International **R**eview of **R**esearch in **O**pen and **D**istributed **L**earning
Volume 19, Number 5

November – 2018

Analysing Structured Learning Behaviour in Massive Open Online Courses (MOOCs): An Approach Based on Process Mining and Clustering



Antoine van den Beemt[1], Joos Buijs[2], and Wil van der Aalst[3]
[1, 2]Eindhoven University of Technology, The Netherlands, [3]RWTH Aachen University, Germany

Abstract

The increasing use of digital systems to support learning leads to a growth in data regarding both learning processes and related contexts. Learning Analytics offers critical insights from these data, through an innovative combination of tools and techniques. In this paper, we explore students' activities in a MOOC from the perspective of personal constructivism, which we operationalized as a combination of learning behaviour and learning progress. This study considers students' data analyzed as per the MOOC *Process Mining: Data Science in Action*. We explore the relation between learning behaviour and learning progress in MOOCs, with the purpose to gain insight into how passing and failing students distribute their activities differently along the course weeks, rather than predict students' grades from their activities. Commonly-studied aggregated counts of activities, specific course item counts, and order of activities were examined with cluster analyses, means analyses, and process mining techniques. We found four meaningful clusters of students, each representing specific behaviour ranging from only starting to fully completing the course. Process mining techniques show that successful students exhibit a more steady learning behaviour. However, this behaviour is much more related to actually watching videos than to the timing of activities. The results offer guidance for teachers.

Keywords: social learning analytics, constructivism, learning analytics, learning behavior, educational data mining, process mining

Introduction

Massive Open Online Courses (MOOCs) (McAuley, Stewart, Siemens, & Cormier, 2010) are usually built in a structured way from modules containing video lectures, quizzes, and discussion forums (Lackner, Kopp, & Ebner, 2014). Collecting and storing all online behaviour in MOOCs results in large amounts of data. Using these "digital traces" (Gillani & Eynon, 2014) about learners and their context to understand and optimize learning and teaching, is known as Learning Analytics (LA; Siemens & Baker, 2012). In the last few years, efforts have been made to relate LA explicitly to learning processes and learning theories (Buckingham Shum & Ferguson, 2012). The focus on enforcing and stimulating learning processes rather than on collecting and analysing large amounts of data, leads to a call for more personal and learner-centric LA (Buckingham Shum & Ferguson, 2012).

This paper explores patterns in students' learning behaviour and learning progress in a MOOC by looking at activity sequences. Learning behaviour in the context of MOOCs refers to how, when, and in what order students watch videos and process other MOOC resources; and when and in what order they make quizzes and assignments. Learning progress consists of results of these efforts in pass or fail of quizzes and in final results of the MOOC or course. The aim of this paper is to describe and explain sequences of learning behaviour, with the purpose of finding indicators for improving the quality of teaching and learning. We intend to increase understandings of how passing and failing students distribute their activities differently along course weeks rather than predict students' grades from their activities. We are looking for learning process models that represent the sequence of students' interactions with MOOC resources in relation to learning progress.

To investigate students' engagement with videos and quizzes, we formulated the following question: What patterns can be found in students' learning behaviour in a MOOC? We answer this question with an exploratory sequence analysis using Process Mining (PM) and hierarchical clustering as methods. Understanding these patterns helps to figure out which students are on a path to passing the course as well as supporting course design for MOOCs. Facts and details about patterns in learning behaviour offer useful tips for students to improve both their learning behaviour and their progress while they follow a MOOC. Furthermore, it can support teachers to make teaching more personal and learner-centric.

Background and Related Work

The fast-developing context of research on LA and MOOCs shows a variety of emerging themes (e.g., Peña-Ayala, 2018; Veletsianos & Shepherdson, 2016) such as MOOC design (Watson et al., 2016), student subpopulations (Kizilcec, Piech, & Schneider, 2013), or student motivation (Koller, Ng, Chuong, & Zhenghao, 2013). However, despite existing research on MOOC design (Watson et al., 2016), it appears that many MOOCs are developed without applying basic instructional design principles (Margaryan, Bianco, & Littlejohn, 2015) as formulated by Reigeluth (2016), for instance.

Furthermore, the "funnel of participation" described as going from awareness to registration, activity, progress, and for some learners, even completion (Clow, 2013), leads to attention toward student dropout (e.g., Kahan, Soffer, & Nachmias, 2017). High dropout numbers are a concern for MOOC providers and educational institutions. Dropout is studied, for instance, by investigating reasons why people are not able to reach their intended goals (Henderikx, Kreijns, & Kalz, 2017). Looking at

behaviour in retrospect helps to cluster students according to actions; and offers insight in how to support future students and avoid dropout.

To successfully obtain personal learning goals in a MOOC (Conijn, Van den Beemt, & Cuijpers, 2018), students need to regulate their learning more compared to traditional, face-to-face education (Hew & Cheung, 2014; Winne & Baker, 2013). Research from the perspective of self-regulated learning (Winne & Hadwin, 1998) indicates that successful MOOC students have high beliefs in their ability to complete academic tasks, and that previous MOOC experiences increase this self-efficacy (Lee, Watson, & Watson, 2019). Furthermore, successful MOOC students are reported to show more self-regulating activities that support them in actively constructing knowledge compared to failing students (Bannert, Reimann, & Sonnenberg, 2014).

Actively constructing knowledge relates to perspectives on learning such as constructivism (e.g., Bruner, 1996). According to constructivist theories, teachers should elicit students' prior conceptions on the topic taught and create a cognitive conflict in students' minds. This conflict confronts students with new phenomena or knowledge; or with conceptions of others (Bächtold, 2013). Science and engineering education literature distinguishes two kinds of constructivism (Loyens & Gijbels, 2008): personal constructivism (PC) or cognitive constructivism, and social constructivism (SC). In this distinction, PC focuses on individual learners; SC focuses on the social relations between teacher and student, or between students. PC considers the process of knowledge construction to be primarily based on interaction between student and learning materials. Both kinds of constructivism are considered complementary because students need guidance in developing an understanding of concepts (PC) before they can incorporate these concepts in other contexts (SC).

In MOOC context, PC activities include replaying videos and watching large(r) proportions of videos, which positively correlates with finishing a course (Sinha, Jermann, Li, & Dillenbourg, 2014). This leads to attention for sequences of student activities, aiming at predictions of student performance or increase of pedagogical quality of MOOCs (see Pena-Ayala, 2017), amongst others. Research analysing sequences of activities indicated that switching assignments, i.e., completing them in an order different from the course content, increases course failure (Kennedy, Coffrin, & De Barba, 2015). Furthermore, Wen and Rosé (2014) describe a case study where passing students showed a bump in engaging with lectures and assignments in the second half of the term, yet where failing students continued at a moderate pace.

Determining patterns from sequences of activities in MOOCs is becoming common to get a better understanding of underlying educational processes (Maldonado-Mahauad, Pérez-Sanagustín, Kizilcec, Morales, & Munoz-Gama, 2018). However, most of the traditional data-mining techniques focus on data dependencies, single events, or simple patterns (Bogarín, Cerezo, & Romero, 2018). This kind of research does not focus on the process as a whole and does not offer clear visual representations of overall learning processes (Trcka, Pechenizkiy, & Van der Aalst, 2011). PM is a robust method that supports the discovery of process models representing sequences of interactions between students and learning materials (Van der Aalst, 2016).

PM applied to raw educational data, with a process-centric approach and focus on sequences of events, is coined Educational Process Mining (EPM, Bogarin, Cerezo, & Romero, 2018). Research in this nascent field often combines PM with clustering techniques, for instance, to identify interaction

sequence patterns and groups of students (Emond & Buffett, 2015), or to optimise comprehensibility of the model obtained (Bogarín et al., 2018). Results of this kind of research show, for instance, that better-graded students have more effortful cognitive activities and use more varied learning strategies in the process of problem solving (Vahdat, Oneto, Anguita, Funk, & Rauterberg, 2015).

Maldonado-Mahauad and colleagues (2018), applying PM from the SRL-perspective, identified three clusters of learners: 1) comprehensive learners, who follow the sequential structure of the materials; 2) targeting learners, who strategically engage with specific course content to pass the assessments; 3) sampling learners, who exhibit more erratic and less goal-oriented behaviour, and underperform compared to the other clusters.

Analysing weekly engagement trajectories of students, Kizilcec, Piech, & Schneider (2013) found four types of learning behaviour:

1) *Completing:* students who at least attempted and completed the majority of assessments in the course,

2) *Auditing:* students who engaged by watching videos rather than assessments,

3) *Disengaging:* students who started off well with assessments but then showed a decrease in engagement generally in the first third of the course, and

4) *Sampling*: students who watched videos for only one or two periods, and then disappeared.

Approach and Added Value

Applying a data-driven approach, we looked at learning behaviour in retrospect to find patterns in knowledge construction resulting from interactions between student and course materials. We focused on sequences of watching videos and submitting quizzes, because interactions between student and medium are considered conceptualisations of higher-order thinking eventually leading to knowledge construction (Chi, 2000). This kind of data-driven approach suits both PM as explorative method, and PC as perspective; and lets us focus on the two variables of watching videos and submitting quizzes with a projection over time. Creating this projection should support teachers and students to improve learning and teaching rather than be a goal per se. The combination of PM and statistics is intended to visualize previously invisible learning processes.

We perceive MOOCs as a step in the development of online learning materials and pedagogies (Bali, 2014); this study is an exploration of how to support learning processes and pedagogies with data from the use of online learning materials. Further, this study is a step from descriptive LA towards explanatory LA (Brooks & Thompson, 2017) by offering a view of loopbacks, deviations, and bottlenecks; including quiz submission behaviour.

