Mining MOOC Clickstreams: On the Relationship Between Learner Video-Watching Behavior and Performance

Christopher G. Brinton, Swapna Buccapatnam, Mung Chiang, and H. Vincent Poor
Department of Electrical Engineering
Princeton University, Princeton, NJ 08544
{cbrinton, swapnab, chiangm, poor}@princeton.edu

ABSTRACT
We study student behavior and performance in two Massive Open Online Courses (MOOCs). In doing so, we present two frameworks 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, we extract recurring subsequences of student behavior, which contain fundamental characteristics such as reflecting (i.e., repeatedly playing and pausing) and revising (i.e., plays and skip backs). We find that some of these behaviors are significantly associated with whether a user will be Correct on First Attempt (CFA) or not in answering quiz questions. With the position-based framework, we then devise models for performance. In evaluating these through CFA prediction, we find that three of them can substantially improve prediction quality in terms of accuracy and F1, which underlines the ability to relate behavior to performance. Since our prediction considers videos individually, these benefits also suggest that our models are useful in situations where there is limited training data, e.g., for early detection or in short courses.

Keywords
Clickstream Data, Video-Watching Behavior, Motif Identification, Data Mining, Performance Prediction, MOOC

1. INTRODUCTION

Over the past decade, technology advances have been influencing the ways we can learn. One such innovation has been the Massive Open Online Course (MOOC), with platforms such as Coursera, edX, and Udacity offering MOOCs that have reached hundreds of thousands of students within single sessions. The low completion rates in these courses, caused in part by small teacher-to-student ratios and the asynchronous nature of interaction, has ignited research interest in understanding how MOOC users learn (e.g., [7, 11, 5]).

A standard MOOC contains three different learning modes for students: video lectures, assessments (e.g., in-video quizzes, homework, and exams), and social networking [6]. Most platforms track student interaction with these forms of learning. For video, this information includes clickstream events, which are generated each time a learner interacts with a video. For assessments, the specific responses to individual questions are tracked, and for the discussion forums, the sequence of posts and comments are stored. This type of data has motivated a number of recent studies focused on understanding how MOOC users learn (e.g., [7, 11, 5]).

What remains understudied, however, is the relationship between these learning modes. In particular, is it possible to associate a student’s behavior with his/her performance in a MOOC? This question has far-reaching implications to methods for improving low completion rates, such as personalized content delivery [6] and instructor analytics [13].

Our work is motivated by this fundamental question. In our investigation, we focus on the video-watching behavior of MOOC students, where users spend the majority of their time learning [7]. These videos are typically equipped with quiz questions, which serve as immediate feedback of the knowledge a student gained from the content in the video. In relating behavior to performance, then, we can consider the clickstreams generated by a user in watching the video associated with a particular quiz, and whether the user was Correct on First Attempt (CFA) or not in answering the given question.

In our investigation, we formalize different ways in which video-watching clickstreams can be represented as compact sequences, and apply the frameworks we develop to meet two objectives:

- **O1. Identifying recurring behaviors of learners**, such as revising content or skipping forward repeatedly.
- **O2. Assessing the impact of behavior on performance**, such as those patterns identified in O1, and the specific positions visited in each video.

In doing so, we employ two datasets coming from two different MOOCs, which contain (after filtering) 315K and 416K clickstream event logs corresponding to 26K and 30K first-attempt quiz submissions.

Previous work [11] has focused on the sequence of events (e.g., play, pause, skip forward) generated by MOOC users in watching videos. In studying O1 and O2, we identify two additional factors that are important to capture: the positions in the video that a user visited, and the duration / length of time between the events and positions. In particular, we first develop an event-based framework to represent clickstreams (Sec. 2), which captures event types and their lengths. Leveraging this framework, we find recurring subsequence *motifs* in our two datasets, for O1 (Sec. 3), as well
as a significant difference in the presence of certain motifs across CFA and non-CFA sequences, for O2. For example, we find that a series of behaviors are indicative of students reflecting on material, and are significantly associated with the CFA sequences in one of the courses. As another example, we identify motifs that are consistent with rapid-paced skimming through the material, and reveal that these are discriminatory in favor of non-CFA in both courses. Incorporating the lengths in addition to the events was essential to these findings, because motif extraction with the events alone does not reveal these insights.

In investigating O2, we also seek to develop models for knowledge gained based on user clicks in a video. The quality of such a model can be evaluated by considering its ability to generalize to incoming samples through prediction. To this end, we will study CFA prediction, which is an important area of research in its own right because it can, for example, help in early detection of struggling and advanced students, and of easy and difficult material [4].

In seeking appropriate models for student performance, we find that while some behavioral patterns of the motifs are significantly associated with performance, their supports and the resulting success estimates are likely not sufficient for prediction. As a result, we propose a second behavioral representation, which uses the sequence of positions visited in a video (Sec. 4). Moreover, in contrast to training over a long course duration as in [4, 8], we consider CFA prediction on a per-video basis, in order to quantify the benefit obtained by the positions in each individual video. We evaluate four different models based on this framework (Sec. 5), and find that three of them obtain substantial improvements in prediction when compared to a baseline that does not use click information. This underscores the ability to relate clicks to knowledge gained, and shows that behavioral information is useful in situations where multiple videos are not available, e.g., in short courses or for detection early in a course.

Summary of contribution. Compared with other work (Sec. 6), we make three contributions. First, we present two novel frameworks for representing clickstream sequences, which we are useful in identifying recurring behavioral patterns and for CFA prediction. Second, we extract recurring subsequences from user clicks using motif identification schemes, and associate these fundamental behaviors with student performance. Third, we show how user click behavior can be used to enhance prediction on a per-video basis.

