JOURNAL OF LEARNING ANALYTICS

Volume 6(1), 16—33. http://dx.doi.org/10.18608/jla.2019.61.2

Discovery and Temporal Analysis of MOOC Study Patterns

Mina Shirvani Boroujeni[1]*, Pierre Dillenbourg[2]

Abstract

The large-scale and granular interaction data collected in online learning platforms such as massive open online courses (MOOCs) provide unique opportunities to better understand individuals' learning processes and could facilitate the design of personalized and more effective support mechanisms for learners. In this paper, we present two different methods of extracting study patterns from activity sequences. Unlike most of the previous works, with post hoc analysis of activity patterns, our proposed methods could be deployed during the course and enable the learners to receive real-time support and feedback. In the first method, following a hypothesis-driven approach, we extract predefined patterns from learners' interactions with the course materials. We then identify and analyze different longitudinal profiles among learners by clustering their study pattern sequences during the course. Our second method is a data-driven approach to discover latent study patterns and track them over time in a completely unsupervised manner. We propose a clustering pipeline to model and cluster activity sequences at each time step and then search for matching clusters in previous steps to enable tracking over time. The proposed pipeline is general and allows for analysis at different levels of action granularity and time resolution in various online learning environments. Experiments with synthetic data show that our proposed method can accurately detect latent study patterns and track changes in learning behaviours. We demonstrate the application of both methods on a MOOC dataset and study the temporal dynamics of learners' behaviour in this context.

Keywords
Learning Analytics, LA, EDM, Sequence mining, Study pattern, Temporal analysis, MOOCs, Markov model, Clustering

Submitted: 05.06.2018 — **Accepted:** 30.10.2018 — **Published:** 15.04.2019

[1] *Computer Human Interaction in Learning and Instruction School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne, Route Cantonale,1015 Lausanne, Switzerland, shirvani.mina@gmail.com*

[2] *Computer Human Interaction in Learning and Instruction School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne, Route Cantonale,1015 Lausanne, Switzerland, pierre.dillenbourg@epfl.ch*

1. Introduction

A large number of learners with different motivation and background knowledge participate in courses offered by online learning environments, such as massive open online courses (MOOCs) and intelligent tutoring systems (ITS). Online learners often have different preferences, constraints, and learning styles. Such individual differences combined with the flexibility of online learning platforms for navigating through the learning materials could result in various engagement patterns. Educational data-mining and learning analytics communities have become more interested in mining sequential interaction logs to identify behavioural patterns of learners in recent years. Using computational methods to discover patterns in students' activity traces could provide insight into their learning strategies and make learning environments more adaptable and more personalized. Different aspects of educational data and interaction logs could be analyzed to shed light on learners' engagement patterns in the online course. In our previous work, we investigate online participation patterns across *time* (Shirvani Boroujeni, Kidzinski, & Dillenbourg, 2016; Shirvani Boroujeni, Sharma, et al., 2016) and *social* (Shirvani Boroujeni et al., 2017) dimensions. In Shirvani Boroujeni, Sharma, et al. (2016), we study patterns in the timing of learners' study sessions and introduce methods of quantifying the level of temporal regularity in terms of following a certain time schedule. Our analysis shows that MOOC learners with a repeating weekly study pattern get better grades in the course. In Shirvani Boroujeni et al. (2017), we study the evolution of social interactions among learners and analyze how learners' roles in MOOC discussion forums change over time. In this work, we focus on a third dimension of educational data, *activity*, and consider the types and sequence of actions performed by learners. This paper is an extended version of our previous work on mining MOOC activity sequences (Shirvani Boroujeni & Dillenbourg, 2018).

Our objective in this work is to investigate MOOC study patterns and perform temporal analysis of learners' longitudinal behaviours over the duration of a course. Previous research on MOOC activity sequence mining often focuses on characterizing relatively short interaction episodes as a composition of learners' interaction with different course materials (Faucon et al., 2016; Geigle & Zhai, 2017). In this work, however, we aim to identify and track learners' study patterns during MOOC assessment periods, that is, their learning sequence from the opening time of an assignment (when it is assigned) until the submission deadline. Most of the previous works in this domain are only applicable once the course is finished because they analyze the full activity sequence. In this work, on the other hand, we provide methods that can be applied during the course to detect changes in behavioural patterns while the course is running. This in turn could enable timely intervention and real-time support to target learners who might require further attention. Moreover, many of the previous studies, such as Faucon et al. (2016), citethansen2017sequence, and citetshih2010unsupervised, overlook the temporal dynamics of students' behaviour because they assume that learners have a fixed study approach. However, this assumption is not necessarily true. Learners might change their study approach during the course depending on the context of the new assignment or, for instance, on whether they find their previous approach ineffective or inefficient. We consider this aspect in our work and perform temporal analysis of interaction patterns to explore the evolution of learning approaches over time. The research questions we aim to answer in this work could be summarized as follows:

- ***What are the different study patterns of learners during MOOC assessment periods, and how do they evolve over time?***

One potential approach to addressing this problem is to define the set of possible study patterns and look for their instances in learners' activity sequences. This approach is referred to as the hypothesis-driven method in the literature and is suitable for examining theoretically grounded learning strategies. However, it relies on human expert knowledge to define all the possible learning approaches, which is not always an easy or even a feasible task. Another possibility is to directly derive the latent study patterns from the data without any prior assumption on how they should look. Such a data-driven approach could possibly enable learning patterns undetected by the first hypothesis-driven approach to be identified.

In this work, we present two methods of answering our research questions. First, in a hypothesis-driven approach, we label students' activity sequences according to predefined patterns and perform clustering to identify prototypical participation trajectories over the duration of the course. Second, we propose a data-driven approach to automatically capture undefined study patterns from interaction sequences. We introduce a complete processing pipeline, which starts by modelling learners' activity sequences, applies clustering to identify common study patterns based on the modelled sequences, and performs cluster matching to enable learning approaches to be tracked over time. We present detailed descriptions of both methods and their results.

The rest of this paper is organized as follows. Section 2 reviews related work, and section 3 presents the dataset. Sections 4 and 5 present our hypothesis-driven and data-driven approaches for analyzing learners' study patterns and the obtained results. Section 6 provides a discussion of the results and concludes the paper.

2. Related Work

Existing methods for extracting learners' study patterns could be classified into two overall categories: *hypothesis-driven* and *data-driven* methods. **Hypothesis-driven** methods aim to detect predefined learning styles from interaction sequences and rely on human expert knowledge to define the set of possible learning approaches. For instance, a hypothesis-driven approach is used in Kizilcec et al. (2013) to classify interaction sequences of MOOC learners into four categories: *on track* (on-time submissions), *behind* (late assignment submissions), *auditing* (watching the videos, without submitting the assignment), and *out* (no participation in the course at all). The hypothesis-driven approach could be used to mine theoretically grounded learning styles. However, due to the complexity of students' behaviour, it is not often feasible to accurately define, a priori, the set of possible learning patterns. **Data-driven** methods could be used to overcome this limitation and enable unsupervised discovery of concrete behavioural patterns from learners' interaction data. In this approach, human intervention in the process is reduced to assessing the validity and utility of the system findings.

Clustering methods and in particular clustering of sequential data have received growing attention in this domain because they allow semi-automatic or open-ended behavioural style detection. In some studies, learners' activity sequences are compared in their original format using sequence similarity measures (Bergner et al., 2014; Desmarais & Lemieux, 2013; Patel et al., 2017; Shen & Chi, 2017), whereas other works use a summarized form or an aggregated representation of the fine-grained activity sequences (Geigle & Zhai, 2017; Klingler et al., 2016; Shih et al., 2010). In this work, we also use summarized forms and provide methods of modelling activity sequences. As opposed to the original sequences, activity models could provide an aggregated view of learners' actions in the course platform and enable comparison of sequences with different lengths. The challenge, however, is to create models that can capture important information from the original sequences and are also easy to interpret. Common techniques for modelling and analyzing activity sequences include sequential pattern mining (Kinnebrew et al., 2013), Markov chain (Faucon et al., 2016; Hansen et al., 2017), hidden Markov models (HMMs) (Geigle & Zhai, 2017; Jeong & Biswas, 2008), and process mining (Trcka et al., 2010). In the following, we provide examples of these techniques applied to educational research.

Sequential pattern-mining methods (Agrawal & Srikant, 1995) seek the most frequent patterns across a set of action sequences. Nesbit et al. (2007) applied this method to study self-regulation behaviours in a multimedia learning environment, and Maldonado et al. (2011) used it to identify frequent interaction sequences that differentiate high- and low-achieving groups in a collaborative tabletop activity. Similarly, Kinnebrew et al. (2013) used sequence-mining techniques in combination with time-series segmentation to identify and compare segments of productive and unproductive learning behaviours. In this work, we use sequence mining to identify and extract different learning behaviours among MOOC learners.

Markov chain representation aggregates sequences of learners' actions into memory-free state transition models, which encode the probability of performing one action type after the other. This common approach is used in educational data-mining research to model activity sequences. Markov transition models partially encode temporal information about the order of the actions in the original sequences and provide visualizations that are often easy to interpret. In Faucon et al. (2016), Markov chains are used to model learners' interaction logs as transition probabilities between different learning activities, and the expectation-maximization (EM) algorithm is used to identify behaviour profiles that characterize groups of similar students. Similarly, students' activities are modelled and clustered based on different similarity measures such as Euclidean distance (Köck & Paramythis, 2011) and Jensen–Shannon divergence (JSD) (Klingler et al., 2016), defined on the transitional probabilities in the Markov chain models. In this work, we also model transition probabilities between different action types, and, through clustering of these transition models, we identify common study patterns among learners (see Section 5.1 for details on our modelling and clustering approach).

HMMs have also been broadly applied to model learning processes in online learning environments. The use of HMM-based clustering techniques for automatic discovery of students' learning strategies in a tutoring system is investigated in Shih et al. (2010). In Li & Biswas (2000), HMMs are used to extract stable groups from temporal data by joint optimization of the model parameters and the cluster count. In Geigle & Zhai (2017), a two-layer HMM is proposed to discover students' behavioural patterns and the transition between them over time. Following this approach, the authors identify four behavioural patterns (states) for MOOC students: *low activity*, *active*, *forum browsing*, and *passive*. By contrasting the state transition of high- and low-performing students, the authors show that high-performing students concentrate for longer on quizzes and participate for longer in forums.

Process mining (Van der Aalst et al., 2004) is another technique that has been applied on educational data to analyze learning processes (Mukala et al., 2015; Bannert et al., 2014). This technique, which originates from the business community, could be adopted to compare students' interaction patterns with predefined models (conformance checking) or to discover the underlying process model from the activity sequences (process discovery). As an example, the compliance between students' video-watching behaviours in a MOOC course and the predefined sequential video-viewing model are assessed in Mukala et al. (2015), revealing that successful students are more likely than unsuccessful students to study sequentially. However, for large-scale unstructured data, such as interaction logs from thousands of students in online courses, the discovered process models are often "spaghetti-like," showing all details and failing to distinguish the important trends (Günther & van der Aalst, 2007). This makes process-discovery methods in their original format inefficient for identifying study patterns in the MOOC context.