Methods

Case Study Descriptives

This study considered student data from the MOOC *Process mining: Data Science in action*, which has been running on Coursera since November 2014. The MOOC consisted of six modules, each counting up to nine videos. Videos were 15 to 27 minutes long, averaging 20 minutes per video. Each module ended with a weekly quiz, and the course concluded with a final quiz. Students that aimed for a "certificate with honour" needed to take a tool quiz (tutorial-style quiz to make them familiar with the tools used); and a peer assignment that asked them to analyse a real-world dataset as well as mimic writing a report by answering several questions. Each new module started one week after the previous, but the weekly quiz could not be submitted until two weeks after the start, to provide students some time to catch-up if needed. Because of a holiday period halfway through the term, the total run time of the MOOC counted 8 weeks. Table 1 shows details about students'and success rates.

Table 1

Global Statistics for the First Run of the MOOC Process Mining: Data Science in Action

Start date	Nov. 14, 2014
Registered	42,480
Visited course page	29,209
Watched a lecture	16.224
Browsed forums	5,845
Submitted an exercise	5,798
Certificates (course/distinction)	1,662
Course certificate	1,019
Distinction certificate	643
End date	Jan. 8, 2015

Note. Statistics were taken from the Coursera course dashboard, except for the "Watched a lecture"; and the final 3 certificate statistics, which were taken from the extracted dataset.

Analysis of learning behaviour was pursued through the stream of click events generated by students on the MOOC's content pages. A "clickstream" is defined as the trail students leave as they browse through video lectures or when they submit quizzes. Learning progress was measured by students' quiz results, final grades, and the type of certificate they get awarded after completing the course (Course Certificate for achieving a grade of 60 or better, or Distinction Certificate for achieving a grade of 90 or better) as well as whether they were on signature track, which means they paid for the course.

Process Mining

PM combines data mining and business process analysis; providing algorithms, tools, and techniques to analyse event data, consisting of traces of observed actions (Van der Aalst, 2016). PM offers three main types of analysis: process discovery, conformance checking, and enhancement (Van der Aalst, 2016). Discovery techniques learn a process model from the provided event log data. Conformance checking attempts to verify conformity of the data to a predefined model and identify deviations, if any; while enhancement provides for models to be improved based on the data in the event logs (Van der Aalst, 2016). Additionally, PM provides techniques and visualizations to further explore and analyse the event log data.

Event Logs

PM needs an event log with student behaviour data. Table 2 shows the minimal columns (case, activity, and timestamp) required as input for PM. PM distinguishes between cases that are following a process. In this paper, each student participating in the MOOC was a case, identified by the field "Student ID". Each student left a trace of observed actions, or events. Each event had (at least) an action that was performed, and a date and time at which it was performed. In Table 2, the lecture watched is the activity, and the timestamp is the time at which the lecture was opened.

Table 2

Example of an Event Log

Student ID	Lecture	Date Time
123456789	Lecture 1.1	2016-01-01 10:00
132456789	Lecture 1.2	2016-01-01 11:00
132456789	Lecture 1.3	2016-01-01 12:00
987654321	Lecture 6.2	2016-01-01 02:00
987654321	Lecture 3.4	2016-01-08 22:00
987654321	Lecture 1.3	2016-01-22 15:00

The event log was constructed using the Coursera data, as shown in Figure 1. For each user, the clickstream is extracted, focusing on lecture watching and quiz results. Note that no in-video action information (e.g., pausing, resuming, in-video quiz interaction) was recorded.

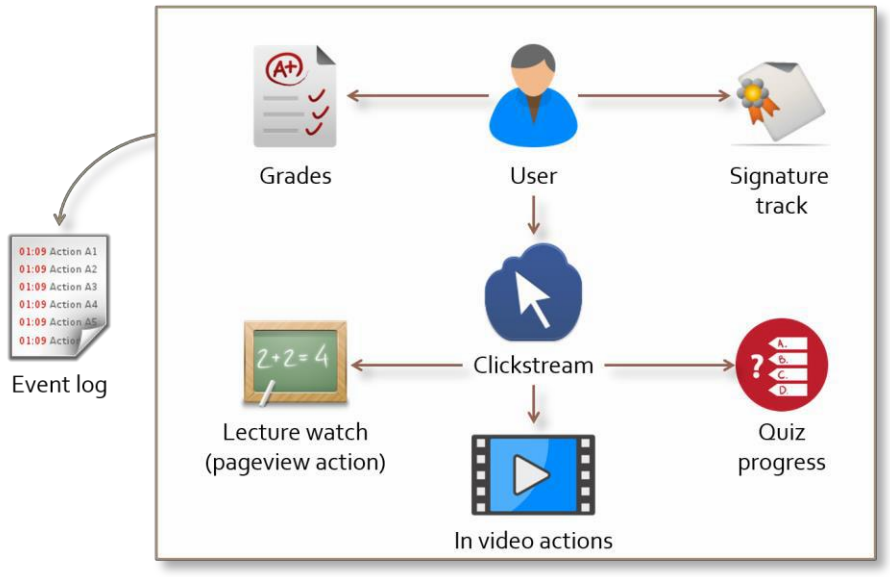

Figure 1. Generalized overview of Coursera data used.

Event Log Description

The extracted event log contains 16,224 students. Students that had no activity observed, belonged to the teaching or Coursera teams, or for which the obtained certificate was not recorded correctly were filtered out. For these students a total of 285,036 events were recorded within the timespan from

November 12, 2014 to January 31, 2015. Events were recorded for each of the 50 lectures, for the submission of weekly quizzes, tool quiz, final quiz, and the two introductory lectures. Figure 2 shows the number of recorded events per activity (e.g., lecture, quiz), sorted by the expected learning sequence. The first "real" lecture is the best-watched item, after which the view counts per video decreases over the course of the MOOC. The weekly quizzes were submitted more often than the average number of times a lecture had been watched in that week, which means that students often made multiple efforts for a quiz.

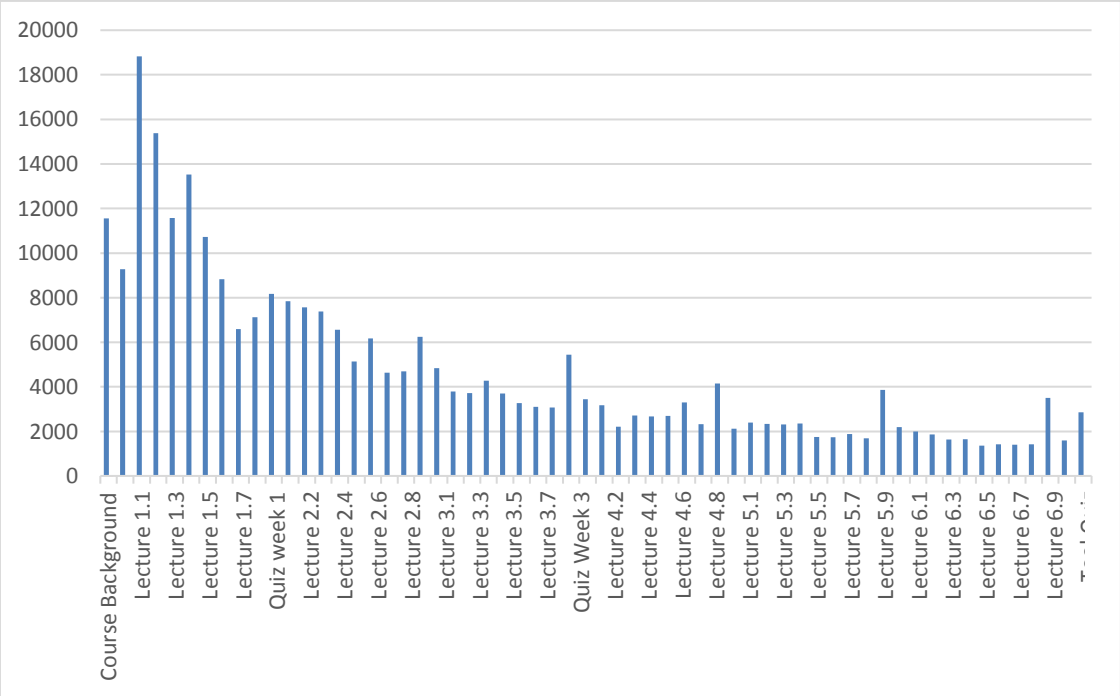

Figure 2. Number of recorded events per activity (e.g., lecture, quiz; sorted by order of expected execution). X-axis items show alternating materials; bars show all consecutive materials.

In Figure 3, the time of execution of the events is visualized using a dotted chart (Song & Van der Aalst, 2007). In a dotted chart, each recorded event is presented as a dot, where each row contain all events of a student; time is recorded on the x-axis. Colour indicates the activity, similar to the x-axis of Figure 2 (e.g., Lecture 1.1, Quiz Week 6). Each dot in the dotted chart of Figure 3 thus represents an observed event for one student, at a particular time. The vertical coloured bars in Figure 3 indicate the weekly rhythm of activities. The arch shows the first activity of students, ordered along the timeline. The density of the dots indicates overall student activity.

