Analysis of Students' Learning Behaviour in Massive Open Online Courses

Presented to :

Roland Tormey

Teaching Advisor

Teaching Support Center

Swiss Federal Institute of Technology

Lausanne

Prepared by:

Alexandru Mocanu,

Timothy Trovatelli,

Valentin Borgeaud,

Zoé Baraschi

Swiss Federal Institute of Technology

Lausanne

31 May, 2019

Executive Summary

Purpose and method of this report

Massive Open Online Courses were created to give access to learning material to an unlimited number of participants through the web. They have become increasingly present at EPFL and the MOOC platform enables us to collect and analyse data about the behaviour of its users. The purposes of this report are to examine the actions surrounding assignment submissions from students in order to:

- verify whether a link between conscientiousness and attainment can be observed.
- extract potential learning patterns such as repetition and strategic approach to learning in the stream of actions of the users.
- find potential collaborations between users taking the same class.

To do so, we'll perform data analysis of the Functional Programming MOOC (2012), which had over 117'000 participants.

Findings and conclusions

From our study, we found that:

- there is a correlation between attainment and conscientiousness.
- the behavior of passing and failing users does not seem to differ significantly.
- the data needed to conclude on the former and on collaboration is insufficient.

Recommendations for further analysis

The data we have studied is pretty old (2012) and is incomplete. Having the content of each submission would be a great help to study collaboration. More details about the user actions in videos would help us extract more meaningful learning patterns. Collecting information about the user's motivations when taking a course would be of useful to better split the users into populations of interest.

Table of Contents

1. Introduction

As the use of technology has become increasingly significant in today's society, the traditional system of education and learning has likewise shifted towards a strong online presence. Through Massive Online Open Courses, or MOOCs, students now receive the opportunity to visualise their lectures, submit their assignments and ask questions to their professors round-the-clock. MOOCs have also enabled an unlimited number of people to sign up and attend free classes based across the globe and thus satisfy their thirst for knowledge.

The impact of these platforms on the research of learning styles and techniques should not be underestimated. With the help of click-stream data, researchers now have the opportunity to record and study every action that students take during their time spent on a learning platform.
Hot topics such as attainment, motivation, performance, regularity and self-regulated learning are the center of numerous past and ongoing studies on the influence of MOOCs on learning.

1.1. MOOCs at EPFL

The Swiss Federal Institute of Technology in Lausanne is a pioneer in MOOCs and is increasingly using this system to teach classes, particularly in classes with a large number of students. Their specialised MOOCs have an international outreach of over 2 million participants [1], and the data recorded from these MOOCs is used and analysed by researchers at the university, with the hope to further understand and improve this platform's influence on learning.

1.2. Purpose of this study

When taking a MOOC, students usually have to submit assignments in order to pass the course. Before the deadline, they have the opportunity to submit the assignment a few times and are able to receive feedback on what they have handed in so far.

Our first goal of the study would be to verify whether the MOOC's data suggests that there is a link between conscientiousness and attainment, as it is usually the case [2], and to quantify how strong this link is. We then would like to investigate whether the actions taken between submissions influence the students' attainment. Finally, we would like to investigate whether identifying collaboration between students is observable by analysing their submission times and locations.

In the class "How People Learn", we have seen that one of the five principles of learning specifies that students learn when they have the opportunity to try, get feedback on errors, and learn again [3]. Thus, by studying students' submissions and their actions in between many attempts for the same assignment (similar to this work [4]), we hope to extract interesting actions sequences suggesting that students are rewatching videos and/or that students are cherry picking video chunks of interest as part of a strategic approach to learning [5]

In regard to collaboration, there are articles discussing about tools provided by MOOCs for collaboration. In [6], they analyse collaborative learning on their MOOC platform by looking at how the tools they provide for collaboration impact the users. They also encourage using group assignments to develop the collaboration skills of the users. In [7], they concentrate on classifying and identifying the role of various collaboration tools in MOOCs in general. They however do not provide statistics to illustrate the impact these tools have, so this study remains at a more descriptive level rather than analytical ones. Our study, on the other hand, has the objective of identifying patterns of collaboration between users by analysing their submission behaviours. Therefore, we are not looking at collaboration tools, like the other articles do, but at how people try collaborating in assignments that would normally be solved individually. This may link to several learning concepts, as we are going to discuss in the “Findings and Discussion” section.

The Center for Digital Education lab - or CEDE - at EPFL has authorized this study and has given us the material necessary to conduct it. We have met with them multiple times to discuss which topics would be of interest to them, as well as to get access to the data to perform the study.
We will report our findings back to them and hopefully give them information on how they can improve the platform to increase student attainment, as well as which information to track in order to have more precise results.

1.3. Scope of this study

This study investigates:

- correlation of conscientiousness of a student with his/her overall attainment.
- learning patterns through user actions sequences and their link to attainment.
- collaborations of users in submissions

We consulted the click-stream data from the Functional Programming MOOC in order to obtain information on participant’s actions during the course. We focused on the data from the course given in 2012. Out of all the participants, we analysed the ones with at least one submission. This gave us a sample size of over 14000 participants.
Note that this course was open to students outside of the university and also the grades of the EPFL students are not available. This means that when analysing attainment we can only refer to the grades that the users obtained in the MOOC and we can not link EPFL students’ grades to our study.

1.4. Sources and methods

Since the data was already available to us, we did not need to design a specific questionnaire or recruit participants. The participants of the study were thus the students taking the online MOOC, who had agreed to the Terms and Conditions [8] of the website Coursera before beginning to use the platform, which indicates that participating institutions of the platform are authorized to use the User Content of the participants taking the classes they provide.

2. Conclusion

Based on the findings of our study, we draw the following conclusions:

- more conscientious users tends to have higher grades than less conscientious ones.
- based on our data, learning patterns such as repetition and strategic approach to learning does not seem to influence the chances of a student passing the course.
- collaboration between users is difficult to study without having access to the submissions content.

3. Recommendations

Here are some recommendations for further analysis:

- collect information about the user’s motivations when taking a class so that we can split the population of users based on their interests and goals, in order to analyze the data more accurately
- save more information about the different kinds of video actions that users are performing, such as rewind and forward jumps

- save the content of the submissions in order to better analyse collaborative behavior between students

4. Findings and discussion

Our research has focused on answering the following questions:

- is there a link between conscientiousness (measured through the lateness of submissions) and attainment?
- is it possible to extract learning patterns using the users' actions done in between submissions and is there a link between the actions performed and attainment?
- do people taking the Functional Programming MOOC collaborate for assignments submissions?