3. Dataset

The dataset used for this study consists of the interaction logs of participants in "Functional Programming Principles in Scala", an undergraduate engineering MOOC produced by EPFL university. The course was composed of seven sets of video lectures and six graded assignments. Course materials were released on a weekly basis. Submissions to each assignment were accepted before the (hard) deadline, and the assessment periods (assignment release to hard deadline) varied between 11 and 18 days. The final grade was computed as the average of the assignment grades, with a passing threshold of 60 out of 100. The dataset includes three categories of events, describing learners' interaction with video lectures (play, pause, download, seek, change speed), assignments (submit), and discussion forums (read, write, vote a message).

3.1 Data Preprocessing

To analyze learners' study patterns during the assessment periods, we split the full sequence of interaction logs into subsequences corresponding to each assessment period. As the assessment periods of assignments might overlap, we refine the resulting subsequences to contain learners' interactions with only the materials of the corresponding week. This process is illustrated in Figure 1. In our analysis, we consider learners who were active in at least three assessment periods. Following these data-preprocessing steps, the final dataset used in this study contains interaction subsequences of 7527 learners during six assessment periods.

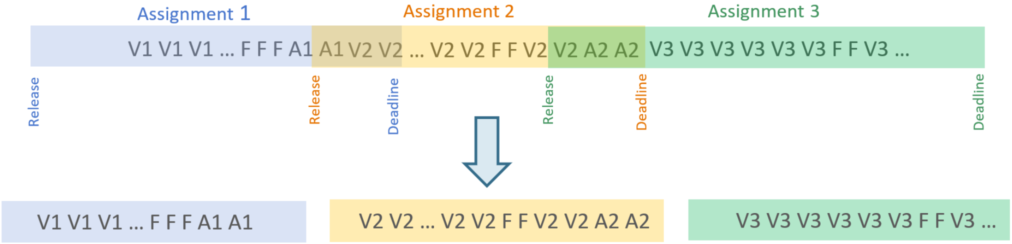

Assignment 1
Assignment 2
Assignment 3
V1 V1 V1 ... F F F A1 A1 V2 V2 ... V2 V2 F F V2 V2 A2 A2 V3 V3 V3 V3 V3 V3 F F V3 ...
Release
Release
Deadline
Release
Deadline
Deadline
V1 V1 V1 ... F F F A1 A1
V2 V2 ... V2 V2 F F V2 V2 A2 A2
V3 V3 V3 V3 V3 V3 F F V3 ...

Figure 1. Data preprocessing: split and refine action sequences into subsequences corresponding to each assessment period.

4. Hypothesis-Driven Approach

4.1 Method

In many of the MOOC courses, the learning material, including video lectures and assignments, is made available on a weekly basis. Learners can freely navigate through these materials and may have different approaches to following the course. For instance, some learners might prefer to watch the videos to learn about the new topics, read previous messages in the discussion forum, ask questions about the difficult concepts, and then try to solve the assignments. Other learners, on the other hand, might prefer go right to the assignments and follow a trial-and-error approach until they succeed. Others might decide to skip the videos because they are already familiar with the topic, or they may prefer to get an idea of what the assignment is about first and then selectively watch the relevant sections of the videos. Learners might also skip doing the assignment or even watching the videos in some periods. To identify these different approaches from learners' interaction traces, we examine their activity subsequences for each assessment period according to the following criteria:

- Does the learner start their learning sequence by watching a video?
- Does the learner submit the assignment before the deadline?
- Does the learner make multiple attempts to solve the assignment?
- For multiple submissions, does the learner review video lectures (instructor's help) or access the discussion forum (social help) after their first attempt?

Based on these criteria and following the structure represented by the decision tree in Figure 2, we label learners' activity subsequence with one of the following study patterns:

- **V_start_one_attempt**: watching the video(s) before submitting the assignment, with only one attempt to solve the assignment.
- **V_start_resubmit_no_help**: watching the video(s) before submitting the assignment, submitting multiple times without accessing videos or forums between the submissions.
- **V_start_resubmit_with_help**: watching the video(s) before submitting the assignment, submitting multiple times with access to the videos or forums after the first attempt.
- **A_start_one_attempt**: starting with the assignment and making only one attempt to solve it.
- **A_start_resubmit_no_help**: starting with the assignment and making multiple submissions without accessing the videos or forums after the first attempt.
- **A_start_resubmit_with_help**: starting with the assignment and making multiple submissions with access to the videos or forums after the first attempt.
- **Audit**: watching the videos but skipping the assignment.
- **Inactive**: skipping both videos and assignments.

Based on the labels that a learner is assigned for each assessment period, we construct a study pattern sequence to describe the learner's engagement over the duration of the course. Once we have the study pattern sequences for all learners in the course, we apply hierarchical agglomerative clustering on the sequences to extract categories of learners with similar study profiles and identify prototypical study pattern sequences over the duration of the course. To determine the optimum number of clusters, we use the *Calinski-Harabasz (CH)* index (Caliński & Harabasz, 1974), which is a well-known method for cluster count estimation. To assess the pairwise distance between study pattern sequences, we use *optimal matching (OM)*, a distance measure for sequence alignment. Following this approach, we determine the degree of dissimilarity between two sequences as the smallest number of edit operations (substitutions) required to turn one sequence into the other.

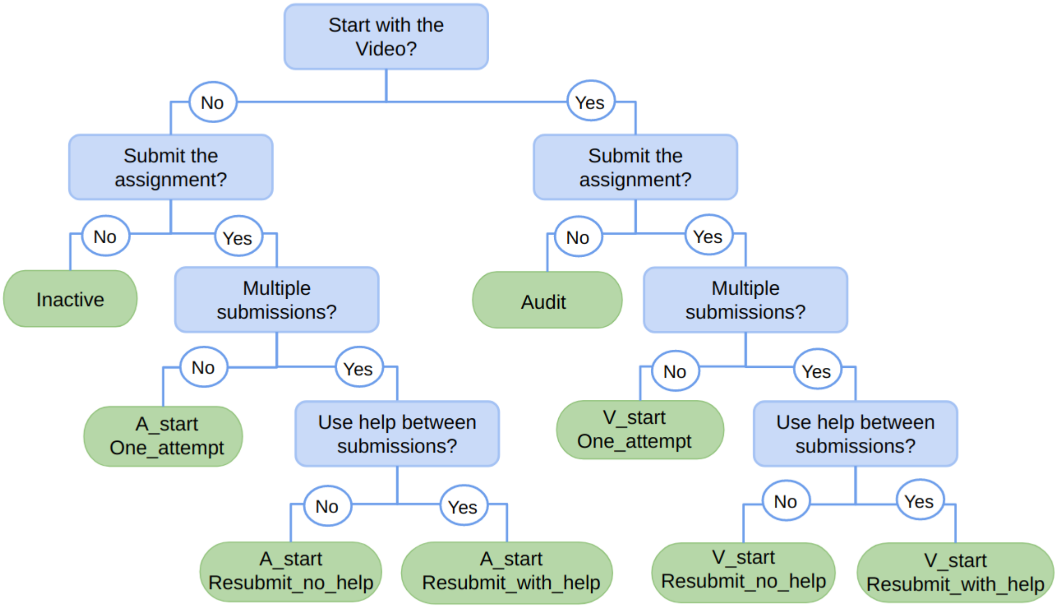


Figure 2. Decision tree for determining study patterns in the hypothesis-driven method. Activity subsequences are examined according to the criteria shown in the tree nodes (blue), and the tree leaves (green) show the study pattern labels.

4.2 Results

4.2.1 Study patterns distribution

Following the described approach, the study patterns of 7527 learners during each of the six assessment periods (total of 45,162 study sessions) were identified. The overall frequency of the resulting patterns and the average grade obtained by learners in each category are represented in Figure 3. As seen in this figure, in the most common case (68% of all study sessions), learners watch videos before submitting an assignment (*V_start*). In nearly half of such sessions, learners make only one attempt to solve the assignment and obtain a high grade (average 9.6 out of 10). Similarly, single submission is the most common approach in *A_start* sessions and is associated with high average grade (9.6). For multiple submissions (35% of all sessions), learners mostly resubmit without any further access to the videos or the discussion forum. In this case, the average grade obtained in the first attempt is lower than with the single-submission patterns.

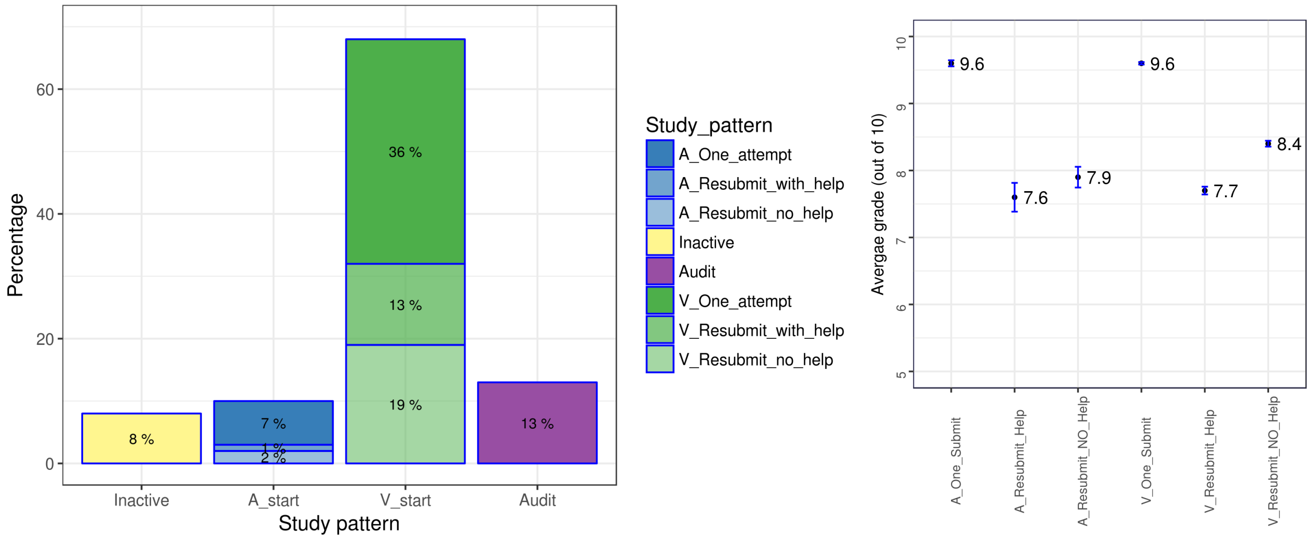
A more detailed comparison of different study patterns shows that learners who start by watching videos start their learning sequence earlier than those who start with the assignment. This is reflected by a significantly longer time between the activity sequence start time and the assignment deadline in the *V_start* and *Audit* approaches compared to the *A_start* approach (8 vs. 4.2 days,$F[1,41575] = 3031, p < .001$). Considering the assignment resubmission behaviours in *V_start* and *A_start* sessions, the χ^2 test shows a significant relation between the number of submissions and the study approach ($\chi^2 = 254, df = 1, p < 0.001$). According to the test residuals, learners in *A_start* sessions are less likely than those in the *V_start* group to make multiple attempts to solve an assignment. However, learners using both approaches perform equally well and get an average score of 8.8 out of 10 ($sd = 2$) in their first attempt. Moreover, as shown in Figure 3, only in a very small proportion of sessions (1%) do learners with the *A_start* approach access the lectures or discussion forum after their first submission of the assignment. These results suggest that learners using the *A_start* approach are likely to have some prior knowledge about the assignment topic, since they achieve a high grade in their first (and often only) attempt, without viewing the course lectures.