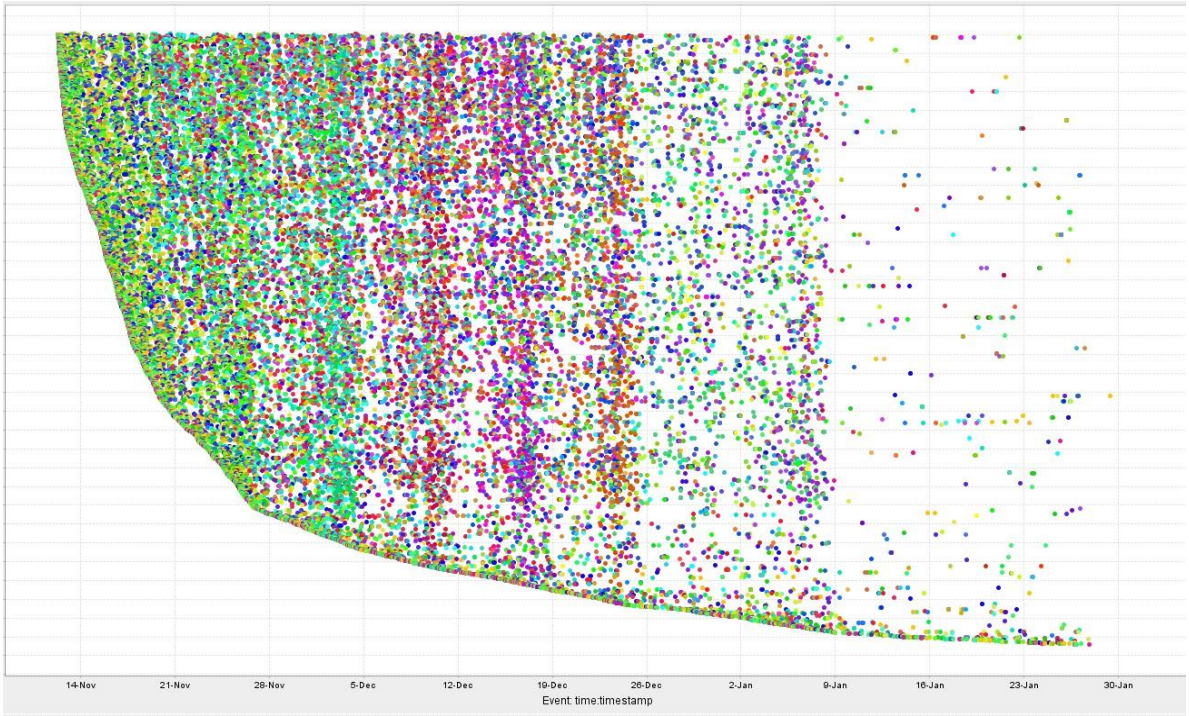

Figure 3. Dotted chart showing the events in the event log over time.

Process Model Discovery

Based on the event data, a process model can be discovered that describes the relation between observed activities, in our case, watching video lectures and submitting quizzes. We focus on the quiz submission process, because including all 60 possible activities would make the process model unreadable. **Error! Reference source not found.** and 5 show the process models that are discovered when considering quiz submission events, split over students that did not obtain a certificate compared to those that did obtain a certificate. Orders of activities followed by large numbers of students would be visualized in the model as sequences of activities, connected by arrows. Because the model immediately starts with a +-sign, indicating parallel execution of the branches, it can be concluded that no clear ordered pattern could be found. The numbers in the model represent numbers of students following a specific path through the model. The ordering from top to bottom has no hierarchical meaning, as all branches are executed in parallel, and is the result of the software package drawing the model.

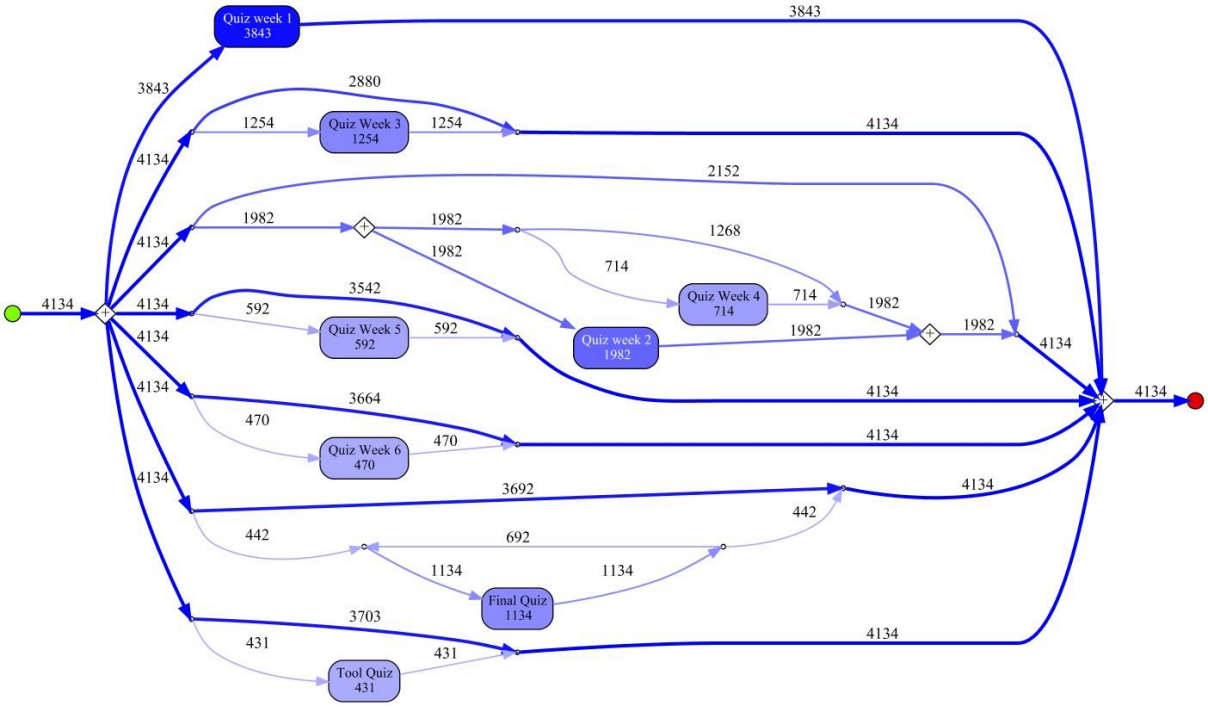

Figure 4. Quiz submission behaviour of students that did not obtain a certificate.

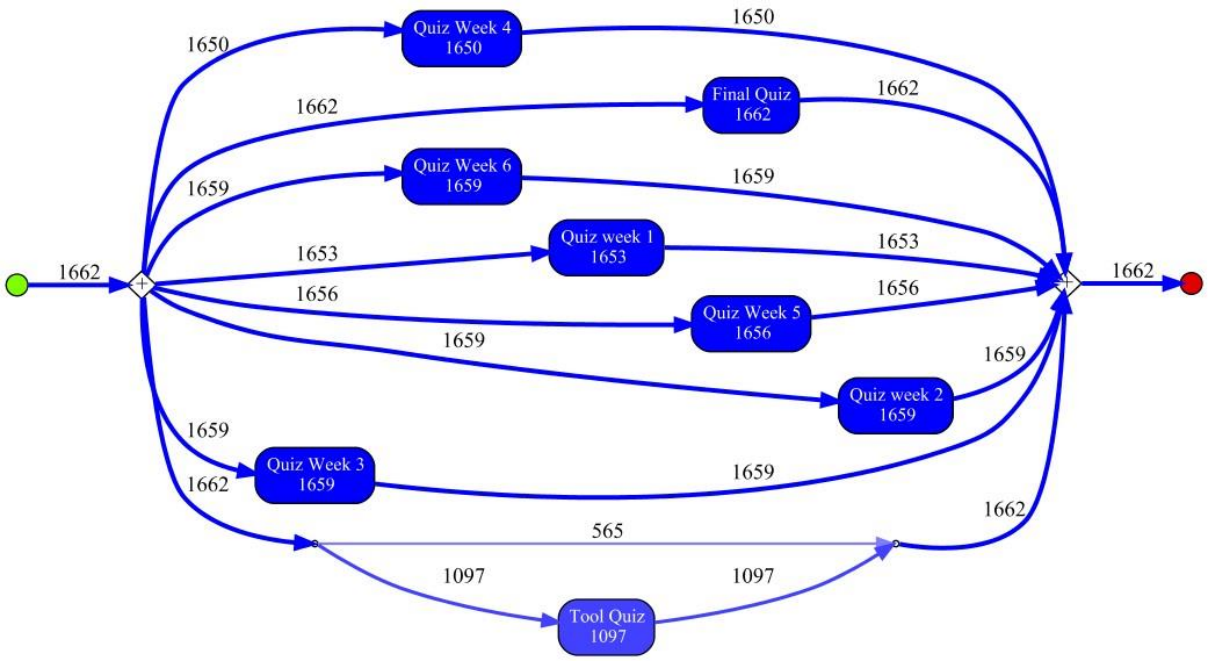

Figure 5. Quiz submission behaviour of students that obtained a certificate.

Conformance Alignments and Learning Behaviour

PM offers techniques to analyse event data in relation to the process model. Conformance checking techniques can verify the conformance of event data on a given process model (Van der Aalst, Adriansyah, & Van Dongen, 2012). In our setting, we can create a normative process model that

describes the ideal sequential study behaviour, i.e., starting at lecture 1, then 2, etc. Then, conformance checking can be applied to compare the actual behaviour with the ideal sequential behaviour.

Alignment-based conformance checking results in an optimal alignment of the observed behaviour of a student, as seen in the event log, and a possible execution of the process model.

