelihood-based algorithms obtain significant improvements in prediction when compared to a baseline that does not use click information, and that a Support Vector Machine (SVM)-based algorithm also obtains improvements over the baseline (though not as substantially). This underscores the ability to relate clicks to knowledge gained, *i.e.,* that video-watching behavior is related to quiz performance, and shows that behavioral information is useful in situations in which multiple videos are not available, *e.g.,* in short courses or for detection early in a course. Further, since the likelihood-based algorithms are directly based on student behaviors (as opposed to the SVM algorithm which learns a more complex function on top of the behaviors), they can generate analytics about content that are interpretable to the instructor.

As stated, the motif identification and early detection methods we develop are aimed to help improve student quiz performance, and ultimately to improve completion rates. For those students who enroll in a MOOC for reasons other than the traditional goal of passing the course, though, quiz performance and completion are not the right measures of efficacy [20]. It is worth mentioning, then, that the scope of our methods are those students aiming to achieve high scores. We will discuss possible extensions in Section VII.

Summary of Contribution. Compared with other work (Section VI), we make three main contributions in this paper:

1) We develop two new frameworks for representing student video-watching behavior as sequences.
2) We extract recurring motifs of student video-watching behavior using motif identification schemes, and associate these fundamental patterns with quiz performance.
3) We demonstrate that video-watching behavior can be used to enhance student performance prediction on a per-video basis, *e.g.,* for earliest detection.

II. Datasets and Clickstreams

In this section, we describe our datasets, and present our first sequence specification based on events and lengths.

A. Our Two MOOCs

Our datasets come from two different courses that we have instructed on Coursera: *Networks: Friends, Money, and Bytes* ('FMB') and *Networks Illustrated: Principles Without Calculus* ('NI').[3] Each of these courses teach networking topics, but 'FMB' explains the mathematical specifics behind the topics, whereas 'NI' is meant as an introduction to the subject (see [2] for more details). We obtained two types of data from Coursera for each of the courses: (i) video-watching clickstreams, which log user interaction with the video player, and (ii) information on the in-video quiz submissions. We will describe the format of the video-watching clickstreams in detail in Section II-B1, which is where we develop our first representation framework.

Course format. The course formats are summarized in Fig. 1. Each is made up of a series of lectures, which are in turn comprised of a set of videos. 'FMB' is a longer course, with 20 lectures, whereas 'NI' only has 6. 'NI' had more, shorter-length videos, with a total of 115 videos and an average (avg.) length of 5.4 min per video, whereas 'FMB' has fewer, longer-length videos, with 93 total and an average length of 16.9 min.

As discussed in Section I-A, for each course, we included in-video quizzes at the end of the videos, which were designed to test a student's recall of the information discussed in the video. Each quiz is a multiple choice question, in radio-response format, with 4-5 possible answer choices. For 'FMB', there was one question at the end of each video, whereas for 'NI', each of the 69 questions was associated with anywhere from 1-4 videos. In mapping videos to quizzes, we refer to "video n" as the contiguous set of videos occurring after question $n-1$ and before question n.

User-Video Pairs. We extract User-Video (UV) Pairs from the data, with two sets of information for video and quiz n:

- *(i) Video-watching trajectory*: The set of video-watching clickstream logs (events) for the user in video n.
- *(ii) CFA result*: Whether the user was CFA or non-CFA on quiz n.

These UV Pairs are how we measure video-watching behavior and quiz performance in each course, as discussed in Section I-A.

In total, for 'FMB' there were 122.5 K UV Pairs with 566 K click events, and for 'NI' these numbers were 149 K and 882 K, respectively. After removing any UV Pair that had at least one null, stall or error contained in its video-watching trajectory, we obtain the totals given in Fig. 1. The avg. CFA score across the UV Pairs are also shown here: 0.663 for 'FMB' $(\text{standard deviation (s.d.)} = 0.473)$, and 0.750 for 'NI' $(\text{standard deviation} = 0.433)$.

The number of observed UV Pairs is large, but also rather sparse if we consider the numbers that we would have to work with if all the students answered all the questions in each course. In particular, only 7.6% and 19.7% of the possible pairs are present in 'FMB' and 'NI,' respectively.[4] The large number of unanswered questions is one of the challenges to MOOC performance prediction in the first place [12]. We will see in Section V that video-watching behavior improves the quality CFA prediction in the presence of this sparsity.

[3] www.coursera.org/course/{friendsmoneybytes,ni}

[4] Subsequent attempts are even more sparse. If we consider those attempts made by a student on a quiz occurring (i) after the first one, (ii) at least three seconds after the previous one (so it is not obviously a random guess), and (iii) at most two minutes after the previous one (so the student did not obviously move away from the question to look for the answer), only 8.2% and 5.5% of the user-quiz pairs in 'FMB' and 'NI' have more than one attempt.

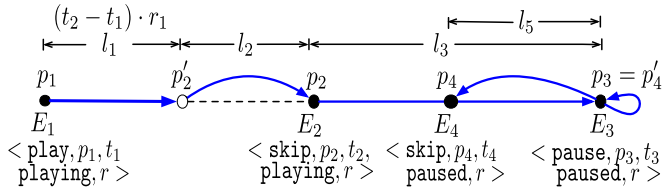


Fig. 2. Illustration of a sequence of clicks E_1 to E_4 on a video, where the horizontal axis denotes the video length. This example will generate 5 events according to the framework proposed in Section II-B based on events and lengths. The length l_j for the events that have this property (not pauses) are depicted above the diagram.

B. Processing Clickstream Events

1) Our Nomenclature for Events: A clickstream log is one of four types: `play`, `pause`, `ratechange`, or `skip`. Each time one of these events is fired, a data entry is recorded that specifies the user and video IDs, event type, playback position, playback speed, and UNIX timestamp for the event.

Formally, let E_i denote the ith click event that occurs while a user is watching a video. We write $E_i = \langle e_i, p_i, t_i, s_i, r_i \rangle$, where e_i is the type of the ith click, p_i is the video position of the player (in seconds) right after E_i is fired, t_i is the UNIX time (in sec) at which E_i was fired, s_i is the state of the video player—either playing or paused—after the click E_i occurs, and r_i is the playback rate (*i.e.,* speed) of the video player resulting from this event. The logs are sequenced chronologically for a UV Pair, *i.e.,* $t_1 < t_2 < \cdots$. Based on the E_i for a UV Pair, we define the following events:

Play (`Pl`): A play event begins at the time when a click event E_i is made for which the state s_i is playing, and lasts until the next click E_{i+1}. It occurs for a duration $d = t_{i+1} - t_i$ and has a length $l = p_{i+1} - p_i$.

Pause (`Pa`): A pause event is defined in the same way as a play event, except it is for which the state s_i is paused, and does not have any length by definition.

Skip back (`Sb`): A skip back (*i.e.,* rewind) event occurs when the type $e_i =$ `skip` and $p_i' > p_i$, where p_i' is the position of the video player immediately before the skip. If $s_{i-1} =$ `playing`, then $p_i' = p_{i-1} + (t_i - t_{i-1}) \cdot r_{i-1}$; if $s_{i-1} =$ `paused`, then $p_i' = p_{i-1}$. The length of the skip is $l = |p_i - p_i'|$, and there is no associated duration.

Skip forward (`Sf`): A skip forward (*i.e.,* fast forward) event is defined as `Sb` is, except it captures the case where $p_i > p_i'$.

Ratechange fast (`Rf`): This occurs when $e_i =$ `ratechange` and the new rate $r_i > 1.0$.[5] There is no duration or length.

Ratechange slow (`Rs`): This occurs when $e_i =$ `ratechange` and $r_i < 1$, again with no duration or length.

Ratechange default (`Rd`): This is when $e_i =$ `ratechange` and $r_i = 1$, *i.e.,* the user is returning to the default speed.

With these definitions, the sequence of events for a UV Pair becomes $(\hat{e}_1, \hat{e}_2, \ldots)$ for $\hat{e}_j \in \mathcal{E} = \{\texttt{Pl}, \texttt{Pa}, \texttt{Sb}, \ldots\}$, $|\mathcal{E}| = 8$. Each $\hat{e}_j$ may have an associated duration parameter d_j and/or length parameter l_j. Fig. 2 shows a schematic for illustration; in this example, the clickstreams would generate: `Pl`, with $l_1 = (t_2 - t_1) \cdot r$ and $d_1 = t_2 - t_1$; `Sf`, with $l_2 = p_2 - p_2'$; `Pl`, with $l_3 = p_3 - p_2$ and $d_3 = t_3 - t_2$; Pa, with $d_4 = t_4 - t_3$; and `Sb`, with $l_5 = p_4' - p_4$. Note that we are inserting `Pl` and Pa events in-between other events, to incorporate the state of the video player during those times. This critical information is not captured through the raw events along, and has been neglected in other work (*e.g.,* in [11]).

De-noising clickstreams. It is important to remove noise in the video-watching trajectories associated with unintentional user behavior. We handle two cases of events separately: *(i) Combining events*: We combine repeated, sequential events that occur within a short duration (5 sec) of one another, since this pattern indicates that the user was adjusting to a final state. This is a common occurrence with forward (`Sf`) and backward (`Sb`) skips, where a user repeats the same action numerous times in a few seconds in seeking the final position; this should be treated as a single skip to the final location. Similarly, a series of rate change (`Rf`, `Rs`, or `Rd`) events may occur in close proximity, indicating that the user was in the process of adjusting the rate to the final value, which should also be treated as a single event. Formally, if there is a sequence of clicks $E_i, E_{i+1}, \ldots, E_{i+K}$ for which $e_i = e_{i+1} = \cdots = e_{i+K}$ and $t_{i+k+1} - t_{i+k} < 5 \ \forall k \in \{0, \ldots, K-1\}$, then we use $E_i' = \langle e_i, p_{i+K}, t_i, s_{i+K}, r_{i+K} \rangle$ in place of $E_i, E_{i+1}, \ldots, E_{i+K}$. *(ii) Discounting intervals*: Clickstream logs are the most detailed accounts of a student's video-watching behavior that are available for online courses today. Even so, it is not possible to determine with complete certainty whether a student is actually watching/focused on the video for the duration of time in-between the occurrence of two events. Still, we can identify two situations in which a `Pl` or Pa event should not be inserted in-between E_i and E_{i+1} to capture the state of the video player. The first situation is if the duration $t_{i+1} - t_i$ is extremely long; in this case, the user was obviously engaging in some off-task behavior during this time. If $s_i =$ `paused`, the threshold on the duration is set to 20 min (as in [21] for web inactivity); if $s_i =$ `play`, then the threshold is set to the length of the video. The second situation is if E_i and E_{i+1} occur on two different videos; here, there is no continuity as the user must have exited the first video and opened the second.

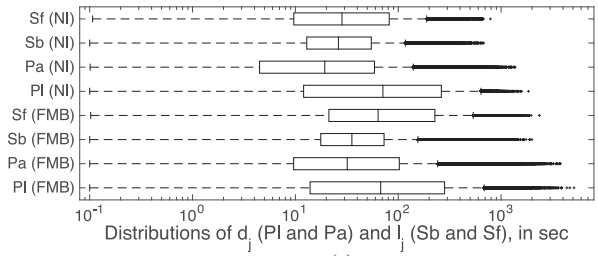
2) Event Lengths: We now look to discretize the length l_j and duration d_j of the events. Fig. 3(a) gives the boxplots of the event distributions from each course. d_j for `Pl` and `Pa` is shown, and we depict l_j for `Sb` and `Sf` (we show only values that are at least 0.1 sec). Basic statistics of each distribution are also given in Fig. 3(b); specifically, the three quartiles Q_1, Q_2, and Q_3 are shown,[6] as are the number of events for each distribution (Size) and the respective fractions (Frac).

We make three high-level observations in comparing the distributions. In each case, we employed a Wilcoxon Rank Sum (WRS) test [22] for the null hypothesis that there was no difference between the distributions for each dataset overall, and report the p-values (p) from those tests:[7] *(i) 'FMB' has longer events*: The distributions for each event (`Pl`, `Pa`, `Sb`, and `Sf`) are shifted to the right for 'FMB' relative to those for 'NI', meaning that 'FMB' tends to have longer events. In each of the four cases, the p-values (p) were highly significant $(p \approx 0)$. The fact that

[5]On Coursera, the default player speed is 1.0. Users can vary this between 0.5 and 2.0, in increments of 0.25.