For each of these questions, we explore our findings in detail and discuss the assumptions made in the analysis, implications and links to concepts related to learning. The methodology used for data exploration and for attaining the results discussed are exposed into much more depth in the appendices.

4.1. Lateness and attainment

The functional programming MOOC requires students to weekly submit a programming homework that is graded automatically. Every homework has a specific open date, soft deadline and hard deadline. Student can submit an unlimited number of time their homework and will not be penalised if they submit after the soft deadline. This homework is at the heart of the learning process in the MOOC since it is the only way for students to get feedbacks on their performance and understanding of the material which is very important in the learning process [9].

In order to investigate the relationship between lateness and attainment, we made the assumption that the time a student submits his or her homework before the deadline is a proxy to his or her conscientiousness. It has been proven that conscientiousness is the factor most often found to be associated with attainment [5]. We thus assumed that highly conscientious students would first attempt a submission shortly after its release and finish it well before the deadline.

For this part of the study, we had a sample of n=2313 participants, consisting of the students having completed the entire set of assignments in the course.

We first created a metric representing how late on average a student submitted his or her homework with respect to the soft deadline[1]. We then linked the score to the student's final grade for the course and computed the correlation. The spearman coefficient is -0.59. This mean there is a negative correlation between the lateness and the overall grade of a student. It suggests that the earlier a student is in average in submitting his reports, the higher his grade is. Using our first hypothesis of lateness being a proxy of conscientiousness, we can confirm that there is a link between submission time and attainment in this MOOC.

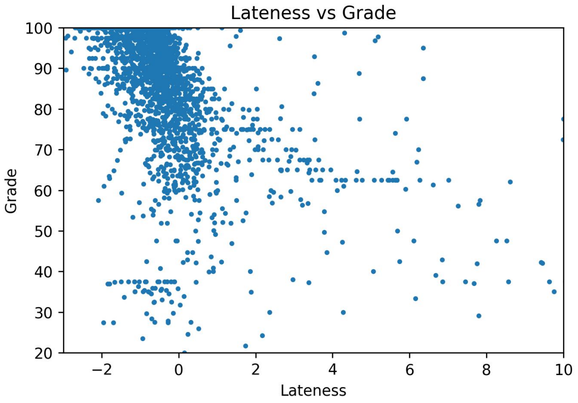

Figure 1: Correlation of Lateness vs Grade

[1] See Appendix - Lateness and attainment

We then had a look at the high achievers of the course, who finished the course with a final grade superior or equal to 90%. We compared the lateness of the latter group with the lateness of passing students having grades between 60% and 90% to see if there are any noticeable differences in the performance of the two groups. We saw that the subset of students with a grade ≥ 90% were more likely to submit their assignments in advance than the rest of the class. The group with a grade lower than 90% also contained almost all participants who handed in their assignments later than the soft deadline. This further proves the link between early assignment submission time and performance.

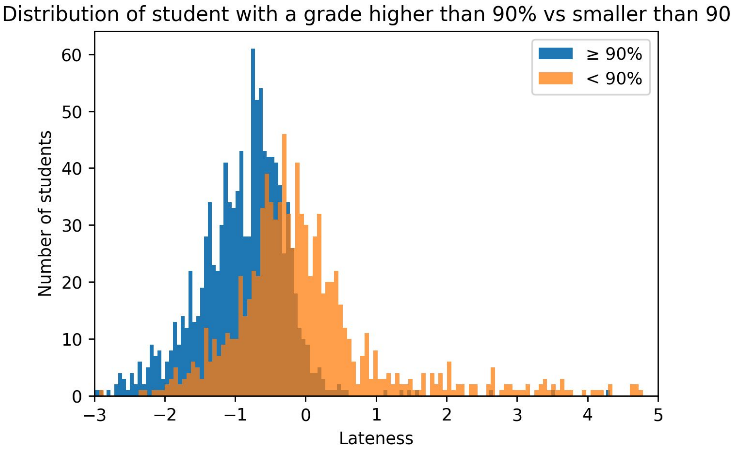

Figure 2: Histograms of Lateness of the two groups

4.2. Actions sequences and attainment

When students use the MOOC platform, all of their actions are recorded in the form of click-stream data. It was important to investigate what exactly the participants were up to between two separate submissions of the same assignment.

In particular, we wanted to find out whether students would go back and watch some video segments in between their submissions, in order to improve their assignment grade with respect to the feedback they received in the previous attempt. We hoped that by examining these actions, we could somehow link them to attainment.

In order to do this analysis, we analyzed the sequences of actions of users in between two submissions for the same problem (assignments and video quizzes) and this for all the problems present in the MOOC. More precisely, we were interested in analyzing the frequency of some specific actions sequences, also called chunks, of users in between submissions. You can find below the bar chart of the most frequent such chunks of 3 consecutive actions.

