Journal of Universal Computer Science, vol. 24, no. 8 (2018), 1131-1150
submitted: 15/1/18, accepted: 5/7/18, appeared: 28/8/18 © J.UCS

User Behavioral Patterns and Early Dropouts Detection: Improved Users Profiling through Analysis of Successive Offering of MOOC

Massimo Vitiello
(Graz University of Technology, Austria
massimo.vitiello@alumni.tugraz.at)

Simon Walk
(Graz University of Technology, Austria
simon.walk@tugraz.at)

Denis Helic
(Graz University of Technology, Austria
dhelic@tugraz.at)

Vanessa Chang
(Curtin University of Technology, Australia
vanessa.chang@curtin.edu.au)

Christian Guetl
(Graz University of Technology, Austria
c.guetl@tugraz.at)

Abstract: Massive Open Online Courses (MOOCs) are one of the fastest growing and most popular phenomena in e-learning. Universities around the world continue to invest to create and maintain these online courses. Reuse of material from previous courses is a shared practice that helps to reduce production costs and enhance future offerings. However, such *re-runs* still experience a high number of users not completing the courses, one of the most compelling issues of MOOCs. Hence, this research utilizes the information from the first run of a MOOC to predict the behavior of the users on a successive offering of the same course. Such information allows instructors to identify users at risk of not to finishing and helps to improve successive offerings. To this end, we analyze two successive offerings of the same MOOC, created by Curtin University on the edX platform. We extract features from the original run of the MOOC and predict dropouts on its re-run. We experiment with a Boosted Decision Tree and consider two different approaches: a varying percentage of users active time and users' first week of interactions with the MOOC. We obtain an accuracy of 0.8 when considering 10% of users active time or the first five days after users initial interaction. We also identify a set of features that are likely to indicate whether users will attrite in the future. Moreover, we discover typical patterns of interactions and notice a first set of tools that account for most interactions and a second one that is practically overlooked by users. Finally, we discover subgroups among the Dropouts characterized by similar behaviors. Such knowledge can be used to shape the structure of courses accordingly.

Key Words: MOOCs, Dropouts prediction, Behavioral patterns
Category: L.3.0, L.3.5, L.3.6

1 Dropouts in MOOCs

This work is an extension of our previous paper, which was presented and published at the international conference MOOC-Maker 2017 [Vitiello et al., 2017b].

Over the past years, the Web became the channel on which a variety of new types of learning methodologies materialized. Massive Online Open Courses (MOOCs) emerged as the natural solution to offer distance education. MOOCs are *Massive*, with unlimited audience; *Online*, as the learning process takes place online, without any geographical barriers; *Open*, since user do not require previous knowledge, nor have enrollment costs; a *Course*, as their structures resemble traditional lectures, with assignments and exams [Rodriguez, 2012].

Despite these advantages, the expectations MOOCs carried with them have not yet been wholly reached. Notably, there are cross-border complexities due to international recognition of credits earned [Taneja and Goel, 2014], lack of meaningful interactions among users leading to a sense of isolation [Rubin, 2013] and difference pacing of users [Bruff, 2013]. Furthermore, nearly all MOOCs suffer from low completion rates (generally lower than 10%) [Jordan, 2014].

High dropout rates are a problem also in successive offerings of the same course, so-called *re-run*. These re-runs have structures, topics, and schedules similar to those of the original course. Therefore, lower efforts are required when organizing successive re-runs, which can be enhanced according to previous users' feedback. Discovering and analyzing patterns of interactions of different groups of users help us to increase our understanding of the learning style of the users and how to best support them, with the explicit goal of mitigating dropout rates.

The research questions for this journal focus on two aspects: (i) can we predict dropouts at an early stage on the re-run of a MOOC using information from the first run of the same MOOC? (ii) can we identify behavioral patterns that characterize the different group of users? Notably, we analyze a MOOCs and one of its re-run offered on edX by Curtin University (Perth, Western Australia).

First, we investigate different percentages of users total active time and also focus on the first week of interactions of each user. We use these two approaches to train a classifier on the first MOOC and detect dropouts on the re-run. This procedure allows us to verify if users' behavior during the initial stage of the course is a reliable indicator of their outcome and if this is also true for re-runs.

Second, we investigate the difference concerning behavioral patterns of Completers and Dropouts. Specifically, we search for common sequences of interactions typical of each class and identify subgroups of Dropouts.

The rest of the paper is organized as follow. In Section 2 we overview relevant literature in the field of dropouts detection and behavioral patterns in MOOCs. In section 3 we detail the datasets and the setup of our experiments. In Section 4 we present and discuss our results. Finally, we highlight our findings and list possible future works in Section 5.

2 Related Work

2.1 Dropouts prediction

[Guetl et al., 2014] analyzed survey answers of users who dropped out, across MOOCs offered by Universidad Galileo. They proposed an Attrition Model for Open Learning Environment Setting (AMOES), which builds up and extends the Funnel of Participation Model [Clow, 2013]. Notably, they identify three healthy attrition subgroups according to users' goals, expectations, and reasons to drop out: *Exploring User*, *Content Learner* and *Restricted Learner*.

[Kloft et al., 2014] experimented with weekly dropouts classification using Support Vector Machines (SVM) on a MOOC with an 81.4% dropout rate offered on Coursera. The authors used cumulative features (number of interaction, number of view of each page of the course) and technical ones (browser, OS, number of screen pixel). They compared a trivial baseline (always predicts one or the other class) to the performance of SVM, which showed higher accuracy.

[Amnueypornsakul et al., 2014] experimented with dropout classification using SVM. The authors used quiz related and activity related features, verifying by ablation analysis that the two sets are both important for the prediction task. Furthermore, they noticed that the class imbalance and the presence of users with few interactions (Inactive) complicate the classification task.

In our previous work, we experimented with dropout prediction across five MOOCs offered by Universidad Galileo on their portal [Vitiello et al., 2016]. We analyzed the MOOCs using SVM and K-Means as classifiers and tested different combinations of features. In our results, K-Means always fell behind SVM, and specific combinations of features improved the accuracy of the prediction. In 2017 we refined our approach and developed a general classifier for dropout detection across different MOOC platforms [Vitiello et al., 2017a]

[Teusner et al., 2015] analyzed three iterations of a MOOC offered on the openHPli platform. While the content of the first two interactions barely differed, the third one was enhanced considering user feedback. The authors concluded that offerings with stable material attracted a wider audience with low effort.