4.2.2 Study patterns over time

In this section, we study the evolution of the overall study patterns (*A_start*, *V_start*, *Audit*) over time to understand whether and how the learners change their approach during the course. As shown by Figure 4, in all assessment periods, around 10% of the learners skip the video lectures and directly submit the assignments (*A_start*). The proportion of learners who watch the videos but do not submit the assignments (*Audit*) gradually increases toward the end of the course (8% vs. 18% in the first and last assignments, respectively). The proportion of *Inactive* students also considerably increases in the last two assessment periods (4% vs. 20%, respectively, in the first and last assignments).

4.2.3 Fixed study pattern

Analysis of individuals' study pattern sequences shows that 47% of the learners change their study approach at least once during the course, whereas the other 53% of learners continue with their initial approach. These learners can be clustered into



(a) *Overall distribution*, **(b)** *Average grade on the first attempt*,

Figure 3. Hypothesis-driven study patterns: overall distribution and average grade obtained by learners in each group

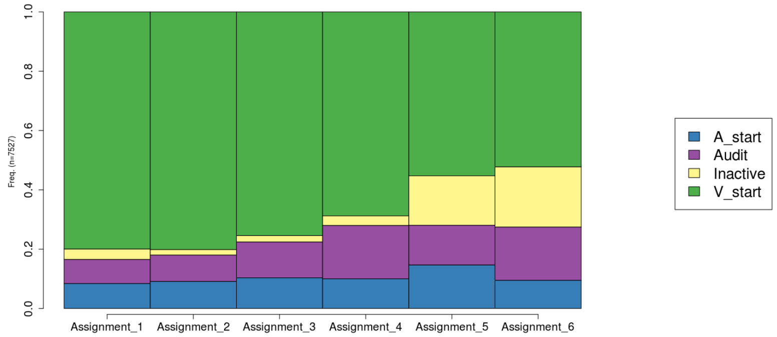


Figure 4. Hypothesis-driven study patterns: distribution over different assessment periods

three categories, represented as Clusters 1 to 3 in Table 1. Learners who follow the *V_start* approach in all the assessment tasks form the largest cluster, **Cluster 1**, making up 44% of participants. This group represents typical MOOC learners who rely on lectures to attain the knowledge required to solve the assignments. On the other hand, 2% of participants, represented by **Cluster 2**, do not spend time watching the videos before submitting any of the assignments. Their high performance level (average grade of 90 out of 100) reflects their proficiency in the course topics. Earning the completion certificate could therefore be one of the main participation motivations for these learners. On the contrary, 7% of learners, **Cluster 3**, do not submit any of the assignments, but they follow most of the video lectures during the course. This group of learners watches the videos as a source of knowledge without intending to receive a certificate. This group can be referred to as *auditing* students, similar to Kizilcec et al. (2013).

4.2.4 Changing study pattern

Unlike the described groups with fixed approaches, 47% of the learners change their study approach at least once during the course. Transition probabilities among study pattern types for this group of learners are represented in Figure 5. Several interesting observations can be made from this diagram. In general, the high probabilities associated with the self-loops suggest that in each assessment period, learners are likely to continue with their previous study approach, especially for the *V_start* and *Inactive* states, which have a stay probability of 0.69. Learners who start by watching videos have a low probability of skipping videos in the next period (transition probability from *V_start* to *A_start* = 0.14), whereas learners with the *A_start* approach show a relatively high probability (0.39) of watching the videos before submitting the next assignment. Students who *audit* the course in one assessment period will most likely continue auditing (probability = 0.44) or go *inactive* (probability= 0.41) in the next period. Once they enter the *inactive* state, participants are not very likely to get engaged in solving the next assignments, but they might continue watching the videos in the next period (transition probability from *inactive* to *audit* = 0.18).

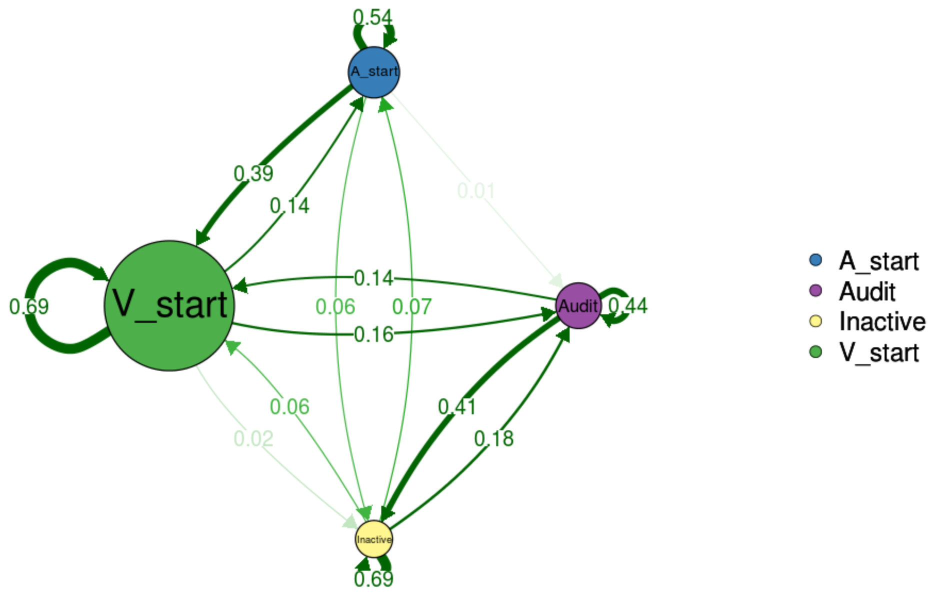


Figure 5. Study pattern transition probabilities for learners who change their approach over time. Node size is proportional to pattern frequency.

4.2.5 Clusters of study pattern sequences

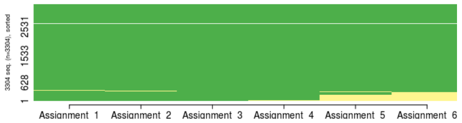
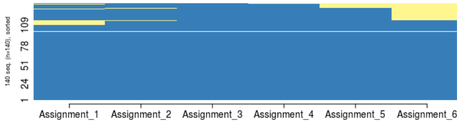
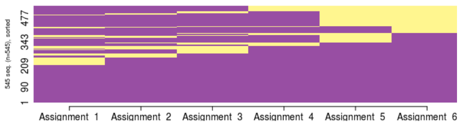
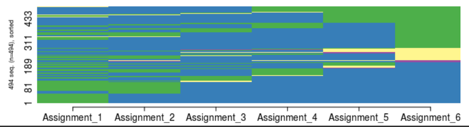
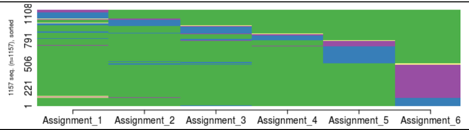
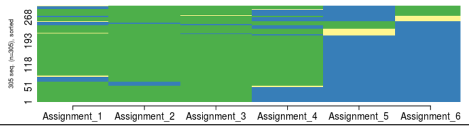
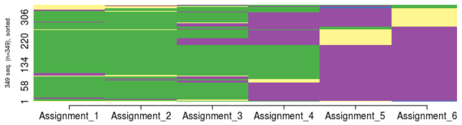
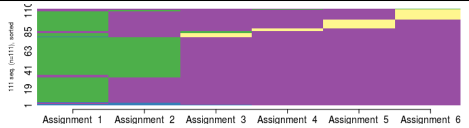
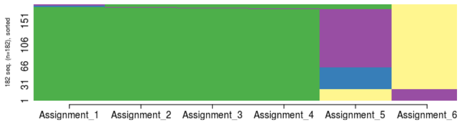
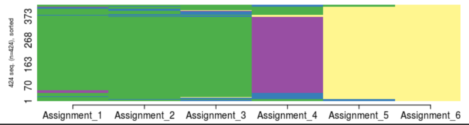
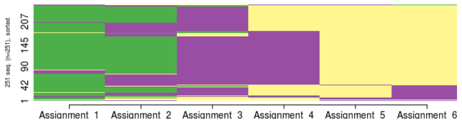
To identify the common study pattern trajectories for learners with a change of approach during the course, we apply clustering to study pattern sequences (hierarchical clustering with OM as the distance measure) and obtain eight clusters, represented as Clusters 4 to 11 in Table 1. We describe the resulting clusters and their attributes, including cluster size, average final grade, and ratio of passed students, and we provide the visualizations of study pattern sequences in each cluster, in the same table.

Learners in **Clusters 4, 5, and 6**, despite having different study pattern profiles, complete the course by submitting almost all of the assignments, and more than 94% of them pass the course with high average grades (above 84). Learners in **Cluster 4** mainly follow the *A_start* approach, but in few assignments, mostly the first or last, do they watch the videos before submitting. **Cluster 5**, on the contrary, comprises learners whose main approach is *V_start*, but they skip the videos in one or two assignments during the course. The start time of the learning sequence for these learners is closer to the assignment deadline in the *A_start* sessions, in comparison with the previous assessment period (4.2 vs. 8.4 days left for the deadline, respectively, in the *A_start* session and the preceding *V_start* session, $t[1,977] = -14.5, p < 0.001$). One possible explanation is that during such periods the learners procrastinate, so the impending deadline makes them temporarily change their study approach and submit the assignments without watching the videos. Learners in **Cluster 6** also mainly prefer to watch the videos first. But in the last two periods, they submit the assignments without watching the videos. These learners achieve nearly a complete grade in the first four assignments (average grade 9.7 out of 10, $sd = 0.8$). Considering that the final grade is calculated based on the assignment grades, such learners are likely to have a high final grade even without receiving a complete score in the remaining assignments. This might be one factor that influences their decision to skip videos in the last periods and directly submit the assignments. However, more information about the learners' experience and conditions is required to precisely determine the factors triggering changes in their study approaches.