2. DATASETS AND CLICKSTREAMS

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

2.1 Two Courses

Our datasets come from two different courses we have instructed on Coursera: Networks: Friends, Money, and Bytes ('FMB') and Networks Illustrated: Principles Without Calculus ('NI'). Each of these courses teach networking, but

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Lectures</th>
<th>Lecture Videos</th>
<th>Video Length (min) avg. (s.d.)</th>
<th>Quizzes</th>
</tr>
</thead>
<tbody>
<tr>
<td>'FMB'</td>
<td>20</td>
<td>92</td>
<td>16.9 (5.96)</td>
<td>92</td>
</tr>
<tr>
<td>'NI'</td>
<td>6</td>
<td>115</td>
<td>5.44 (2.17)</td>
<td>69</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Users Clickstream Events</th>
<th>User-Video Pairs</th>
<th>CFA Score avg. (s.d.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3770</td>
<td>314,632</td>
<td>0.663 (0.473)</td>
</tr>
<tr>
<td>20</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>115</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Basic information on the two datasets. The values in the right column are the final numbers after data filtering.

'FMB' delves into mathematical specifics, whereas 'NI' is meant as an introduction to the subject (see [6] for details). We obtained two types of data from each course: video-watching clickstreams, which log user interaction with the video player, and quiz submission information.

Course format. The course formats are summarized in Fig. 1. Each contains a series of lectures, 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 and an average (avg.) length of 5.4 min per video, whereas 'FMB' has a total of 93 with an avg. length of 16.9 min.

Each course has in-video quizzes, in the form of single multiple choice questions with four choices each. For 'FMB', there was one question at the end of each video. For 'NI', each of the 69 questions was associated with anywhere from 1 to 4 videos. As a result, in mapping videos to quizzes, we will refer to "video X" as the contiguous set of videos occurring after question X – 1 and before question X.

User-Video Pairs. Given our goal of relating user behavior to performance, we extract User-Video (UV) Pairs from the data, which consisting of two sets of information:

- Video-watching trajectory: The set of clickstream logs (events) for the user in the video.
- CFA result: Whether the user was Correct on First Attempt (CFA) or not for the corresponding quiz.

In total, there were 122.5K UV Pairs for 'FMB', with 566K click events. For 'NI', these numbers were 149K and 882K, respectively. After removing any UV Pair that had at least one null, stall or error recorded, we were left with the numbers given in Fig. 1 for each course. The avg. CFA score across the UV Pairs was 0.663 for 'FMB' (standard deviation (s.d.) = 0.47), and 0.750 for 'NI' (s.d. = 0.43).

2.2 Clickstream Events

2.2.1 Our nomenclature for events

Clickstream logs are generated as one of four types: play, pause, ratechange, and 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 $i$th click event that occurs while a user is watching a video. We write $E_i = (e_i, p_i, t_i, s_i, r_i)$, where $e_i$ is the type of the $i$th click, $p_i$ is the video position (in sec) 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 — as a result of $E_i$, 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 these $E_i$, 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.

1. www.coursera.org/course/{friendsmoneybytes,ni}
We do not insert Sb, with rate change and the rate \( r \). We combine repeated, sequential events to our online technical report [2] for more details. Associated duration \( d \) for \( \hat{r} \) (\( R_f \)): This event occurs when \( \hat{r} \) is rate change and \( r \leq 1 \). There is no associated duration.

**Skip forward (Sf):** A skip forward (\( r \), fast forward) event is defined as Sb, except it captures the case where \( p_i > p' \).

**Rate change fast (RF):** This event occurs when \( E_i \) is a rate change and the rate \( r_i > 1.0 \). There is no \( d \) or \( l \).

**Rate change slow (RS):** This occurs when \( e_i \) is rate change and \( r_i < 1 \), again with no duration or length.

**Rate change default (RD):** This occurs when \( e_i \) is rate change and \( r_i = 1 \), \( i.e. \), returning to the default.

The sequence of events for a UV Pair then becomes \( \hat{e}_1, \hat{e}_2, ... \) for \( \hat{e}_j \in \mathcal{E} = \{ \text{Pl, Pa, Sb, ...} \} \), \( |\mathcal{E}| = 8 \). Each \( \hat{e}_j \) may have an associated duration \( d_j \) and/or length \( l_j \). Fig. 2 shows a schematic to illustrate this; the clickstream logs here would generate: Pl, with \( l_1 = (t_2 - t_1) \cdot r_1 \) and \( d_1 = t_2 - t_1 \); Sf, with \( l_2 = p_2 - p_1 \); Pl, with \( l_3 = d_3 \); Pa, with \( d_4 = t_4 - t_3 \); Sb, with \( l_5 = p_4 - p_1 \). 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 only the events in the raw data, and has been captured in other work (\( e.g. \), in [11]).

**Denosing clickstreams.** In order to remove noise associated with unintentional user behavior, we handle two cases of events separately. We state those here and refer the reader to our online technical report [2] for more details:

**Combining events:** We combine repeated, sequential events \( E_i \) and \( E_{i+1} \) that occur within a short duration (5 sec) of one another, since this indicates that the user was adjusting to the final state (\( e.g. \), skipping back to find a specific location).

**Discounting intervals:** We do not insert play and pause events in-between \( E_i \) and \( E_{i+1} \) if (i) they occur on two different videos (\( e.g. \), the user closed one video and opened another) or (ii) the duration \( d_i \) is extremely long (\( e.g. \), the user was engaging in some off-task behavior). For (ii), if \( s_i = \text{pause} \), the threshold is 20 min (as in [14] for web inactivity), and if \( s_i = \text{play} \), it is the length of the video.

### 2.2.2 Event lengths

We now look to discretize the length \( l \) and duration \( d \) of the events for comparative purposes. To this end, 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. Also, in these plots, 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, as are the number of events for each distribution (Size) and the respective fractions (Frac).

**Event intervals.** We see from Fig. 3 that \( l \) and \( d \) can vary substantially between events and datasets (see our technical report [2] for comparative details). 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:

1. \( \hat{e}_j \in \{ \text{Sb, Sf} \} \): 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 \( l_j \in [Q_{q_j-1}, Q_{q_j}] \), 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 case, this would be mapped to Sb2.

2. \( \hat{e}_j = \text{Pa} \): In this case, the mapping works the same as the previous, except \( q_j \) is chosen based on \( d_j \) instead.