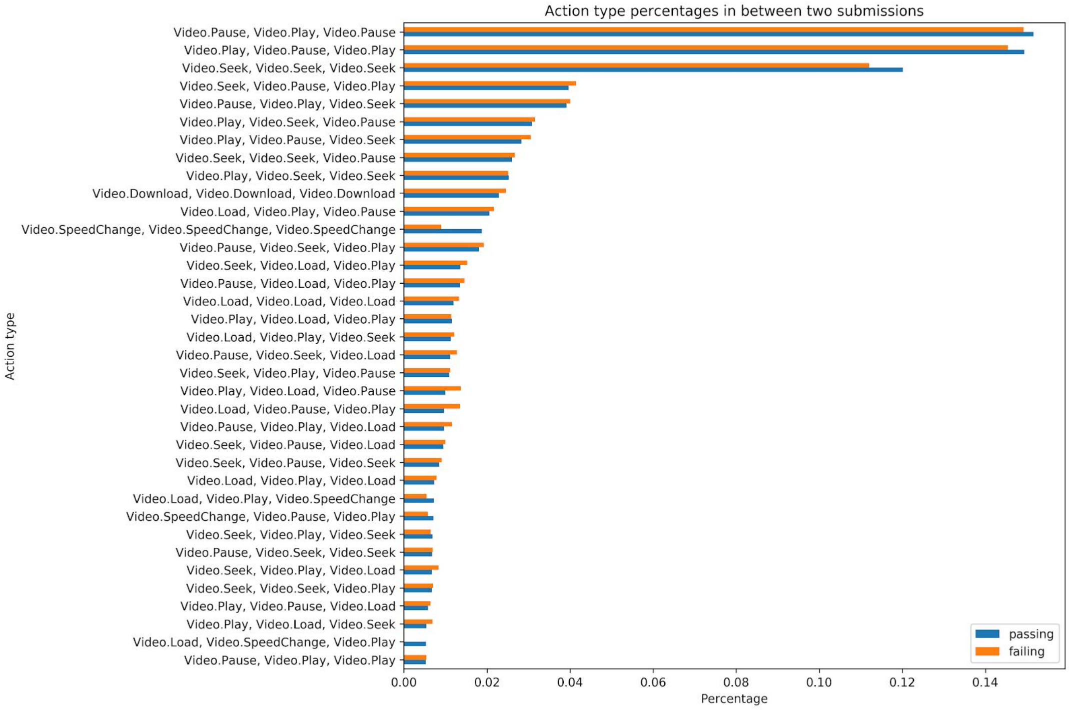

Figure 3: Top action chunks between submissions

We have differentiated the chunks for actions of passing (blue) and failing (red) students. We can see that the actions of both categories of students are almost equal in quantity. This means that according to video events, both failing and passing users adopt a similar behavior when watching videos in between submissions and therefore video events cannot explain why a user passes or fails the class. The graph also highlights some important chunks of actions that we can try to bind to multiple learning concepts:

- [PAUSE, PLAY, PAUSE] / [PLAY, PAUSE, PLAY]: Those chunks are the two most present ones in the sequences of actions. We have two hypotheses explaining why. The first one is because people are sometimes interrupted and need to pause the video consequently. The second hypothesis is that users are taking notes about the video they are watching, which is great to enhance learning (even greater if they are taking manuscript notes [10])
- Pretty much all chunks containing video seeks ([SEEK]) can be assimilated as being action chunks representing either repetition where the user goes back to watch again some content or either strategic approach to learning where the user only focus his resources on the material he doesn't know and has never seen before for example [5].

After doing this part of the study, we decided to compare it to all of the action sequences of all users. Similarly to the previous step, each action is represented by a chunk of events of size 2 or 3. Note that in this step, we are solely interested in the entire set of actions of the users and not their actions in between two submissions of the same assignment. We have plotted a bar chart of the chunks, sorted by their number of occurrences, for both passing and failing students.

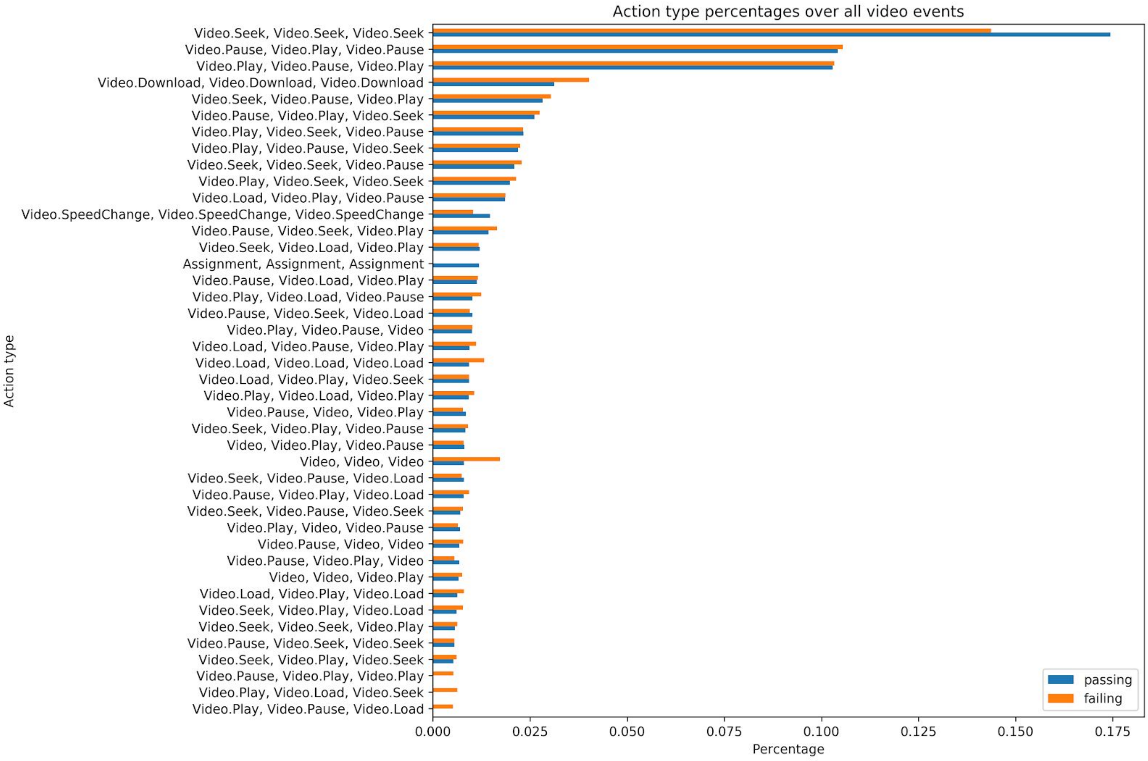

Figure 4: Overall top action chunks

We notice that the dominant chunks of actions done by users are almost the same as in the previous study. We observe that the top chunk of actions is now [SEEK, SEEK, SEEK] instead of [PAUSE, PLAY, PAUSE] in the previous study, that is now in second place. Users passing the course tend to have more chunks of this type than failing users. We don't know for sure why it is the case, but we have two hypotheses:

- The first one is that, as most of the MOOC have a completion rate below 13% [11], the users not passing the class tend to watch less videos and therefore the proportion of chunks [SEEK, SEEK, SEEK] is lower than for passing students.
- The second hypothesis we have is that passing users have a bigger proportion of [SEEK, SEEK, SEEK] chunks because they tend to rewatch

more videos or watch videos more strategically by jumping to parts of the videos in which they have more interest than failing users[5]

If the last hypothesis was verified, this would show that there is a difference in the behavior that failing and passing users adopt when watching videos and that users adopting repetition or strategic approach to learning strategies would have higher chances of passing the class. However, we do not have enough information to verify this claim.

For other top chunks such as [PAUSE, PLAY, PAUSE] / [PLAY, PAUSE, PLAY] and top chunks containing SEEK actions among PLAY and/or PAUSE actions, the same reasoning and explanation as given for the analysis of actions sequences between submissions can be given: The chunks [PAUSE, PLAY, PAUSE] and [PLAY, PAUSE PLAY] can either be assimilated as a user being interrupted in its watch time for some real life reasons or as a user taking notes about the videos; the chunks containing SEEK actions among PLAY and/or PAUSE actions can be assimilated as chunks representing either that the user is rewatching the videos for repetition or not well understood part or that the user is doing forward jumps in the videos to focus only on the parts of the video that he is interested in or on parts that he has never heard of. These chunks can be therefore assimilated to strategic approach to learning.