Table 3 shows an example of such an alignment between the observed trace <Lect 1.1, Lect 1.2, Lect 1.6, Lect 1.4> and a possible run of the process model <Lect 1.1, Lect 1.2, Lect 1.3, Lect 1.4, Lect 1.5, Lect 1.6>. An alignment consists of a sequence of moves. Each move is either an activity recorded in the trace (but not executed in the model; move on log only), an activity that was executed by the process model but that did not occur in reality (move on model only), or a combination of both model and event log (synchronous move).

Table 3

Example of an Alignment Calculated between an Observed Trace and an Expected Sequence From the Process Model

Trace	Lect 1.1	Lect 1.2	Lect 1.6	>>	Lect 1.4	>>	>>	Lect 1.3
Model	Lect 1.1	Lect 1.2	>>	Lect 1.3	Lect 1.4	Lect 1.5	Lect 1.6	>>
Move type	Synchronous move		Move on Log only	Move on model only	Synchronous move	Move on model only		Move on Log only
Watch type	Watched regularly		Watched early	Not watched	Watched regularly	Not watched	Watched early	Watched late

Based on the order information, videos can be assigned any of these labels: *Watched Early*, *Watched Regularly, Watched Late, New or Not Watched* (bottom row

Table 3). By aggregating these labels, student behaviour and level of commitment in the MOOC can be defined; providing insights regarding PC. These labels were used to perform cluster analysis.

Cluster and Means Analysis

To explore a pattern of related watching behaviour among MOOC users, cluster analysis on the cases was applied. The independent variables consisted of mean scores of watching videos per week. Because there was no a priori classification scheme, hierarchical agglomerative cluster analysis was applied instead of discriminant or assignment methods (Everitt, Landau, Leese, & Stahl, 2011). To minimize the variance within clusters, Ward's method was applied with squared Euclidian distance. However, because this measure is affected by variables with large size or dispersion differences, z-scores were applied as well. Within each cluster mean scores were computed for watching behaviour per week and for quiz scores.

Results

For each respondent included in the analysis, at least one video watching instant was available. The general image of student activities (Figure 3) shows a weekly rhythm, but also late entrance of some students, sometimes even after all deadlines had passed. It also appears from Figure 3 that the overall activity of students decreased during the run of the MOOC. However, some students who started early on, still showed observed activity after all the deadlines had passed.

Our data showed that certificate students did not necessarily exhibit structured learning behaviour (Figure 4). Non-certificate students showed even less structured behaviour with some quiz submissions being skipped. Similar results were found for video watching, where non-certificate students skipped many videos, and certificate students did so in the last few weeks of the course.

Cluster analysis resulted in four clusters describing types of learning behaviour (see Table 4). Different cluster solutions did not result in comprehensive groupings of MOOC students in relation to the weekly mean scores of watching videos. Cluster 1 (N = 11,875; 73.2%) represents students who enrolled in the course yet quickly dropped off. Cluster 4 (N = 1,293; 8.0%) represents students who enrolled and showed, on average, steady learning behaviour and progress, resulting in high pass rates. Cluster 2 (N = 1,795; 11.1%) represents students who enrolled, did an attempt to watch videos and submit quizzes, yet failed to continue their learning behaviour. Cluster 3 (N = 1,261; 7.8%) represents students who enrolled and did a serious attempt to watch videos and submit quizzes, yet often failed to continue their learning behaviour as well, albeit at a later stage during the course. The difference between clusters 2 and 3 was also found in the achievement levels that show, for cluster 3, an increase in Course Certificate and Distinction Certificate level. The small percentage of passing students in clusters 1 and 2 represents learning behaviour that resembles failing students' sequences of activities, yet turns out to be successful.

Table 4

Cluster Size and Achievement Level

	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	Count	%	Count	%	Count	%	Count	%
Fail	11,487	96.8	1,694	94.4	1,062	84.2	306	23.7
Course certificate	237	2.0	67	3.7	131	10.4	584	45.2
Distinction certificate	138	1.2	34	1.9	68	5.4	403	31.2
Total	11,875	100	1,795	100	1,261	100	1,293	100

Learning Behaviour and Learning Progress Within Each Cluster

The four clusters were compared for differences between mean scores on video watching, quiz submission; and mean scores for weekly quiz results (see Table 5 and Table 6). To compute mean scores on video watching and quiz submission, timestamps were translated into a 4-point scale. Video watching was computed as an average per student per week and subsequently for each cluster per week. Quiz submission and scores were computed as an average for each cluster per quiz. Members of cluster 1, on average, never watched videos regularly or early, nor submitted quizzes on time, and passed no single quiz. Some students in this cluster started watching videos and made an effort for the quizzes;

however, they dropped off massively after week 1. Cluster 2, on average, started watching late, and increasingly procrastinated watching videos and submitting quizzes. Quiz scores (see Table 6) for cluster 2 start off reasonably well, with many students passing at least the first quiz. However, results soared rapidly after quiz 2. Cluster 3, on average, also started watching late, however, students in this cluster prolonged their watching behaviour further, and made a greater effort to submit quizzes throughout the course. This cluster shows better average results up to week 3. Cluster 4 showed steady watching behaviour, although most videos on average were watched at a rather late point in time. The standard deviation (SD)of cluster 4 shows a wider dispersion of watching behaviour and quiz submission towards the end of the course, indicating less coherence in this cluster. Watching videos late, and submitting quizzes non-ordered however, did not negatively influence quiz results because cluster 4 shows, on average, steady quiz results; leading to passing the Tool Quiz and Final Quiz and eventually passing the complete course.

Table 5

Mean Video Watching Scores and Quiz Submission Scores (and SD) per Week

	Cluster							
	1		2		3		4	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Video watching								
Week1	.40	.45	1.88	.68	2.09	.69	2.24	.55
Week2	.05	.18	.82	.69	2.21	.74	2.60	.54
Week3	.01	.08	.16	.37	1.15	.98	2.55	.54
Week4	.01	.06	.06	.22	0.46	.73	2.44	.65
Week5	.01	.05	.04	.20	0.26	.58	2.29	.78
Week6	.01	.05	.04	.20	0.23	.61	2.07	.98
Quiz submission								
Quiz 1	.40	.93	1.51	.1.26	2.04	1.03	2.35	.74
Quiz 2	.14	.50	.33	.74	.95	1.00	1.15	.99
Quiz 3	.19	.71	.40	.98	1.26	1.40	2.67	.81
Quiz 4	.16	.64	.29	.86	.74	1.22	2.45	.93
Quiz 5	.14	.62	.26	.81	.62	1.17	2.45	1.00
Quiz 6	.14	.61	.23	.76	.59	1.14	2.40	1.05
Tool quiz	.08	.40	.13	.51	.31	.74	1.17	1.00
Final quiz	.16	.67	.23	.79	.60	1.19	2.47	1.14

Note. Scores on 4-point scale; 0 = not watched/submitted, 1 = watched/submitted late, 2 = watched/submitted regularly, 3 = watched/submitted early.

Table 6

Mean (and SD) Quiz Scores and Grades

			Cluster							
			1		2		3		4	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
Quiz1			.65	1.49	2.53	2.08	3.56	1.71	4.21	1.14
Quiz2			.30	.98	.75	1.40	2.25	1.70	3.36	1.15
Quiz3			.19	.76	.37	1.00	1.27	1.55	2.91	1.24
Quiz4			.19	.82	.33	1.02	.88	1.54	3.30	1.41
Quiz5			.19	.82	.34	1.07	.81	1.56	3.39	1.41
Quiz6			.19	.85	.33	1.13	.86	1.69	3.55	1.56
Tool quiz			.30	1.60	.52	2.10	1.31	3.20	5.32	4.67
Final quiz			.71	3.25	1.11	4.02	3.10	6.40	13.43	6.71
Distinction grade			3.33	12.70	7.36	15.48	17.66	23.14	55.73	28.68
Course grade	5.09	16.98	12.13	20.84	27.96	29.10	75.19	26.14		

Note.The maximum score for the week quizzes is 5 points, 10 for the tool quiz and 20 for the final quiz. The course and distinction grades are on a scale between 0 and 100.

Cluster Analysis by Activity Frequency

6 shows how often each video or quiz was accessed, split per cluster. This basic analysis suggests that students in cluster 1 mainly accessed materials in week 1. Cluster 2 students also mainly accessed materials in week 1 and 2, but submitted quizzes after week 2 as well. Cluster 3 students seemed to be active until week 3, sometimes week 4, but afterwards showed less activity, except for quizzes. Students in cluster 4 showed activity throughout the whole course.