[6]By definition, quartiles separate data in increments of 25%.

[7]We use the WRS test because Shapiro-Wilk tests detected significant departures from normality for each of the distributions.



(a)

Dataset	Event	Size	Frac	Q_1	Q_2	Q_3
'FMB'	Pl	112.7K	53%	13.9	67.5	282.4
	Pa	51.2K	24%	9.6	31.9	102.4
	Sb	29.4K	14%	17.7	35.4	72.7
	Sf	18.2K	8.6%	21.2	63.7	227.2
'NI'	Pl	103.5K	58%	12.0	71.0	262.6
	Pa	46.4K	26%	4.5	19.3	58.8
	Sb	17.8K	10%	12.9	26.2	54.7
	Sf	10.7K	6.0%	9.6	28.4	81.7

(b)

Fig. 3. Distribution of the lengths for four events across both 'NI' and 'FMB'. For Pl and Pa, this represents the time elapsed before the next event, and for Sb and Sf, this is the distance of the skip. (a) Boxplots of the distributions for each dataset. (b) Tabulated statistics for the distributions.

Pa is longer for 'FMB' is consistent with this subject material being more difficult. *(ii)* Sf *is longer than* Sb: The distribution of Sf is shifted to the right relative to Sb for both 'FMB' and 'NI' ($p < 1\text{E}-6$). This indicates that when students skip forward, they tend to pass more material than they review when skipping back. Sb also occurs more frequently than does Sf for both courses. *(iii)* Pl *is longer than* Pa: The distributions for Pl and Pa in both datasets indicate that users tend to stay in the playing state longer than in the paused state ($p \approx 0$). This effect is stronger in the case of 'NI', which is again consistent with the fact that the 'FMB' material is more difficult.

Event intervals. Clearly, l_j and d_j can vary substantially between events and datasets. To account for this relative variation, we will use the four intervals in-between the three quartiles for each event (given in Fig. 3(b)) to discretize the lengths. We specify three cases: *(i)* $\hat{e}_j \in \{\texttt{Sb}, \texttt{Sf}\}$: When the event is a skip, we map it to $\hat{e}_j q_j$, where $q_j \in \{1,2,3,4\}$ is chosen such that $l_j \in \left[Q_{q_j-1}, Q_{q_j}\right)$, with $Q_0 = 0$ and $Q_4 = \infty$. For example, suppose that event E_i is such that $\hat{e}_j = \text{Sb}$ and $l_j = 20$ sec. In either course, this would be mapped to Sb2. *(ii)* $\hat{e}_j = \texttt{Pa}$: The mapping here works the same as in the first case, except q_j is chosen based on d_j instead. *(iii)* $\hat{e}_j = \texttt{Pl}$: Two long duration play events could still have different qualitative interpretations.[8] To account for this, when $\hat{e}_j = \text{Pl}$, we map it to $\hat{e}_j q_{j,1}\, \hat{e}_j q_{j,2} \ldots \hat{e}_j q_{j,K}$, where $q_{j,k} \in \{1,2,3\}$ for $k = 1, \ldots, K$ is chosen according to

$$q_{j,k} = \begin{cases} 3, & d_j - \delta_{j,k} > Q_3 \\ \arg\min_{q_{j,K}} \left(d_j - \delta_{j,K} \leq Q_{q_{j,K}}\right), & \text{otherwise,} \end{cases}$$

with $\delta_{j,k} = \sum_{k'=1}^{k-1} Q_{q_{j,k'}}$ at each step. For example, suppose an event is Pl with $d_j = 550$ sec. For the quartiles in 'NI', this would be mapped to Pl3 Pl3 Pl2.

[8]The other events do not have this issue since they are not related to new, incoming information.

3) Event-Type Sequence Specification: Let $\mathcal{S} = \{\texttt{Pl1}, \texttt{Pl2}, \texttt{Pl3}, \texttt{Pa1}, \ldots, \texttt{Pa4}, \texttt{Sb1}, \ldots, \texttt{Sb4}, \texttt{Sf1}, \ldots, \texttt{Sf4}, \texttt{Rf}, \texttt{Rs}, \texttt{Rd}\}$ be the set of 18 events with quantized lengths. For each UV Pair, we encode the clickstream log $E_1, \ldots, E_n$ as $S = (s_1, s_2, \ldots, s_{n'})$, where each $s_j \in \mathcal{S}$ is chosen according to the specifications in Section II-B2. As we will see in Section III, using this alphabet that incorporates event types and lengths allows us to obtain insights that cannot be gleaned with events alone.

For comparison, we will refer to an event with length 1 as "short," 2 as "medium," 3 as "medium-long," and 4 as "long."

III. Motifs of Video-Watching

Using the event-type specification, we identify short, recurring sub-sequences within user video-watching behavior, *i.e.,* behavioral *motifs*. As we will see in Section III-B, these motifs capture fundamental characteristics such as reflecting on or reviewing material. We will also see that some of these motifs are significantly associated with student CFA scores.

A. Motif Extraction

We make use of the MEME Suite software package [23] for motif extraction. MEME has been applied in bioinformatics for motif identification in sequences of nucleotides and amino acids.

Model and algorithm. The underlying algorithm is based on a probabilistic mixture model, where the key assumption is that each subsequence is generated by one of two components: a position-dependent motif model, or a position-independent background model. Under the motif model, each position j in a motif is described by a multinomial distribution, which specifies the probability of each character (*i.e.,* each $s \in \mathcal{S}$ from Section II-B3) occurring at j. The background model is a multinomial distribution specifying the probability of each character occurring, independent of the positions; we employ the standard background of a 0-order Markov chain. A latent variable is assumed that specifies the probability of a motif occurrence starting at each position in a given sequence [23].

Motif extraction is formulated as maximum likelihood estimation over this model, and an expectation-maximization (EM) based algorithm is used to maximize the expectation of the (joint) likelihood of the mixture model given both the data (*i.e.,* the sequences) and the latent variables. We use the standard Dirichlet prior based on character frequencies for EM.

Extraction. Each UV Pair's clickstream sequence is encoded using the 24-character protein alphabet [23]. To do this, we choose the first 18 non-ambiguous characters $\mathcal{F}$, and then specify a 1:1 mapping $\mathcal{S} \leftrightarrow \mathcal{F}$. Whereas other work has focused on a single motif width (*e.g.,* at 4 in [11]), we extract those of widths $w \in \{4, \ldots, 10\}$ from our datasets, with E-values (defined below) at most 0.05. We will see that both long and short motifs can be insightful (see Fig. 7).

For each motif, we obtain its E-value, and its position specific probability matrix (PSPM): (i) *E-value*: The E-value judges overall significance. It is defined as the fraction of motifs (with the same width and occurrences) that would have higher log likelihood ratio if the sequences had been generated according to the background model. (ii) *PSPM*: This gives the fraction of times that each character appears in each position of the

motif, taken over all sightings of the motif in the dataset. In the following, denote the PSPM for a motif by $\mathbf{P} = [p_{i,j}]$, where $p_{i,j}$ is the fraction of times event j occurs at position i.

Representation. At each position i, we consider all events j with $p_{i,j} \geq 0.25$.[9] Formally, let $\mathcal{A}_i$ be the sequence of indices into the event set $\mathcal{S}$ for i, arranged such that $p_{i,\mathcal{A}_i(k)} \geq p_{i,\mathcal{A}_i(k+1)}$ and $p_{i,\mathcal{A}_i(k+1)} \geq 0.25 \ \forall k$. Then, there are three cases on the way i is represented:

1) If $|\mathcal{A}_i| > 1$, i is represented as $\left[\mathcal{S}_{A_i(1)}\ \mathcal{S}_{A_i(2)}\ \cdots\right]$.
2) If $|\mathcal{A}_i| = 1$, then the square brackets are omitted, with just $\mathcal{S}_{A_i}$ displayed.
3) If $A_i = \varnothing$, then i is displayed as '$\star$' to indicate that this position was taken by a variety of events, none of which occurred even 25% of the time.

For example, the sequence [Pl2 Pl3] Pa1 $\star$ [Sf1 Sf2 Sf4] is of length 4, with the first position being either Pl2 or Pl3 at least 50% of the time (Pl2 at least as often as Pl3), the second position being Pa1 at least 25% of the time, the third position being any event, and the last being either Sf, Sf2, or Sf4 at least 75% of the time.

Motif support. For each motif, we obtain the fraction of sequences (FS) in which it occurs, *i.e.,* its support across sequences, as well as the number of videos it appears in. We also obtain FS0 and FS1 as the fraction of non-CFA and CFA sequences in which the motif appears, respectively.

Mixed-effects modeling. To relate video-watching behavior to quiz performance, we seek to quantify the effect that each motif has on whether a UV Pair will have a CFA or a non-CFA response. In fitting such a model, it is important to account for the fact that the individual students and videos can affect the CFA result, since each appears multiple times in the dataset. Hence, for each course, we fit a logistic mixed-effects model [24] to CFA score, with the frequencies of the motifs treated as fixed effects and the specific user and video IDs as random effects. We denote the fixed-effects matrix is $\mathbf{A} = [a_{s,m}]$, with $a_{s,m}$ as the number of times motif m appears in sequence s.

If the p-value (p) for a motif in this model is low, we can conclude that the motif has a significant effect on the CFA score independent of specific videos and students. To obtain a measure of effect size for each motif, we convert their coefficients from the fitted model (which are expected changes in the log-odds of CFA) to the expected change in CFA probability (Δc) for each additional occurrence of the motif.

B. Results

We obtained 87 and 123 motifs from 'FMB' and 'NI', respectively, which are the subject of the following analysis.

1) Motif Supports: We first analyze how the motif supports vary across sequences, videos, and students. We find that the supports are reasonably high across videos and students, but that each individual UV Pair tends to not exhibit many motifs.

Sequences. In Fig. 4, we plot the Empirical CDF (ECDF) of the fraction of sequences that each motif appears in, for both CFA and non-CFA, considering all sequences with at least one motif. In each course, the supports are similar: for 'FMB', each motif appears in 7.1% of the non-CFA sequences on average, and 7.7% of the CFA; for 'NI', each appears in 5.9% for both CFA and non-CFA. Considering the fixed-effects matrices $\mathbf{A}$, then, less than 8% of their entries are non-zero. In both courses, the motifs with largest support (first row in Fig. 7(a) and (b)) appear in $> 25\%$ of the sequences.

[9]With 18 different events, a threshold of 25% is roughly 5 times the expected occurrence from a uniform random selection of events.

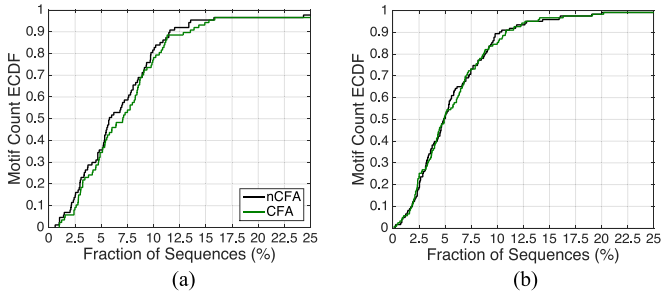


Fig. 4. ECDFs of the number of sequences that each motif appears in, for both CFA and non-CFA. The supports are consistent between the CFA and non-CFA groups in each course. (a) 'FMB', (b) 'NI'.

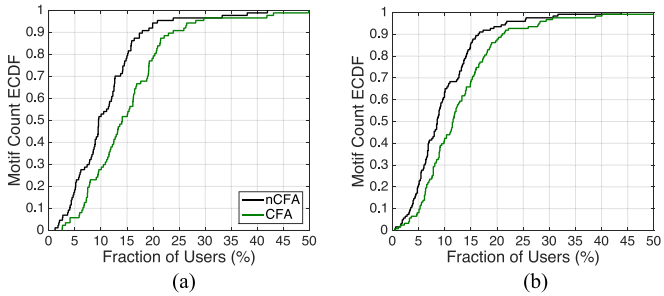


Fig. 5. ECDFs of the number of students that exhibit each motif, across both CFA and non-CFA sequences. In both courses, CFA sequences have a higher support for motifs across users. (a) 'FMB', (b) 'NI'.