3. \( \hat{e}_j = \text{Pl} \): Two long duration play events could still have different qualitative interpretations. To account for this, when \( \hat{e}_j = \text{Pl} \), we map it to \( (\hat{e}_j, q_j) \), where \( q_j \in \{1, 2, 3\} \) for \( k = 1, ..., n \) is chosen as: 3 if \( d_j > \delta_j > Q_3 \); and as \( \text{arg max}_{s_{i,j}} (d_j - \delta_i \leq Q_{s_{i,j}}) \) otherwise, with \( \delta_j = \sum_{k=1}^{j-1} Q_{q_k} \). For example, suppose an event is Pl with \( d_j = 550 \) sec. For the quartiles in ‘NI’, this would be mapped to Pl13 P13 P12.

### 2.2.3 Event-type sequence specification

Let \( S = \{P1, P12, P13, Pa1, ..., Pa4, Sb1, ..., Sb4, Sf1, ..., Sf4, Rf, Rs, Rd\} \), with \( |S| = 18 \). For each UV Pair, we encode the clickstream log \( E_1, ..., E_n \) as \( S = (s_1, s_2, ..., s_n) \) where each \( s_j \in S \) is chosen according to the specifications in Sec. 2.2.2. As we will see in Sec. 3, using an alphabet that incorporates both event types and lengths allows us to obtain insights that are difficult to glean with events alone.

For purpose of comparison, we will refer to a length of

---

3 By definition, quartiles separate data in increments of 25%.

4 These standard quartiles will lead to interesting insights in Sec. 3. More generally, we can optimize the way the distributions are divided, which we leave for future work.

5 The other events do not have this issue since they are not related to processing new information.
3. MOTIFS OF VIDEO-WATCHING

Using the event-type specification, we identify short, recurring sub-sequences within user behavior, i.e., behavioral motifs. As we will see in Sec. 3.2, these motifs capture fundamental video watching characteristics such as reflecting or revising material. We will also see that some of these motifs are significantly associated with performance.

3.1 Motif Extraction

In order to find the motifs in our datasets, we make use of a well-known, open-source software package called MEME Suite [1]. MEME has been widely applied in bioinformatics, for motif identification in sequences of nucleotides and amino acids. The algorithm underlying MEME 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. We find this model, and the underlying expectation-maximization algorithm of MEME, to be applicable to our setting. For more details on MEME and our encoding of events for processing in the software, see our technical report [2].

Through MEME, we obtain the E-value and the position specific probability matrix (PSPM) for each identified motif. An E-value judges the overall significance of the motif; more precisely, it is the number of subsequences that would have equal or higher likelihood of being generated according to the position-independent background distribution. Typically, only motifs that have E-value ≤ 0.05 are considered.

**Representation.** The PSPM gives the fraction of times that each character appears in each position of a motif, taken over all sittings of the motif. Denote the PSPM for a motif by \( P = [p_{ij}] \), where \( p_{ij} \) is the fraction of times event \( j \) occurs at position \( i \). At each position, we consider all events \( j \) with \( p_{ij} \geq 0.25 \). Formally, let \( A_i \) be the sequence of indices into the event set \( S \) for \( i \), arranged such that \( p_{i,A_i(k)} \geq p_{i,A_i(k+1)} \) and \( p_{i,A_i(k+1)} \geq 0.25 \). There are three cases on the way \( i \) is represented. If \( |A_i| = 1 \), then the square brackets are omitted, with just \( S_{A_i} \) displayed. If \( A_i = \emptyset \), then \( i \) is displayed as \( * \) 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 \( [P12 P13] Pa1 \times [Sf1 Sf2 Sf4] \) is of length 4, with the first position being either P12 or P13 at least 50% of the time (P12 at least as often as P13), the second position being Pa1 at least 25% of the time, the third position being any event, and the last being either Sf1, Sf2, or Sf3 at least 75% of the time.

**Support.** For each motif, we obtain the fraction of sequences (FS) in which it occurs, i.e., its support, as well as the number of videos it appears in (see our technical report [2] for the procedure). In order to see how the supports vary between correct and incorrect submissions, we do this for CFA and non-CFA sequences separately. Then, for each motif, the estimated probability of success \( \hat{p} \) (i.e., of a CFA submission) related to a sequence containing this motif is \( 0.5 + FS1 - FS0 \), where FS1 and FS0 are the fraction of sequences in CFA and non-CFA for which the motif appears.

\(^8\)With 18 different events, a threshold of 25% is roughly 5 times the expected occurrence from random selection.

In order to test whether \( \hat{p} \) is significant, we run a two-sample test for proportions [10] for the null hypothesis that there is no difference between FS1 and FS0. If the \( p \)-value \( (p) \) for this test is low enough (≤ 0.05), then there is a large enough difference given the supports to conclude \( \hat{p} \) is significant.

3.2 Results

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

**Motif overview.** We first analyze how the motif supports vary across sequences and videos. Overall, we find that the motifs are reasonably supported across sequences and videos on average, for both CFA and non-CFA in each course.

**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. The supports are similar across these groups: for ‘FMB’, each motif appears in 5.9% of the non-CFA sequences on average, and 6.5% of the CFA; for ‘NI’, this is 5.8% for CFA, and 4.2% of the non-CFA. In both courses, the motifs with largest support (first row in Fig. 6(a) and (b)) appear in > 25% of the sequences.

**Videos:** Fig. 5 gives the ECDF of the number of videos that each motif occurred in at least once and at least 10 times. Here, CFA has higher support than non-CFA. We also see that the supports decrease for higher thresholds.\(^7\) For ‘FMB’ in (a), we can see that while the top 20% of the motifs appear in at least 67 videos for CFA (50 for non-CFA), this number is only 18 considering at least 10 occurrences (8 for non-CFA). For ‘NI’ in (b), the top 20% appear in at least 64 videos for CFA (44 for non-CFA), and this number drops to 14 for at least 10 occurrences (2 for non-CFA).