4.3. Finding collaborations

We associate collaborations with a strategic approach to learning, learning in groups with peers and self-efficacy beliefs. Solving assignments from scratch, especially when having a small amount of prior knowledge, can be very time consuming. Collaboration can therefore be seen as a strategic way of learning, as getting advice and explanations from someone who already solved or who is very advanced in solving the assignment can be very useful in understanding the material and in spending less time on it. Learning in groups with peers implies collaborating with peers, therefore the collaborations may hint towards such a way of learning. Another motivation for the collaboration between users is that some may have weak self-efficacy beliefs, thus seeking the help of others in solving the assignments or simply confirming that their results and approach are appropriate.

When looking for collaborations, we assume that people are more prone to collaborate if they are close in location, so we group the data by cities. Within these clusters, we assume that a collaboration takes place if two people have many submissions that are close to each other in time.

Following the study, we have concluded that the signs of collaboration in the Functional Programming MOOC are pretty weak in general, with a total of 146 collaborations among 44755 students. If we concentrate only on the Lausanne area, we get a more convincing result, with 39 collaborations among 219 students. The highest submission overlap count is also obtained for the Lausanne area, with a value of 71, which indicates a pretty clear collaboration. This may be due to the users actually being EPFL students taking this course in their study programme, so that they are more motivated to get good grades and pass the class.

The results of the collaboration study may have been more relevant if the contents of the submissions were available as well, so as to confirm the similarity between the submissions. Using only the times of the submissions, we can very well get noise due to coincidental overlaps.

5. References

[1] S. Ch, “EPFL online courses attract more than 2 million students,” *SWI swissinfo.ch*, 14-Feb-2018. [Online]. Available: https://www.swissinfo.ch/eng/society/e-learning_epfl-online-courses-attract-more-than-million-students/43899142. [Accessed: 31-May-2019].

[2] S. A. Woods, F. C. Patterson, A. Koczwara, and J. A. Sofat, “The value of being a conscientious learner,” *Journal of Workplace Learning*, vol. 28, no. 7. pp. 424–434, 2016.

[3] R. Tormey, “Evidence on teaching and learning,” *How People Learn course*. Sep-2018.

[4] J. Maldonado-Mahauad, M. Pérez-Sanagustín, R. F. Kizilcec, N. Morales, and J. Munoz-Gama, “Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses,” *Computers in Human Behavior*, vol. 80. pp. 179–196, 2018.

[5] R. Tormey, “Notes on Personality and Learning Styles,” *How People Learn course”*, Sep. 2018.

[6] T. Staubitz, T. Pfeiffer, J. Renz, C. Willems, and C. Meinel, “Collaborative learning in a MOOC environment,” in *Proceedings of the 8th annual international conference of education, research and innovation*, 2015, pp. 8237–8246.

[7] “AN INSIGHT TO COLLABORATION IN MOOC,” *International Journal of Advance Engineering and Research Development*, vol. 4, no. 07. 2017.

[8] “Terms,” *Coursera*. [Online]. Available: https://www.coursera.org/about/terms/revisions. [Accessed: 31-May-2019].

[9] “SAGE Journals: Your gateway to world-class research journals,” *SAGE Journals*. [Online]. Available: https://journals.sagepub.com/doi/abs/10.3102/003465430298487. [Accessed: 31-May-2019].

[10] R. Tormey, *Lecture 1 Video A*. ,”*How People Learn course*”, Sep. 2018.

[11] D.F.O.Onah, J.Sinclair, R.Boyatt, “DROPOUT RATES OF MASSIVE OPEN ONLINE COURSES: BEHAVIOURAL PATTERNS,” The University of Warwick (UNITED KINGDOM).

[12] MetaHG, “MetaHG/SHS-HPL,” *GitHub*. [Online]. Available: https://github.com/MetaHG/SHS-HPL. [Accessed: 31-May-2019].

6. Appendices

6.1. Study Participants

The study participants are people taking the Functional Programming MOOC offered by EPFL on Coursera. When taking the course, they have agreed that their interaction with the platform could be monitored and used for analysis studies such as the current one.

6.2. Materials Used

The materials used consist of an SQL database of grades, locations and actions of the users that took part in the MOOC.

6.3. What participants were asked to do

The participants were not asked to do anything specific. They voluntarily involved in studying the content of the MOOC and solving various assignments and quizzes in the process.

6.4. Design of the Study

6.4.1. Finding collaborations

6.4.1.1. Setup and assumptions

In order to determine whether people collaborate in the MOOC assignments, we first made the assumption that collaborations would take place between people that are closeby geographically and whose submissions are close to each other in time. Therefore, we grouped submissions by city and analysed collaborations within these clusters. We considered that two people collaborated if they had some number of submissions, greater than a threshold, in the same time windows. We used time windows of 1 hour, 30 minutes, 10 minutes and 5 minutes.

6.4.1.2. Precautions and observations

To account for the possibility of an overlap in submissions due to the deadline being close, we considered submissions for a given time before the deadline, namely 12 hours before the soft deadline. To also take care of the case of coincidental submission overlaps, we considered that two users collaborate if they have a number of overlaps above a given threshold, namely 15.

The distributions of submission overlap counts is the same when considering submissions for some time before the hard deadline or when considering all the submission overlaps. Therefore, we consider only submissions for some time before the soft deadline as having potential signs of collaboration.

6.4.1.3. Results

For a time window of one hour, we detected a number of 146 collaborations. Figure 5 shows a histogram for the maximum number of collaborations in each city and a plot of the distribution of submission overlaps. We selected the threshold based on this distribution of submission overlaps plot.

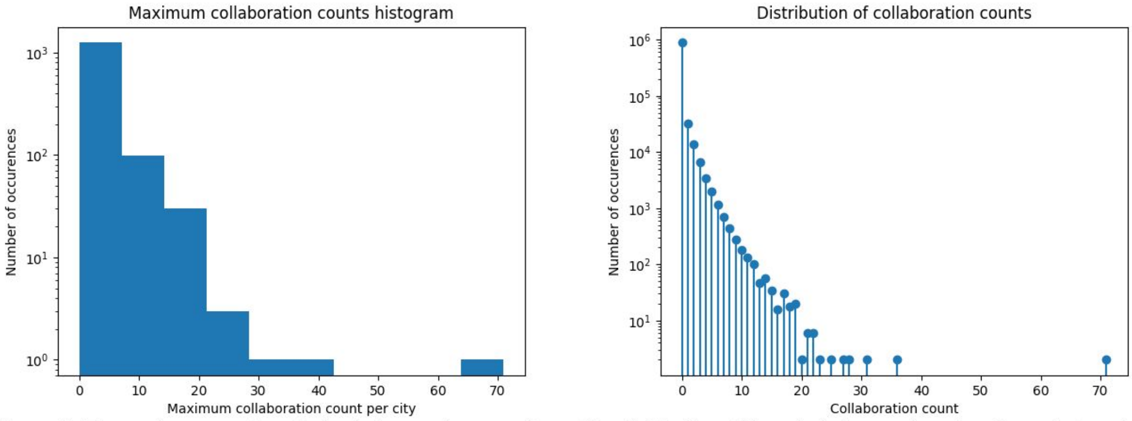

Figure 5: The maximum number of submission overlaps per city and the distribution of the submission overlaps for a time windows of one hour

For a time window of 30 minutes, we get only 36 collaborations.