2.2 Behavioral patterns

[Mukala et al., 2015] studied the behavior of users of a MOOC offered on Coursera. The authors used clickstreams information and compared the interactions of several groups of users. The authors discovered a structured and sequential learning approach is a typical trait of successful students and pointed out a direct relation between engaging with videos and obtaining a certificate of completion.

[Gelman et al., 2016] analyzed how the behavior of the users changes through time in a set of four MOOCs. The authors extracted a set of features and applied non-negative matrix factorization (NMF) to represents the user behavior transitions over time. Their results showed a total of eight unique categories of behaviors being persistent during the courses.

[Sinha et al., 2014] investigated how users engage with videos of a Coursera MOOC. Specifically, the authors derived an Information Processing Index (IPI) describing users interaction with videos. Their results imply the existence of groups of users characterized by similar IPI.
To our best knowledge, there are no research on behavioral patterns on the re-run of MOOCs.

3 Materials and Methods

3.1 Datasets

Our dataset consists of the original offering of a MOOC, referred to as *MOOCC1*, and the first of its re-runs, coded as *Re-Run1*. The original offering *MOOCC1* was available online during the second semester of 2015, while the re-run *Re-Run1* was available online between April and May 2016. Both offers had no entry prerequisites. The courses included independently created video content and regular activities, such as polls, questions, and discussion board tasks.

The course syllabus consisted of a total of four modules, each estimated to require a time commitment of two hours per week. An extra introductory module and a course wrap up module completed the course calendar. To complete each of the four main modules, participants needed to complete an activity and a quiz, with the quizzes being an extension of activities. Therefore, engaging in the activities helped participants to answer the questions in the quizzes. Each quiz accounted for 25% of the final grade, with a Certificate of Achievement issued to participants with an overall score of equal or greater than 70%.

The two courses were for the larger part similar to each other regarding contents and activities, with *Re-Run1* undergoing only some minor changes. Table 1 reports a summary of the enrollments and completion of the original MOOCs and its re-run. As the column *Enrollments* reports, *MOOCC1* has a total of 21,948 enrolled users and its re-run, *Re-Run1*, counts 10,368 enrolled

Table 1: Summary of the MOOC and the re-run. The table includes the number of users for each class, the number of dropouts and the dropout rates calculated in relation to both the *Active* users and the number of *Enrollments* (in brackets).

MOOC	Enrollments	Active	Inactive	Completers	Dropouts	Dropouts Rate
MOOCC1	21948	13396	8552	1500	11896 (20448)	89% (93%)
Re-Run1	10368	5932	4436	208	5724 (10160)	96% (98%)

users. Within the enrolled users, we distinguish users that enroll and leave the MOOCs without engaging any further (i.e., *Inactive* users), from those who have more than the simple enrollment interaction (i.e., *Active* users). *Completers* are users that completed the MOOCs, while *Dropouts* are those who failed to do so. Overall, the dropout rates are never lower than 89%.

Both offers are structured in a self-paced manner and are organized in two phases. During the first phase, users can only access the course main page and enroll and the course's material is not available. This initial phase lasts for roughly two months for both MOOCs. At the beginning of the second phase, the course material is uploaded all at once, and users can engage at their own pace. Enrollment is possible during the second phase as well. After the official end of the MOOCs, users still can register and interact, but, in this case, they can not obtain a certificate as the course is already officially over. Due to these settings, we consider only enrollments that took place before the official end of the MOOC. Furthermore, we also discard interactions occurring before the course's official start, as the course material is not available yet at this point. Both MOOCs also included a course forum where users can post and discuss.

The edX platform offers a standard set of tools that administrators of MOOCs can use and combine to shape the structure of their courses. Correctly, this MOOC included six specific tools[1]; *LMS* (Learning Management System) consists of the learning contents of the course and allow users to navigate through it; *Video* identify the interaction of the users who watch the videos developed for the course; *Problem* is the tool the users interact with to submit solutions and answers to the tests and assignments of the MOOCs; *Poll & Survey* can be used to get users opinions about specific topic; *Bookmark* lets users organize and mark the material of the course for a more personalization of their learning; *Forum* is the social part of the MOOCs where users and educators can communicate and exchange ideas, thoughts or questions about the course and the platform.

The actions that users performed while interacting with one of these tools are further described in more details in the logs. Specifically, each tool consists of a set of events that identify the action performed. For example, interactions with the tool *Video* are detailed by events such as *VideoPlayed*, *VideoPaused* and so on. We use this high details of the logs to create a large set of features for the dropouts prediction experiments as outlined in the next section.

3.2 Experimental Setup

3.2.1 Early dropout detection on MOOC re-runs

We extract a set of features to describe each user in our dataset. First, we calculate a set of time-based features that build upon the concept of sessions. A

[1] The complete list of edX's events is available at http://edx.readthedocs.io

session is a set of chronologically ordered interactions, in which each interaction happens within a certain timespan from the previous and the next one. Notably, we use a threshold of 30 minutes. Following this concept, we define the following features: *Sessions* as the total number of users' sessions; *Requests* as the total number of interactions per user; *Active Time* is the total time users interacted with the MOOC (sum of all sessions' duration); *Days* as the number of days during which users interacted at least once with the MOOC.

Furthermore, we compute 4 averaged features; *Timespan Clicks* is the average timespan between two consecutive clicks in the same session (averaged over all sessions); *Session Length* as *Active Time* divided by *Sessions*; *Session Requests* as *Requests* over *Sessions*; *Day Requests* is *Requests* divided by *Days*. Moreover, we exploit the detailed edX logs to identify the type of event triggered and the particular tool each interaction refered to. Specifically, we consider the events included in the six tools available for MOOCs in this setting (*LMS*, *Video*, *Problem*, *Poll & Survey*, *Bookmark* and *Forum* as described in section 3.1) and create a feature for each event by counting the number of interactions of users.

Table 2 report the list of events for each tool, together with our session related features. The *Video* tool, is the only one that allows distinction between Browser and Mobile (through the edX mobile application). To categorize these two sources, we create an additional tool, filtering out Mobile interactions from *Video*, and we name this new tool *Video Mobile*. We create a feature for each event, counting the number of times users' interactions triggered each of these and define two different approaches and calculate the features according to these.