3.2.1 Individual motifs

We now list the most significant motifs, listed in Fig. 6. To generate this table, we applied the following procedure

\(^7\)We see the same trends for thresholds besides 10 as well.
### (a) 19 Motifs for ‘FMB’.

<table>
<thead>
<tr>
<th>Group</th>
<th>Motif</th>
<th>E-value</th>
<th>FS</th>
<th>F50</th>
<th>FS1</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pa</td>
<td>P12 Pa4 Pa5 Pa4</td>
<td>1.8E-06</td>
<td>13.2</td>
<td>13.2</td>
<td>50.1</td>
<td>0.991</td>
</tr>
<tr>
<td>H</td>
<td>P12 Pa4 Pa5 Pa4</td>
<td>3.6E-108</td>
<td>5.7</td>
<td>5.7</td>
<td>10.0</td>
<td>8.08</td>
</tr>
<tr>
<td>H</td>
<td>P12 Pa4 Pa5 Pa4</td>
<td>6.8E-40</td>
<td>10.0</td>
<td>10.0</td>
<td>50.0</td>
<td>3.6E-108</td>
</tr>
<tr>
<td>H</td>
<td>P12 Pa4 Pa5 Pa4</td>
<td>3.8E-85</td>
<td>3.6</td>
<td>3.6</td>
<td>10.0</td>
<td>8.08</td>
</tr>
</tbody>
</table>

### (b) 21 Motifs for ‘NI’.

<table>
<thead>
<tr>
<th>Group</th>
<th>Motif</th>
<th>E-value</th>
<th>FS</th>
<th>F50</th>
<th>FS1</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pa</td>
<td>P12 Pa4 Pa5 Pa4</td>
<td>3.8E-105</td>
<td>9.3</td>
<td>9.3</td>
<td>50.3</td>
<td>0.665</td>
</tr>
<tr>
<td>H</td>
<td>P12 Pa4 Pa5 Pa4</td>
<td>8.84E-07</td>
<td>7.1</td>
<td>7.1</td>
<td>10.0</td>
<td>8.84</td>
</tr>
<tr>
<td>H</td>
<td>P12 Pa4 Pa5 Pa4</td>
<td>3.8E-85</td>
<td>3.6</td>
<td>3.6</td>
<td>10.0</td>
<td>8.08</td>
</tr>
<tr>
<td>H</td>
<td>P12 Pa4 Pa5 Pa4</td>
<td>3.8E-205</td>
<td>6.88</td>
<td>6.88</td>
<td>10.0</td>
<td>8.84</td>
</tr>
</tbody>
</table>

Figure 6: Identified motifs for each course. Each motif is grouped by the events it contains outside of Pl. FS is the fraction of success (CFA) if a sequence contains the motif, and the p-value (p) is the significance of the motif. The motif with the highest support in ‘FMB’ – Pa I – can be viewed as a long sequence of medium to long pauses followed by short to long pauses (Pa1 – Pa4).

The motif with the highest support in ‘FMB’ – Pa I – can be viewed as a long sequence of medium to long pauses followed by short to long pauses, which is also characteristic of Pa IV in ‘FMB’. This behavior occurs more often in CFA class in both cases (p ≤ 0.01). Motifs Pa III and Pa IV in ‘FMB’ as well as Pa V in ‘NI’ are long sequences too, but consist of short to medium pauses followed by short to medium pauses and do not distinguish between CFA and non-CFA (p > 0.07). Motifs Pa II in ‘FMB’ and Pa I in ‘NI’ are also shorter sequences, but with medium to medium-long pauses followed by long pauses, and also do not significantly differentiate between CFA and non-CFA (p > 0.3).

Note that in ‘NI’, the Pa group exhibits less discriminative capability. For this reason, we do not draw conclusions on differences between the classes from these sequences.

### Revising (Sb, Sb+P4).

From the six motifs in the Sb group, we identify two interesting, recurring subsequences: P12 Sb3 P12 Sb3 (Sb I and II for ‘FMB’, and Sb I for ‘NI’), and P12 Sb2 P12 Sb2 (Sb III and IV for ‘FMB’ and Sb II for ‘NI’). Roughly speaking, each of these is associated with playing for a length of video, and then revising some or all of that content. To see this, consider the ranges of Pl and Sb from Fig. 3 associated with these subsequences: P12 covers 14 to 68 sec for ‘FMB’, 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, which are local considering the video lengths. This is further justified by the fact that Sb, a long skipback, does not appear in these subsequences. Note that 5 of the 6 motifs containing these behaviors are significantly associated with CFA ($p \leq 0.01$).

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 the video. In particular, out of all Sb events, those originating within 10 sec of the videos’ end constitute 16% and 13% of the total for non-CFA and CFA in ‘FMB’. This, combined with the motifs suggesting revision when Sb occurs, implies that those students who are revising multiple times before answering a quiz have a higher chance of success.

Consistent with the observation for Sb, 4/5 of the motifs in Sb+Pa are significantly associated with CFA ($p \leq 0.02$).

**Skimming (Sf, Sf+Sb).** 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 $Q_1$ to $Q_{j+1}$ for Sf is always larger than $Q_{j-1}$ to $Q_1$ for Pl. This recurring behavior in the Sf group can then be interpreted as skimming through the material quickly with less exposure to the material. We find that 3 of these 4 motifs are significant in favor of non-CFA ($p \leq 0.03$). We contrast this to a finding in [4], where the total number of skip forwards in a sequence was not found to be significantly associated with either CFA or non-CFA. This underscores the utility of considering the clickstream sequences, rather than computing aggregate quantities to summarize them.

While Sb+Sf can possibly be interpreted as skipping forward with caution, we find that this is also close to being significant in favor of non-CFA ($p \leq 0.06$).

**Speeding.** Referring to Rf+Rd in ‘FMB’, motifs I and III indicate that viewing the material at a faster than default rate, i.e., speeding, is more often associated with the CFA class than not ($p < 0.03$). With these motifs, learners also return to the default rate, indicating they are slowing down for important content. To this point, in ‘FMB’, we see no significant motifs for slower than default rates; however, one does exist in ‘NI’ (Rf+Rd+Rs). Also, Rf II in ‘NI’ is more significantly associated with non-CFA ($p = 0.05$), which could indicate that a faster rate is harmful in this course.