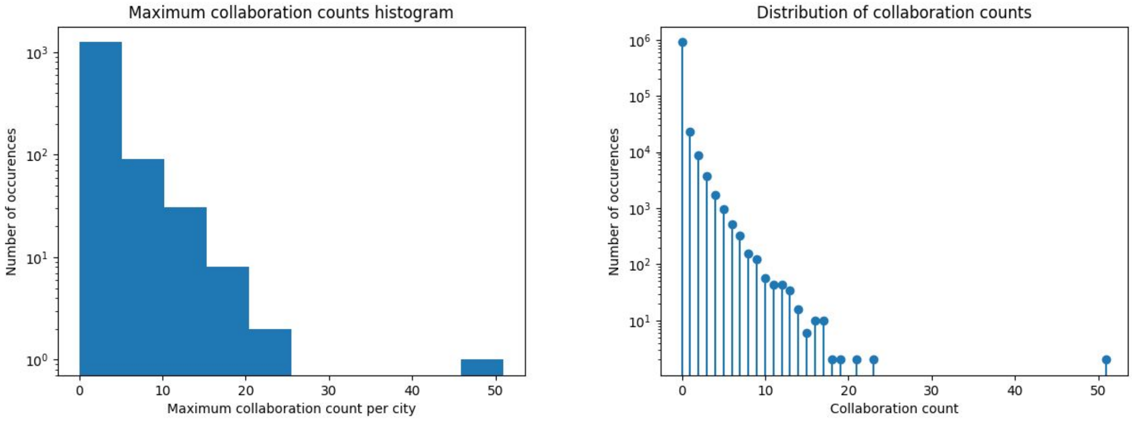

Figure 6: The maximum number of submissions overlaps per city and the distribution of the submission overlaps for a time window of 30 minutes

For 10 and 5 minutes, we drop to only 2 collaborations. The total number of users is 44755. Considering that two people who collaborate may not necessarily make submissions in a very small time window, we consider that the one hour time window is the most appropriate one.

In Lausanne, in particular, we detect a number of 34 collaborations, for one hour time windows, out of 219 students, which is a pretty high number, especially compared to the average collaboration count. This could be motivated by the fact that there may be EPFL students who actually took the course in their programme of study and would be more motivated to get a good grade than other MOOC users.

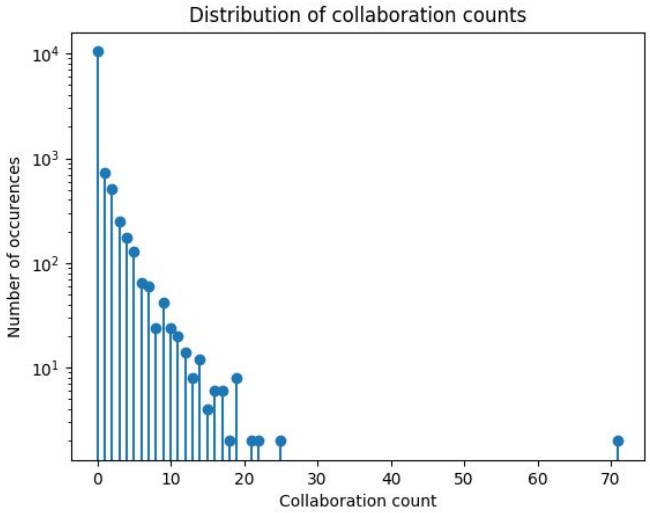

Figure 7: The submission overlap distribution for users in Lausanne for one hour time window

6.4.1.4. Remarks

It is hard to say when we have a collaboration and when not based only on submission times and not on the content of those submissions as well. Considering the high number of users and the small number of collaborations detected, even if collaborations take place, they are actually some outliers.

6.4.2. Lateness and attainment

The goal of this part was to find out if there is a link between submissions' earliness and attainment. Each student has to submit homework periodically during the MOOC. These assignments are mandatory and a student need an average ≥ 60% to pass the class. We made the assumption that how early on average a student is in his/her submission is a proxy to his/her conscientiousness. We thought that a conscientious student would be more prone to begin an assignment as soon as it is released and finish it early. We'll now describe how we crafted the lateness metrics and our methodology.

6.4.2.1. The data

First it is important to know the data that we're are treating. For each homework, there is an "Open time", "Soft deadline" and a "Hard Deadline".
The open time is the date at which the student can start to send submit submissions to the grading system. Soft deadline is the official shown deadline on the MOOC, after this deadline, students are actually still able to submit their homework for 97h (~4 days). It appears that if they submit in this interval of time they don't have any score penalty.

We have at the beginning a total of 234'171 submissions. After cleaning the data by deleting degenerated submissions (no grade, submitted before open time) we end up with 195'171 samples. We also chose to delete the first assignment which was optional and only for the user to understand how the submission system works. We're now left with 174'507 samples.

6.4.2.2. Lateness and course grade

Lateness metrics

We'll now describe how we found our results regarding the link between the lateness of a user and its course grade. In this part we only consider people having handled in all the assignments. This leaves 2313 users.
We now want to study if there is a correlation between the student final grade and his lateness in submitting the assignments.
We first define the average submissions timestamp per user per homework:

$$A(u,h) = \frac{1}{n} \sum_{s \in sub(u,h)} s_{ts}$$

Where u is a user and h is a specific homework. $sub(u,h)$ return all the submissions of a user for a specific homework. s_{ts} is the timestamp of a submission and $n = |sub(u,h)|$

We then define the normalized lateness metric by homework and by user. This metric represents how late a user is in his submission for a specific homework normalized by the length of time submission window.

$$L_{norm}(u,\ h) = \frac{A(u.h) - h_{softclose\ time}}{h_{hardclose\ time} - h_{opentime}}$$

We now only need a lateness score per user over all assignments. Let us define the lateness normalized per user as a simple sum over the lateness of all homework:

$$\tilde{L}_{\ norm}(u) = \sum_{h \in H} L_{norm(u,h)}$$

Where H is the set of all homework.