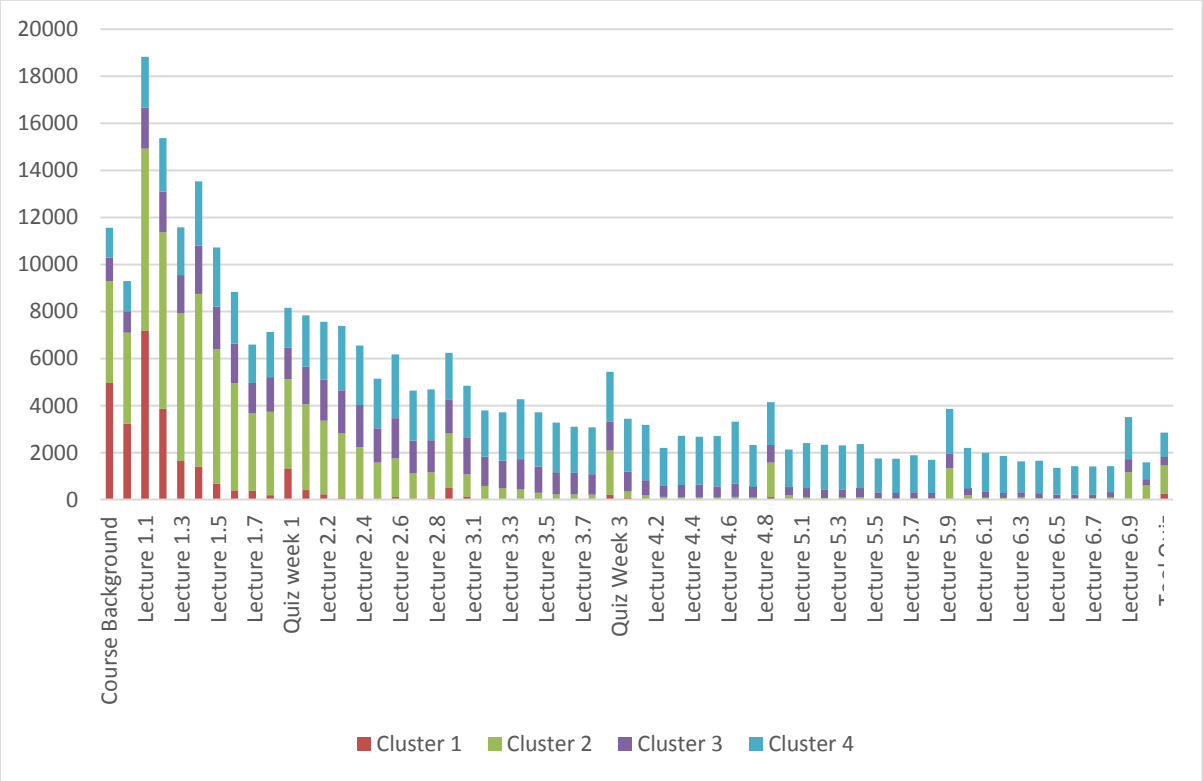

Figure 6. Activity frequency counts split per cluster.

Cluster Analysis Using Dotted Charts

The dotted chart visualization shows for cluster 1 (Figure 7) three types of students: those that started before the deadline of week 1, students that only have observations between the week 1 and 6 deadlines, and those students that started after the week 1 deadline but continued after the week 6 deadline still.

Figure 8 (cluster 2), indicates that more students started before deadlines, and fewer after. Students in cluster 2 also have, on average, 18 activities observed, indicating they were also more active. Figure 9 shows that, in cluster 3, even more students started before deadlines. This cluster represents students with on average 53 observed activities. Figure 10 shows for cluster 4 a pattern resembling cluster 3. Cluster 4 contains students with on average 98 observed activities.

The differences in density indicate that students in cluster 1 watched the least videos of all clusters. However, the colours show that mainly videos from the first module were watched. Each cluster shows a more dense distribution of dots, indicating that not only more videos were watched, but also closer after another. Furthermore, cluster 3, and especially cluster 4, have very little observed activity after the final course deadline.

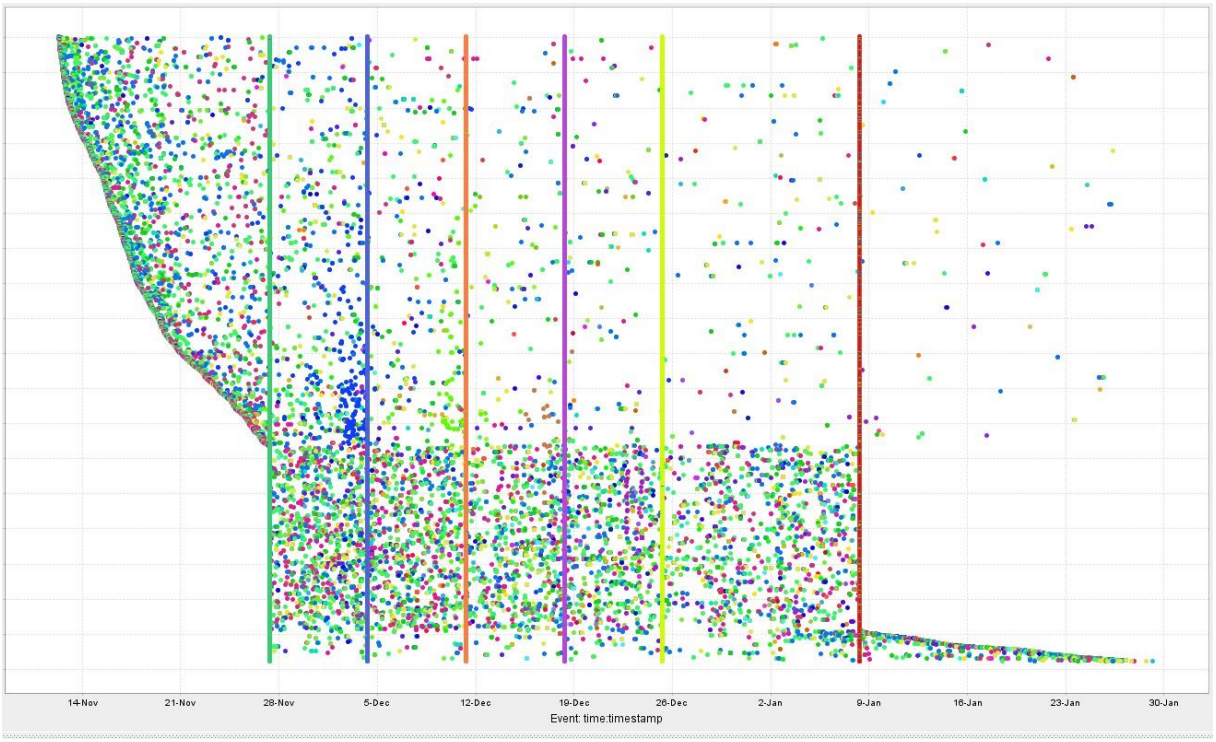

Figure 7. Video watching trends for cluster 1.

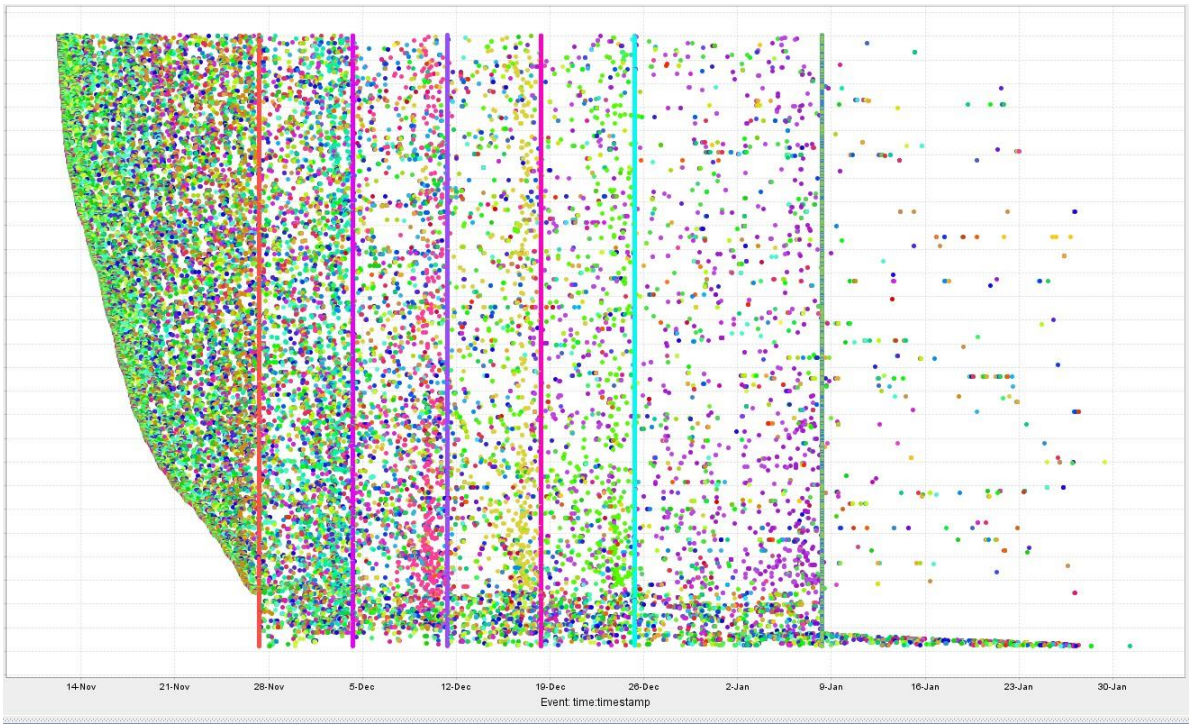

Figure 8. Video watching trends for cluster 2.