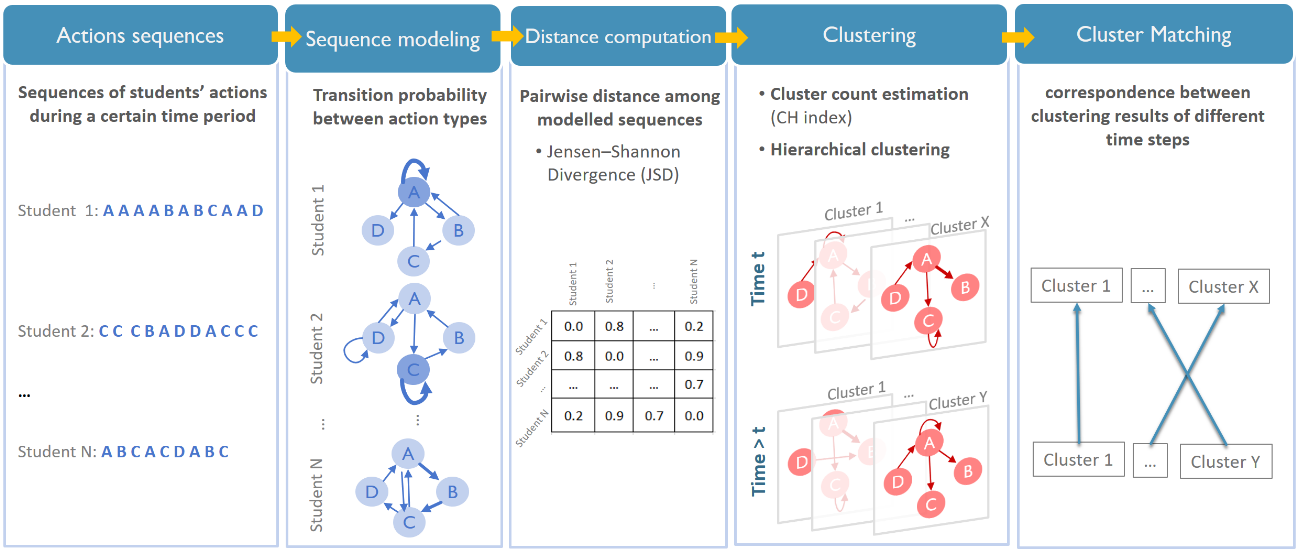
The last five clusters (Clusters 7 to 11) show learners who start the course with an active approach as they get engaged both in watching the videos and in submitting the assignments (*V_start*), but their engagement level decreases over the duration of the course. Learners in **Clusters 7 and 8** remain engaged until the end of the course. However, over time they lose motivation for doing the assignments and continue watching the course lectures without making any submissions. Learners in Cluster 7 submit nearly half of the assignments, whereas those in Cluster 8 only submit the first one or two, before switching to the auditing state. **Clusters 9, 10, and 11** demonstrate profiles of disengaging learners or dropouts. The dominant pattern in the learners' study profiles in these clusters is to start with the *V_start* approach, change to the *Audit* state (stop submitting the assignments), and finally stop watching the videos and drop out. The three clusters differ in the point at which learners' engagement level decreases. Participants in Cluster 9 submit the first four assignments, and 66% of them acquire enough points to pass the course, whereas those in Clusters 10 and 11 stop doing assignments after one to three weeks and eventually drop out about a week later.

The identified engagement profiles could inform the design of intervention mechanisms to support learners who might be facing problems in completing the assignments (e.g., learners in Clusters 7 to 11). An example could be suggesting

Table 1. Clusters of study pattern sequences extracted using hierarchical clustering. Clusters 1 to 3 represent learners with a fixed study approach during the course. Clusters 4 to 11 represent categories of role sequences for learners who change their approach over time. The vertical axis in the pattern sequence charts represents students in each cluster, and the horizontal axis represents assignments. Note that the chart height is **not** proportional to the cluster size. Other columns represent cluster size, average final grade of cluster members, ratio of passed students in each cluster, and description of the study pattern profiles.

		Study pattern sequences	**Size (%)**	**Final grade**	**Pass %**	**Description**
		🟦 A_start 🟨 Inactive 🟪 Audit 🟩 V_start				
Fixed approach	**Cluster 1**		3304 (44%)	92 (sd:13)	96%	**Submit all, *V_start***: watch videos before submitting all the assignments
	Cluster 2		140 (2%)	90 (sd:15)	94%	**Submit all, *A_start***: submit all the assignments without having watched the videos
	Cluster 3		545 (7%)	0 (sd:0)	0%	**Auditing**: watch most videos without submitting any assignments
Changing approach	**Cluster 4**		494 (7%)	88 (sd:15)	94%	**Submit all, mainly *A_start***: submit all the assignments, occasionally watch videos before submission
	Cluster 5		1157 (15%)	84 (sd:14)	95%	**Submit all, mainly *V_start***: submit all the assignments, skip videos before only one or two assignments
	Cluster 6		305 (4%)	89 (sd:12)	99%	**Submit all, *V_start* then *A_start***: watch videos before submission at the beginning, skip them for the final assignments
	Cluster 7		349 (5%)	44 (sd:15)	20%	**Complete,*V_start* then *Audit***: stop submitting the assignments after the first half of the course, continue watching videos
	Cluster 8		111 (1%)	19 (sd:10)	0%	**Complete,*V_start* then *Audit***: only submit the first one or two assignments, continue watching videos without submitting
	Cluster 9		182 (2%)	63 (sd:11)	66%	**Disengage at the end** : start with the *V_start* approach, in the last two weeks switch to audit and then drop out
	Cluster 10		424 (6%)	47 (sd:11)	12%	**Disengage in the middle**: start with the *V_start* approach, switch to audit in the second half of the course and eventually drop out
	Cluster 11		251 (3%)	23 (sd:10)	0%	**Disengage at the beginning**: start with the *V_start* approach, switch to audit after only one or two weeks and then drop out

supplementary learning materials to such learners or connecting them to the well-performing learners through the discussion forums.


Actions sequences
Sequence modeling
Distance computation
Clustering
Cluster Matching
Sequences of students' actions during a certain time period
Student 1: A A A A B A B C A A D
Student 2: C C C B A D D A C C C
...
Student N: A B C A C D A B C
Transition probability between action types
Student 1
Student 2
Student N
A
B
C
D
Pairwise distance among modelled sequences
• Jensen–Shannon Divergence (JSD)
• Cluster count estimation (CH index)
• Hierarchical clustering
Time t
Time > t
Cluster 1
...
Cluster X
Cluster Y
correspondence between clustering results of different time steps
Cluster 1
...
Cluster X
Cluster 1
...
Cluster Y

	Student 1	Student 2	...	Student N
Student 1	0.0	0.8	...	0.2
Student 2	0.8	0.0	...	0.9
...	...	...	...	0.7
Student N	0.2	0.9	0.7	0.0

Figure 6. Data-driven method pipeline

5. Data-Driven Approach

5.1 Method

In this section, we introduce an unsupervised processing pipeline to discover and track latent study patterns from students' interaction sequences. The proposed pipeline, as illustrated in Figure 6, consists of four steps: **(1)** activity sequence modelling, **(2)** distance computation, **(3)** clustering, and **(4)** cluster matching. The method receives as input the action sequences extracted from learners' log data; transforms them into probability distributions, which model transitions between different action types; computes pairwise dissimilarities between the modelled sequences; estimates the optimal number of clusters; and performs clustering to identify groups of learners with similar study patterns in each time period. At each time step t, matching clusters with those at times $t-1$ and earlier are identified. This enables us to track learners' study patterns over time and capture changes in their study approaches, which is an advantage of our proposed method over a recent clustering method proposed in Klingler et al. (2016). Because the only input to our method is the sequence of learners' activities, it can be used to model and track learners' interaction patterns at different levels of action granularity or time resolution. Moreover, our clustering pipeline can automatically capture changes in the number and size of clusters and can be used to detect cluster evolution events such as cluster form, dissolve, split, and merge.

5.1.1 Activity sequence modelling

Let $A = \{a_1, a_2, ..., a_k\}$ be the set of possible actions in a platform and $S^t = (s_1, s_2, ..., s_n)$, $s_i \in A$ be the sequence of actions performed by a learner during time period t. We model the learner's action sequence as a matrix $F_{k \times k}$, where f_{ij} represents the frequency of observing action a_i right before a_j in S^t. We then transform F into a normalized vector P by normalizing frequencies to represent proportions. Since the entries in P sum to one, we can consider P as a probability distribution. P provides an aggregated view of the original sequence, encoding probabilities of transitions between different action types. Unlike Markov chain models (as used in Klingler et al. (2016) and Köck & Paramythis (2011)), our representation can directly reflect the frequent transitions in learners' action sequences. Hereafter, we refer to P as the learner's activity model.

5.1.2 Distance computation

To perform clustering on the modelled sequences, a dissimilarity measure needs to be defined to compare learners' activity models. Since the introduced models are in the form of probability vectors, we can use the JSD (Lin, 1991), a distance metric designed to compare probability distributions. The JSD for two probability distributions is bounded in $[0, 1]$, and the value of zero denotes identical distributions.

5.1.3 Clustering

Learners' activity models can then be clustered in order to identify individuals with similar study patterns. For this purpose, we apply hierarchical agglomerative clustering method on the extracted pairwise dissimilarity matrix between learners' activity models. In cluster analysis, determining the optimal number of clusters is a major challenge. Several methods have been proposed in the literature to automatically estimate the number of clusters based on the information intrinsic to the data. The *Calinski–Harabasz index* (Caliński & Harabasz, 1974) and the *silhouette coefficient* (Rousseeuw, 1987) are among the most well known methods for this purpose. Such methods in general measure compactness of clusters (similarity between points in the same cluster) and separateness between different clusters (how far apart points in different clusters are). In our pipeline, we use the *Calinski–Harabasz* index to estimate the number of clusters based on the distance matrix between activity models. Since learners might change their behaviour, some clusters might disappear or new clusters might emerge over time. Therefore, we separately compute the number of clusters in every time period.

5.1.4 Cluster matching

After extracting clusters of activity models in each time period, cluster matching is required to identify the correspondence between clusters in the most recent time step and those of previous steps. In social network analysis, cluster matching is often employed for group evolution discovery (Bródka et al., 2013) or tracking dynamic communities over time (Greene et al., 2010). In this context, the overlap between cluster members is a criterion considered to compute cluster similarity. However, in our processing pipeline, this step aims to identify corresponding study patterns in the clustering results of different time periods. Therefore, the similarity of activity pattern models should be taken into account for assessing cluster similarity. We apply a method, similar to the Ward method (Murtagh & Legendre, 2014), for computing the similarity of activity pattern clusters. The Ward method is used in hierarchical agglomerative clustering to select the clusters to be merged in each step. According to this method, the most similar clusters are the ones that minimize the increase in the sum of squared errors (Euclidean distance) after the merge. Inspired by this approach, we define the distance, d, between two clusters, C_i and C_j, as the amount of increment in the sum of errors (JSD distances in our case) when they are combined:

$$d(C_i, C_j) = SE_{C_{ij}} - (SE_{C_i} + SE_{C_j}), \tag{1}$$

where C_{ij} is the union of the two clusters, C_i and C_j, and SE_C is the sum of errors for cluster C defined as

$$SE_C = \sum_{x_i \in C} JSD(x_i, m_C), \tag{2}$$

where m_C represents the centroid of cluster C, defined as the mean vector, and JSD refers to the Jensen–Shannon divergence.