### 3.2.2 Key messages

Overall, we make a few conclusions from our analysis:

**Reflecting.** Pausing to reflect on material repeatedly is the most commonly occurring behavior. If the time spent reflecting is not too long, but longer than the time spent watching, then a positive outcome is most likely (in ‘FMB’).

**Revising.** Repeated revision of the material suggests students will gain a better understanding of the content.

### 3.2.3 Overall analysis

We begin with a few definitions. Let $v \in V$ denote video $v$ in the set of videos $V$ for a course, indexed chronologically (i.e., by release date of the videos). Also, let $c \in C$ denote class $c$ in the set of binary classes $C = \{0, 1\}$, where $c = 0$ indicates a non-CFA submission and $c = 1$ is CFA. With $u \in U$ as user $u$ in the set of all users $U$, we let $U^c \subseteq U$ be the set of users who have a UV Pair for $v$, and $U^c \subseteq U$ be those who fall into class $c$ with respect to their answer submission. For evaluation in Sec. 5, we will generate training ($U^c_{tr}$) and test ($U^c_{te}$) sets as subsets of $U^c$ through a procedure described in [4]; $U^c_{tr}$ and $U^c_{te}$ are always chosen such that $U^c_{tr} \cap U^c_{te} = \emptyset$.

#### 4.1 Modeling Framework

We 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 $N(w_v) = [h_v/w_v]$ uniform intervals, such that interval $i \in P'(w_v) = \{1, \ldots, N(w_v)\}$ spans the range $\{(i-1)\cdot w_v, i\cdot w_v\}$. For each UV Pair, we can then model the behavior as a sequence of positions $p^{v,u,v} = (p_1, p_2, \ldots, p_{n-1})$, where $p_c \in P'(w_v)$ is the index of the nth position visited.

In generating these positions, we first apply the same de-noising procedure described in Sec. 2.2.1 to each event $E_i$. Then, for each UV Pair, we do the following, starting with $p$ as an empty sequence:

1. **Skimming.** Skimming through material quickly, even with caution, is costly in terms of knowledge gained.

2. **Speeding.** Students who watch the videos at a faster than default rate may already be familiar with the material, leading to a correct answer (in ‘FMB’). They also may slow to the default if they sense unfamiliar material.

Based on these conclusions, we emphasize the importance of having included the lengths, in addition to the events, in this framework. As an example, the sequence $P_1 Sb P_1 Sb$ identified in [11] cannot be associated with revising, because it is not clear how far back the student has skipped relative to having played. In the same way, $P_1 Sf P_1 Sf$ cannot be concluded as skimming.

From these findings, it is clear that clickstream motifs are useful in studying behavior, and that they can be significantly related to performance. In terms of using them to model behavior for CFA prediction, however, there are two drawbacks. First, while the supports are reasonable considering these are rather long subsequences, none of the motifs appear in a majority of the sequences (max 28.5%). Second, none of the $p$ success estimates deviate substantially from 50% (max 3.3%). Hence, we will now turn to an alternate clickstream sequence representation which is more applicable to CFA prediction. Nonetheless, some of the conclusions made from this analysis will guide our modeling choices.

#### 4.1.1 Position-based sequence specification

We 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 $N(w_v) = [h_v/w_v]$ uniform intervals, such that interval $i \in P'(w_v) = \{1, \ldots, N(w_v)\}$ spans the range $\{(i-1)\cdot w_v, i\cdot w_v\}$. For each UV Pair, we can then model the behavior as a sequence of positions $p^{v,u,v} = (p_1, p_2, \ldots, p_{n-1})$, where $p_c \in P'(w_v)$ is the index of the nth position visited.

In generating these positions, we first apply the same de-noising procedure described in Sec. 2.2.1 to each event $E_i$. Then, for each UV Pair, we do the following, starting with $p$ as an empty sequence:

1. **Recall from Sec. 2.1 that we consider a “video” to be all videos for a given quiz.
2. **We choose the last interval to be of variable size extending to the end of the video.
3. **For brevity, we will typically refer to $p^{v,u,v}$ as just $p$, with the understanding that it refers to the UV Pair in question.
(1) For \( E_1 \), add \( |p_1|/w_0 \) to \( p \).

(2) Consider each sequential pair of events \( E_i, E_{i+1} \), \( i \geq 1 \). If \( s_i = \text{pause} \), then only \( |p_{i+1}|/w_0 \) is added to \( p \). But if \( s_i = \text{play} \), then: if \( e_i \notin \{\text{Skip}, \text{ng}, \text{p}, \text{ng}+1, \ldots, |p_{i+1}|/w_0 - 1 \} \), \(|p_i|/w_0 \) is added to \( p \); if \( e_i = \text{Skip} \), then the addition is instead \(|p_i|/w_0 + 1, \ldots, |p_{i+1}|/w_0 - 1, |p_i|/w_0, |p_{i+1}|/w_0 \).

For example, suppose \( b_0 = 300, w_0 = 15 \), and a user generates \( E_1 = (\text{Play}, 0, 0, \text{playing}, 1, 0) \), \( E_2 = (\text{skip}, 200, 50, \text{playing}, 1, 0) \), \( E_3 = (\text{ratechange}, 230, 50, \text{playing}, 1, 25) \), and \( E_4 = (\text{pause}, 300, 127, \text{paused}, 1, 25) \) on the video. Then, \( p = (0, 1, 2, 3, 13, 14, 15, 16, \ldots, 20) \).

### 4.1.2 Model factors

In general, there are (at least) three types of information for each \( p^{v,c} \) that could have an effect on performance:

**Positions.** The number of times a given \( p_i \) was visited. One would expect these to be different between CFA and non-CFA, because certain parts of the videos will be more important to the questions than others. To see this, we can refer to two types of motifs which were identified as associated with CFA: reflecting, which indicates that these sequences may have more visits to important positions through pausing, and revising, which suggests that these sequences may have more visits to positions associated with the questions through repeated revision before answering the question. Further, the skimming motif suggests that non-CFA sequences will have less visits to important positions.