Linking lateness with grades

We now just have to merge the course grade of each user with its lateness. Once we do that we can try to look for a correlation. We removed user that have an overall grade of 0 since it is either inactive user or student that submitted random things to assignments. The Spearman correlation between the Grade and the Lateness normalized is -0.59. The p-value of the correlation is < 0.001. The null hypothesis is having a uncorrelated system.

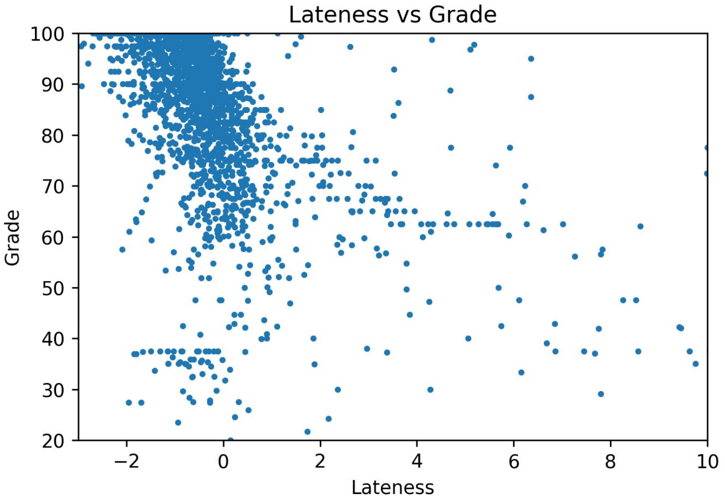

Figure 8: Correlation between Lateness and Grade

We can indeed see on this scatter plot the negative correlation between Lateness and grades. It suggests that indeed the earliness in handing in the assignments and the grade is correlated. With our hypothesis of lateness being a proxy of conscientiousness, we can confirm that there is a link between conscientiousness and performance.

6.4.2.3. Lateness of different group

Now let's have a closer look at the people that performed very well with an overall grade of 90%. We want to see if there is a significant difference between the conscientiousness of student with a grade higher than 90% (1111 students) and with student with a grade lower than 90% but that passed the class (946 students). To do so, we split the data in two sets and compare their statistics. Using a t-test we can assert that the mean lateness of the two populations is not the same (p-value < 0.001).

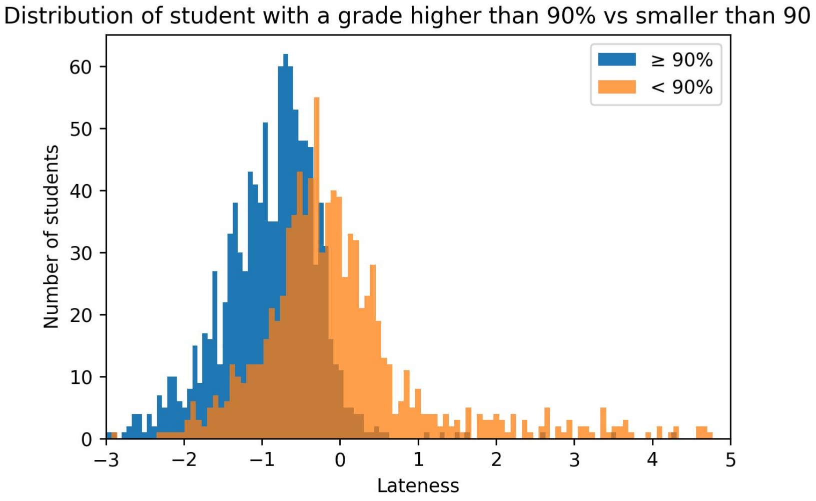

Figure 9: Histograms of lateness of two groups

We can clearly see on the graph that the two means are different. Student with a grade ≥ 90% submit have a Lateness metric of -0.86 while the other group has a Lateness metric of 0.17. The < 90% group submit later in average and this group contains most of the users that are very late; almost all of the user with a lateness metric bigger than 0.5 are in the < 90% group.

This tells us again that conscientiousness is very well correlated with student's performance.

6.4.3. Actions sequences and attainment

In this section, we describe the methodology used to extract the actions sequences between two submissions for the same homework for a user. We also describe how we measured the similarity or differences of actions between succeeding and failing users. A brief description of the data available is also presented.

6.4.3.1. Available data

To study action sequences, we need to extract the time intervals between two submissions for the same homework. For this, we use the data available in a table containing problem events information such as the problem ID, the problem type, the timestamp of the event and the user ID. We also need to extract the actions done by the users during those intervals. For this we use a table containing video events information such as the event type which is one of the following [PLAY, SEEK, PAUSE, SPEED_CHANGE, LOAD, DOWNLOAD, ERROR, STALLED], the timestamp of the event and the user ID.

6.4.3.2. Extracting users' action sequences

First, we extract the intervals of time that we want to study for each user, i.e. the intervals of time which correspond to intervals in between two submissions for each homework. Then, for each user, we find all the video actions that he did during each interval. Note that intervals and therefore actions as well are ordered in increasing order. Next, we filter all the students having less such intervals than some threshold for all homework taken together and we only keep the ones having more. We explain below how this threshold was chosen. Finally, we filter all students that don't have any video events during any of these intervals and only keep the ones that have at least one video event in any of these intervals. The students that are removed because of the last filtering phase correspond to students who submitted their solution to their homework multiple times but didn't do any apparent video actions in between.

The selected threshold for the first filtering step is 25 (intervals) and was chosen based on the computation resources we had available. Indeed, by choosing 25 as a threshold, the number of users resulting from the filtering is 7798 which is not much. However, the number of video events they represent is around 5.7 millions, which is more than half the total number of video events for this MOOC. Also, there are 7 assignments and 21 quizzes in the MOOC for a total of 28 different problems. By choosing 25 as a threshold, the selected users will ever have done around two submissions per problem or a lot more than two submissions for some problems and only one or zero for others. In any case, this sample of user should be interesting to study as they probably rewatched videos between two submissions for the same homework.