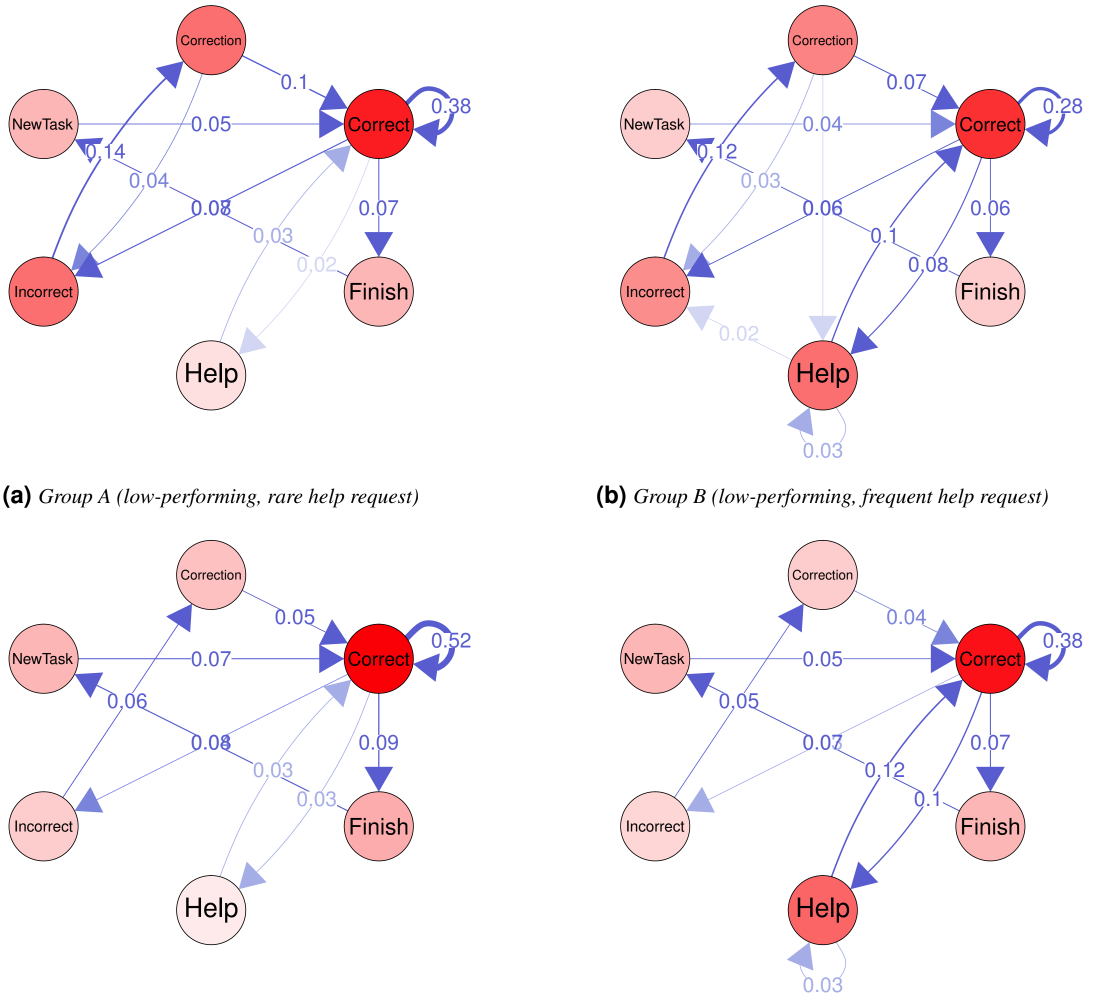
Based on the defined cluster distance measure, for each cluster C_i at each time step, we identify the closest one to it, C_j, from the set of clusters obtained in previous time steps. For multiple candidates for the closest cluster, we choose the most recent one. If the distance between C_i and C_j is smaller than a threshold (95% quantile of the set of distances between candidate matching clusters), we consider the two clusters to be matching and assign the same labels to them. Otherwise, we consider C_i as a new cluster and associate a new label with it.

5.2 Results

Given the unsupervised nature of our data-driven approach, it is important to validate the method before applying it to our MOOC dataset. A synthetic dataset could be used for this purpose. To assess all of the different components of our proposed pipeline, we require a dataset that simulates different learning behaviours, including changes in individuals' behaviour over time and also in the number of existing categories (clusters). The synthetic dataset presented in Klingler et al. (2016) matches all of these requirements and provides a perfect way to evaluate our data-driven pipeline. This dataset simulates students' action sequences in an ITS and consists of different scenarios that involve changes in learners' activity patterns. It therefore enables us to evaluate the ability of our proposed pipeline in modelling learners' activity sequences and tracking their behaviour over time. In this section, we first demonstrate the application of our proposed pipeline on the mentioned dataset. This experiment could also provide evidences of the applicability of our method in different educational contexts and platforms other than MOOCs. Then we employ the proposed method to capture and analyze learners' study patterns during the assessment periods in our MOOC dataset (described in Section 3).

5.2.1 Simulated study

The synthetic data in Klingler et al. (2016) simulate action sequences of 80 learners over 50 training sessions in a tutoring system. In each session, students needed to complete 20 tasks. Each task was composed of eight steps, and students had to correctly solve all the steps in order to finish the task. Learners' abilities θ were sampled from a normal distribution with mean $\mu = 0$ and variance $\sigma = 1$, and task difficulties d were sampled uniformly from the $[-3, 3]$ range. The probability of



(a) *Group A (low-performing, rare help request)* **(b)** *Group B (low-performing, frequent help request)*

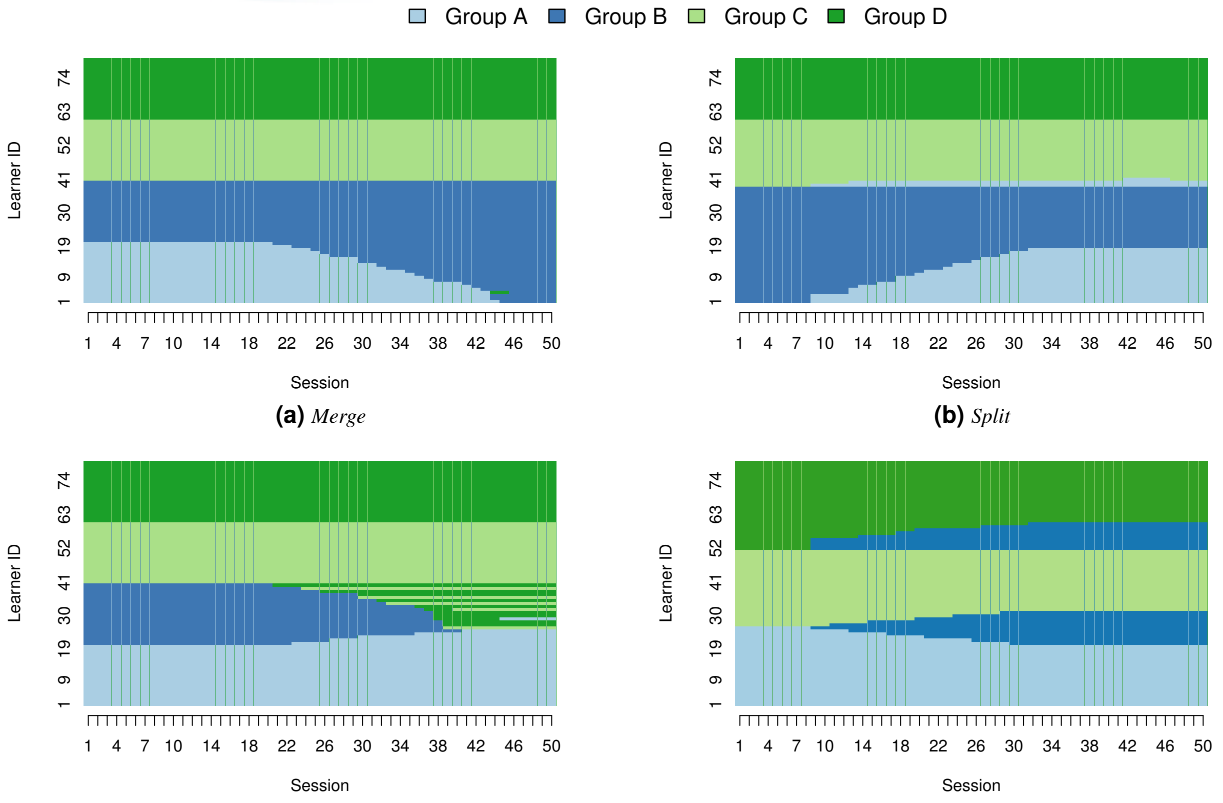
(c) *Group C (high-performing, rare help request)* **(d)** *Group D (high-performing, frequent help request)*

Figure 7. The average activity models for the four resulting clusters in the first simulated scenario (cluster merge). In the transition diagrams, node colour intensity is proportional to the action probability and edge thickness is proportional to the transition probability. The resulting clusters correctly capture the four simulated behaviours.

correctly solving a task for each student was given as $p(y) = (1+e^{-(\theta-d)})^{-1}$. Students could ask for help at any point during the training session, with a probability of p_H. Six types of actions were considered for the learners: $S = \{$*new task*, *correct*, *incorrect*, *correction*, *help*, and *finish*$\}$. In the simulated data, high-performing students were modelled by setting $\theta = 1$ and low-performing learners were simulated by setting $\theta = -1$. Moreover, normal help-seeking behaviour was modelled by a small probability-of-help request $p_H = 0.05$, whereas frequent help-seeking behaviour (help abuse) was simulated by a large probability of asking for help $p_H = 0.2$. Following this approach, four groups of learners with different behaviours were simulated:

- **Group A**: low-performing learners with rare help requests
- **Group B**: low-performing learners with frequent help requests
- **Group C**: high-performing learners with rare help requests
- **Group D**: high-performing learners with frequent help requests

In the synthetic dataset in Klingler et al. (2016), four artificial scenarios are considered, simulating different cluster evolution events, including cluster merge, split, dissolve, and form. The **first** scenario simulates merging clusters. In this scenario, after about 20 sessions, low-performing learners with rare help requests (group A) start abusing the help, and eventually group A completely merges into group B. The simulation in the **second** scenario starts with three groups, B, C, and D. Over time, some



(a) *Merge*

(b) *Split*

(c) *Dissolve*

(d) *Form*

Figure 8. Sequences of the learners' interaction patterns over 50 sessions in the 4 simulated scenarios. In the sequence charts, each horizontal line represent the interaction pattern sequence of one learner. Our proposed pipeline correctly captures the changes in cluster count and size and detects cluster merge, split, dissolve, and form.

of the low-performing students with frequent help calls (group B) stop abusing the help and consequently group B splits into groups A and B. In the **third** scenario, which simulates a dissolving cluster, learners in group B switch to the other approaches and eventually group B completely dissolves into the other three groups. Finally, the cluster-forming event is simulated in the **fourth** scenario. In this case, the simulation starts with three groups, A, C, and D. Over time a fourth group, B, is formed, which gradually absorbs students from the other groups until all four groups have equal size.