First, we consider various percentages of users total active time. Notably, we consider all interactions within the first 1% to 100% of the total active time (per user) and call this setting *Scaled Time*. Second, we calculate features considering the first seven days after a users' first interaction. We name this setting *Days*. To overcome the class imbalance problem [Guo et al., 2008], we adjust the class distribution of our MOOCs by randomly oversampling the smaller class. We randomly pick and add examples to the smaller class until we have the same number of samples in both classes. For each approach we run a prediction experiment using Boosted Decision Trees, an ensemble classifier combines a set of decision tree into a single classifier. For each model, the misclassified examples get a higher weight, so the next decision tree focuses more on correctly predicting these.

We run classification experiments using two different set of users as input. First, we consider the all *Enrollments* and then only the *Active* users, as indicated in Table 1. We evaluate our experiments using accuracy, calculated as fraction of correctly predicted examples. Therefore, this measure assumes values between 0 and 1; a value of 0 indicates that our model classified all examples incorrectly, while a value of 1 means a correct prediction of all the considered examples.

As a final analysis, we investigate the importance of our features in the classification task. Boosted Decision Trees also provide a weight for each used feature, representing the number of times a feature is used to split the data across every single decision tree. Thus, the higher the weight, the more precise the obtained split. We explore the ranking of the features for both metrics when the input consists only of *Active* users.

3.2.2 Identification of user behavioral patterns

With this type of analysis, we aim at discovering how the users engage with the tools available in our MOOCs. Particularly, we want to know if there is a distinction between the tools used by the Dropouts compared to those that the Completers engage with. Furthermore, we want to investigate which sequences of interactions (tools) most characterize users of one or the other class. To this end, we consider all the available tools reported in Table 2 and use them to calculate the chronological sequence of interactions of each user. From these sequences, we create a transition matrix P describing the probability of users navigating among the set of tools T. Therefore, we indicate with P_{ij} the probability of interactions from tool i to tool j, such that $\sum_{i,j} P_{ij} = 1$ [Gagniuc, 2017].

To account for users that enroll but never interact any further, we extend the set of available tools S with an extra one called X_NULL. Therefore, users

Table 2: Summary of Tools and their events. The first column lists the name of the tools and the second one reports its list of events.

Tools	Events
Session Related	Sessions, Requests, Active Time, Days, Timespan Clicks, Session Length, Session Requests, Day Requests
Main Page Links	About, Faqs, Home, Instructor, Progress, StudyAtCurtin
LMS	TabSelected, PreviousTabSelected, NextTabSelected, LinkClicked, OutlineSelected
Video	CaptionHidden, CaptionShown, LanguageMenuHidden, LanguageMenuShown, Loaded, Paused, Played, PositionChanged, SpeedChanged, Stopped, TranscriptHidden, TranscriptShown
Video Mobile	CaptionHiddenM, CaptionShownM, LanguageMenuHiddenM, LanguageMenuShownM, LoadedM, PausedM, PlayedM, PositionChangedM, SpeedChangedM, StoppedM, TranscriptHiddenM, TranscriptShownM
Problem	Check, CheckFail, FeedbackHintDisplayed, Graded, HintDisplayed, Rescore, RescoreFail, Reset, ResetFail, Save, SaveFail, SaveSuccess, Show, ShowAnswer
Poll & Survey	PollSubmitted, PollViewResults, SurveySubmitted, SurveyViewResults
Bookmark	Accessed, Added, Listed, Removed
Forum	CommentCreated, ResponseCreated, ResponseVoted, Searched, ThreadCreated, ThreadVoted

that register and then drop out without any other interactions have a transition *Enrollment-X_NULL*. Moreover, we do not consider the *Session Related* tool (see Table 2) because this is an artificial tool comprising temporal features that we created for the dropouts prediction experiment and, therefore, it does not represent any of the edX initially available tools. We calculate the set of tool transitions for each class separately and use this information to construct two matrixes for each category of users and then calculate their differences (see 3).

In the first matrix, we report the number of transitions for each pair of tools and divide each entry by the total number of transitions of the class. In this way, we obtain a transition matrix that includes the probabilities of each set of transitions for each type of users [Asmussen, 2008, Gagniuc, 2017]. Further, we subtract the matrix calculated for the Dropouts to the one of the Completers, obtaining a difference matrix that we call *Transition probability difference*.

We create a second matrix similarly, but we divide each entry by the total number of users of the class. Therefore, we are calculating the per-user average number of transitions for each pair of tools. We subtract the Dropouts matrix from the Completers one and obtain a second difference matrix that we refer to as *Per-user average transition differences*.

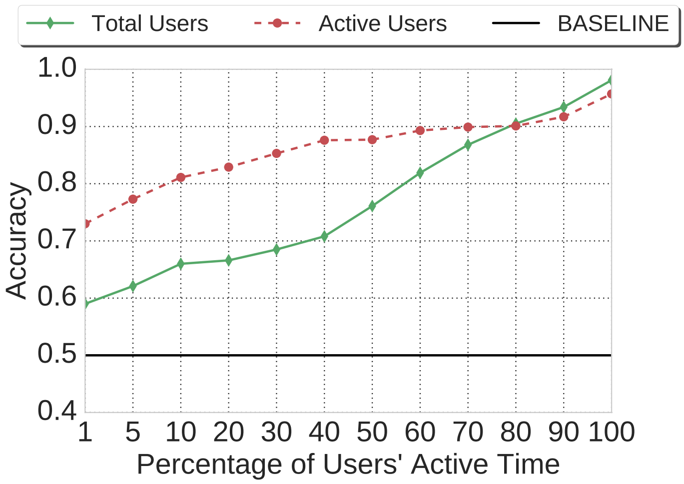
We verify the dissimilarity in behavior using these matrices and visualize these graphically through heatmaps [Wilkinson and Friendly, 2009].