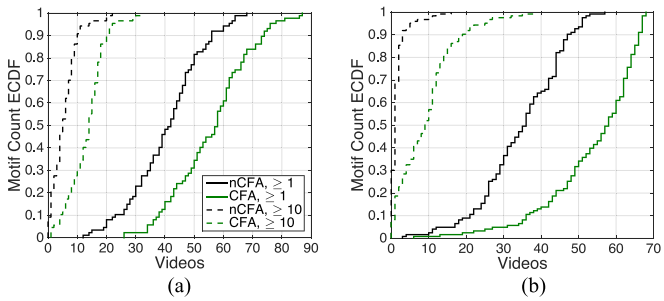


Fig. 6. ECDFs of the number of videos that each motif appears in, across both CFA and non-CFA sequences. In both courses, CFA sequences have a higher support for motifs across videos. (a) 'FMB', (b) 'NI'.

Students. Fig. 5 gives the ECDF of the fraction of students who trigger each motif at least once (*i.e.,* across all videos the student watched), over students who trigger at least one motif. We see that more users exhibit more motifs in CFA than in non-CFA, for both courses: for 'FMB' (resp. 'NI'), each motif is on average triggered by 15.2% (12.8%) of users in the CFA sequences, and only 11.1% (9.9%) in non-CFA.

Videos. In Fig. 6, we show the ECDF of the number of videos that each motif occurred in at least once and at least 10 times (*i.e.,* across all students who saw the video). Overall, CFA has higher support than non-CFA over videos. We also see that the supports decrease for higher thresholds, *e.g.,* for 'FMB', while

Group		Motif	E-value	FS(%)	FS0(%)	FS1(%)	Δc(%)	p-value
Pa	I	[Pl2 Pl3] [Pa4 Pa3] [Pl2 Pl1] [Pa2 Pa3] Pl2 Pa3 Pl2 [Pa2 Pa3] Pl3	5E-64	28.5	26.2	29.5	+0.68	0.624
	II	Pl2 Pa4 Pl2 Pa4	2E-06	13.2	13.2	13.3	-1.81	0.145
	III	[Pl1 Pl3 Pl2] Pa4 Pl2 Pa3 [Pl2 Pl1] ⋆ Pl2 ⋆ [Pl3 Pl1] Pa3	2E-15	10.9	9.33	11.6	+4.89	0.0128*
	IV	Pl1 Pa1 Pl1 Pa1 Pl1	7E-13	7.75	6.31	8.38	+5.80	6.2E-3**
Sb	I	Sb3 [Pl2 Pl1] [Sb2 Sb3] Pl2 Sb2 Pl2 [Sb2 Sb3] [Pl3 Pl2]	0	10.2	8.86	10.8	+0.32	0.872
	II	Sb4 Pl1 Sb4 Pl2 [Sb3 Sb4]	8E-05	4.64	3.75	5.03	+7.52	6.8E-3**
	III	Sb3 Pl2 Sb3 [Pl3 Pl2]	0	3.54	3.08	3.74	+4.67	0.0638.
Sf	I	Pl2 Sf3 [Pl1 Pl2] Sf2 [Pl1 Pl2] Sf1 [Pl2 Pl1] [Sf2 Sf1]	0	9.50	9.56	9.47	+1.29	0.591
	II	[Sf4 Sb4] [Pl1 Pl2] [Sf3 Sf4] [Pl1 Pl2] [Sf3 Sf4]	0	7.59	7.61	7.58	-3.36	0.130
	III	[Pl2 Pl1] [Sf2 Sf3] Pl1 [Sf3 Sf2]	2E-37	7.25	7.12	7.30	-4.75	9.6E-3**
Rf	I	Pl3 [Rf Rd] [Pl2 Pl1] Rf [Pl3 Pl2] Rf	0	4.55	3.89	4.84	+1.62	0.539
	II	Rf Rd [Pl1 Pl2] Rf Pl3	1E-70	1.77	1.22	2.00	+9.30	0.039*

(a)

Group		Motif	E-value	FS(%)	FS0(%)	FS1(%)	Δc(%)	p-value
Pa	I	[Pl2 Pl3] Pa4 [Pl2 Pl3] Pa4 Pl3	2E-81	26.8	27.6	26.5	-2.39	0.162
	II	Pl2 Pa4 Pl2 Pa4	2E-44	14.3	15.9	13.7	-2.71	0.105
	III	Pl1 Pa1 Pl1 Pa1 Pl1 Pa1 [Pl1 Pl3]	0	11.7	12.7	11.4	-5.29	0.036*
Sb	I	[Sb3 Sb4] [Pl2 Pl3] [Sb3 Sb2] Pl2 [Sb3 Sb2] [Pl3 Pl2]	0	9.17	9.47	9.07	+0.74	0.787
	II	Sb2 Pl2 Sb3 Pl2 [Sb2 Pa3]	6E-63	6.05	5.76	6.15	-5.93	0.065.
	III	Sb2 [Pl2 Pl1] [Sb2 Pa2] Pl2 [Sb2 Sb3] [Pl2 Pl3]	0	5.74	4.76	6.07	+9.63	7.8E-3**
Sf	I	Pl2 [Sf3 Sf2] Pl1 [Sf3 Sf2]	0	7.71	9.68	7.04	-5.97	0.024*
	II	[Pl3 Pl1] [Sf3 Sf4] [Pl1 Pl2] [Sf4 Sf3 Sf2] [Pl1 Pl2] [Sf3 Sf4]	0	7.76	8.95	7.36	+6.45	0.085.
	III	[Sf3 Sf4] [Pl2 Pl1] [Sf4 Sf3] Pl1 [Sf3 Sf4]	8E-88	5.31	6.75	4.83	-7.14	0.066.
Rf	I	[Pl1 Pl2] Rf Pl2 Rf Pl2 Rf	0	9.21	8.58	9.43	+8.07	9.8E-3**
	II	[Pl1 Rd] Rf Pl1 Rf Rd [Pl1 Pl2] Rf Pl3	0	3.86	4.61	3.61	-10.8	0.016*
	III	[Rf Rs] [Pl2 Pl1] Rd [Pl2 Pl3] Rf	7E-16	2.53	3.14	2.33	-8.11	0.089.

(b)

Fig. 7. Representative sample of motifs identified for each course. Each motif is grouped by the dominant event it contains outside of Pl. FS is the fraction of sequences over both CFA and non-CFA, while FS0 and FS1 are for the separate cases. Δc is the estimated change in the probability of success (CFA) for every additional occurrence of the motif, and the *p*-value (p) is the significance of Δc (a . indicates $p \leq 0.1$, a * indicates $p \leq 0.05$, and a ** indicates $p \leq 0.01$). (a) Motifs for 'FMB'. (b) Motifs for 'NI'.

the top 20% of motifs appear in at least 67 videos for CFA, this drops to only 18 videos considering at least 10 occurrences.

We will turn now to analyze the specific motifs, and identify patterns associated with quiz performance. Given that the matrix $\mathbf{A}$ is sparse (even among those sequences that have at least one motif), note that we will move to a more appropriate model for online CFA prediction in Section IV.

2) Individual Motifs: We inspect patterns in the most significant of the 210 extracted motifs. This list is obtained by applying the following procedure. First, noticing that all motifs contain play (Pl) events, we group them into categories based on the most recurring alternate event, leading to four groups. Then, within each category, we consider each motif that either (i) has one of the top-5 highest supports or (ii) has a significant p-value (≤ 0.1) returned from the mixed-effects model. Finally, if one motif is a subsequence of another, then we remove the one that has lower support or is less significant.

In Fig. 7, we give the representative sample of these motifs that are mentioned in the following discussion, by group. Each motif is assigned an ID consisting of its group and number (*e.g.,* Pa II in 'FMB' is motif `Pl2 Pa4 Pl2 Pa4`). In Fig. 8, we visualize the key properties exhibited by each group.

Overview. The motifs exhibit many similar structural attributes, which occur in spite of the fact that the encoding quantiles are different for each event and course (see Fig. 3). Also, since MEME finds ungapped motifs (*i.e.,* those existing as exact matches in the data, without a separate layer of similarity matching), these identified behaviors exist exactly in the sequences, contrary to other work [11] which has resorted to approximate string searching. The motifs in the `Pa` (pause) group have the largest supports (FS) overall ($\geq 10\%$ mostly), which is consistent with the fact that there are fewer `skip` and `ratechange` events in the datasets (see Fig. 3(b)).

We present our most interesting observations for each group:

Reflecting (`Pa`): The occurrence of play together with pause indicates that lectures are generally thought-provoking, causing students to *reflect on material they just saw* (see Fig. 8(a)). In both courses, the events forming the motifs in this group cover the entire range from short to medium-long plays (`Pl1–Pl3`) interspersed with short to long pauses (`Pa1–Pa4`).

The motifs with highest supports in 'FMB' and 'NI'—Pa I—can be viewed as sequences of medium to medium-long plays with medium-long to long pauses in-between. This behavior is not significantly associated with CFA or non-CFA in either case, though $(p > 0.1)$. Motif Pa III in 'FMB' is different from these in that it has a short play interspersed too, and it *is* significantly correlated with an increase in the chance of CFA $(p < 0.02, \Delta c = +4.89\%)$. This may indicate that a student pausing longer relative to the plays in-between is an effective strategy in 'FMB'. Motifs Pa II in 'FMB' and 'NI', with medium plays followed by long pauses, also do not differentiate between the groups $(p > 0.1)$.

The comparison between Pa IV in 'FMB' and Pa III in 'NI' is particularly interesting. Both of these motifs are short pauses and plays interspersed, indicating a tendency to reflect frequently on a small amount of material at a time. While in 'FMB' it is significantly associated with an improvement in CFA $(p < 0.01, \Delta c = +5.80\%)$, in 'NI', it is associated with a decrease in the chance of CFA $(p < 0.05, \Delta c = -5.29\%)$. Short pauses in 'NI' may be a sign of unresolved confusion.

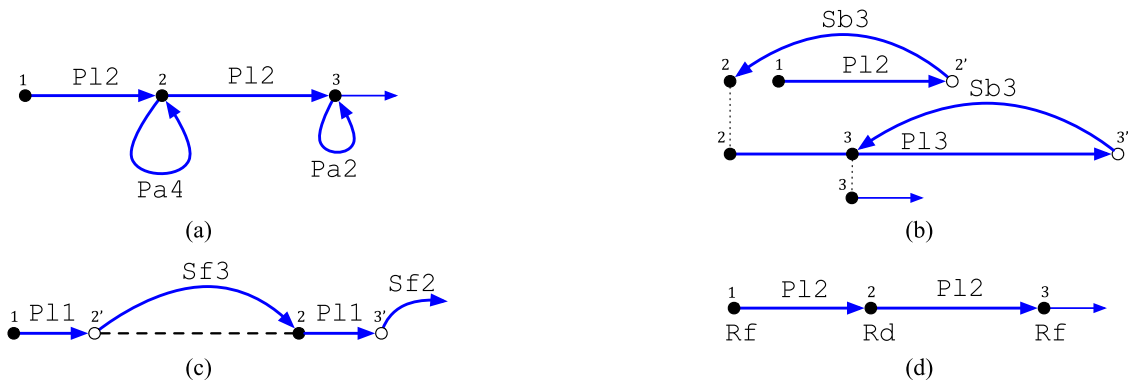


Fig. 8. Illustration of the key video-watching behaviors exhibited by each of the four motif groups in Fig. 7. For brevity, we omit variations that exist between the individual motifs within the groups. In each case, the horizontal axis represents the video length, as in Fig. 2. For (b) and (c), horizontal jumps represent skips with lengths relative to the distance, whereas the vertical jumps in (b) just indicate continuity. (a) Reflecting (Pa): Pl2 Pa4 Pl2 Pa2, (b) Reviewing (Sb): Pl2 Sb3 Pl3 Sb3, (c) Skimming (Sf): Pl1 Sf3 Pl1 Sf2, (d) Speeding (Rf): Rf Pl2 Rd Pl2 Rf.