For all the described scenarios, we used our processing pipeline to model and cluster learners' action sequences in each session and identified corresponding clusters in different sessions using the cluster matching step. Using this approach, four clusters of interaction patterns were identified in all of the scenarios. Figure 7 represents the average activity models for the resulting clusters in the first scenario (similar results were obtained for the other scenarios). According to the transition probabilities between different action types, the resulting clusters clearly correspond to the four simulated learner groups. The two clusters in Figures 7a and 7b depict low-performing learners as reflected by the relatively high transition probabilities between *incorrect* and *correct* actions. Such learners therefore make more mistakes than the high-performing learners in Figures 7c and 7d. Regarding the help-seeking patterns, the help-abusing behaviour is reflected by the frequent transitions between *help* and *correct* actions in Figures 7b and 7d, whereas such transitions are quite rare for learners with normal help-seeking behaviour in Figures 7a and 7c.

Based on the clusters that students were assigned to in each session, we build their interaction pattern sequences in the four simulated scenarios. According to the resulting sequences, illustrated in Figure 8, our method successfully captures the described cluster evolution events in all of the scenarios. Furthermore, comparison of our clustering results with the truth on the ground confirms the high accuracy of our method in labelling learners' study sessions. Our proposed method achieves 95% accuracy in the first and third scenarios (clusters merging and forming) and 93% accuracy in the second and forth scenarios (clusters splitting and dissolving).

Overall, the presented experiments with simulated data demonstrate that our processing pipeline is able to detect different interaction patterns among learners and provides models that are easy to interpret. The validity of the clustering and cluster-matching steps is also confirmed by these results, showing that our method correctly captures changes in the number and size of clusters and is able to detect changes in learners' behaviours over time.

5.2.2 MOOC study patterns

As mentioned, our processing pipeline can be used to model learners' activity patterns at different levels of granularity. To employ the described data-driven approach to model study patterns in MOOC, we choose daily granularity of actions. We label each day according to the type of activities (regardless of their order) performed by the learner (***V***ideo access, ***F***orum access, and ***A***ssignment submission), with one of the following states: $\{A, F, V, AF, AV, FV, AFV, inactive\}$. We then describe individuals' daily state sequences as the list of daily states between their first and last activity day during each assessment period. As an example, if a learner starts the learning sequence by watching the videos on two successive days, does not perform any action on the next three days, and accesses the forum (read/write) and submits the assignment the day after, their daily state sequence would be $\{V, V, Inactive, Inactive, Inactive, AF\}$.

Following this data preparation procedure, we construct the daily state sequences for all six assessment periods for a randomly selected sample of 2000 learners. Sampling was done due to high memory requirement for pairwise distance computation on the full dataset. The set of state sequences is then provided as input to the described clustering pipeline. For each assessment period, learners' activity models, which in this case represent transition probabilities between different daily states, are constructed and clustered according to the estimated number of clusters. Cluster labels are then refined based on the cluster-matching results. The centre of each cluster (average vector) is considered as the representative study pattern for the learners in each cluster.

Following this approach, we identified 13 different study patterns (clusters) from learners' interaction logs. Table 3 summarizes the results, including visualization of the study patterns and the most frequent daily state sequences in each cluster. Table 3 also describes the study patterns and their attributes (size and average cluster errors). The low average cluster errors (0 to 0.1) reflect the accuracy of the clustering results.

The resulting clusters capture meaningful patterns in learners' study sequences. According to the state transition diagrams and their descriptions in Table 3, the extracted study patterns differ in the duration of study sequences and also in the daily activity types. In most cases, learners work on the materials for a week during one or multiple consecutive days. For instance, learners with *patterns 8*, *10*, and *12* have a single activity day. In *pattern 10*, learners directly submit the assignments without accessing any other course materials, whereas in *pattern 8*, they watch the videos and submit the assignment, and in *pattern 12*, they also access the discussion forum. Learners in *patterns 4*, *5*, *6*, and *9* study during two or more successive days, whereas, in *patterns 3*, *7*, and *13*, learners have multiple inactive days during their learning sequences.

Table 2 provides an overview of the estimated number of clusters and the list of detected study patterns at each assessment period. As reflected by the cluster counts, a higher variability is observed in learners' study approaches at the beginning of the course. Most of the detected patterns remain present in learners' interaction sequences over the duration of the course (*patterns 1*, *2*, *4*, *7*, *8*, *10*, and *11*), whereas some other patterns, such as *patterns 5* and *9*, disappear over time. During the second and fourth assessment periods, two new study patterns are formed (*patterns 12* and *13*), both of which dissolve into other patterns after only one or two periods.

Table 2. Estimated number of clusters and list of detected study patterns (see Table 3) in each assessment period. New patterns in each period are highlighted.

Assignment	# clusters	Cluster list (study patterns)
1	11	Pattern **1**, **2**, **3**, **4**, **5**, **6**, **7**, **8**, **9**, **10**, **11**
2	12	Pattern 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, **12**
3	10	Pattern 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
4	10	Pattern 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, **13**
5	8	Pattern 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
6	8	Pattern 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13

Figure 9 depicts the transition probabilities between different study patterns, extracted from learners' study profiles over the duration of the course. According to the self-loop probabilities, *patterns 10* and *7* are the most stable study patterns in the sense that learners following these approaches are likely to continue with the same approach in the next assessment period. *Pattern 11*, which represents inactive learners during a period, is also associated with high stay probability, suggesting that inactive learners in one period would remain inactive in the next period, with a probability of 0.6. *Pattern 10* represents learners with only one activity day on which they submit the assignment and do not access any videos or the discussion forum. This pattern could represent a similar approach as A_start, described in Section 4. *Pattern 7*, which is the most frequent study pattern, represents students who watch the lectures, and after few inactive days they either continue watching the videos or submit the assignment. This pattern receives relatively strong connections from the other nodes (except *pattern 10*), suggesting that learners with other approaches might adopt this study pattern for the next assessment period, with probabilities between 0.2 and 0.4.

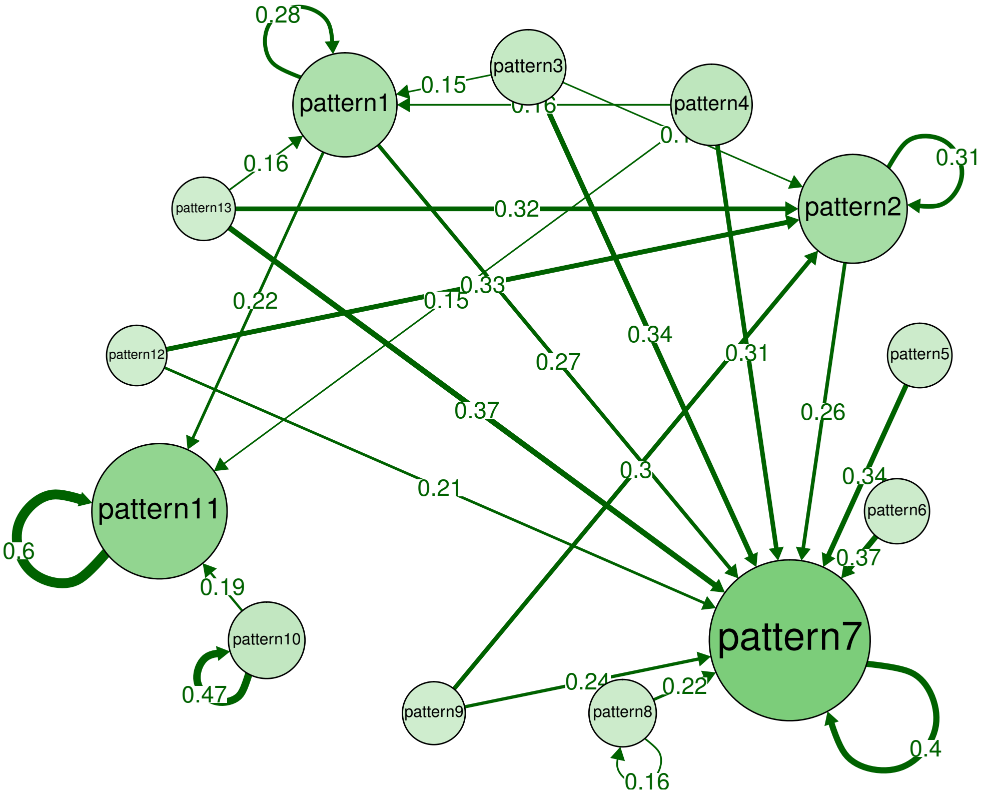


Figure 9. Transition probabilities between data-driven study patterns. Node size is proportional to pattern frequency, and edge thickness is proportional to the transition probability.