4 Results and Discussion

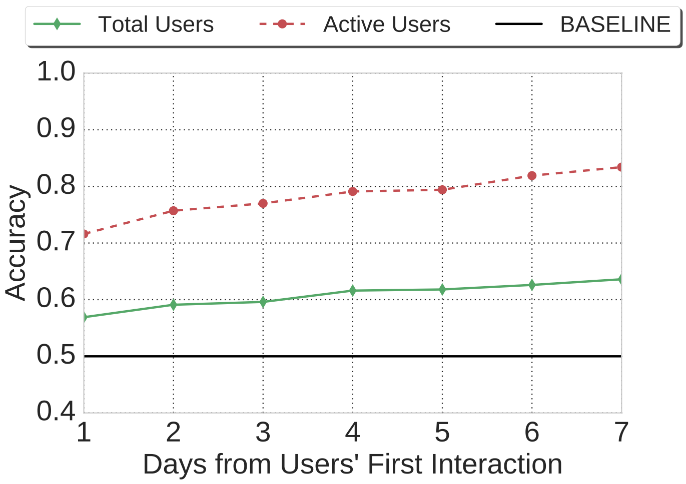
4.1 Early dropout detection on MOOC re-runs

Figures 1(a) and 1(b) report the results of the classification for the two approaches. The x-axes indicate the days after a user's first interaction for the Days experiment and the percentage of a users' active time for the Scaled Time experiment. For both figures, the y-axes indicate accuracy and are bounded between 0.4 and 1. The solid black horizontal line at 0.5 is the baseline, a lower bound representing the performances of a classifier that randomly predicts a class. Therefore, classifiers with accuracy under the baseline are no better than random prediction.

For the Days experiment reported in Figure 1(b), the accuracy, when we consider only the active users, is always increasing the more days we inspect. Notably, the accuracy is never lower than 0.7 and increases over 0.8 when we experiment with all seven days after users' first interaction. These results indicate how the first week of user interactions already represent a good indication of which users will eventually drop out. The accuracy, when considering all enrolled users, floats around 0.6. Therefore, the feature set does not characterize the two classes in this case. It is likely that a lot of users register for the MOOCs at an early stage, or at least more than a week before the material becomes available.



(a) Scaled Time



(b) Days

Figure 1: Dropouts prediction results on Re-Run1. Figure 1(a) reports the results for the Scaled Time approach and Figure 1(b) the ones for the Days approach. Full green lines are experiments considering all users (cf. Enrollments in Table 1), while dashed red ones are experiments with only the active users (cf. Active in Table 1). Considering only active users always yields the highest accuracy, except when the considered percentage of a user's active time is larger than 80%.

At this time, fewer interactions are possible, and users are likely to come back once the material is available. Thus, this approach has a lower overall accuracy.

The results of the Scaled Time experiment are plotted in Figure 1(a). Also, with this setting, considering only the active users is the approach that produces the highest accuracy. Again, the accuracy is never lower than 0.7 and gets as high as 0.96 when using the whole users' active time. When we take into account all enrolled users, the accuracy ranges from 0.59 to 0.98, constantly increasing as the percentages of users' active time get higher. Both settings have a similar profile when these percentages get higher than 80%.

Table 3 lists the best performing features for the two approaches when considering the active users. The first column contains the Tool and the second its specific features. The remaining columns report the weights for the features of the Days and the Scaled Time experiments respectively. For reasons of space, we report only some values for both approaches. Notably, we show day 1, 4 and 7 for the First 7 Days approach and 5%, 50% and 100% of users' active time. The weights highlighted boldly are the highest for that particular experiment.

We see that *Progress* is always one of the features with the highest weight for

Table 3: Feature Scores of the Days and Scaled Time Experiments. The features with the highest scores for both approaches are boldfaced. *Progress* is always among the features with the highest scores, while *ProblemCheck* scores increase the more days after users' first interaction and active time per user we consider.

		Days			Scaled Time		
Tool	Feature	1	4	7	5%	50%	100%
Session Related	Timespan Clicks	**62.3**	**53.9**	**57.4**	**63.1**	28.6	10.7
	Active Time	**44.6**	**39.3**	32.9	40.6	25.2	16.4
	Session Length	36.3	38.2	28.2	**59.5**	**39.3**	**56.1**
	Requests Active Day	24.6	18.1	17.0	21.8	29.7	28.8
	Session Requests	34.0	36.0	35.8	23.5	36.0	22.7
Problem	ProblemCheck	21.4	32.2	**41.0**	28.6	**47.5**	**59.8**
	ProblemGraded	3.1	7.6	4.0	1.1	13.3	42.7
	ProblemShow	18.0	16.8	20.2	18.4	11.6	35.8
Main Page Links	Home	28.0	28.8	22.6	28.9	16.5	12.5
	Progress	**50.6**	**54.2**	**54.7**	**55.1**	**84.6**	**93.8**
	StudyAtCurtin	22.0	18.8	11.0	9.9	2.6	7.7
LMS	NextSelected	22.1	13.2	11.2	21.8	29.7	14.6
	TabSelected	26.9	18.3	17.4	29.2	8.8	22.2
Video	VideoLoaded	24.6	20.3	17.7	22.4	11.4	9.0
	VideoPlayed	16.6	15.2	9.6	22.4	14.2	11.1

both experiments. These interactions refer to users accessing a dedicated page to track their scores for single problems and the current overall course grade. Particularly, this page includes reports of the obtained scores on each graded assignment in the form of a bar chart. The weight of this feature increases with the days and time percentage. If we extract the number of interactions of this type for both classes, we obtain a total of $17,240$ for the Completers and of $30,916$ for the Dropouts for *Re-Run1*. For *MOOCC1* we have $121,228$ interactions for the Completers and $54,768$ for the Dropouts. The class average yields 82.88 and 80.98 interactions for the Completers, and 6.88 and 5.40 for the Dropouts of *Re-Run1* and *MOOCC1* respectively. Besides, if for the Dropouts we also include the users with only the enrollment action, the averages get as low as 3.04 for *Re-Run1* and 2.68 for *MOOCC1*.

Hence, continual monitoring of the personal progress strongly indicates whether a user will drop out. Other studies reported similar correlations between checking the progress and the probability of dropping out from MOOCs offered on edX [Balakrishnan and Coetzee, 2013]. Similarly, *ProblemCheck* becomes more significant, the more days and higher percentages of interactions are analyzed. This action describes a problem being correctly checked by the system after users submitted an answer to it. The high scores of this feature come as no surprise, as users are likely to solve problems only after they study and learn from the course's material, that is, at a later stage during the MOOC.

Tools such as *Video Mobile* and *Forum* never obtain significant weights. In the case of *Video Mobile*, this situation indicates that users mostly interact with the MOOCs using a desktop machine rather than the edX mobile application. *Poll & Survey* and *Bookmark* are rarely used, either due to being poorly advertised or to users not regarding them as particularly useful to complete MOOCs.