Reviewing (Sb). From the six motifs in the Sb group, we identify two interesting, recurring subsequences: Pl2 Sb3 Pl2 Sb3 (Sb I and III in 'FMB', and Sb I in 'NI'), and Pl2 Sb2 Pl2 Sb2 (Sb III in 'NI'). Roughly speaking, each of these is associated with *playing for a length of video, and then reviewing some or all of that content* (see Fig. 8(b)). To see this, consider the ranges of Pl and Sb from Fig. 3 associated with these subsequences: for 'FMB', Pl2 covers 14 to 68 sec, and Sb2 to Sb3 covers 18 to 73 sec; for 'NI', these ranges are 12 to 71 sec and 13 to 55 sec. The play and skip ranges are closely overlapping in each case. Taking the extreme ends of each range, they are associated with skipping back anywhere from 1 min below the starting play point to 50 sec after it,[10] which are local considering the video lengths.

Note that 2 of the 4 motifs containing these behaviors are significantly correlated with an increase in CFA probability ($p < 0.07, \Delta c > +4.0\%$). The fact that Sb II of 'FMB' has the highest $\Delta c = +7.52$ may also indicate that *reviewing more than what was just played* can further increase the chance of success, given the presence of long skip backs (Sb4) here.

We also considered the number of skip backs originating at each video position across all UV Pairs. We find that the largest origination point of these events is at the end of videos. In particular, out of all Sb events, those originating within 10 sec of the videos' end constitute roughly 15% of the total in 'FMB' for both non-CFA and CFA sequences.[11] This, combined with the motifs suggesting improvement when reviewing occurs, implies that those students who are reviewing *multiple times* before answering a quiz have more success.

The notable exception to this is Sb II in 'NI'. Here, revision is associated with a *decrease* in the chance of CFA ($\Delta c = -5.93\%$). Similar to Pa III discussed before, this may be an indicator of excessive confusion in this course.

Skimming (Sf). In both of the courses, the motifs in the Sf group are primarily medium to long skips forward with short to medium plays in-between. Further, the skips are longer than the plays occurring before and after; comparing the lengths of Pl and Sf events in Fig. 3, we see that for both courses, range Q_j to Q_{j+1} for Sf is always larger than Q_{j-1} to Q_j for Pl. This recurring behavior can then be interpreted as *skimming through the material quickly* with less exposure to the material (see Fig. 8(c)). We find that 3 of these 6 motifs are significant in favor of non-CFA ($p < 0.07, \Delta c < -4.0\%$), in contrast to the work in [12] where the total number of skips forward in a sequence was not found to be correlated with the CFA probability. This difference underscores the utility of considering the clickstream sequences, rather than computing aggregate quantities to summarize them.

Also, Sb and Sf occurring together in a motif (*e.g.,* Sf II in 'FMB') can possibly be interpreted as *skipping forward with caution*. Still, we find that this is close to being significant in favor of non-CFA ($p = 0.13$). There is an exception to the generally negative correlations in the Sf group, though: Sf II in 'NI', where skimming is associated with an increase in CFA probability ($p < 0.1, \Delta c = +6.45$). With short plays in-between long skips forward, this is *fast* skimming, and can be explained by the fact that in some cases, a student will already be familiar with the more basic 'NI' material.

Speeding (Rf). Rf I in 'NI' indicates that viewing the material at a faster than default rate, *i.e., speeding*, is significantly associated with an increase in CFA probability ($p < 0.01, \Delta c > +8.0\%$). The other motifs making up the Rf groups for each course have students returning to the default rate (Rd), indicating they are *slowing down for important content* (see Fig. 8(d)). In 'FMB', this is positively associated with CFA in one case (Rf II, with $\Delta c = +9.3\%$), whereas in 'NI', it is significantly associated with a decrease in CFA score in both cases (Rf II and Rf III, with $\Delta c < -8.0\%$). In 'NI', slowing the rate back down to the default could be a sign that a student saw something confusing, but did not take the time to *e.g.,* reflect or review to clear up the confusion.

3) Key Messages: Overall, we draw a few conclusions.

Motif groups. There are four groups of motifs: *(i) Reflecting*, *i.e.,* pausing to reflect on the video material repeatedly. If the time spent reflecting is not *too* long relative to the time spent watching, this tends to be correlated with a higher chance of success on the quiz. At the same time, if the pausing is very short, it could indicate unresolved confusion. *(ii) Reviewing*, *i.e.,* repeated review of the video content just watched. This tends to be correlated with an increase in the chance of success.

[10] We assume a default playback rate as an approximation.

[11] If we take the highest location of Sb for each video outside of the last 10 sec, these positions contain roughly 4% of the total for non-CFA and CFA.

(iii) Skimming, *i.e.,* skipping through video material quickly. This tends to be associated with a lower chance of success, even when done with caution. *(iv) Speeding*, *i.e.,* watching the video at a faster than default rate and slowing down at certain times. Different variations are associated with different impacts on the chance of success.

Importance of CFA associations. Some motifs are significantly correlated with substantial changes in the probability of CFA, independent of the specific videos and/or students (from Δc in Fig. 7, the increases can be as high as 9%, and the decreases as low as 10%). For each motif, the direction of the association is particularly important, because in many cases either would be intuitive. For example, a reviewing motif could presumably come from a student reinforcing material in the video prior to taking the quiz (in line with an increase in CFA probability) or from excess confusion caused by the material in the video (in line with a decrease in CFA probability), but the results indicate the former tends to be more likely in these courses. As another example, skimming could come from a student believing confidently that he/she is already familiar with the content in a video, which could intuitively be either a correct (increase in CFA probability) or an incorrect (decrease in CFA probability) perception, but results favor the latter.

Importance of lengths/durations. We emphasize the importance of having included the lengths/durations, in addition to the events, in our framework from Section II-B3 in order to make these conclusions. For instance, the sequence `Pl Sb Pl Sb` identified in [11] cannot be associated with reviewing, because it is not clear how far back the student has skipped relative to having played in-between. In the same way, `Pl Sf Pl Sf` cannot be concluded as skimming, because the lengths of play and skip are not indicated in the model. Also, even small changes in the motif lengths can affect significance (*e.g.,* in 'FMB', while Pa I is associated with CFA, Pa II is not).

IV. Model of Position Sequence

In this section, we will formalize our second sequence representation, which factors in the location in the videos that a student visited. Then, we will present CFA models based on this framework, which will be evaluated in Section V.

A. Modeling Framework

1) Definitions: Let $v \in \mathcal{V}$ denote video v in the set of videos $\mathcal{V}$ for a course, indexed chronologically (*i.e.,* by release date of the videos).[12] Also, let $c \in \mathcal{O}$ denote class c in the set of binary classes $\mathcal{O} = \{0, 1\}$, where $c = 0$ indicates a non-CFA submission and $c = 1$ is CFA. With $u \in \mathcal{U}$ as user u in the set of all users $\mathcal{U}$, we let $\mathcal{U}^v \subset \mathcal{U}$ be the set of users who have a UV Pair for v, and $\mathcal{U}^{v,c} \subset \mathcal{U}^v$ be those who fall into class c with respect to their answer submissions. For evaluation in Section V, we will generate training ($\mathcal{U}^v_T$) and test ($\mathcal{U}^v_E$) sets as subsets of $\mathcal{U}^v$; $\mathcal{U}^v_T$ and $\mathcal{U}^v_E$ are always chosen such that $\mathcal{U}^v_T \cap \mathcal{U}^v_E = \varnothing$.

2) Position-Based Sequence Specification: We will divide each video into a number of intervals. Let h_v be the length (in sec) of v. We define w_v to be the width that partitions v into $N(w_v) = \lfloor h_v / w_v \rfloor$ uniform intervals, such that interval $i \in \mathcal{P}^v(w_v) = \{1, \ldots, N(w_v)\}$ spans the range $[(i-1) \cdot w_v, i \cdot w_v]$. For each UV Pair, we can then model the video-watching behavior as a sequence of positions $\mathbf{p}^{u,v} = (\rho_1, \rho_2, \ldots, \rho_n, \ldots)$, where $\rho_n \in \mathcal{P}^v(w_v)$ is the index of the nth position visited.[13]

To generate these sequences, we first apply the same denoising procedure described in Section II-B1 to each event E_i. Then, for each UV Pair, starting with $\mathbf{p} = ()$ we do the following:

1) For E_1, append $\lfloor p_1 / w_v \rfloor$ to $\mathbf{p}$.
2) Consider each sequential pair of events $E_i, E_{i+1}, i \geq 1$. If the state $s_i =$ `paused`, then only $\lfloor p_{i+1}/w_v \rfloor$ is added to $\mathbf{p}$. But if $s_i =$ `playing`, then:
   a) If the event $e_i \neq$ `skip`, then $(\lfloor p_i/w_v \rfloor + 1, \ldots, \lfloor p_{i+1}/w_v \rfloor - 1, \lfloor p_{i+1}/w_v \rfloor)$ is added to $\mathbf{p}$.
   b) If $e_i =$ `skip`, then $(\lfloor p_i/w_v \rfloor + 1, \ldots, \lfloor p'_i/w_v \rfloor - 1, \lfloor p'_i/w_v \rfloor, \lfloor p_{i+1}/w_v \rfloor)$ is appended instead.[14]

For example, suppose $h_v = 300, w_v = 15$, and a user generates $E_1 = \langle$`play`$, 0, 0,$ `playing`$, 1.0\rangle$, $E_2 = \langle$`skip`$, 200, 50,$ `playing`$, 1.0\rangle$, $E_3 = \langle$`ratechange`$, 230, 80,$ `playing`$, 1.25\rangle$, and $E_4 = \langle$`pause`$, 300, 127,$ `paused`$, 1.25\rangle$ on the video. Then, $\mathbf{p} = (0, 1, 2, 3, 13, 14, 15, 15, 16, \ldots, 20)$.

3) Model Factors: There are (at least) three types of information for each $\mathbf{p}^{u,v}$ that could have an effect on performance:

(1) Positions. First is the number of times a given position $i \in \mathcal{P}^v(w_v)$ was visited. One would expect these to differ between CFA and non-CFA, because certain parts of videos will be more important to questions. We can see this by referring to the motifs that had correlations with increases or decreases in CFA probability. CFA sequences with reviewing motifs may have more visits to positions associated with the questions through repeating. On the other hand, non-CFA sequences with skimming motifs may have less visits to these important positions. Sequences with reflecting motifs may have more visits to important positions through pausing, too.

(2) Transitions. Second is the number of transitions between the positions, *i.e.,* the number of times a given tuple (i, j) is a subsequence of $\mathbf{p}^{u,v}$. Considering each tuple (ρ_n, ρ_{n+1}):

1) If $\rho_{n+1} < \rho_n$, then the user had skipped back. We call this a backward transition.
2) If $\rho_{n+1} > \rho_n + 1$, then the user had skipped over the material in (ρ_n, ρ_{n+1}). This is a forward transition.
3) If $\rho_{n+1} = \rho_n + 1$, then the user moved directly to the next position. This is a direct transition.
4) If $\rho_{n+1} = \rho_n$, then the user had some event within the current position. This is a repeat transition.

We say that direct and repeat transitions are local, whereas backward and forward are non-local. As with positions, the transition factors can capture the motif behavior associated with changes in CFA probability, except in terms of sequences of visits, *e.g.,* backward transitions capture `Sb` in a reviewing motif, and forward transitions capture `Sf` in a skimming motif.

(3) Time spent. Third is the amount of time spent at the different positions. One would expect these times to be indicative of CFA/non-CFA in a similar manner to visit frequencies.

[12] Recall from Section II-A that we define a "video" to be all videos for a quiz.

[13] For brevity, we will typically refer to $\mathbf{p}^{u,v}$ as just $\mathbf{p}$, with the understanding that it refers to the UV Pair in question.