**Transitions.** The number of transitions between the positions, i.e., the number of times \( p_i, p_j \) is a subsequence of \( p^{v,c} \), across each pair \( i, j \). This is related to the type of event that was executed between \( i \) and \( j \):

- If \( p_j < p_i \), then the user had skipped back. We call this a backward transition.
- If \( p_j > p_i + 1 \), then the user had skipped over the material in \((p_i, p_j)\). We call this a forward transition.
- If \( p_j = p_i + 1 \), then the user moved directly to the next position. This is a direct transition.
- If \( p_j = p_i \), 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, but in terms of sequences of visits.

**Time spent.** The amount of time spent at the different positions. One would expect these times to be discriminatory in a similar manner to visit frequencies.

In order to evaluate the benefit of including each of these factors, we will consider four prediction models: Discrete time Positions (DP), which incorporates the number of visits to each position; Continuous time Positions (CP), which uses the time spent at each position; Discrete time Transitions (DT), which models transitions between positions; and Continuous time Transitions (CT), which factors in inter-arrival times between positions. Each of these models are trained on each video separately, which allows us to compare prediction quality on a per-video basis in Sec. 5.4.2

### 4.2 Position Models

For the DP and CP models, video positions are treated as independent events.

**Discrete Time Positions (DP).** Let \( f^{v,c} = [f_i]^{v,c} \in [0, 1]^{|N(w)|} \) be the probability distribution of visit frequency across positions \( i \in P^v(w) \). This is estimated over the UV Pairs in the training set \( \mathcal{U}^{v,c} \) as

\[
\hat{f}_i^{v,c} = \frac{O_i^{v,c}}{\sum_j O_j^{v,c}},
\]

where \( O_i^{v,c} \) is the number of occurrences of \( p_i \) over sequences in \( \mathcal{U}^{v,c} \), i.e.,

\[
O_i^{v,c} = \sum_{u \in \mathcal{U}^{v,c}} I_{\{p_i \in u\}}.
\]

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

\[
L(p | f^{v,c}) = g^{v,c} \cdot \prod_{j} f_j^{v,c},
\]

where \( g^{v,c} = [\mathcal{U}^{v,c}]/|\mathcal{U}| \) is the estimated class bias for video \( v \). Then, the prediction \( \hat{c} \), which could have an effect on performance:

**Continuing Time Positions (CP).** In this model, we consider the holding times within each position. Let \( r^{v,c} = [r_i]^{v,c} \in \mathcal{R}^{|N(w)|} \) be the vector of the total time spent by \( U^{v,c} \) in state \( i \). This is estimated as

\[
r_i^{v,c} = \sum_{u \in \mathcal{U}^{v,c}} \sum_n \xi_{(p_{n-1} = i)} \cdot d_n,
\]

where \( d_n \) is the duration of event \( n \) in \( p \) (see Sec. 2.2.1).

We will assume that the holding times at each position are exponentially distributed. The reason for this is to have a direct comparison with the CT model in Sec. 4.3, where the interarrival times also have this distribution. The estimates in (4) then become the means of the distributions, and the likelihood of \( p \) for a UV Pair in \( \mathcal{U}^{v,c} \) is

\[
L(p | r^{v,c}) = g^{v,c} \cdot \prod_{i} (1/r_i^{v,c}) \exp \left( -T_i / r_i^{v,c} \right),
\]

where

\[
T_i = \sum_n \xi_{(p_{n-1} = i)} \cdot d_n
\]

is the total time spent by sequence \( p \) at position \( i \). The MAP for CP is the same as (3), except substituting (5) for (2).

### 4.3 Transition Models

In modeling transitions between positions, we will only consider one-step transitions. This is common in webpage clickstream analysis (e.g., [14]), and will be useful here since the state spaces we consider can be large, depending on \( w_0 \).

**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 \( j \) [9]. 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.

\[14\]This may not be ideal because unlike sequences of webpage, 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.
To see this, we inspect the $p^{v,c}$ for varying $w_v$. For each $v$, we find the total of the four types of transitions for each position, sum over positions, and then average. We do this for all $w_v \in \{5, 10, \ldots, 600\}$ (i.e., through 10 min), and then average across $v$. By comparing the resulting fractions across $w_v$ for each course, we find that the vast majority of transitions are local, i.e., either repeat or direct. In particular (see our technical report [2] for a plot), the largest fraction of backward transitions is 2.3% ($w_v = 120$) for ‘FMB’ and 1.5% for ‘NI’ ($w_v = 60$), and that for forward transitions is 2.4% (for $w_v = 150$) for ‘FMB’ and 1.2% for ‘NI’ ($w_v = 70$).

As a result, our models will aggregate all observed forward transitions to form a single, uniform probability at each position, and likewise for backward. To this end, we define $I_{i,k} = \{1, \ldots, i-1\}$ for $k = 1; \{i\}$ for $k = 2; \{i+1\}$ for $k = 3; \{i + 2, \ldots\}$ 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 $F^{v,c} = \{f_{i,k}\}^{v,c} \in [0,1]^{N(w_v),4}$ be the matrix of transition probabilities, where $f^{v,c}_{i,k}$ is the probability that the next position will be in $I_{i,k}$ given the current is $i$. We also assume that the transitions are homogeneous, i.e., independent of time $n$.