It also appears that interactions within the *Forum* barely relate to Completers or Dropouts. First, it is possible that the course's structure does not require users to engage with the forum. This situation could be due to unchallenging classes or, more likely, due to the self-paced setting of the MOOCs. Users participate at their own pace and confront the same challenges at different times. As a consequence, the role of the forum as a real-time communication channel and as the first source of help might be limited.

4.2 Identification of user behavioral patterns

We obtain very similar results for *MOOCC1* and *Re-Run1* and, therefore, in this section we only focus on the result for *MOOCC1*.

In Tables 4(a) and 4(b) we report the *Transition probability difference* and the *Per-user transition difference* of *MOOCC1*. Specifically, Table 4(b) lists the differences of the per-user average number of transitions among each pair of tools between Completers and Dropouts. In Table 4(a) we report the differences between Completers and Dropouts concerning the probability of transition of every pair of tools. The first column of both tables reports the starting tools of a transition, while the first row lists the final tools of a transition. Furthermore, we also have the extra tool X_NULL to catch single-interaction users. Positive values identify transition with higher probability or average number per-user for the Completers. Negative entries represent the opposite situation. Transitions with a percentage or a per-user average number of transitions lower than 0.001, are simply listed as < 0.001 with the relative signs (minus or none). Moreover, we visualize these matrixes graphically as heatmaps in Figure 2(a) for the *Transition probability difference* and in Figure 2(b) for the *Per-user transition difference.*

Analyzing these matrixes and heatmaps, we discover several aspects. Specifically, we notice that the tool *Bookmark* is never used by both classes of user. Furthermore, *Video Mobile* also accounts for a very limited number of transitions. Moreover, transitions *Enrollment-*X_NULL are the ones with the more significant negative difference in both matrixes. We already expected this negative values since these transitions are only found in the Dropouts class and identify users enrolling but never engaging in the MOOC. From the heatmaps reported in Figure 2(a) and 2(b) we see that most transitions are concentrated along three directions. First, many transitions start and end in the *LMS* tool.

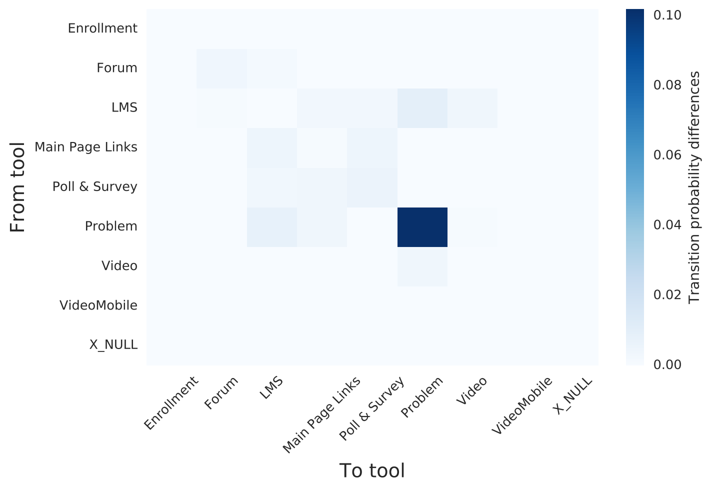
Table 4: Transition matrix difference for MOOCC1. Table 4(a) refers to the difference of the average number of transition per-user between Completers and Dropouts. Table 4(b) reports the difference of the probabilities of transitions between Completers and Dropouts.

(a): Per-user average transition differences

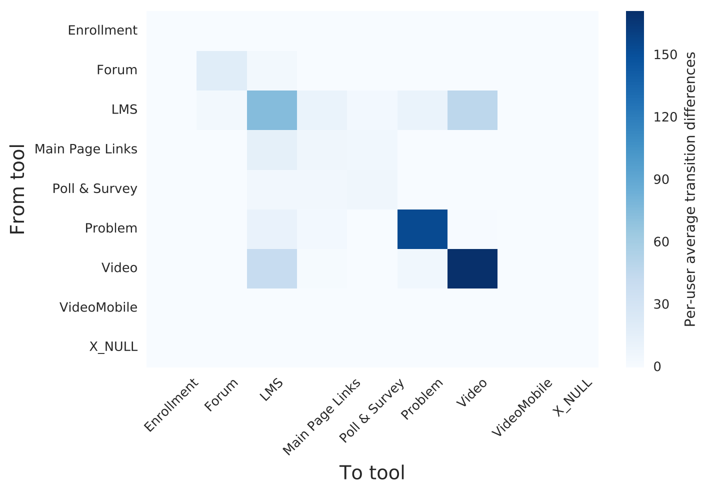
	Bookmark	LMS	Enrollment	Forum	Main Page	Poll & Survey	Problem	Video	VideoMobile	X_NULL
Bookmark	0	0	0	0	0	0	0	0	0	0
LMS	0	74.851	0.006	4.702	11.687	4.271	12.042	48.153	0.077	0
Enrollment	0	0.005	0.019	<0.001	0.562	0	0	0.005	-0.002	-0.418
Forum	0	5.210	0.001	19.570	1.168	0.009	0.063	0.258	0.004	0
Main Page	0	15.943	0.028	0.991	7.854	6.689	0.271	1.139	0.013	0
Poll & Survey	0	5.513	0	0.003	5.548	7.516	0.188	0.760	0.001	0
Problem	0	12.572	0.001	0.440	4.151	0.064	154.064	1.555	0.008	0
Video	0	41.755	0.001	0.579	2.368	0.978	6.217	171.275	0.022	0
VideoMobile	0	0.039	-0.001	0.004	0.070	0.002	0.008	0.017	1.044	0
X_NULL	0	0	0	0	0	0	0	0	0	0

(b): Transition probability differences

	Bookmark	LMS	Enrollment	Forum	Main Page	Poll & Survey	Problem	Video	VideoMobile	X_NULL
Bookmark	0	0	0	0	0	0	0	0	0	0
LMS	0	-0.017	-0.001	0.002	0.004	0.003	0.010	0.006	<-0.001	0
Enrollment	0	<-0.001	-0.007	<-0.001	-0.008	0	0	<-0.001	<-0.001	-0.009
Forum	0	0.002	<-0.001	0.005	-0.001	<0.001	<0.001	<-0.001	<-0.001	0
Main Page	0	0.005	<-0.001	-0.001	0.002	0.005	<0.001	-0.005	<-0.001	0
Poll & Survey	0	0.004	0	<0.001	0.005	0.007	<0.001	<0.001	<-0.001	0
Problem	0	0.009	<-0.001	<0.001	0.005	<0.001	0.103	0.001	<0.001	0
Video	0	-0.002	<-0.001	<0.001	<-0.001	<0.001	0.005	-0.119	<-0.001	0
VideoMobile	0	<-0.001	<-0.001	<0.001	<-0.001	<0.001	<0.001	<-0.001	-0.012	0
X_NULL	0	0	0	0	0	0	0	0	0	0



(a) Transition probability difference



(b) Per-user average transition difference

Figure 2: Heatmap for MOOCC1. Start tools of transitions are listed on the rows and the final tools are reported in the columns. The darker the colours of the cells the higher the differences between the Completers and Dropouts matrixes.