[14] Recall from Section II-B1 that when E_i is a skip event, p'_i is the position of the video player immediately before the skip.

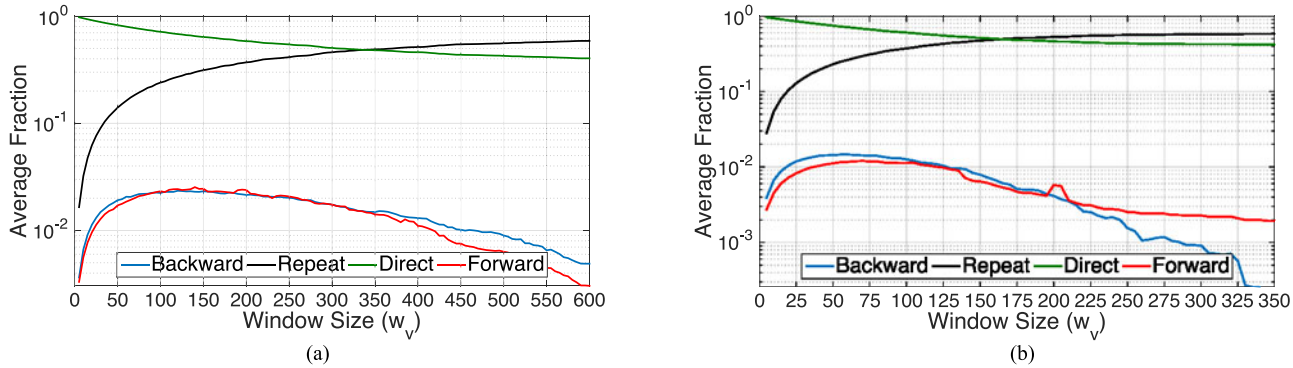


Fig. 9. Plot of the fraction of local (repeat and direct) and non-local (backward and forward) transitions for each window size w_v, averaged over all UV Pairs for each position and video v, for each dataset. The fraction of non-local transitions is very low in each case. (a) 'FMB', (b) 'NI'.

We will consider four prediction models based on these factors. Three of them are likelihood-based: Discrete time Positions (DP), which incorporates the number of visits to each position; Discrete time Transitions (DT), which models transitions between positions; and Continuous time Transitions (CT), which factors in inter-arrival times between positions. For comparison, we include a standard SVM predictor that uses position counts as features. In practice, the major advantage of the likelihood-based schemes over SVM is that its feature space is directly interpretable, leading *e.g.,* to content analytics. Each model will be tested on each video separately, allowing us to compare results on a per-video basis in Section V.

B. Position-Based Modeling

Discrete Time Positions (DP). For the DP model, video positions are treated as independent events. Let $\mathbf{f}^{v,c} = [f_i]^{v,c} \in [0,1]^{N(w_v)}$ be the probability distribution of visit frequency across positions $i \in \mathcal{P}^v(w_v)$. This is estimated over the UV Pairs in the training set $\mathcal{U}_T^{v,c}$ as

$$f_i^{v,c} = O_i^{v,c} \Big/ \sum_j O_j^{v,c}, \tag{1}$$

where $O_i^{v,c}$ is the number of occurrences of p_i over sequences in $\mathcal{U}_T^{v,c}$. In other words, $O_i^{v,c} = \sum_{u \in \mathcal{U}_T^{v,c}} O_{u,i}^{v,c}$, where $O_{u,i} = \sum_n \mathbb{I}_{\{\rho_n = i\}}$ is the number of times student u was at i.

We test the ability of this model to identify which class each $u \in \mathcal{U}_E^v$ belongs to. For this purpose, we compute the likelihood of observing $\mathbf{p}$ on video v to be in c, given $\mathbf{f}^{v,c}$, as

$$L(\mathbf{p} \,|\, \mathbf{f}^{v,c}) = g^{v,c} \cdot \prod_n f_{\rho_n}^{v,c}. \tag{2}$$

Then, the prediction $\tilde{c} \in \{0,1\}$ of the class for $\mathbf{p}$ is determined by application of the Maximum a Posteriori Probability (MAP) decision rule. But recall that there is a bias towards $c = 1$ for each course (see Fig. 1). As a result, we introduce a term $b_v \geq 0$ into MAP, which will be tuned through the cross validation procedure described in Section V-A:

$$\tilde{c} = \begin{cases} 1 & g^{v,1} L\left(\mathbf{p} \,|\, \mathbf{f}^{v,1}\right) > g^{v,0} L\left(\mathbf{p} \,|\, \mathbf{f}^{v,0}\right) + b_v \\ 0 & g^{v,1} L\left(\mathbf{p} \,|\, \mathbf{f}^{v,1}\right) < g^{v,0} L\left(\mathbf{p} \,|\, \mathbf{f}^{v,0}\right) + b_v \\ \mathbb{I}_{\{U \geq g^{v,0}\}} & \text{otherwise} \end{cases}, \tag{3}$$

where $g^{v,c} = |\mathcal{U}_T^{v,c}| / |\mathcal{U}_T^v|$ is the estimated class bias for video v, and U denotes a random number drawn from $[0,1]$.

Support Vector Machine (SM). Let $\mathbf{O}_T^v = [O_{u,i}]_T^v$ be the user-position matrix consisting of all $u \in \mathcal{U}_T^v$, and let $\mathbf{c}_T^v$ be the vector of corresponding CFA scores. We fit $\mathcal{M} : \mathbf{O}_T^v \rightarrow \mathbf{c}_T^v$ as an SVM [25] over the training set, and then test the algorithm by comparing $\mathbf{c}_E^v$ to $\mathcal{M}(\mathbf{O}_E^v)$ for the users $u \in \mathcal{U}_E^v$ in the test set. We use the standard linear kernel, with the standard regularization parameter C_v tuned through cross validation.

C. Transition-Based Modeling

In modeling transitions between positions, we will only consider one-step transitions. This is common in webpage clickstream analysis (*e.g.,* [21]), and will be useful here since the state spaces we consider can be large, depending on w_v.[15]

1) Aggregating Non-Local Transitions: The cohort estimator for a Markov chain model uses the fraction of transitions from state i to j in estimating the probability of transitioning from i to j [26]. We found this model not appropriate here, because the number of transitions between two non-local positions is rather sparse, implying that there is not enough data to estimate these specific transitions.

To see this, we inspect the sequences $\mathbf{p}^{v,c}$ for varying w_v. In particular, for each position in video v, we first find the total number of times each type of transition from Section IV-A3 occurs, aggregated across the UV pairs. Then, we sum these totals over all positions, and find the fraction of each type of transition. We repeat this for each $w_v \in \{5, 10, \ldots, 600\}$ (*i.e.,* through 10 min), and then average across the videos v for each w_v. Fig. 9 shows the result for each course, from which we make two observations for local and non-local transitions: *(i) Tradeoff between local transition types*: As w_v increases, the percentage of repeat transitions increases monotonically (from roughly 2% to 60% in each course), while the percentage of direct transitions decreases monotonically (from roughly 98% to 40% in each course). This is to be expected, since each position is increasing in size with w_v. *(ii) Infrequency of non-local transitions*: The majority of transitions are local. For example, the largest fraction of backward transitions is just over 2% in 'FMB', occurring at $w_v = 120$.

[15]This may not be ideal because unlike sequences of webpages, learning builds on itself. It is harder to estimate higher order transitions due to position-specific data sparsity. We still see substantial benefit with a one-step model.

As a result of the second observation, the models that follow will aggregate all observed forward transitions to form a single, uniform probability at each position, and likewise for backward transitions. To this end, we define $\mathcal{I}_{i,k} = \{1, \ldots, i-1\}$ for $k = 1$; $\{i\}$ for $k = 2$; $\{i+1\}$ for $k = 3$; and $\{i+2, \ldots\}$ for $k = 4$ to be the set of states constituting a backward $(k = 1)$, repeat $(k = 2)$, direct $(k = 3)$, and forward $(k = 4)$ transition at position i.

Discrete Time Transitions (DT). In this model, we discretize time, discounting the interarrival times. Let $\mathbf{F}^{v,c} = [f_{i,k}]^{v,c} \in [0,1]^{N(w_v),4}$ be the matrix of transition probabilities, where $f_{i,k}^{v,c}$ is the probability that the next position will be in $\mathcal{I}_{i,k}$ given the current is i. We also assume that the transitions are homogeneous, *i.e.,* independent of time n.

Considering the sequences of positions $\mathbf{p}$ across users $u \in \mathcal{U}_T^{v,c}$, we obtain the number transitions from i to k as

$$O_{i,k}^{v,c} = \sum_{u \in \mathcal{U}_T^{v,c}} \sum_n \mathbb{I}_{\{\rho_n = i,\, \rho_{n+1} \in \mathcal{I}_{i,k}\}}. \tag{4}$$

From (4), we estimate $f_{i,k}^{v,c} = O_{i,k}^{v,c} / \sum_j O_{i,j}^{v,c}$, and the likelihood of $\mathbf{p}$ from user $u \in \mathcal{U}_E^v$ on video v is

$$L\left(\mathbf{p} \,|\, \mathbf{F}^{v,c}\right) = f_{\rho_1}^{v,c} \cdot \prod_n f_{\rho_n, \rho_{n+1}}^{v,c}, \tag{5}$$

where $f_{p_1}^{v,c}$ is the distribution at the initial position ρ_1 of $\mathbf{p}$, obtained from (1). The MAP decision rule for DT is the same as in (3), except with (5) in place of (2).

Continuous Time Transitions (CT). This model incorporates the interarrival times between transitions. Rather than computing the time-varying transition probabilities, we instead work with the transition rates [26]. To this end, we define $\mathbf{Q}^{v,c} = [q_{i,k}]^{v,c} \in \mathcal{R}^{N(w_v),4}$ as the transition rate matrix for the model, where $q_{i,k}, k \neq 2$ represents the rate of departure from position i and arrival at a position in $\mathcal{I}_{i,k}$.

Let $\mathbf{r}^{v,c} = [r_i]^{v,c} \in \mathcal{R}^{N(w_v)}$ be the vector of the total time spent by $\mathcal{U}_T^{v,c}$ in state i. These terms are estimated as

$$r_i^{v,c} = \sum_{u \in \mathcal{U}_T^{v,c}} \sum_n \mathbb{I}_{\{\rho_n = i\}} \cdot d_n, \tag{6}$$

where d_n is the duration of event n in $\mathbf{p}$ (see Section II-B1). In estimating the $q_{i,k}$, we must also obtain the number of transitions from i to k over users $u \in \mathcal{U}_T^v$, *i.e.,* the $O_{i,k}^{v,c}$ from (4); with this, the $q_{i,k}^{v,c}$ terms are estimated as

$$q_{i,k}^{v,c} = \begin{cases} O_{i,k}^{v,c} / r_i^{v,c} & k \neq 2 \\ -\sum_{k \neq 2} q_{i,k}^{v,c} & k = 2 \end{cases}. \tag{7}$$

Finally, the likelihood of sequence $\mathbf{p}$ for $u \in \Omega_E^v$ is computed via

$$L\left(\mathbf{p} \,|\, \mathbf{Q}^{v,c}\right) = \prod\nolimits_{i,k;k \neq 2} \left(q_{i,k}^{v,c}\right)^{o_{i,k}} \exp\left(-q_{i,k}^{v,c} \cdot T_i\right), \tag{8}$$

where $o_{i,k} = \sum_n \mathbb{I}_{\left\{\rho_n = i,\, \rho_{n+1} \in \mathcal{I}_{ik}\right\}}, k \neq 2$ is the number of transitions from i to k for the sequence $\mathbf{p}$, and $T_i = \sum_n \mathbb{I}_{\{\rho_n = i\}} \cdot d_n$ is the time spent by $\mathbf{p}$ in i. Once again, the MAP decision rule is as in (3), except with (8) in place of (2).

We also considered another position-based model, Continuous Time Positions (CP), which used the time spent at each position in likelihood computation. We omit it because its results were strictly lower than these three likelihood-based models.