The overall resulting population from this extraction is a population of users having done at least 26 submissions and therefore at least 25 submissions (worst case is having done 26 submissions for the same homework which results in 25 intervals of potential interest) with at least one video event in between one of these intervals. The user population size after the last filtering step is 1669. Among these users, 1285 (77%) successfully passed the course and 384 (23%) failed the course. From those numbers, we could think that users having done a lot of submission overall and having rewatched or strategically watch the videos in between two submissions tend to finish and pass the course. However, this hypothesis is quite far-fetched. Indeed, it could just be that among the total population of failing users of the dataset, only a few of them tend to do 26 or more submissions. This explanation seems more reasonable as other studies have shown that the completion for most MOOC is below 13% [11]. To verify this potential explanation, we looked at how many different problems both succeeding and failing users submitted a solution. For passing students the median is 13 problems. For failing students, the median is 8 problems. Now if we only look at assignments, which is a subset of all problems, the median is 5 for succeeding students and 3 for failing students. Figure 10 shows the distribution of users having done some specific number of assignments. From this, we can clearly see that failing students tend to participate less to problems and assignments than succeeding students, most probably because they stop to follow the course after a few weeks. Sadly, the data does not allow us to identify which problems belong to which week in the course and we therefore cannot be more precise and confident in our explanation.

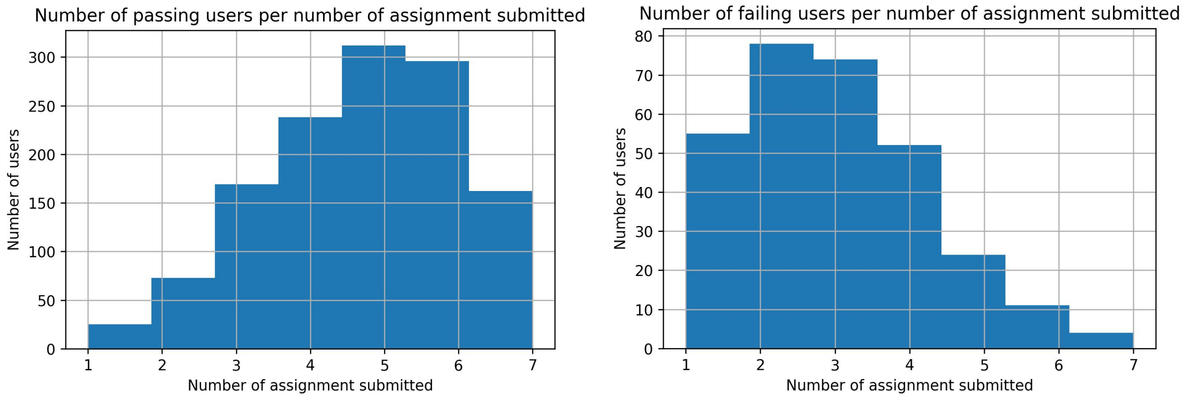

Figure 10: Number of users per number of assignment submitted

6.4.3.3. Building action chunks

Now that we have extracted the users intervals with their corresponding actions, we need to analyze the sequences of actions. To do this, sequences need to be split into multiple (overlapping) chunks of identical size. The chunk size is arbitrary, although if we want to be able to identify potential repetition patterns or strategic approach to learning patterns, it needs to be at least of size 2 or 3. This way, we can plot a barchart to find out the overall count of each chunk type (see figure 11). Nonetheless, this only gives us an idea about what kind of video actions sequences are the most present. The data we were given have some fields describing the old time and new time of the video after each video action made by the user. However, those two attributes are always equal in our dataset and a significant amount of them are null. Therefore those two attributes do not bring us any information to distinguish whether a user is doing a rewind of the video or if he is skipping some part of the video. Consequently, when analyzing chunks such as [SEEK, PLAY] or [PLAY, SEEK, PLAY], we are not able to distinguish between the user going forward or backward. Having this information would have helped us distinguish a repetition pattern from a strategic approach to learning pattern where the user only looks for what he is the most unclear about.

6.4.3.4. Results

Action sequences between submissions

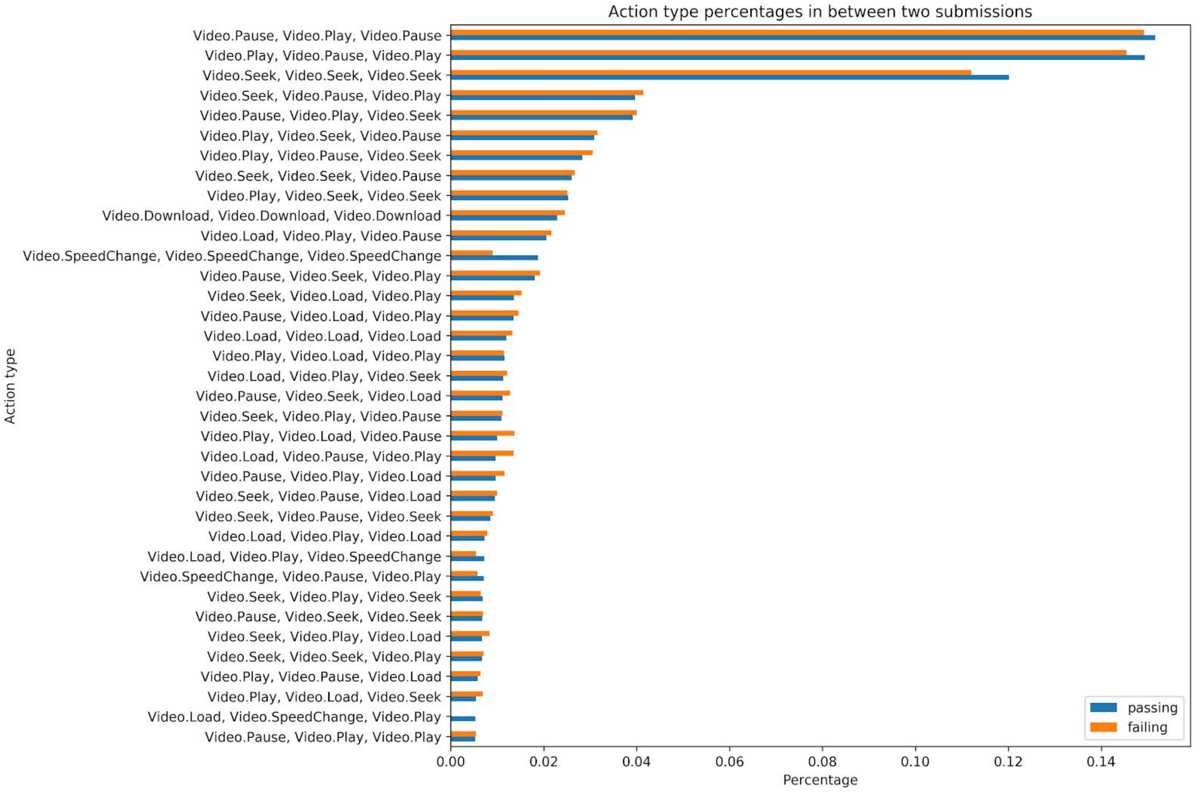

Figure 11: Top action chunks between submissions

One interesting thing we can notice on the graph (figure 11) though is that the kind of action chunks that failing users and passing users (that have at least 26 submissions as a reminder) are almost identical in quantity. This means that according to video events, both failing and passing users adopt a similar behavior when watching videos and therefore video events cannot explain why a user fails or passes the class.