This tool includes interactions to navigate through the materials and lectures of the MOOCs and, therefore, it is a central tool that users have to interact with.

Second, there is a concentration along the main diagonal of the matrixes, representing transitions within the same tool (self-loop). Specifically, the *Problem-*

Problem and *Video-Video* loop are the transitions with the highest difference between Completers and Dropouts. Other transitions from and to the same tool also display a specific difference between the two classes.

Interestingly, the tool *Forum*, has low values of difference for all its transitions. These low values are due to general restricted use of this tool by both Completers and Dropouts. This outcome is somehow in contrast with other results found in the literature, where social interactions are a reliable indicator of users engagement and success in the course [Rosé et al., 2014, Yang et al., 2013, Yuan et al., 2013]. Our results reflect the specific course design and activities proper of the MOOCs in our dataset and, therefore, we would need to increase the number and the type of courses to generalize this particular finding.

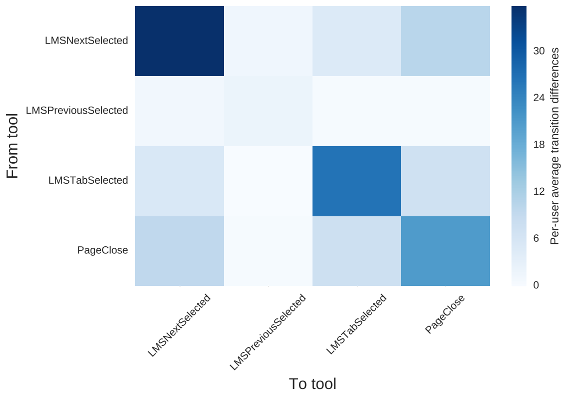
We also analyzed this situation in more details with transitions taking place among the tools with the higher difference between Completers and Dropouts: *LMS*, *Video* and *Problem* tools. We construct the transition matrixes using the events of each of these tools (see Table 2). To this end, we calculate the differences in the percentage and the per-user average number of transition between Completers and Dropouts, considering only the total number of transitions and the total number of users that interacted with these tools.

We show the results for the per-user average number of transition case for *MOOCC1* in Figure 3(a), 3(b) and 3(c) for the *LMS*, *Video* and *Problem* tools respectively. To improve the readability of these plots, we omit events with entry equal to 0. These entries represent either Completers and Dropouts having the same number of average transition per-user (which is practically never the case) or transitions that never took place for both classes, which is the most common case. Therefore, the omitted events are the ones the users never engage with.

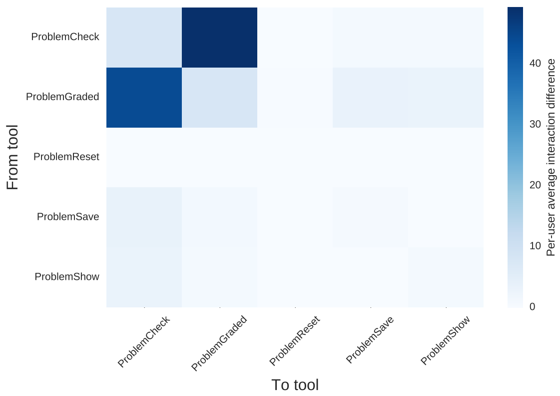
For *LMS*, depicted in Figure 3(a), the events *LMSLinkClicked* and *LMSOutlineSelected* are never used. Both these events identify the opening of a new tab in the web browser, a link extern to the MOOC in the case of *LMSLinkClicked* and a link within the course for *LMSOutlineSelected*. Users engage with the *LMS* tool mostly by navigating between subsections (*LMSNextSelected* and *LMSPreviousSelected*) or to a particular unit within a subsection (*LMSTabSelected*).

Most of the transitions within the *Problem* tool, as shown in Figure 3(b), happen between the *ProblemCheck* and the *ProblemGraded* events. Users submit the answer for a problem (*ProblemCheck*) and, if the answer is correct, the event *ProblemGraded* is triggered. The remaining events of this tool rarely happen: users seem to never request hints (*HintDisplayed* and *FeedbackHintDisplayed*) and only seldom use *ShowAnswer* to see the correct answer to a question. However, it is not clear whether users are not aware of the possibility of visualize hints or if the problems within this MOOC are not particularly demanding.

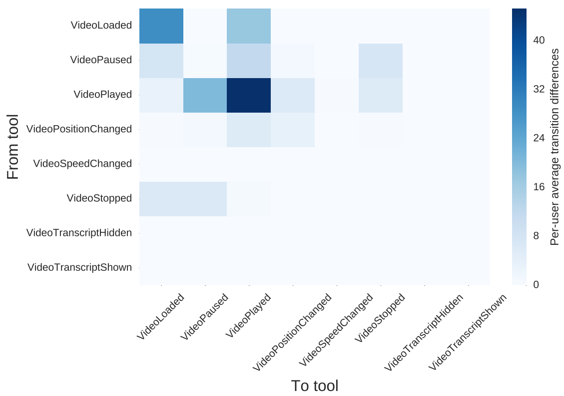
From the transitions of the *Video* tool showed in Figure 3(c), we see that the most common pattern of users is to play videos one after another (transition



(a) LMS



(b) Problem



(c) Video

Figure 3: Per-user average transition heatmaps of LMS, Problem and Video.