V. Prediction Evaluation

In this section, we evaluate the performance of the models described in Section IV. We pose the following questions:

1) How beneficial is it to include video-watching positions and transitions for CFA prediction on individual videos?
2) How do the likelihood-based models compare against the SVM-based model?
3) Is one of position or transition-based model clearly better than the other, or would some combination be the best?
4) Is it beneficial to include position durations?

Skewed-Random (SR). To answer the first question, we will consider an algorithm that does not make use of clickstream data, to act as a baseline for evaluating the gain from incorporating video-watching behavior. SR finds the CFA bias $g^{v,1}$ over the training set $\mathcal{U}_T^v$, and predicts $c = 1$ $g^{v,1}$ of the time (similar to the baseline used in [12]). Note that in our application of CFA prediction for individual videos, more sophisticated baselines that would leverage similarities across users and/or quizzes without behavioral data (*e.g.,* collaborative filtering like in [18] and [27]) are not applicable.

A. Procedure

Metrics. Let TP, FP, TN, and FN be the number of true and false positives, and true and false negatives obtained by a model on an evaluation set. The first metric we consider is *accuracy* (Acc), *i.e.,* $(\mathrm{TP} + \mathrm{TN}) / (\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN})$. Since the quizzes are biased towards CFA (see Fig. 3), we found that unconstrained maximization of accuracy during the tuning procedure (described below) led to high recall (rec), *i.e.,* $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ but low precision (prec), *i.e.,* $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$. To avoid this, we will subject tuning to the constraint that the chosen parameters have at least 25% of the truly negative samples predicted negative, and likewise for the positives. To this end, the second metric we consider is the standard (balanced) *F1 score*, obtained as $2 \cdot (\mathrm{prec} \times \mathrm{rec}) / (\mathrm{prec} + \mathrm{rec})$ [25]. As the harmonic mean of precision and recall, F1 is limited by the minimum of the two, capturing the tradeoff between them that is induced by this constraint.

Even a few percent improvement in these metrics can be a substantial benefit for CFA prediction. To see this, we can take an example of the first video in 'FMB', which is the earliest point of application of these algorithms, and also the point at which the dropoff before the next video is the highest. Assuming that the total number of incorrect responses (roughly 1150) stayed the same, then for every 1% improvement in prediction accuracy, we could identify another 12 students who would get the question incorrect. Further, if the dropoff rate (roughly 25%) were to stay constant among the incorrect responses, then each 1% improvement we give the chance to detect three more students who would otherwise drop off.

Training and testing. For each algorithm and each video, we obtain the accuracy and F1 metrics over N evaluation iterations. In each iteration, we use the following procedure:

	w_v		b_v (C_v)		Acc		F1	
	avg	s.d.	avg	s.d.	avg	s.d.	avg	s.d.
SR	-	-	-	-	0.510	0.073	0.573	0.109
SM	62	37	2.8E5	4.0E5	0.545	0.064	0.583	0.138
DP	176	116	4.9E-5	1.3E-4	0.569	0.080	0.645	0.132
DT	263	109	3.5E-5	1.0E-4	0.572	0.084	0.614	0.165
CT	212	99	2.1E-6	3.7E-6	0.558	0.085	0.619	0.162

(a)

	w_v		b_v (C_v)		Acc		F1	
	avg	s.d.	avg	s.d.	avg	s.d.	avg	s.d.
SR	-	-	-	-	0.531	0.069	0.607	0.108
SM	30	15	1.8E5	3.6E5	0.563	0.077	0.627	0.122
DP	75	35	3.2E-4	7.6E-4	0.589	0.093	0.654	0.176
DT	105	72	3.7E-3	7.8E-3	0.587	0.099	0.652	0.152
CT	71	38	1.6E-5	3.3E-5	0.587	0.097	0.661	0.165

(b)

Fig. 10. Summary of the tuned parameters and quality metrics obtained across the videos for each course. b_v applies to the likelihood-based algorithms, while C_v is for SM. The avg and s.d. are taken first over the 10 evaluation sets for each video, and then over all the videos. (a) 'FMB', (b) 'NI'.

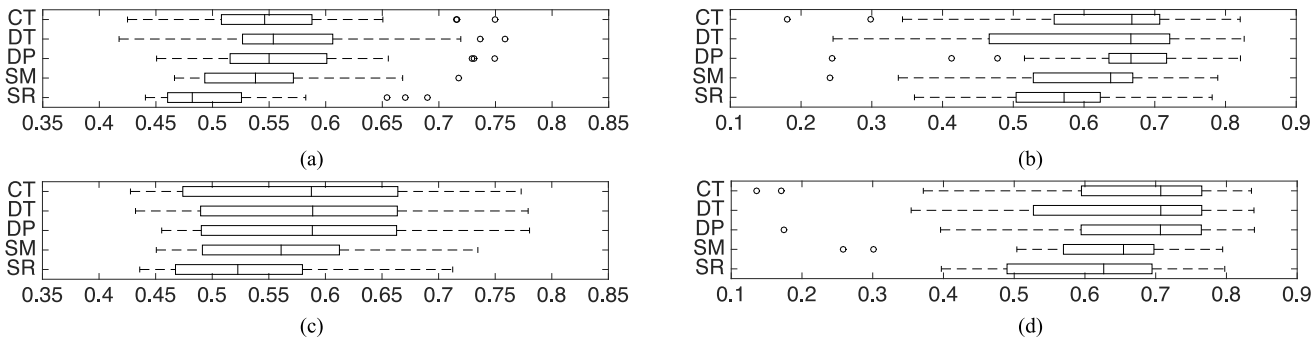


Fig. 11. Boxplots of CFA prediction quality across both courses, considering accuracy and F1. Each datapoint is that measured on one of the videos considered. Qualitatively, we see that (i) the likelihood-based algorithms (DP, DT, and CT) outperform SR for both metrics, and (ii) the SVM-based (SM) algorithm also outperforms SR, but not as substantially. (a) 'FMB', accuracy, (b) 'FMB', F1, (c) 'NI', accuracy, (d) 'NI', F1.

1) Divide the elements of $\mathcal{U}^v$ into K disjoint folds $\mathcal{U}_1^v, \mathcal{U}_2^v, \ldots, \mathcal{U}_K^v$. In doing so, randomly allocate samples of CFA and non-CFA to folds, ensuring that the number of class instances is equal across folds (*e.g.,* $|\mathcal{U}_k^{v,c}| = |\mathcal{U}_l^{v,c}| \, \forall k, l$).
2) Set $\mathcal{U}_E^v = \mathcal{U}_K^v$ and $\mathcal{U}_T^v = \mathcal{U}^v \setminus \mathcal{U}_K^v$.
3) Tune the algorithm parameters w_v and b_v (for likelihood-based) or C_v (for SVM) over the training set $\mathcal{U}_T^v$, through the **parameter tuning** procedure described below.
4) With the tuned parameters, compute the features for each algorithm over $\mathcal{U}_T^v$, and evaluate the fitted models on $\mathcal{U}_E^v$.

The obtained metrics are averaged over the N iterations. In our evaluation, we set $N = 10$ and $K = 5$.

Parameter tuning. Each algorithm has two parameters that must be tuned. Let $\mathcal{W}, \mathcal{B}$, and $\mathcal{C}$ be sets of potential values for the video width $w_v \in \mathcal{W}$, the likelihood bias $b_v \in \mathcal{B}$, and the regularization control $C_v \in \mathcal{C}$.[16] To tune these parameters for an algorithm, we apply a standard Cross-Validation (CV) procedure over the training set [25], which reduces to the following. First, for each CV iteration $k \in \{1, \ldots, K-1\}$:

1) Set $\mathcal{U}_C^v = \mathcal{U}_k^v$ and $\mathcal{U}_R^v = \mathcal{U}_T^v \setminus \mathcal{U}_k^v$.
2) For each pair $(w_v, b_v) \in \mathcal{W} \times \mathcal{B}$ or $(w_v, C_v) \in \mathcal{W} \times \mathcal{C}$, find the result from training on $\mathcal{U}_R^v$ and testing on $\mathcal{U}_C^v$.

Then, average the accuracy values for each pair of parameters over the $K-1$ CV iterations. The pair which yields the highest average accuracy is selected, subject to the constraint described with the metrics above.

[16]We set $\mathcal{W} = \{5, 10, \ldots, 20, 30, \ldots, 600\}, \mathcal{B} = \{0, 2^{-60}, 2^{-58}, \ldots, 1\}$, and $\mathcal{C} = \{10^{-7}, 10^{-6.5}, \ldots, 10^{7}\}$. For each parameter, these choices ensured that most selections across videos did not lie on one of the endpoints.

B. Results and Discussion

Since there is a sharp decline in quiz submissions over time, we only consider those for which there are at least 100 samples of both CFA and non-CFA instances, so that there at least 20 samples from each group in each of the $K = 5$ folds. We evaluate on the 24 videos for 'FMB' and the 32 for 'NI' that satisfy this criteria, which is a total of 56.

1) Overview of Results: Summary information on the tuned w_v and b_v (or C_v) parameters, as well as the two performance metrics—Accuracy (Acc) and F1—can be found for each algorithm and each course in Fig. 10. Here, we give the average (avg) and standard deviation (s.d.) of these values, taken across evaluation iterations for each video, and then across videos. The distribution of the performance values are plotted for each course in Fig. 11; in each box, the performance on one video is one data point.

From Fig. 11, we can see immediately that (i) the likelihood-based (DP, DT, and CT) algorithms perform substantially better than SR overall, and (ii) the SVM-based (SM) algorithm outperforms SR overall, but not as much as do the likelihood-based methods. Also, the improvement is higher for accuracy than it is for F1, which is expected since the tuning procedure monitors accuracy. In order to test for significance in the performance differences between each pair of models, we run a WRS test (as in Section II-B) for the null hypothesis that there is no difference between the distributions in Fig. 11. The resulting *p*-values (p) from these tests are tabulated in Fig. 12, and verify our qualitative assessment from the boxplots.

2) Comparing Individual Algorithms: We now answer the specific questions posed at the beginning of Section V.

1: Benefit of clickstream data. We quantify how beneficial clickstream data can be for prediction. To do this, we compare

	SR	SM	DP	DT	CT
SR	–	0.013*	2.5E-3**	2.2E-3**	0.018*
SM	0.013*	–	0.25	0.20	0.63
DP	2.5E-3**	0.25	–	0.75	0.72
DT	2.2E-3**	0.20	0.75	–	0.28
CT	0.018*	0.63	0.72	0.28	–

(a)

	SR	SM	DP	DT	CT
SR	–	0.36	0.014*	0.16	0.065.
SM	0.36	–	0.078.	0.15	0.15
DP	0.014*	0.078.	–	0.85	0.77
DT	0.16	0.15	0.85	–	0.98
CT	0.065.	0.15	0.77	0.98	–

(b)

	SR	SM	DP	DT	CT
SR	–	0.092.	8.0E-3**	0.019*	0.015*
SM	0.092.	–	0.25	0.32	0.30
DP	8.0E-3**	0.25	–	0.91	0.79
DT	0.019*	0.32	0.91	–	0.94
CT	0.015*	0.30	0.79	0.94	–

(c)

	SR	SM	DP	DT	CT
SR	–	0.34	0.012*	0.045*	6.3E-3**
SM	0.34	–	0.052.	0.12	0.039*
DP	0.012*	0.052.	–	0.90	0.99
DT	0.045*	0.12	0.90	–	0.86
CT	6.3E-3**	0.039*	0.99	0.86	–

(d)

Fig. 12. p-values (p) from applying pairwise WRS tests to the boxplots in Fig. 11. (a) 'FMB', accuracy, (b) 'FMB', F1, (c) 'NI', accuracy, (d) 'NI', F1.

the DP algorithm (which, from Figs. 10 and 11, appears to have the highest overall quality) to the SR baseline.