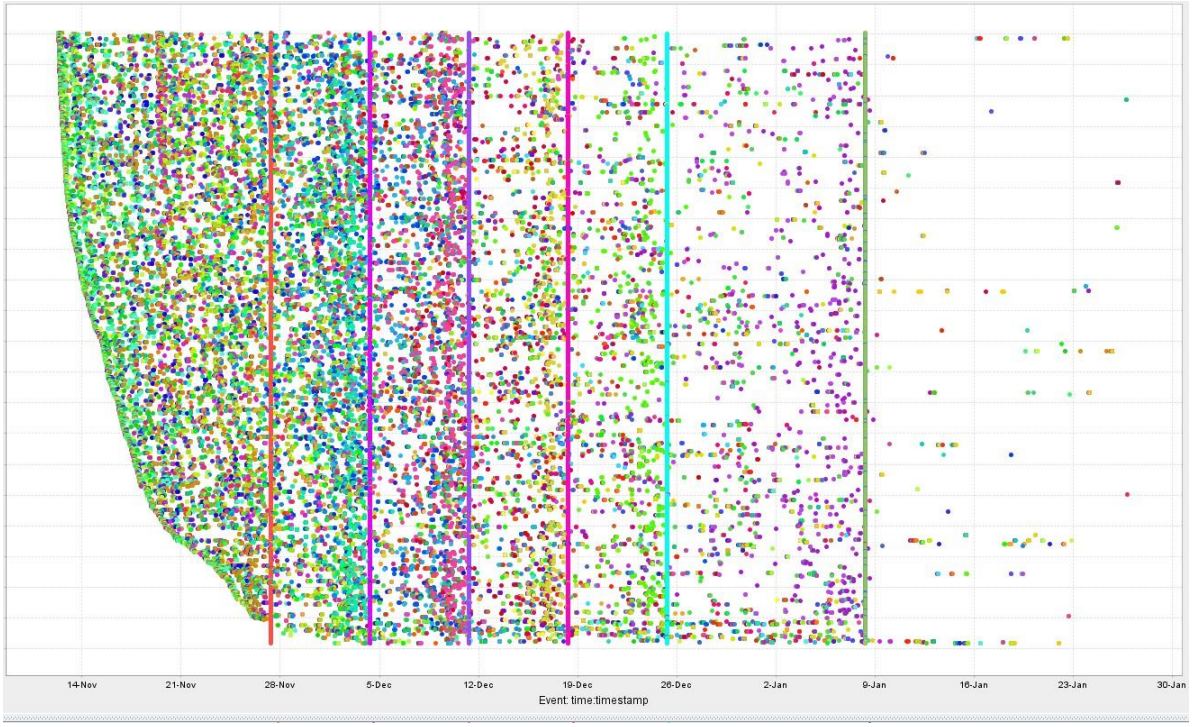

Figure 9. Video watching trends for cluster 3.

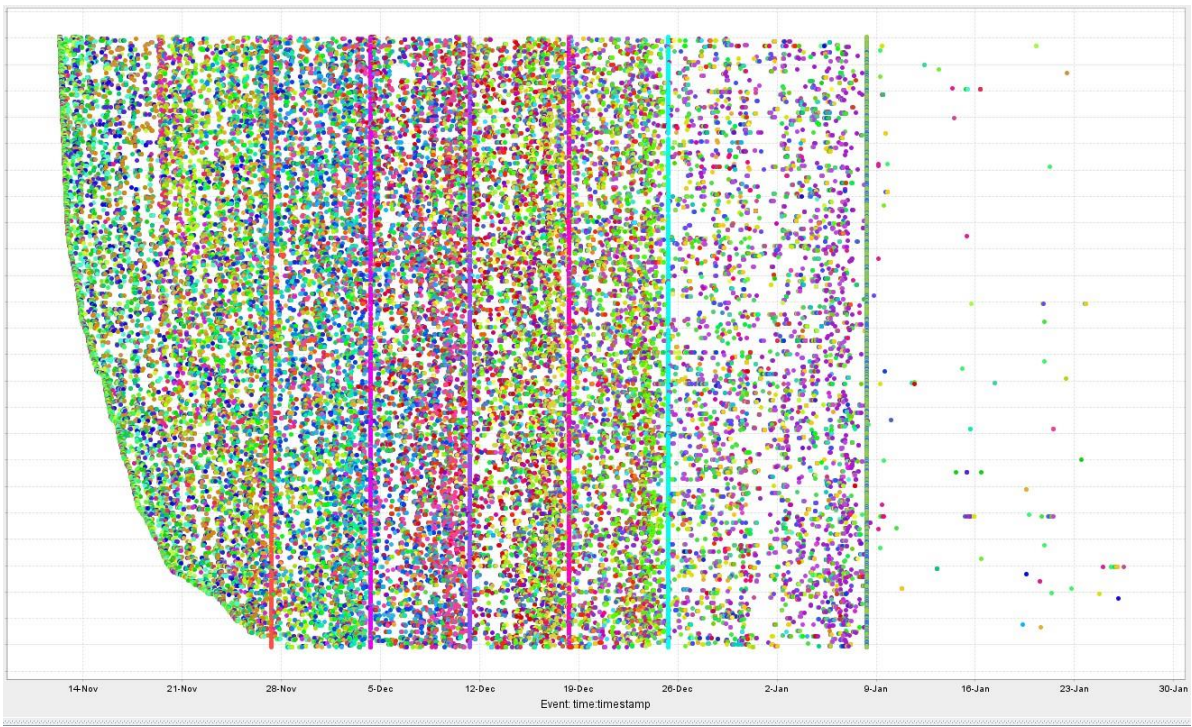

Figure 10. Video watching trends for cluster 4.

Process Models Discovered per Cluster

Guided by the clusters, the quiz submission process can be discovered. Figure 11 shows this process for cluster 1. The process model starts with submission of quiz 1, which was observed 932 times for 1127 students. Increasingly more students skipped the submission of other weeks (from 807 students not

submitting the week 2 quiz, to 1,101 students not submitting the quiz for week 6). Furthermore, the position of the final quiz – in-between quiz week 2, and quiz week 3 - is interesting with 246 students who tried it (less than the number of students attempting quizzes for week 3 or later).

Students in cluster 2 (Figure 12) submitted quizzes in a rather random fashion, with rapidly dropping submission numbers. The quiz for week 1 was submitted by 2,560 students out of 2,645; 687 students attempted the quiz for week 6; 643 students tried the final quiz. The quiz submission for cluster 3 (Figure 13) shows that the quizzes for weeks 1 and 2 are ordered, but during the course the quizzes were submitted without a clear order. Out of 894 students, 343 students submitted the week 6 quiz, while 318 submitted the final quiz. 14 shows the behaviour of cluster 4 regarding quiz submissions. Of the 1130 students, most submitted all week quizzes and the final quiz (1,033), while the tool quiz was only submitted by 725 students. The order of quizzes, however, was not very structured after week 1.

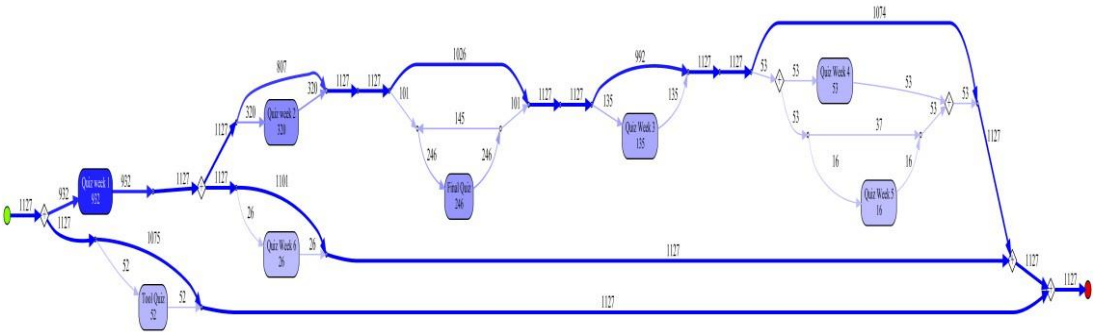

Figure 11. Quiz submission process model for cluster 1.

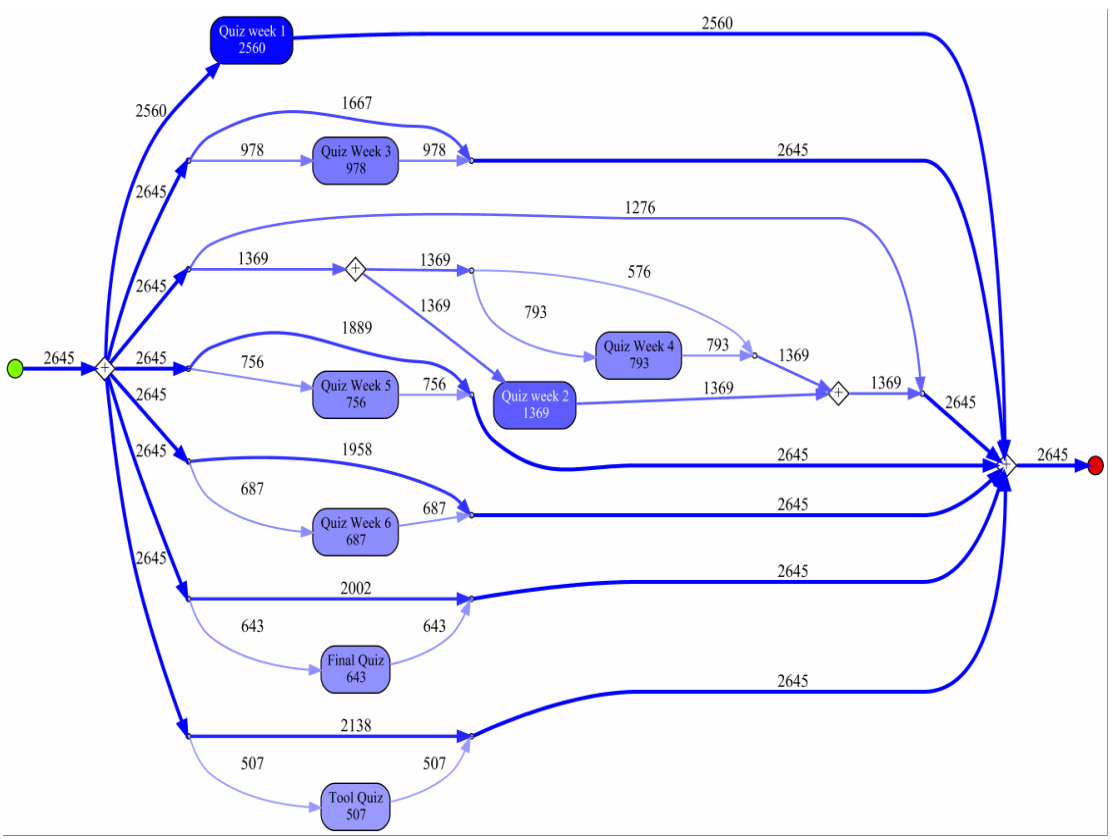

Figure 12. Quiz submission process model for cluster 2.