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

$$O^{v,c}_{i,k} = \sum_{w\in U^{v,c}_r} \sum_n I_{\{\rho_n=i, \rho_{n+1}\in I_{i,k}\}}.$$ 

From this, we estimate the $f^{v,c}_{i,k}$ as

$$f^{v,c}_{i,k} = \frac{O^{v,c}_{i,k}}{\sum O^{v,c}_{i,j}},$$

and the likelihood of $p$ from user $u \in U^{v,c}_r$ on video $v$ is

$$L(p | F^{v,c}) = g^{v,c} \cdot f^{v,c}_{p_1} \prod_{j=2}^{n} f^{v,c}_{p_{j-1},p_{j}},$$

where $f^{v,c}_{p_1}$ is the distribution at the initial position $p_1$ of $p$, obtained from (1). The MAP for DT is again as same as in (3), except with (8) 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 [9]. To this end, let $Q^{v,c} = \{q^{v,c}_{i,k}\}^{v,c} \in \mathcal{R}^{N(w_v),4}$ be the transition rate matrix for the model, where $q^{v,c}_{i,k} \neq 2$ represents the rate of departure from position $i$ and arrival at a position in $I_{i,k}$. In estimating the $q^{v,c}_{i,k}$, we first obtain the number of transitions from $i$ to $k$ over users $u \in U^{v,c}_r$ as

$$O^{v,c}_{i,k} = \sum_{w\in U^{v,c}_r} \sum_n I_{\{\rho_n=i, \rho_{n+1}\in I_{i,k}\}}, k \neq 2.$$ 

Then, the $q^{v,c}_{i,k}$ terms are computed as

$$q^{v,c}_{i,k} = \begin{cases} 
O^{v,c}_{i,k} / r^{v,c}_{i,k} & k \neq 2 \\
\sum_{k \neq 2} q^{v,c}_{i,k} & k = 2,
\end{cases}$$

where $r^{v,c}_{i,k}$ is the total time spent in $i$, defined in (4). From this, the likelihood of sequence $p$ for $u \in U^{v,c}_r$ is obtained:

$$L(p | Q^{v,c}) = g^{v,c} \cdot \prod_{i,k \neq 2} \left( q^{v,c}_{i,k} \cdot \exp\left(-\frac{p^{v,c}_{i,k}}{T_i}\right) \right),$$

where $T_i$ is the time spent by $p$ in $i$, as defined in (6), and $n_{i,k} = \sum_n I_{\{\rho_n=i, \rho_{n+1}\in I_{i,k}\}}, k \neq 2$ is the number of transitions from $i$ to $k$ for the sequence $p$. Once again, MAP is as in (3), except with (10) in place of (2).

### 5. Prediction Evaluation

In this section, we evaluate the performance of the models described in Section 4. We pose the following questions:

1. How beneficial is it to include positions and transitions for CFA prediction on individual videos?
2. Is one of position or transition-based model clearly better than the other, or would some combination be the best?
3. Is it beneficial to include position durations?

#### Skewed-Random (SKR).

In our investigation, we also consider one algorithm that does not make use of clickstream data, to act as a baseline for comparison [4]. It finds the CFA bias $g^{v,c} \cdot F^{v,c}_{i-1}$, and predicts $c = 1$ $g^{v,c}$ of the time.

### 5.1 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 through evaluation over $U^{v,c}_r$. In selecting the model parameters $w_v$ and $b_v$ over $U^{v,c}_r$, we choose those leading to highest accuracy, i.e., $(TP + TN)/(TP + FP + TN + FN)$. Since the quizzes are biased towards CFA (see Fig. 3), we found that unconstrained maximization of accuracy led to parameters with high recall (rec), i.e., $(TP/(TP + FN))$ but low precision (prec), i.e., $(TP/(TP + FP))$. To avoid this, we subject the search to the constraint that the chosen parameters have at least 25% of the truly negative samples predicted negative, and likewise for the positives, averaged over the training evaluations. To this end, we will also consider the standard (balanced) F1 score for each model, obtained as $2 \cdot (prec + rec)/(prec + rec)$. As the harmonic mean of prec and rec, F1 is limited by the minimum of the two.

#### Cross Validation (CV).

As stated in Sec. 4.1, $U^{v,c}_r$ and $U^{v,c}_E$ are always chosen such that $U^{v,c}_r \cap U^{v,c}_E = \emptyset$. To do this, we divide $U^{v,c}$ into K-folds $U^{v,c}_1, U^{v,c}_2, \ldots, U^{v,c}_K$, and average the results of training the model on $U^{v,c}_k = U^{v,c} \setminus U^{v,c}_k$ and testing the model on $U^{v,c}_k = U^{v,c}_k$. This K-fold CV process is repeated N times, and the results averaged over the N trials. In generating $U^{v,c}_k \forall k$, we ensure that the number of positive and negative samples is consistent across each fold. To do this, for each CV iteration we randomly allocate $|U^{v,c}_k|/K$ elements from $U^{v,c}_E$ and $|U^{v,c}_k|/K$ elements from $U^{v,c}_1$ into $U^{v,c}_k$ (without replacement). For our experiments, we set $N = 10$ and $K = 5$, for a total of 50 runs in each CV iteration.

#### Parameter tuning.

Each algorithm has two parameters that must be tuned: the video width $w_v$, and the likelihood bias $b_v$. We treat both of these as discrete parameters, with a fine granularity of search. In particular, in each CV iteration, we choose a different pair $(w_v, b_v) \in \{5, 10, \ldots, 20, 30, 45, \ldots, 600\} \times \{0, 2^{-56}, 2^{-57.5}, \ldots, 1\}$, for a total of 182 iterations for each $v$. In the end, we select the combination which yields the highest accuracy, subject to the constraint stated above. For $w_v$, we choose this set of values since (i) 5 sec corresponds to the threshold of combining repeat events (see Sec. 2.2.1), and (ii) 600 is close to the minimum video length in both courses. For both parameters, these choices were seen to ensure that most selections across videos did not lie on one of the grid endpoints.

### 5.2 Results and Discussion

After tuning, we evaluate each algorithm through CV. Since there is a sharp dropoff in quiz participation over time, we only consider those for which there are at least 100 samples of both CFA and non-CFA classes, so that there are at least 20 samples from each class in each of the 5 CV folds. This
leaves a total of 24 videos for ’FMB’ and 32 for ’NI’.