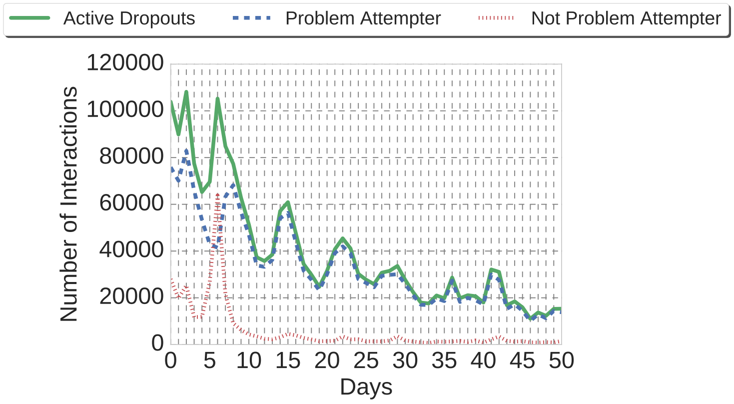
VideoPlayed-VideoPlayed). The transition *VideoLoaded-VideoLoaded* is the second most common one, which takes places when multiple videos are available on a page and, therefore, are loaded one after another upon page loading. In general, users interact almost exclusively with a "Play-and-Pause" approach and only rarely engage with other events of the *Video* tool. Change of the language of the video (*VideoLanguageMenuShown* and *VideoLanguageMenuHidden*) are never triggered, but that is no surprise as the MOOCs was available only in English. *VideoCaptionShown* and *VideoCaptionHidden* are also never triggered, while *VideoTranscriptShown* and *VideoTranscriptHidden* are mostly paired in the same transition and rarely present in transition with any of the other events.

Overall, we conclude that there is a well-defined subset of tools that includes the majority of users interactions. Moreover, there are also clear patterns of navigation within this subset of tools that describe the way users engage in MOOCs. Furthermore, specific events of the tools are never triggered by either Completers nor Dropouts.

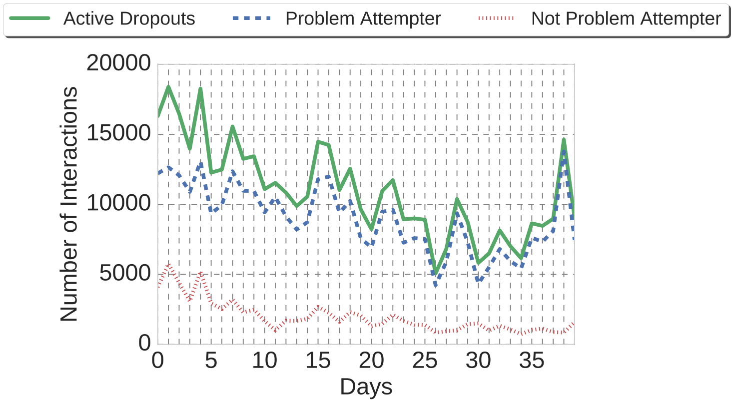
4.3 Dropouts patterns

We investigate in more details the dropouts class. As previously mentioned (see section 3.1) we had a first subgroup of dropouts who enrolled and then never further engaged with the MOOCs. We referred to this subgroup as *Inactive* in Table 1. The focus remains with the Dropouts, especially those that after enrolling interacted with the MOOC.

During the first phase of the MOOC, users can only access the course main page and enroll, but no further interactions are possible. Afterward, in the second phase, the course material is online, and users can actively interact. We analyze the distribution of interactions of the Dropouts during both phases and discover that there is a set of Dropouts that do not interact during the second phase. These users already abandon the MOOC before its official start. The remaining part of the Dropouts is also active in the second phase, and we refer to them as *Active Dropouts*. This latter group of users mostly interact with the *LMS*, *Problem* and *Video* tools as do the Completers. However, we find a subgroup of users that never interact with the *Problem* tools and only engaged with watching videos and navigating through the course material. We name this subgroup *Not Problem Attempter* in contrast to the remaining Dropouts that we call *Problem Attempter*. Figure 4(a) and 4(b) depict the distribution of interactions of the *Problem Attempter* (dashed blue lines), the *Not Problem Attempter* (dotted red lines) and the *Active Dropouts* (full green lines) for *MOOCC1* and *Re-Run1* respectively. Comparing our results with other literature's findings, we notice that our *Not Problem Attempter* shares some similarities with the *Auditing* group as described in [Kizilcec et al., 2013]. Notably, this later analysis identified groups of users employing a clustering approach on datasets from a different domain.



(a) MOOCC1



(b) Re-run1

Figure 4: Distribution of interactions of Dropouts. Figure 4(a) refers to *MOOCC1* and Figure 4(b) to *Re-run1*. The x-axes report the duration of the MOOCs (second phase) in number of days, the y-axes specify the number of interactions. We plot the *Problem Attempter* as blue dashed lines, the *Not Problem Attempter* as dotted red lines and the sum of the two as *Active Dropouts* as full green lines.

As we were able to discover a similar class of users, there are certain interdomain similarities that are discoverable despite the particular approach adopted.

We can see a clear difference in the distributions of the two subgroups in both figures. In Figure 4(a) we see a spike at day five in the distribution of the *Not Problem Attempter* of *MOOCC1*. We see a similar spike around the same date for the *Problem Attempter* also and we speculate a modification of the already uploaded course material or video as a possible explanation. For *Re-*

run1, depicted in Figure 4(b), we notice a spike in the distribution of the *Problem Attempter* towards the final day of the course. This spike might reflect a final rush of the users to try to obtain a certificate before the end of the course. In fact, after the official end of the course, users can still interact with the MOOC, but they can't earn a certificate even upon completion of the course. The non-presence of a final spike for *MOOCC1* in Figure 4(a) might be due to the longer duration of the course and a better organization of the learning by the users.

5 Conclusion & Future Work

In this work, we experimented with dropout detection on a MOOC re-run, analyzed the pattern of interactions of the users and verified the existence of subgroups of Dropouts. Specifically, we posed two research questions.

(i) Can we predict dropouts at an early stage on the re-run of a MOOC using information from the first run of the same MOOC? Our results indicated that in the first week of users' interactions, the features we created from the available information strongly predicted is users will complete the re-run. Furthermore, we evaluated the importance of each of the features used for the classification task. We discovered that the frequency users check their progress and correctly solve problems within a short period after their first interacting with a MOOC, strongly correlate to users' probability of completing the MOOC. We also noted that users barely engage with specific tools and found that the benefits of social tools (*Forum*), appear to be related to the way MOOCs are organized (i.e., limited benefits for self-paced MOOCs) and on the efforts each course required.