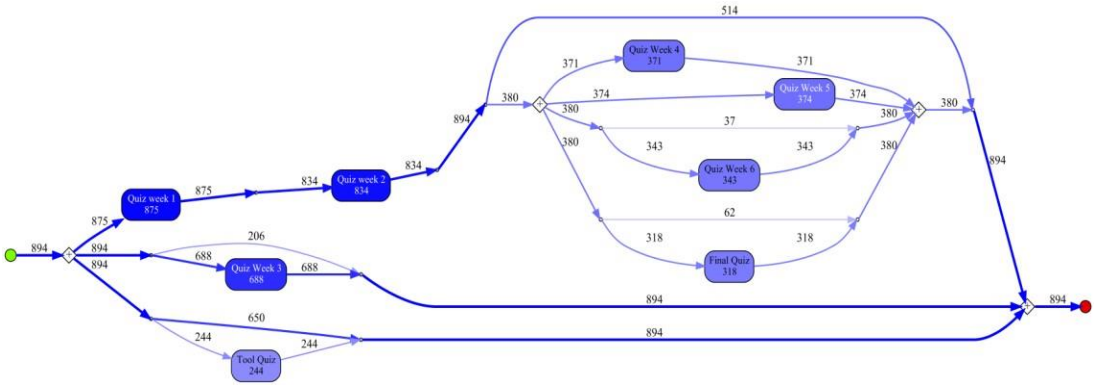

Figure 13. Quiz submission process model for cluster 3.

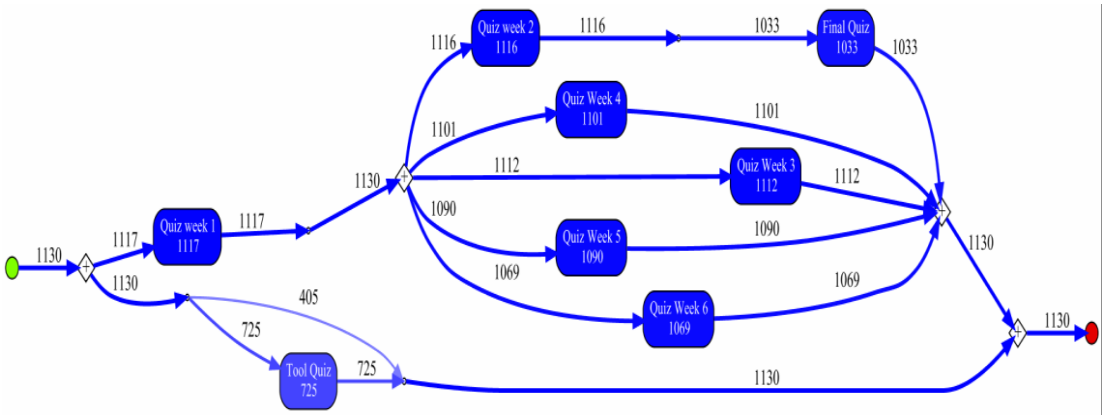

Figure 14. Quiz submission process model for cluster 4.

Discussion

To answer our research question, the results of cluster analysis in relation to learning behaviour indicate that regularly watching successive videos in batches leads to the best learning outcome. However, the results indicate that this behaviour is much more related to the order of watching videos than to the actual timing. This procrastinating behaviour is also found in other studies (e.g., Wen & Rosé, 2014). Passing a course obviously requires good quiz scores. However, the results do not confirm that refraining from "assignment switching" (Kennedy et al., 2015) leads to better results. Regarding quiz submission behaviour, both non-certificate students and certificate students showed irregular patterns, with non-certificate students more often skipping quizzes.

Students in cluster 1, represent the top part of the "funnel of participation" (Clow, 2013), because they were aware, had registered, had watched one or two videos, and then dropped off. Putting teacher effort into students of cluster 1 would result in little effect. The few students in this cluster who made quizzes, did this in a relatively ordered way. Following Kizilcec and colleagues (2013), and Madonado-Ahauad and colleagues (2018), these students can be labelled as "Samplers."

Cluster 2 students enrolled, made an effort, and then dropped off. A large number of these students submitted quizzes, but in a rather random order. Learning results soared rapidly from the second week on. This cluster had a large group of students that started early and dropped off gradually before the

end of the course, and a smaller group that started late and then gradually dropped off. This cluster can be labelled “Disengagers,”, in line with Kizilcec and colleagues (2013). Because these students passed at least the first quiz, they showed a certain level of understanding of the course topic. Therefore, giving them support in active knowledge construction could be beneficial.

Students in cluster 3 made a serious effort, however appeared to fail halfway through the course and onwards. They started watching videos late, but showed a more steady learning behaviour and continued to submit quizzes, on average,with better quiz results compared to cluster 2. However, the patterns of video watching and quiz submission were disordered after week 2, with quickly dropping submission numbers. Their behaviour looks like ‘Targeting” (Madonado-Ahauad et al., 2018), but our students often appeared unable to successfully complete the course. Interpreting this behaviour from perspectives such as the behaviour intention theory might shed light on student motives for enrolling and dropping off (Henderikx et al., 2017; Yang & Su, 2017). clusters 2 and 3 appear to be nuances of the disengaging learners (Kizilcec et al., 2013), with cluster 3 showing somewhat more ordered learning behaviour, and better learning results. Cluster 3 represents a serious endeavour and can be labelled “Venturers.” To decrease chances of failure, these students might benefit from guidance focused on personal learning goals (Conijn et al., 2018) and Self Regulated Learning (SRL) (Hew & Cheung, 2014), because their first quiz results suggest sufficient capacities to complete the course. Analyzing their progress through the stages of self-regulated learning could clarify how their goals were defined and might also shed light on how these changed according to their progress (Winne & Baker, 2013). Specific support in submitting quizzes regularly or early might improve their results.

Cluster 4 consists of students who showed high pass rates. They watched videos late, but steadily in batches of related videos and in course order. Towards the end of the course, less coherence was found in both watching and quiz submission. These are Madonado-Ahauad’s (2018) comprehensive learners, and Kizilcec’s (2013) completing learners; however, our students appeared to work increasingly disordered as time progressed. We labelled cluster 4 “Accomplishers” to reflect success with an effort. These students watched fewer videos in the final stage of the course, because they were at that time submitting and repeating quizzes, which might also be a cue for teacher guidance. Because they passed the course, these students showed the best self-regulating learning behaviour (Bannert et al., 2014).

Although cluster 4 has the largest number of passing students, teacher’s efforts might still be beneficial. Students in this cluster tried hard, yet could still improve their learning progress by focusing on less varied learning strategies (Vahdat et al., 2015). Furthermore, most rows in the dotted chart of cluster 4 end with purple dots, indicating activities in week 6. This suggests that most people made it to week 6 (irrespective of whether they obtained a certificate or not). This is also visible in the process models discovered for the four clusters.

Looking at learning behaviour within clusters and between clusters can inform teachers about locations for possible improvements in course materials and support for specific students. For instance, quiz results for clusters 2 and 3 in week 2 and 3 show remarkable differences. This can be compared with information about watching behaviour in those weeks, to find possible experienced difficulties in the materials. With added data such as whether in-video quizzes were passed on a first attempt, or whether videos were repeatedly watched, these analyses can be refined in future research.

Conclusion

Process mining, combined with traditional statistics, applied from a perspective of personal constructivism showed a fruitful approach to investigate learning behaviour and learning progress in MOOCs. It can be used to describe over time how student activities are ordered into patterns and to what results these sequences lead. This description in turn can inform teachers to improve course (re)design, and to support them in engaging students in the course. With an understanding of sequences of learning activities of groups of students, teachers can evaluate the content and the order of lectures and videos within a certain lecture. However, although analysing patterns can show where improvements in course materials are needed, pedagogical knowledge is necessary to indicate how to improve these materials.

This study knows limitations that also can serve as starting point for future research. First, our study was limited by distal data about video watching timings and order, and quiz submission and results. Additional proximal data about, for instance, personal learning goals

help to better understand the 4 clusters and underlying motivations. Personal learning goals could be measured with a pre-questionnaire (Henderikx et al., 2017) or during the MOOC, given that learning objectives might change over time. With personal learning goals specified, it could be determined how these goals influence the behaviour in MOOCs, which in turn can be used for more personalized improvements and student support.

Access to learning analytics data is usually restricted to teachers or management. With access to their own data, students can evaluate their learning process, which in turn supports self-directed learning and self-efficacy. This study suggests that LA can provide critical insights related to students' overall learning behaviour and its impact on performance.