**Overview of results.** Summary information on the tuned $w_v$ and $b_v$ values, as well as the two performance metrics – Accuracy and F1 – can be found for each course in Fig. 7, where we give the mean and s.d. across the tested videos. We expect the performance to be higher for accuracy than for F1 score, because we chose parameters based on this metric through tuning. The distribution of these values across videos are also plotted for each course in Fig. 8; in each box, the performance on one video is one data point.

From this, we can see immediately that the DP, DT, and CT algorithms perform the best overall, and that CP is comparable to SKR.\textsuperscript{15} A larger window has a denoising effect, because it aggregates more information within each position and transition; the fact that CP chooses the largest window size on average in both courses (over 5.5 min) is an indication that the holding times alone can be noisy.

In order to test for significance in the performance differences between each pair of models, we run a Wilcoxon Rank Sum (WRS) test \cite{10}, for the null hypothesis that there is no difference between the distributions in Fig. 8.\textsuperscript{16} The p-values from these tests are tabulated in our technical report \cite{2}, and verify the observed differences.

**1: Benefit of clickstream data.** We assess how beneficial the position and transition information is for prediction.

**Accuracy.** Considering accuracy first, we refer to Fig. 8.\textsuperscript{(a&c)} Here, we see that the DP, DT, and CT models are clearly shifted to the right relative to SKR, indicating higher quality. This difference is also statistically significant or each algorithm across both courses ($p \leq 2.0E-3$). For both ’FMB’ and ’NI’, the average improvement of these three algorithms relative to SKR is roughly 14% in each case. Further, each algorithm outperforms SKR in each individual video. CP, on the other hand, only has a slight edge over SKR, with an average improvement of less than 2% for each course.

**F1:** For F1, we refer to Fig. 8.\textsuperscript{(b&c)} Again, we see that DP, DT, and CT are shifted to the right relative to SKR overall, but not as substantially. This is especially true for DT, which has a high range of F1 scores (0.63 and 0.53 for ’FMB’ and ’NI’ excluding outliers). For DP and CT, the average improvements in F1 of roughly 16% for ’FMB’ and 10% for ’NI’ are again significant ($p \leq 4.0E-3$); for CP, the differences of roughly 9% for both courses are also significant but not as substantially ($p \leq 0.04$). For individual videos, DP and CT both outperform SKR in roughly 90% of the cases; for DT, 76% for ’FMB’ and 88% for ’NI’. CP performs roughly the same as SKR, with the average slightly lower.

**2: Positions vs. transitions.** We compare DP to DT, and CP to CT, case by case.

DP vs. DT: In terms of accuracy, DP and DT are comparable for both courses. As for F1, DP has modestly better performance, especially for ’FMB’ where it has an improvement over DT of roughly 6%. Additionally, the lower quartiles for DT are shifted to the left relative to DP. When considering individual videos, however, the results are more mixed: for each course, DT and DP each perform better in roughly 50% of the cases. Overall, the differences between DT and DP are not statistically significant ($p \geq 0.78$).

CP vs. CT: CT is superior to DP for each metric and dataset, with the results significant in all cases ($p \leq 4.3E-3$).

**3: Discrete vs. continuous.** Finally, we compare discrete to continuous. For brevity, we only present CT and DT (since CT outperforms CP). In terms of accuracy, both algorithms are comparable. As to F1, CT is modestly better for ’FMB’, with an improvement of roughly 4%. However, the differences are not statistically significant. CT outper-
forms DT in half of videos for ‘FMB’, and 59% for ‘NI’.

**Key messages.** Many aspects of position-based video behavior are useful for CFA prediction: the frequency of visits to each position (DP), the frequency of transitions between positions (DT), and transitions incorporating holding times (CT). These benefits are also measured on individual videos, which 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. The holding times alone (CP), however, were seen to be too noisy. Both positions and transitions can be useful; DP, DT, and CT are comparable overall, performing better on different sets of videos, suggesting an ensemble may work best.

### 6. RELATED WORK

We discuss recent, key works on MOOC, and student video-watching analysis and CFA prediction. For a more comprehensive discussion, refer to our technical report [2].

**MOOC studies.** There were a number of recent analytical studies on MOOC platforms, some focusing on specific learning modes (see [3] for a survey). Our work is fundamentally different from these in that it explores the association between behavior with two modes: videos and assessments.

**Video-watching analysis.** Existing work on video watching behavior [7, 4] has focused session-level user characteristics (e.g., rewatching sessions), rather than click-level information. The work in [11] is most similar to ours, 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 in two important ways: (i) rather than assuming a predefined set of behavioral actions, we extract the most prominent motifs using motif identification algorithms, and (ii) we are concerned with mapping motifs to efficacy, in contrast to [11] where the objective is to predict engagement, next click, and dropout.

**Performance prediction.** Researchers have developed predictors for whether a student will be CFA or not in traditional education settings. Recently, [8] developed SPARFA-Trace, a framework which traces a learner’s knowledge evolution through the sequence of material accessed and questions answered. Our work is unique in that (i) it focuses on relating click-level data – video-watching behavior – to performance, and (ii) it focuses on prediction within single videos. In general, there has been a lack of work on CFA prediction for MOOC, where the fraction of assessments a user completes can be much less [5]. The recent work of [4] studied the predictive capability of session-level video-watching quantities computed from clickstreams (e.g., fraction of the video watched), considering multiple users and videos in the course simultaneously. Focusing on individual videos, our models are position-dependent, and the improvements in accuracy relative to the baseline that we obtain are higher (14% vs. 9.5% for the same benchmark).

**Webpage clickstream analysis.** While webpage clickstream analysis [14, 12] is an active area of research, video-watching clickstreams are fundamentally different than these applications, which concern transitions between webpages rather than behavior within a single window.

### 7. CONCLUSION

In this work, we studied student video-watching behavior, performance, and their association in MOOC. In doing so, we 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 accomplished two goals: (1) we 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; (2) we 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.

There are a number of next steps we are investigating, e.g., to use the identified motifs for user and content analytics; to optimize the selection of quantiles used divide the event lengths based on the resultant motifs; to consider position durations under a non-exponential assumption; and to see whether prediction improvement can be obtained through higher order transitions.

### 8. REFERENCES