The graph also highlights some important chunks of actions that we can try to bind to learning concepts:

- [PAUSE, PLAY, PAUSE] / [PLAY, PAUSE, PLAY]: Those chunks are the two most present ones in the sequences of actions. We have two hypotheses explaining why. The first one is because people are sometimes interrupted and need to pause the video consequently. The second hypothesis is that users are taking notes about the video they are watching, which is great to enhance learning (even greater if they are taking manuscript notes [10])
- Pretty much all chunks containing video seeks can be assimilated as being action chunks representing either repetition where the user goes back to watch again some content or either strategic approach to learning where the user only focus his resources on the material he doesn't know and has never seen before for example[5], [10] .

General action sequences

We also studied the action sequences as a whole for each user. We used the same methodology that was used to study the actions in specific intervals, that is actions sequences are splitted into chunks of size 2 or 3 and we then count of number of occurrences of each chunk type. However, note that this time we don't care about intervals and just analyze all the user actions in time. The graph below (figure 12) presents the comparison between failing and passing students actions. We notice that the dominant chunks of actions done by users are almost the same as when studying chunks of actions in submission intervals. We still observe that the top chunk of actions is now [SEEK, SEEK, SEEK] instead of [PAUSE, PLAY, PAUSE] as seen previously that is now ranked in the second place. Users passing the course to have to have more chunks of this type than failing users. We don't know for sure why it is the case, but we have two hypotheses. The first one is that the dropout rate is usually very high for MOOCs [11], therefore failing students tend to watch less videos and consequently the proportion of chunks [SEEK, SEEK, SEEK] is lower than for passing students. The second hypothesis we have is that passing users have a bigger proportion of [SEEK, SEEK, SEEK] chunks because they tend to rewatch more videos or watch videos more strategically by jumping to parts of the videos in which they have more interest than failing users [5]. If the last hypothesis was verified, this would show that there is a difference in the behavior that failing and passing users adopt when watching videos and that users adopting repetition or strategic approach to learning strategies would have higher chances of passing the class. However, we do not have enough information to verify this claim.

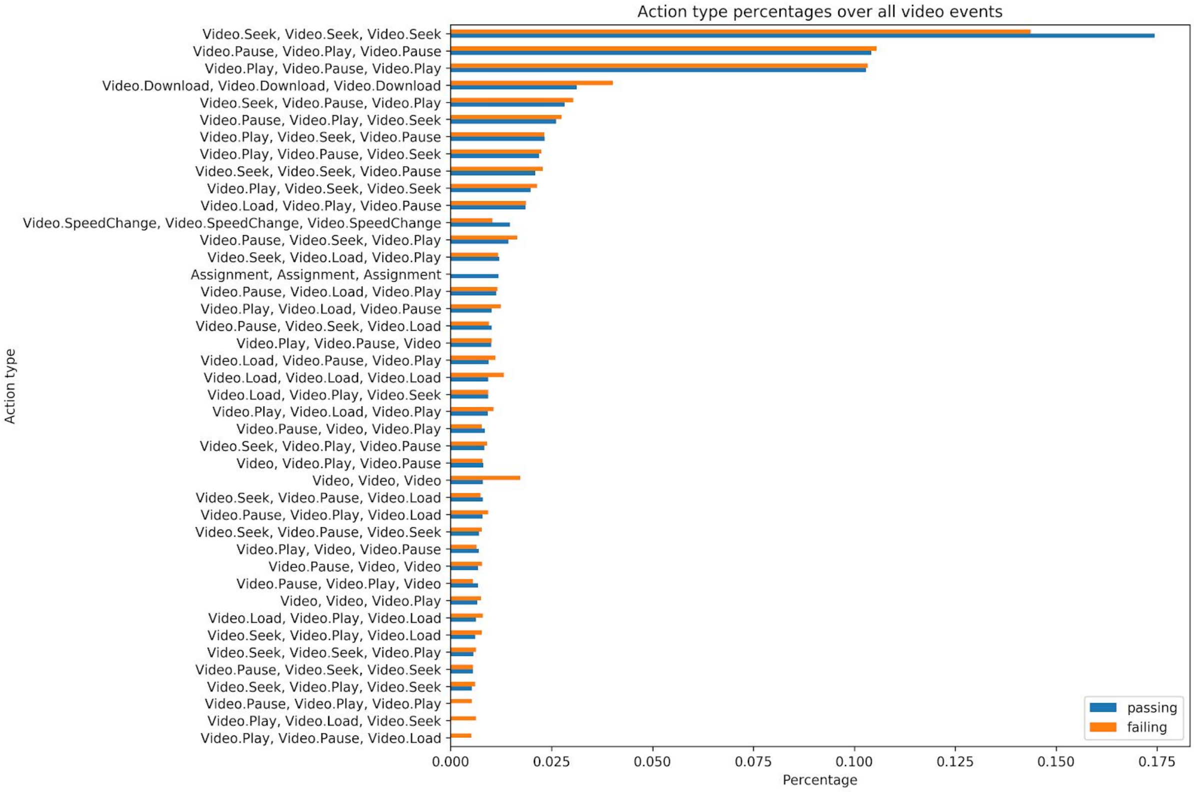

Figure 12: Overall top action chunks

For other top chunks such as [PAUSE, PLAY, PAUSE] / [PLAY, PAUSE, PLAY] and top chunks such as containing SEEK actions among PLAY and/or PAUSE actions, the same reasoning and explanation as given for the analysis of actions sequences between submissions can be given: The chunks [PAUSE, PLAY, PAUSE] and [PLAY, PAUSE PLAY] can either be assimilated as a user being interrupted in its watch time for some real life reasons or as a user taking notes about the videos; The chunks containing SEEK actions among PLAY and/or PAUSE actions can be assimilated as chunks representing either that the user is rewatching the videos for repetition or not well understood part or that the user is doing forward jumps in the videos to focus only the parts of the video that he is interested in or on parts that he has never heard of. These chunks can be therefore assimilated to strategic approach to learning.

6.4.3.5. Code

If you would like to see our code, you can find it in the following Github repository [12]. There is one notebook per section of our study.