Considering accuracy first, refer to Fig. 11(a)&(c). DP is clearly shifted to the right relative to SR: for 'FMB', the shift in the mean of DP relative to SR is roughly 12%, and for 'NI', the improvement is roughly 11%. From Fig. 12, we see that this difference is also highly significant in both courses $(p < 0.01)$. For F1, refer to Fig. 11(b)&(d): again, we see that DP is shifted to the right relative to SR overall, though not quite as substantially. The increase in means of roughly 13% for 'FMB' and 8% for 'NI' are both significant ($p < 0.02$ from Fig. 12).

For further analysis of the differences, in Fig. 13 we show the comparison between DP and SR across the individual videos. The difference in the metrics obtained are shown for each video (specifically, DP minus SR). For accuracy in (a&c), we see that DP outperforms SR (*i.e.,* has a positive difference) for 98% of the videos across both datasets (all except one of the 56). For F1 in (b&d), this drops to 88% (all except seven).

Note further that in Fig. 11(b), the magnitudes on negative videos are substantially smaller than the magnitudes on the highest positive cases. In (d), however, two of the videos (3 and 27) have very high drops. One would expect that these would be instances where SR already had high performance due to a high bias (skew) in favor of either CFA or non-CFA (*e.g.,* a video with an easy or a hard quiz). Surprisingly, the opposite is true: the CFA biases are close to 0.5 (roughly 0.46 in both cases), and there are videos with smaller and larger biases for which DP outperforms SR substantially.

2: Likelihood vs. SVM. For this, we compare DP with SM. In Fig. 11, we see that DP is shifted to the right for each course and metric. While the average improvements in accuracy of roughly 4% in both courses are not statistically significant $(p > 0.1)$, the improvements in F1 of 11% for 'FMB' and 4% for 'NI' are significant $(p < 0.1)$.

In Fig. 13, we plot the difference between DP and SM for individual videos (as is done for DP and SR). The comparison here is consistent with the observations from the boxplots: the F1 improves in 71% of the videos in (b&d), while the accuracy only improves in 66% of the cases in (a&c). Note also that, contrary to the comparison between DP and SR above, for each of the metrics and datasets, the videos in which DP has the highest gain over SM show significantly larger improvements than those in which SM shows the highest gains over DP. In other words, the improvement is less variable.

So, DP does outperform SM, but not as substantially as it outperforms SR. In fact, SM has a statistically significant gain over the baseline SR in terms of accuracy ($p < 0.1$ in Fig. 12).

3: Positions vs. transitions. For this, we compare DP to DT. In terms of accuracy, in Fig. 11(a)&(c) we see that the algorithms are comparable for both courses. As for F1 in Fig. 11(b)&(d), DP is modestly better on average, especially for 'FMB' where it has an improvement of roughly 5%. DT has a higher range in each case (excluding outliers), with generally lower performance than DP below quartile Q2 (*e.g.,* in F1 for 'FMB') but, in accuracy for 'FMB', also higher above Q2. When considering individual videos,[17] DT and DP each perform better on roughly half of the videos in each course, with the exception of accuracy in 'FMB' for which DT has higher quality the majority of the time. Overall, the differences between DT and DP are not statistically significant for either course or metric ($p > 0.1$ in all cases).

4: Discrete vs. continuous. Finally, we compare DT to CT. In Fig. 11, first consider accuracy. For 'FMB', DT is shifted to the right by roughly 3% relative to CT, whereas for 'NI', the algorithms are comparable. As to the F1-score, while DT and CT are comparable overall, the distribution for CT is slightly shifted to the right for both courses. Considering individual videos, DT outperforms CT on more videos for each dataset and metric. In particular, in Fig. 13(a)&(c), it has higher accuracy in 71% of the cases, and higher F1 in 61% of the cases. Still, overall, the differences are not statistically significant for either course or metric ($p > 0.1$ in all cases).

3) Key Messages: Many aspects of position-based video behavior are useful for CFA prediction: the frequency of visits to each position (DP and SM), the frequency of transitions between positions (DT), and transitions incorporating holding times (CT). Each of these algorithms obtained higher quality than the SR baseline on both metrics and datasets tested, with statistically significant gains in most cases. Overall, the likelihood-based algorithms obtained the highest quality (with a slight edge given to DP), while the SVM-based algorithm forms a middle tier, and the SR baseline at the lowest.

The likelihood-based algorithms employ feature spaces that are representing user behavior directly; namely, positions visited and transitions. A significant advantage of this is that it leads to natural interpretations in terms of learner actions, which can be related to CFA scores for *e.g.,* content analytics. Though the

[17]We omit bar graphs over videos as in Fig. 13 for brevity.

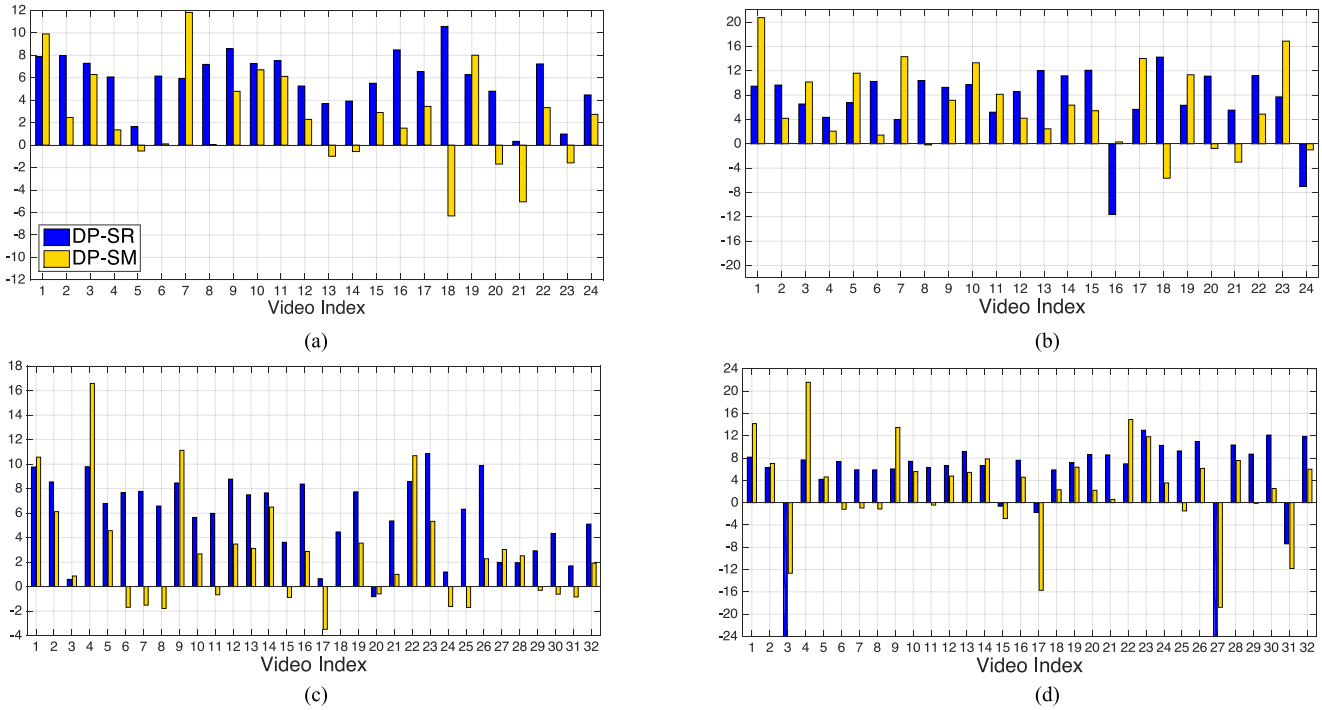


Fig. 13. Difference in quality (in percent) between DP and SR (*i.e.,* DP minus SR) and between DP and SM (DP minus SM) across individual videos, for each course and each metric. A positive bar indicates that DP performed better than the algorithm, while a negative bar indicates it performed worse. DP outperforms both algorithms in the majority of cases, but it outperforms SR in more cases than it does SM. (a) 'FMB', Accuracy, (b) 'FMB', F1, (c) 'NI', Accuracy, (d) 'NI', F1.

SVM-based model we tested is more complicated, and hence not as easily interpretable, it was still seen to obtain lower quality results than did the likelihood-based algorithms, *i.e.,* the latter has better interpretability and better quality. Still, we believe that higher quality predictions could be obtained by passing these features through algorithms with higher complexity (*e.g., kernel-based* SVM, rather than linear) to learn over higher dimensional spaces. The downside is that this would eliminate interpretability entirely, which is important for this application. Related to this, an interesting avenue of future work would be to use the position and transition matrices inferred over the CFA classes to generate recommendations guiding learner behavior to different locations in real time.

Also, in our evaluation, CFA prediction was done on a per-video basis. This underscores the applicability of these models to situations where there is not a lot of information across multiple lectures, *e.g.,* for quick detection early in a course, or for short courses that have few videos to start with.

VI. Related Work

We discuss recent, key works on MOOCs, student video-watching analysis, and student performance prediction.

MOOC studies. With the proliferation of MOOCs in recent years, there have been a number of analytical studies on these platforms. Some have focused on a more general analysis of all learning modes, *e.g.,* [7] and [20] studied learner engagement variation over time and across courses. Others have focused on specific modes, *e.g.,* in terms of forums, [5] analyzed the decline in participation over 73 courses. There has also been work on identifying taxonomies of student motivation for enrolling in MOOCs, through *e.g.,* designing and administering surveys [9] and interviews [8], and on studying how intention is predictive of course behavior. Our work is fundamentally different from these works in that it (i) studies low level video-watching behavioral details, and (ii) explores the *association* between behavior with two modes: video and assessment.

Video-watching analysis. Most existing works on learner video-watching behavior [10], [12], [13], [28] have focused on session-level user characteristics (*e.g.,* re-watching sessions) or aggregate quantities (*e.g.,* number and duration of pauses). The works of [12] and [13] identified ways in which these types of quantities are correlated with student performance on quizzes. Our work is fundamentally different from these because it represents behavior as sequences. The work in [11] is perhaps the most similar to ours in this regard, since it is also concerned with recurring patterns in clickstream sequences for MOOC users. The authors define a mapping of subsequences of events to predefined behavioral actions (*e.g.,* skipping, slow watching) and perform approximate string search to locate these behaviors in clickstreams. Our work on motif identification differs from this in two important ways: (i) rather than assuming a predefined set of actions, we extract the recurring sequences directly using motif identification algorithms, and (ii) we are concerned with mapping motifs to assessment performance, in contrast to engagement.

Grade prediction. Researchers have developed predictors for how students will perform on assessments (*e.g.,* [12], [17], [18], [27], [29]) and for what their final grades will be (*e.g.,* [14], [30], [31]) in courses, some with application to traditional education settings and others for distance learning. Several techniques have been applied for this purpose, such as collaborative filtering algorithms [18], [27], support vector machines [12],

[31], and probabilistic graphical models [29], [30]. Recently, [17] developed SPARFA-Trace, which traces a learner's knowledge through the sequence of material accessed and questions answered. [14] proposed an algorithm to predict the final grade of each student after each assessment in a class based on the past history of students' performance. Compared with these works, ours is unique in that (i) it focuses on relating click-level data—video-watching behavior—to assessment performance, and (ii) it focuses on prediction within single videos for earliest detection. In this regard, [12] studied the predictive capability of session-level video-watching quantities computed from clickstream data (*e.g.,* the fraction of the video watched and the number of rewinds), considering multiple users and videos in the course simultaneously. Focusing on individual videos, our models are instead position-dependent, and the improvements in accuracy relative to the baseline that we obtain are strictly higher than those cited here (3% increase to the same baseline). Overall, we emphasize that the models used in each of these other works are not readily applicable to our setting, because we focus on the case of individual videos where similarities among users/quizzes is not available.

Webpage clickstream analysis. Webpage clickstream analysis [21], [32], [33] remains an active area of research. Video-watching clickstreams are fundamentally different from these applications, which are concerned with transitions between webpages rather than behavior within a single window.