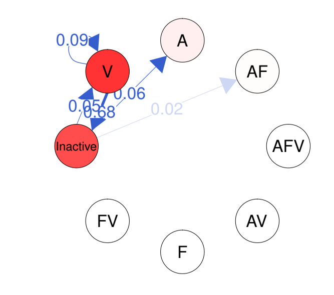
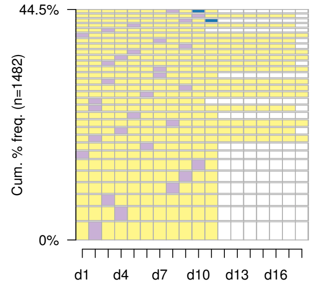
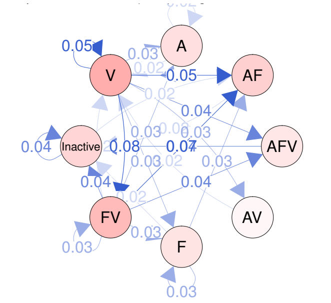
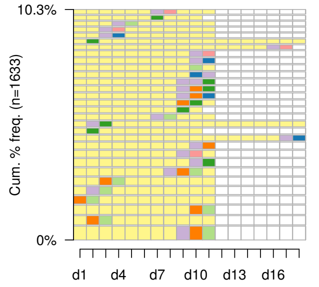
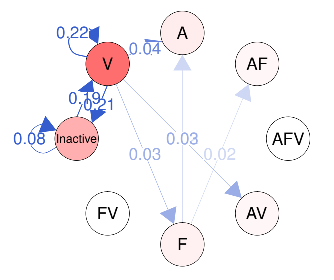
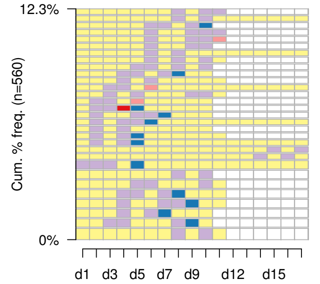
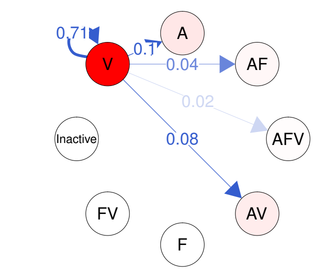
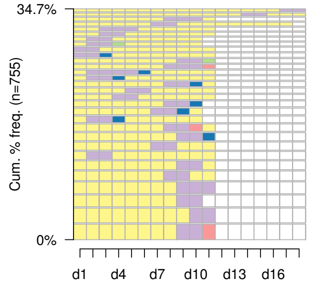
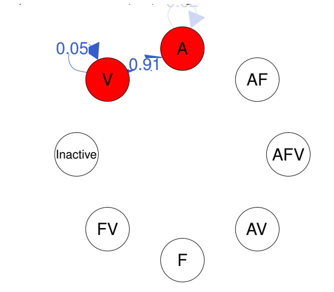
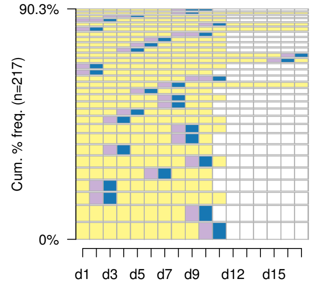
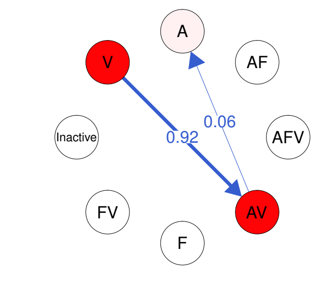
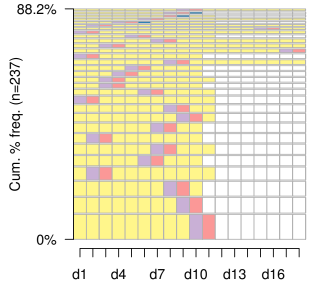
6. Discussion

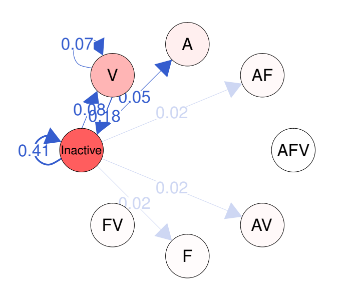
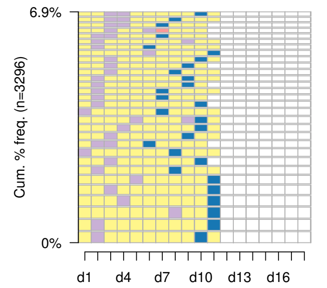
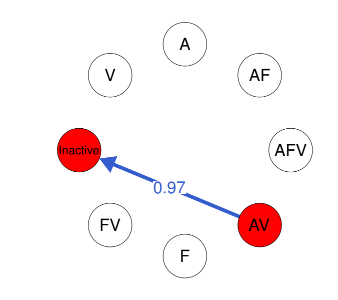
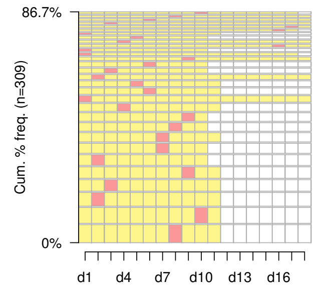
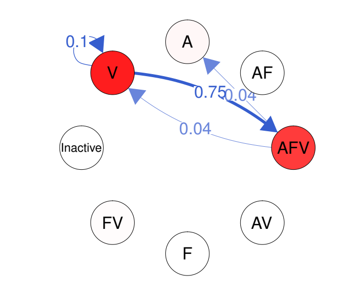
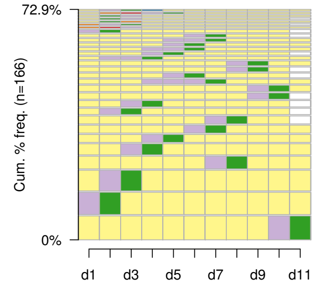
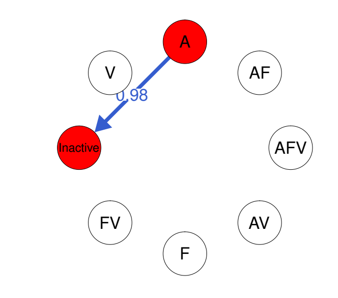
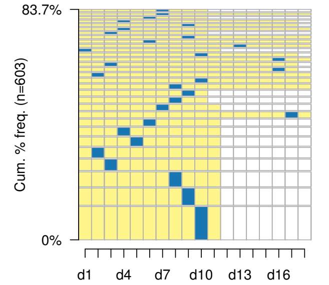
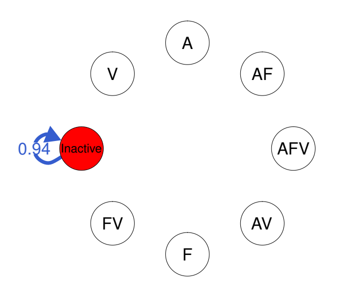
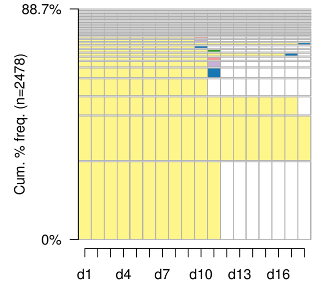
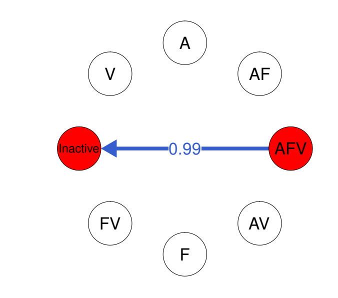
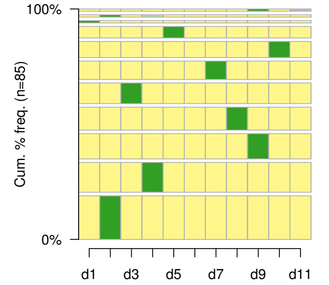
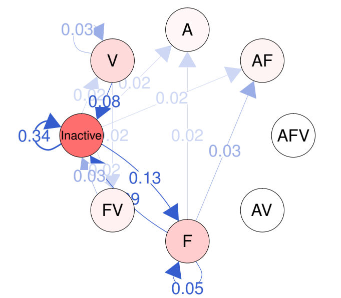
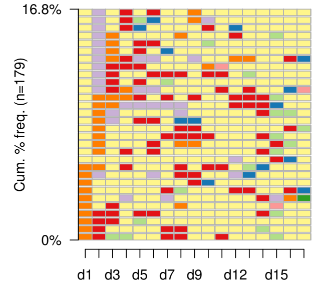
In this work, we investigated learners' study approaches during MOOC assessment periods. To answer our research question, we proposed two different methods for detecting and tracking latent study patterns over time. First, we introduced a hypothesis-driven approach to label learners' activity sequences based on interactions with lectures and assignments. We found that about 44% of the learners in our MOOC dataset watch lecture videos prior to submitting each assignment. On the other hand, about 2% of learners skip the videos in all assessment periods. Through unsupervised categorization of study pattern sequences, we identified different longitudinal engagement profiles among learners. We showed that some learners temporally change their study approach during a few periods and then continue with their initial approach, whereas others permanently switch to a new approach. Changing the study approach could be an indicator of facing difficulties in the learning process. Detecting such events could therefore enable instructors to provide personalized support to the learners.

In the second method, we proposed a processing pipeline for unsupervised discovery and temporal analysis of interaction patterns from sequential activity logs. The proposed method is general and requires only the collection of action sequences as input. It can therefore be used to model and analyze interaction patterns in various online learning environments, including MOOCs and ITSs. The cluster-matching step in our pipeline makes it possible to run the algorithm at different steps during the course period and identify changes in learning behaviours. This feature could in turn enable the development of an alert system for providing real-time feedback and support for the instructors and learners. Moreover, the presented pipeline allows for the analysis of interaction patterns at different levels of granularity and time resolution. Through experiments with simulated data, we showed that our pipeline enables us to detect learners' behavioural patterns, provides interpretable models describing them, and captures temporal dynamics of learning behaviours. We further applied our pipeline to a MOOC dataset to explore learners' study patterns in this context. Using this approach, we identified 13 different study patterns and investigated transitions among them.

As mentioned before, a very important feature of the introduced methods in this work is the real-time support and feedback that they could enable once deployed during the course. Unlike most existing research with post hoc analysis of learners' activity traces, our proposed methods could be applied to analyze engagement patterns while the course is running, which could enable instructors to identify and provide timely support to learners with difficulties in following the course. A possible extension to our work could be to integrate an overview of the captured behaviours and study pattern sequences in analytics dashboards. This information could help the teaching team improve the course design, for instance by identifying the factors that trigger course-wide drifts in learners' engagement patterns. Moreover, the extracted learning behaviours could be used to improve personalization of online learning platforms and ITSs. In future work, it would also be interesting to study the relation between learning strategies and performance and also the influence of participants' backgrounds, educational contexts, and demographics on their study approaches.

Table 3. Data-driven study patterns extracted from learners' interaction logs. For each pattern, transition diagrams (left) show the average activity model, and grid charts (right) show the 20 most frequent daily state sequences for each study pattern. The horizontal axes in the sequence charts represent days in the assessment period, and rows represent sample sequences (row height is proportional to sequence frequency). In the pattern description, *N* represents the frequency of each pattern and *AE* is the average error (average distance between activity models and the cluster mean vector).