Further research should focus on early detection of clusters of students based on their learning behaviour, combined with, for instance, personal learning goals. The purpose would be to replicate in other MOOCs the drop-off patterns found here, and to examine which effort is needed to keep cluster 2, and especially cluster 3 students on board. Furthermore, future research could also offer a better understanding of how students can be engaged in the course, with the purpose to increase a MOOC's success rate. Ideally, in educational research, these types of analysis are accompanied by controlling for background characteristics such as age, gender, level of knowledge, or motivation. This is also a consideration for future research for which process mining techniques provide a solid base.

Acknowledgements

This work has been supported by the European Data Science Academy (EDSA) project, and by EIT digital, the knowledge and innovation community of the European Institute of Innovation & Technology.

References

Bächtold, M. (2013). What do students "construct" according to constructivism in science education? *Research in Science Education*, *43*(6), 2477–2496. https://doi.org/10.1007/s11165-013-9369-7

Bali, M. (2014). MOOC pedagogy: Gleaning good practice from existing MOOCs. *MERLOT Journal of Online Learning and Teaching*, *10*(1), 44–56. Retrieved from http://jolt.merlot.org/vol10no1/bali_0314.pdf

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

Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(1). https://doi.org/10.1002/widm.1230

Brooks, C., & Thompson, C. (2017). Predictive modelling in teaching and learning. In C. Lang, G. Siemens, A. F. Wise, & D. Gasevic (Eds.), *Handbook of Learning Analytics* (pp. 61-68). DOI 10.18608/hla17

Bruner, J. (1996). *Toward a theory of instruction*. Cambridge: Harvard University Press.

Buckingham Shum, S., & Ferguson, R. (2012). Social learning analytics. *Educational Technology & Society*, *15*(3), 3–26. Retrieved from https://www.j-ets.net/ETS/journals/15_3/2.pdf

Chi, M. (2000). Self-explaining: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology*. Mahwah, NJ: Lawrence Erlbaum Associates.

Clow, D. (2013). MOOCs and the funnel of participation. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK '13* (p. 185). Leuven, Belgium: ACM Press. https://doi.org/10.1145/2460296.2460332

Conijn, R., Van den Beemt, A., & Cuijpers, P. (2018). Predicting student performance in a blended MOOC. *Journal of Computer Assisted Learning*, *34*(5), 615–628. https://doi.org/10.1111/jcal.12270

Emond, B., & Buffett, S. (2015). Analyzing student inquiry data using process discovery and sequence classification. In: *Proceedings of the 8th International Conference on Educational Data Mining* (p. 412-415). Madrid, Spain. Retrieved from http://www.educationaldatamining.org/EDM2015/proceedings/edm2015_proceedings.pdf

Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis*. London: Wiley.

Gillani, N., & Eynon, R. (2014). Communication patterns in massively open online courses. *Internet and Higher Education*, *23*, 18–26. https://doi.org/10.1016/j.iheduc.2014.05.004

Henderikx, M. A., Kreijns, K., & Kalz, M. (2017). Refining success and dropout in massive open online courses based on the intention–behavior gap. *Distance Education*, *38*(3), 353–368. https://doi.org/10.1080/01587919.2017.1369006

Hew, K. F., & Cheung, W. S. (2014). Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges. *Educational Research Review*, *12*, 45–58. https://doi.org/10.1016/j.edurev.2014.05.001

Kahan, T., Soffer, T., & Nachmias, R. (2017). Types of participant behavior in a massive open online course. *The International Review of Research in Open and Distributed Learning*, *18*(6). https://doi.org/10.19173/irrodl.v18i6.3087

Kennedy, G., Coffrin, C., & De Barba, P. (2015). Predicting success: How learners' prior knowledge, skills and activities predict MOOC performance. In *Proceedings of the Fifth International conference on Learning Analytics And Knowledge* (pp. 136–140). New York: ACM Press.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK '13* (p. 170). Leuven, Belgium: ACM Press. https://doi.org/10.1145/2460296.2460330

Koller, D., Ng, A., Chuong, D., & Zhenghao, C. (2013). Retention and intention in massive open online courses. *Educause Review*, (May/June), 62–63. Retrieved from https://er.educause.edu/~/media/files/article-downloads/erm1337.pdf

Lackner, E., Kopp, M., & Ebner, M. (2014, April 24 - 25). How to MOOC?–A pedagogical guideline for practitioners. In *Proceedings of the 10th International Scientific Conference eLearning and Software for Education,* (pp. 215–222). Bucharest: Editura Universitatii Nationale de Aparare "Carol I".

Lee, D., Watson, S. L., & Watson, W. R. (2019). Systematic literature review on self - regulated learning in massive open online courses. *Australasian Journal of Educational Technology*, *35*(1), 28–41. https://doi.org/10.14742/ajet.3749

Loyens, S. M. M., & Gijbels, D. (2008). Understanding the effects of constructivist learning environments: Introducing a multi-directional approach. *Instructional Science*, *36*(5–6), 351–357. https://doi.org/10.1007/s11251-008-9059-4

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., & Munoz-Gama, J. (2018). Mining theory-based patterns from big data: Identifying self-regulated learning strategies in massive open online courses. *Computers in Human Behavior*, *80*, 179–196. https://doi.org/10.1016/j.chb.2017.11.011

Margaryan, A., Bianco, M., & Littlejohn, A. (2015). Instructional quality of massive open online courses (MOOCs). *Computers and Education*, *80*, 77–83. https://doi.org/10.1016/j.compedu.2014.08.005

McAuley, A., Stewart, B., Siemens, G., & Cormier, D. (2010). The MOOC model for migital practice. Retrieved from http://davecormier.com/edblog/wp-content/uploads/MOOC_Final.pdf

Peña-Ayala, A. (Ed.). (2017). *Learning analytics: Fundaments, applications, and trends* (Vol. 94). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-52977-6

Peña-Ayala, A. (2018). Learning analytics: A glance of evolution, status, and trends according to a proposed taxonomy. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(3), 1–29. https://doi.org/10.1002/widm.1243

Reigeluth, C. M. (2016). Instructional theory and technology for the new paradigm of education. *Revista de Educación a Distancia*, (50), 1–18. http://dx.doi.org/10.6018/red/50/1b

Siemens, G., & Baker, R. S. J. d. (2012). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12* (pp. 252-254). Vancouver, British Columbia, Canada: ACM Press. https://doi.org/10.1145/2330601.2330661

Sinha, T., Jermann, P., Li, N., & Dillenbourg, P. (2014). Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3–14. Doha, Qatar: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-4102

Song, M., & Van der Aalst, W. M. P. (2007). Supporting process mining by showing events at a glance. In K. Chari & A. Kumar (Eds.), *Proceedings of 17th Annual Workshop on Information Technologies and Systems (WITS 2007*; pp. 139–145). Montreal.

Trcka, N., Pechenizkiy, M., & Van der Aalst, W. M. P. (2011). Process mining from educational data. In C. Romero, S. Ventura, M. Pechenizkiy, & R. Baker (Eds.), *Handbook of educational data mining* (pp. 123–142). Boca Raton: CRC Press.

Vahdat, M., Oneto, L., Anguita, D., Funk, M., & Rauterberg, M. (2015). A learning analytics approach to correlate the academic achievements of students with interaction data from an educational simulator. In G. Conole, T. Klobučar, C. Rensing, J. Konert, & E. Lavoué (Eds.), *Design for teaching and learning in a networked world* (pp. 352–366). Cham: Springer International Publishing.

Van der Aalst, W., Adriansyah, A., & Van Dongen, B. (2012). Replaying history on process models for conformance checking and performance analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(2), 182–192. https://doi.org/10.1002/widm.1045

Van der Aalst, W. M. P. (2016). *Process Mmining: Data science in action*. Heidelberg: Springer.

Veletsianos, G., & Shepherdson, P. (2016). A systematic analysis and synthesis of the empirical MOOC literature published in 2013–2015. *The International Review of Research in Open and Distributed Learning*, *17*(2). https://doi.org/10.19173/irrodl.v17i2.2448

Watson, S. L., Loizzo, J., Watson, W. R., Mueller, C., Lim, J., & Ertmer, P. A. (2016). Instructional design, facilitation, and perceived learning outcomes: An exploratory case study of a human trafficking MOOC for attitudinal change. *Educational Technology Research and Development*, *64*(6), 1273–1300. https://doi.org/10.1007/s11423-016-9457-2

Wen, M., & Rose, C. P. (2014). Identifying latent study habits by mining learner behavior patterns in massive open online courses. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14* (pp. 1983–1986). Shanghai, China: ACM Press. https://doi.org/10.1145/2661829.2662033

Winne, P. H., & Baker, R. S. J. d. (2013). The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *JEDM | Journal of Educational Data Mining*, *5*(1), 1–8. Retrieved from: Retrieved from https://jedm.educationaldatamining.org/index.php/JEDM/article/view/28

Winne, P. & Hadwin, A. (1998). Studying as self-regulated learning. In D.J. Hacker, J.E. Dunlosky, & A.C. Graesser (Eds.), *Metacognition in educational theory and practice*. (pp. 277-304). Mahwah: Lawrence Erlbaum Associates.

Yang, H.-H., & Su, C.-H. (2017). Learner behaviour in a MOOC practice-oriented course: in empirical study integrating TAM and TPB. *The International Review of Research in Open and Distributed Learning*, *18*(5). https://doi.org/10.19173/irrodl.v18i5.2991