VII. Conclusion and Extensions

In this work, we have studied student video-watching behavior, quiz performance, and their association in MOOC. In doing so, we have formalized two frameworks for representing user clickstreams: one based on sequences of events with discretized lengths, and one based on sequences of positions visited. With datasets from two MOOCs encoded in these frameworks, we have accomplished two main goals: (i) we have mined the sequences to identify recurring motifs in user behavior, and discovered that some of these characteristics are significantly associated with CFA and non-CFA quiz submissions; and (ii) we have proposed models for relating user clickstreams to knowledge gained, and showed how multiple aspects of this behavior can improve CFA prediction quality on individual videos.

The models we have proposed here can be extended for future work in several ways. For one, the event-based sequence representation in Section II can be generalized to optimize the selection of quantiles used to divide the event lengths. Also, the position-based models in Section IV can be extended to consider higher-order transitions and durations under a non-exponential assumption, to see whether the prediction quality in Section V can be improved further.

More generally, recall that students can have different motivations for taking MOOCs in the first place. In this work, we have limited our scope to those students who are interested in answering questions, which leaves an important future step to account for differing motivations in the behavioral analysis. The most objective way for this to be done is perhaps to release a questionnaire at the beginning of the course asking students to indicate their intentions for enrolling. Then, behavioral characteristics could be identified for each group separately, and the results compared.

Related to this is the fact that we have focused on *video-watching* behavior and *in-video quiz* performance in this work. In accounting for different motivations, the definition of performance can be adjusted depending on what the student hopes to achieve, like high quiz scores, a broadened social network, or some combination of different metrics. Similarly, behaviors from different learning modes can be incorporated; in MOOCs, this includes forum discussions, and in settings beyond MOOCs, it can include behavior exhibited on any type of content integrated into the course.

Finally, we remark that the true test of the methods we have developed here is their impact on instructor interventions. We have shown that certain video-watching motifs are correlated with quiz scores, and quantified the quality of behavior-based CFA prediction, but a larger question still remains: *How will an instructor make use of the motifs and predictions?* To investigate this, we are currently working with the learning technology company Zoomi Inc.[18] to integrate these algorithms into an instructor dashboard that is being deployed to various learning scenarios. With these analytics in the hands of course instructors, we will determine their effects by monitoring overall changes in student performance based on the interventions that are made as a result.

Acknowledgment

The authors thank the three anonymous reviewers for their valuable comments.

References

[1] C. G. Brinton and M. Chiang, "Social learning networks: A brief survey," in *Proc. Conf. Informat. Systems Sci.*, Princeton, NJ, 2014, pp. 1–6.

[2] C. G. Brinton, R. Rill, S. Ha, M. Chiang, R. Smith, and W. Ju, "Individualization for education at scale: MIIC design and preliminary evaluation," *IEEE Trans. Learn. Technol.*, vol. 8, no. 1, pp. 136–148, 2015.

[3] D. Clow, "MOOCs and the funnel of participation," in *Proc. 3rd ACM Conf. Learning Analyt. Knowledge*, Leuven, Belgium, pp. 185–189, 2013.

[4] D. Yang, T. Sinha, D. Adamson, and C. P. Rose, "Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses," in *Proc. NIPS Data-Driven Educ. Workshop*, 2013, pp. 14.

[5] C. G. Brinton, M. Chiang, S. Jain, H. Lam, Z. Liu, and F. M. F. Wong, "Learning about social learning in MOOCs: From statistical analysis to generative model," *IEEE Trans. Learn. Technol.*, vol. 7, no. 4, pp. 346–359, 2014.

[6] K. Jordan, "Initial trends in enrollment and completion of massive open online courses," in *Int. Rev. Res. Open Distrib. Learn.*, pp. 1–28. vol. 15, no. 1, 2014.

[7] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Engaging with massive online courses," *WWW. ACM*, pp. 687–698, 2014.

[8] S. Zheng, M. B. Rosson, P. C. Shih, and J. M. Carroll, "Understanding student motivation, behaviors and perceptions in MOOCs," *ACM Comput. Support. Cooperat. Work Social Comput.*, pp. 1882–1895, 2015.

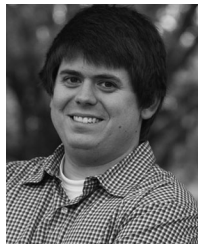
[9] R. F. Kizilcec and E. Schneider, "Motivation as a lens to understand online learners: Toward data-driven design with the OLEI scale," in *ACM Trans. Computer-Human Inter.*, vol. 22, no. 6, pp. 1–24, 2015.

[10] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller, "Understanding in-video dropouts and interaction peaks in online lecture videos," in *Proc. 3rd ACM Conf. Learning@Scale*, Edinburgh, Scotland. pp. 31–40, 2014.

[11] T. Sinha, P. Jermann, N. Li, and P. Dillenbourg, "Your click decides your fate: Inferring information processing and attrition behavior from MOOC video clickstream interactions," in *Proc. 2014 Conf. Empirical Methods in Natural Language Processing*, Doha, Qatar, pp. 3–14, 2014.

[18] www.zoomiinc.com

[12] C. G. Brinton and M. Chiang, "MOOC performance prediction via clickstream data and social learning networks," in *Proc. IEEE Conf. Computer Communications (INFOCOM)*, Hong Kong, China. 2015.
[13] N. Li, Ł. Kidziński, P. Jermann, and P. Dillenbourg, "MOOC video interaction patterns: What do they tell us?," *Design Teach. Learn. Networked World*, pp. 197–210, 2015, Springer.
[14] Y. Meier, J. Xu, O. Atan, and M. van der Schaar, "Predicting grades," *IEEE Trans. Signal Process.*, vol. 64, no. 4, pp. 959–972, 2015.
[15] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, "Tuned models of peer assessment in MOOCs," in *Proc. 6th Int. Conf. Educational Data Mining*, pp. 153–160, 2013.
[16] K. Stephens-Martinez, M. A. Hearst, and A. Fox, "Monitoring MOOCs: Which information sources do instructors value?," in *Proc. 3rd ACM Conf. Learning@Scale*, Edinburgh, Scotland, pp. 79–88, 2014.
[17] A. S. Lan, C. Studer, and R. G. Baraniuk, "Time-varying learning and content analytics via sparse factor analysis," *Proc. 20th SIGKDD Inter. Conf. Knowledge Discovery and Data Mining*, New York, NY, pp. 452–461, 2014.
[18] A. Toscher and M. Jahrer, "Collaborative filtering applied to educational data mining," in *Proc. Knowledge Discovery and Data Mining Cup 2010: Educational Data Mining Challeng.*, [Online] Available: http://pslcdatashop.org/KDDCup/workshop/papers/KDDCup2010_Toescher_Jahrer.pdf 2010.
[19] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon, "Individualized Bayesian knowledge tracing models," *Artificial Intelligence in Education*. New York, NY, USA: Springer, 2013, pp. 171–180.
[20] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses," in *Proc. 3rd ACM Conf. Learning Analytics and Knowledge*, Leuven, Belgium, pp. 170–179, 2013.
[21] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao, "You are how you click: Clickstream analysis for sybil detection," in *Proc. 22nd USENIX Security Symp.*, Washington, DC, pp. 241–256, 2013.
[22] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, FL, USA: CRC Press, 2003.
[23] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, "MEME suite: Tools for motif discovery and searching," *Nucleic Acids Res.*, p. gkp335, 2009.
[24] D. Bates, M. Maechler, B. Bolker, and S. Walker, "lME4: Linear mixed-effects models using eigen and S4," *R Package Version*, vol. 1, no. 7, 2014.
[25] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.
[26] J. R. Norris, *Markov Chains*. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[27] Y. Bergner, S. Droschler, G. Kortemeyer, S. Rayyan, D. Seaton, and D. E. Pritchard, "Model-based collaborative filtering analysis of student response data: Machine-learning item response theory," in *Proc. 5th Inter. Conf. Educational Data Mining*, pp. 95–102, 2012.
[28] J. M. Aiken, S.-Y. Lin, S. S. Douglas, E. F. Greco, B. D. Thoms, M. D. Caballero, and M. F. Schatz, "Student use of a single lecture video in a flipped introductory mechanics course," 2014, DOI: arXiv 1407.2620.
[29] Z. Pardos and N. Heffernan, "Using HMMs and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset," in *Proc. Knowledge Discovery and Data Mining Cup 2010: Educational Data Mining Challenge.* [Online] Available: http://pslcdatashop.org/KDDCup/workshop/papers/pardos_heffernan_KDD_Cup_2010_article.pdf, 2011.
[30] E. Osmanbegović and M. Suljić, "Data mining approach for predicting student performance," in *J. Econom. Rev.*, vol. 10, no. 1, pp. 3–12, 2012.
[31] C. Romero, M.-I. López, J.-M. Luna, and S. Ventura, "Predicting students' final performance from participation in online discussion forums," *Elsevier Comput. Educ.*, vol. 68, pp. 458–472, 2013.
[32] S. Gündüz and M. T. Özsu, "A web page prediction model based on clickstream tree representation of user behavior," in *Proc. Ninth SIGKDD Inter. Conf. Knowledge Discovery and Data Mining*, Washington, DC., pp. 535–540, 2003.
[33] M. Speiser, G. Antonini, A. Labbi, and J. Sutanto, "On nested palindromes in clickstream data," in *Proc. 18th SIGKDD Inter. Conf. Knowledge Discovery and Data Mining*, Bejing, China. pp. 1460–1468, 2012.



Christopher G. Brinton (S'08) received the PhD Degree in EE from Princeton University in May 2016, the Master's Degree in EE from Princeton in May 2013, and the BSEE from The College of New Jersey in May 2011. Currently, he is the Head of Advanced Research at the learning technology company Zoomi Inc. His research focus is in big learning data analytics, social learning networks, and integrated and individualized courses. He is a MOOC co-instructor, having reached out to over 250,000 students since 2012, and co-author of the book *The Power of Networks: Six Principles that Connect our Lives*.

Swapna Buccapatnam is a postdoctoral researcher at the IBM T. J. Watson Research Center, Yorktown Heights, NY, USA. Prior to this, she was a postdoctoral research associate in the Department of Electrical Engineering at Princeton University. She received her Ph.D. in electrical and computer engineering from the Ohio State University in 2014 and her undergraduate degree in electrical engineering from the Indian Institute of Technology, Madras, India, in 2008. Her research interests lie in stochastic modeling and analysis, machine learning, data analytics, and optimization.

Mung Chiang (S'00–M'03–SM'08–F'12) is the Arthur LeGrand Doty Professor of Electrical Engineering at Princeton University, Princeton, NJ. His research on communication networks received the 2013 Alan T. Waterman Award from the U.S. National Science Foundation, the 2012 Kiyo Tomiyasu Award from IEEE, and various young investigator awards and paper prizes. A TR35 Young Innovator Award recipient, he created the Princeton EDGE Lab in 2009 to bridge the theory-practice divide in networking by spanning from proofs to prototypes, resulting in several technology transfers to industry and two startup companies. He is the Chairman of the Princeton Entrepreneurship Advisory Committee and the Director of the Keller Center for Innovations in Engineering Education. His MOOC in social and technological networks reached about 200,000 students since 2012 and lead to two undergraduate textbooks and he received the 2013 Frederick E. Terman Award from the American Society of Engineering Education. He was named a Guggenheim Fellow in 2014.

H. Vincent Poor (S'72–M'77–SM'82–F'87) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990 he has been on the faculty at Princeton, where he is the Michael Henry Strater University Professor of Electrical Engineering and Dean of the School of Engineering and Applied Science. Dr. Poor's research interests are in the areas of stochastic analysis, statistical signal processing, and information theory, and their applications in wireless networks and related fields such as social networks and smart grid. Among his publications in these areas is the recent book *Mechanisms and Games for Dynamic Spectrum Allocation* (Cambridge University Press, 2014). Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and a foreign member of Academia Europaea and the Royal Society. He is also a fellow of the American Academy of Arts and Sciences, the Royal Academy of Engineering (U.K.), and the Royal Society of Edinburgh. He received the Technical Achievement and Society Awards of the IEEE Signal Processing Society in 2007 and 2011, respectively. Recent recognition of his work includes the 2014 URSI Booker Gold Medal, the 2015 EURASIP Athanasios Papoulis Award, the 2016 John Fritz Medal, and honorary doctorates from several universities in Europe and Asia.