Study pattern	Sample state sequences	Study pattern	Sample state sequences
🟦 A 🟩 AF 🟩 AFV 🟥 AV 🟥 F 🟨 Inactive 🟧 FV 🟪 V			
Pattern 1, N=1482 (12%), AE=0.04 Watching videos on a single day, with a low probability (0.06) of submitting the assignment after one inactive day		**Pattern 2, N=1633 (14%), AE=0.13** No dominant pattern, similar probability of transition between different daily states	
 0.09 V A AF 0.06 0.95 0.68 0.02 Inactive AFV FV AV F		 0.05 0.03 A V 0.02 0.05 AF 0.02 0.04 0.04 Inactive 0.08 0.07 0.03 0.03 AFV 0.03 0.02 0.03 0.04 0.04 0.02 AV FV 0.03 0.03 F 0.03	
Pattern 3, N=560 (5%), AE=0.06 Multiple non-consecutive days of video watching, submitting the assignment after the last video-watching day		**Pattern 4, N=755 (6%), AE=0.04** Successive days of video watching, followed by a submission day on which learners are also likely to access the forum or videos	
 0.22 A 0.04 V AF 0.19 0.21 0.08 Inactive 0.03 0.03 0.02 AFV FV AV F		 0.71 A 0.1 V 0.04 AF 0.02 Inactive 0.08 AFV FV AV F	
Pattern 5, N=217 (2%), AE=0.01 Single day of video watching, followed by a pure submission day (without any access to the forum or videos)		**Pattern 6, N=237 (2%), AE=0.01** Video-watching day, followed by watching videos and submitting the assignment on the next day	
 0.05 A V 0.91 AF Inactive AFV FV AV F		 A V AF 0.06 Inactive 0.92 AFV FV AV F	

Study pattern	Sample state sequences	Study pattern	Sample state sequences
🟦 A 🟩 AF 🟩 AFV 🟥 AV 🟥 F 🟨 Inactive 🟧 FV 🟪 V			
Pattern 7, N=3296 (27%), AE=0.05 Video-watching day followed by several inactive days and then continue watching or submitting the assignment		**Pattern 8, N=309 (3%), AE=0.01** Watching videos and submitting the assignment on a single day	
 A V AF AFV AV F FV Inactive 0.07 0.05 0.08 0.18 0.02 0.41 0.02 0.02			
Pattern 9, N=166 (1%), AE=0.04 Watch videos on one day, continue watching, and access the forum and submit the assignment on the next day		**Pattern 10, N=603 (5%), AE=0** Single day of activity with submission to the assignment, without any access to the videos or forum	
 0.1 A V AF 0.75 0.04 0.04 Inactive AFV FV AV F			
Pattern 11, N=2478 (21%), AE=0.01 Inactive, no access to the materials of the corresponding assignment before the deadline		**Pattern 12, N=179 (1%), AE=0.05** Single day of activity, with watching the videos, accessing the forum, and submitting the assignment	
			
Pattern 13, N=1482 (12%), AE= 0.04 Forum oriented, watching videos followed by multiple non-consecutive forum access days, finish with submitting the assignment			
 0.03 A V AF 0.02 0.08 0.02 0.34 Inactive 0.02 0.02 AFV 0.03 0.03 0.13 FV AV F 0.05			

References

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the 11th International Conference on Data Engineering* (ICDE1995), 6–10 March, 1995, Taipei, Taiwan (pp. 3–14). Piscataway, NJ, USA: IEEE.

Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and Learning*, *9*(2), 161–185. https://doi.org/10.1007/s11409-013-9107-6

Bergner, Y., Shu, Z., & von Davier, A. (2014). Visualization and confirmatory clustering of sequence data from a simulation-based assessment task. In J. Stamper, Z. Pardos, M. Mavrikis, & B. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (EDM2014), 4–7 July 2014, London, UK (pp. 177–184). International Educational Data Mining Society.

Bródka, P., Saganowski, S., & Kazienko, P. (2013). GED: The method for group evolution discovery in social networks. *Social Network Analysis and Mining*, *3*(1), 1–14. https://doi.org/10.1007/s13278-012-0058-8

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics — Theory and Methods*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Desmarais, M., & Lemieux, F. (2013). Clustering and visualizing study state sequences. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (EDM2013), 6–9 July 2013, Memphis, TN, USA (pp. 224–227). International Educational Data Mining Society/Springer. http://www.educationaldatamining.org/EDM2013/papers/rn$_{p}aper_{3}3.pdf$.

Faucon, L., Kidzinski, L., & Dillenbourg, P. (2016). Semi-Markov model for simulating MOOC students. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th International Conference on Educational Data Mining* (EDM2016), 29 June–2 July 2016, Raleigh, NC, USA (pp. 358–363). International Educational Data Mining Society. https://pdfs.semanticscholar.org/4714/4a7b8d84915b807282149693e2a3dfda6bdc.pdf.

Geigle, C., & Zhai, C. (2017). Modeling MOOC student behavior with two-layer hidden Markov models. In *Proceedings of the 4th ACM Conference on Learning @ Scale* (L@S 2017), 20–21 April 2017, Cambridge, MA, USA (pp. 205–208). New York: ACM. https://jedm.educationaldatamining.org/index.php/JEDM/article/view/211.

Greene, D., Doyle, D., & Cunningham, P. (2010). Tracking the evolution of communities in dynamic social networks. In *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining* (ASONAM2010), 9–11 August 2010, Odense, Denmark (pp. 176–183). IEEE. https://doi.org/10.1109/ASONAM.2010.17

Günther, C., & van der Aalst, W. (2007). Fuzzy mining — Adaptive process simplification based on multi-perspective metrics. *Business Process Management*, *4714*, 328–343. https://doi.org/10.1007/978-3-540-75183-0$_2$4

Hansen, C., Hansen, C., Hjuler, N., Alstrup, S., & Lioma, C. (2017). Sequence modelling for analysing student interaction with educational systems. *arXiv preprint arXiv:1708.04164*.

Jeong, H., & Biswas, G. (2008). Mining student behavior models in learning-by-teaching environments. In R. S. J. de Baker, T. Barnes, & J. E. Beck (Eds.), *Proceedings of the 1st International Conference on Educational Data Mining* (EDM'08), 20–21 June 2008, Montreal, QC, Canada (pp. 127–136). International Educational Data Mining Society.

Kinnebrew, J. S., Loretz, K. M., & Biswas, G. (2013). A contextualized, differential sequence mining method to derive students' learning behavior patterns. *JEDM — Journal of Educational Data Mining*, *5*(1), 190–219. https://pdfs.semanticscholar.org/70fc/fbd0b95bb61e4142275c869619997c5a187d.pdf.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 170–179). New York: ACM. https://doi.org/10.1145/2460296.2460330

Klingler, S., Käser, T., Solenthaler, B., & Gross, M. H. (2016). Temporally coherent clustering of student data. In *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 102–109). New York: ACM. https://pdfs.semanticscholar.org/de63/dfee95651eaad731b435b09064559f3223e7.pdf.

Köck, M., & Paramythis, A. (2011). Activity sequence modelling and dynamic clustering for personalized e-learning. *User Modeling and User-Adapted Interaction*, *21*(1), 51–97. https://doi.org/10.1007/s11257-010-9087-z

Li, C., & Biswas, G. (2000). A Bayesian approach to temporal data clustering using hidden Markov models. In *Proceedings of the 17th International Conference on Machine Learning* (ICML2000), 29 June–2 July 2000, Stanford, CA, USA (pp. 543–550). San Francisco, CA: Morgan Kaufmann Publishers. https://pdfs.semanticscholar.org/50b9/6db8ef7d550ce72630cd1a6d196c7b311815.pdf.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151. https://doi.org/10.1109/18.61115

Maldonado, R. M., Yacef, K., Kay, J., Kharrufa, A., & Al-Qaraghuli, A. (2011). Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, & J. Stamper (Eds.), *Proceedings of the 4th Annual Conference on Educational Data Mining* (EDM2011), 6–8 July 2011, Eindhoven, Netherlands (pp. 111–120). International Educational Data Mining Society.

Mukala, P., Buijs, J., & Van Der Aalst, W. (2015). *Exploring students' learning behaviour in MOOCs using process mining techniques.* https://research.tue.nl/en/publications/exploring-students-learning-behaviour-in-moocs-using-process-mini. Eindhoven University of Technology, BPM Center Report.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification*, *31*(3), 274–295.

Nesbit, J. C., Zhou, M., Xu, Y., & Winne, P. (2007). Advancing log analysis of student interactions with cognitive tools. In *Proceedings of the 12th Biennial Conference of the European Association for Research on Learning and Instruction* (EARLI), 28 August–1 September 2007, Budapest, Hungary. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.150.6671.

Patel, N., Sellman, C., & Lomas, D. (2017). Mining frequent learning pathways from a large educational dataset. *arXiv preprint arXiv:1705.11125*. https://arxiv.org/pdf/1705.11125.pdf.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Shen, S., & Chi, M. (2017). Clustering student sequential trajectories using dynamic time warping. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), *Proceedings of the 10th International Conference on Educational Data Mining* (EDM2017), 25–28 June 2017, Wuhan, China (pp. 266–271). International Educational Data Mining Society. http://educationaldatamining.org/EDM2017/proc$_{f}iles/papers/paper_{9}4.pdf$.

Shih, B., Koedinger, K. R., & Scheines, R. (2010). Unsupervised discovery of student strategies. In R. S. de Baker, A. Merceron, & P. I. P. Jr. (Eds.), *Proceedings of the 3rd International Conference on Educational Data Mining* (EDM2010), 11–13 June 2010, Pittsburgh, PA, USA (pp. 201–210). International Educational Data Mining Society. http://educationaldatamining.org/EDM2010/uploads/proc/edm2010$_{s}ubmission_{5}5.pdf$.

Shirvani Boroujeni, M., & Dillenbourg, P. (2018). Discovery and temporal analysis of latent study patterns in MOOC interaction sequences. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (LAK '18), 5–9 March 2018, Sydney, NSW, Australia (pp. 206–215). New York: ACM. https://doi.org/10.1145/3170358.3170388

Shirvani Boroujeni, M., Hecking, T., Hoppe, H. U., & Dillenbourg, P. (2017). Dynamics of MOOC discussion forums. In *Proceedings of the 7th International Conference on Learning Analytics and Knowledge* (LAK '17), 13–17 March 2017, Vancouver, BC, Canada (pp. 128–137). New York: ACM. https://doi.org/10.1145/3027385.3027391

Shirvani Boroujeni, M., Kidzinski, Ł., & Dillenbourg, P. (2016). How employment constrains participation in MOOCs. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th International Conference on Educational Data Mining* (EDM2016), 29 June–2 July 2016, Raleigh, NC, USA (pp. 376–377). International Educational Data Mining Society. https://infoscience.epfl.ch/record/218785/files/EDM16$_{employment_{pOSTER.pdf}}$.

Shirvani Boroujeni, M., Sharma, K., Kidziński, Ł., Lucignano, L., & Dillenbourg, P. (2016). How to quantify student's regularity? In K. Verbert, M. Sharples, & T. Klobucar (Eds.), *Proceedings of the 11th European Conference on Technology Enhanced Learning* (EC-TEL 2016), 13–16 September 2016, Lyon, France (pp. 277–291). Springer. https://doi.org/10.1007/978-3-319-45153-4$_{2}$1

Trcka, N., Pechenizkiy, M., & van der Aalst, W. (2010). *Process Mining from Educational Data*. Boca Raton, FL: Chapman & Hall/CRC.

Van der Aalst, W., Weijters, T., & Maruster, L. (2004). Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, *16*(9), 1128–1142. https://doi.org/10.1109/TKDE.2004.47