(ii) Can we identify behavioral patterns that characterize the different group of users? Through transition matrixes, we discovered that most transitions are found as self-loop on the *LMS*, *Problem* and *Video* tools. We investigated each of these tools with finer granularity and verified that users interact with a precise subset of the events available. Moreover, we also noted that both classes of users never use certain components. Furthermore, we discovered groups of users among the Dropouts with similar characteristics. Specifically, we found out that certain users only interact during the first phase of MOOCs and we showed that among the Dropouts that engage after the start of the course a subgroup never attempts to solve problems and quizzes (*Problem* tool). Instead, this subgroup watches videos and mostly browse through the course material. These results can help to improve the structure and the way MOOC's content is offered.

Verification as to why interactions with the *Forum* do not influence the probability of users completing the MOOCs, is a planned future work. Our analysis excluded interactions of instructors and only focused on the users. Evaluating the impact of instructors reaching out to users through tools such as forum and discussion boards may help us to assess if improvement of completion rates can also be achieved by using social engagement. We would also like to analyze how users engage with their peers, verify if different levels of engagement exist and check whether interactions among users improve their overall engagement and, thus, help to lower the dropout rate of MOOCs.

Analyses at tool level can represent a valuable next step. Abstracting from the particular event that took place, might help to differentiate more precisely the types of tools used and to confirm our current findings. Analogously to interaction and click-pattern mining approaches from other domains [Walk et al., 2017, Walk et al., 2015, Walk et al., 2014], we plan on identifying interaction types of users by clustering users of MOOCs according to their click- and interaction patterns to improve dropout detection. Besides, we also plan to explore further and detail the Dropouts concerning subgroups of users with similar behavior.

Acknowledgments

This work is in part supported by the Graz University of Technology, Curtin University, and the MOOC Maker Project (http://www.moocmaker.org/, Reference: 561533-EPP-1-2015-1-ES-EPPKA2-CBHE-JP).

References

[Amnueypornsakul et al., 2014] Amnueypornsakul, B., Bhat, S., and Chinprutthiwong, P. (2014). Predicting attrition along the way: the uiuc model.

[Asmussen, 2008] Asmussen, S. (2008). *Applied probability and queues*, volume 51. Springer Science & Business Media.

[Balakrishnan and Coetzee, 2013] Balakrishnan, G. and Coetzee, D. (2013). Predicting student retention in massive open online courses using hidden markov models. *Electrical Engineering and Computer Sciences University of California at Berkeley.*

[Bruff, 2013] Bruff, D. (2013). Lessons learned from vanderbilts first moocs.

[Clow, 2013] Clow, D. (2013). Moocs and the funnel of participation. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 185–189. ACM.

[Gagniuc, 2017] Gagniuc, P. A. (2017). *Markov Chains: From Theory to Implementation and Experimentation.* John Wiley & Sons.

[Gelman et al., 2016] Gelman, B. U., Revelle, M., Domeniconi, C., Veeramachaneni, K., and Johri, A. (2016). Acting the same differently: A cross-course comparison of user behavior in moocs. In *EDM*, pages 376–381.

[Guetl et al., 2014] Guetl, C., Chang, V., Hernández Rizzardini, R., and Morales, M. (2014). Must we be concerned with the massive drop-outs in mooc? an attrition analysis of open courses. In *Proceedings of the International Conference Interactive Collaborative Learning, ICL2014.*

[Guo et al., 2008] Guo, X., Yin, Y., Dong, C., Yang, G., and Zhou, G. (2008). On the class imbalance problem. In *Natural Computation, 2008. ICNC'08. Fourth International Conference on*, volume 4, pages 192–201. IEEE.

[Jordan, 2014] Jordan, K. (2014). Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distributed Learning*, 15(1).

[Kizilcec et al., 2013] Kizilcec, R. F., Piech, C., and 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*, pages 170–179. ACM.

[Kloft et al., 2014] Kloft, M., Stiehler, F., Zheng, Z., and Pinkwart, N. (2014). Predicting mooc dropout over weeks using machine learning methods.

[Mukala et al., 2015] Mukala, P., Buijs, J., and Van Der Aalst, W. (2015). Exploring students learning behaviour in moocs using process mining techniques.

[Rodriguez, 2012] Rodriguez, C. O. (2012). Moocs and the ai-stanford like courses: Two successful and distinct course formats for massive open online courses. *European Journal of Open, Distance and E-Learning*, 15(2).

[Rosé et al., 2014] Rosé, C. P., Carlson, R., Yang, D., Wen, M., Resnick, L., Goldman, P., and Sherer, J. (2014). Social factors that contribute to attrition in moocs. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 197–198. ACM.

[Rubin, 2013] Rubin, B. (2013). Online courses: possibilities and pitfalls (letter to the editor). *The New York Times*, 28.

[Sinha et al., 2014] Sinha, T., Jermann, P., Li, N., and Dillenbourg, P. (2014). Your click decides your fate: Inferring information processing and attrition behavior from mooc video clickstream interactions. *arXiv preprint arXiv:1407.7131*.

[Taneja and Goel, 2014] Taneja, S. and Goel, A. (2014). Mooc providers and their strategies. *International Journal of Computer Science and Mobile Computing*, 3(5):222–228.

[Teusner et al., 2015] Teusner, R., Richly, K., Staubitz, T., and Renz, J. (2015). Enhancing content between iterations of a mooc–effects on key metrics.

[Vitiello et al., 2017a] Vitiello, M., Walk, S., Chang, V., Hernandez, R., Helic, D., and Guetl, C. (2017a). Mooc dropouts: A multi-system classifier. In *European Conference on Technology Enhanced Learning*, pages 300–314. Springer.

[Vitiello et al., 2017b] Vitiello, M., Walk, S., Helic, D., Chang, V., and Guetl12, C. (2017b). Predicting dropouts on the successive offering of a mooc.

[Vitiello et al., 2016] Vitiello, M., Walk, S., Hernández, R., Helic, D., and Gütl, C. (2016). Classifying students to improve mooc dropout rates. *Research Track*, page 501.

[Walk et al., 2017] Walk, S., Espín-Noboa, L., Helic, D., Strohmaier, M., and Musen, M. A. (2017). How Users Explore Ontologies on the Web: A Study of NCBO's BioPortal Usage Logs. pages 775–784.

[Walk et al., 2015] Walk, S., Singer, P., Noboa, L. E., Tudorache, T., Musen, M. A., and Strohmaier, M. (2015). Understanding how users edit ontologies: Comparing hypotheses about four real-world projects. In *International Semantic Web Conference*, pages 551–568. Springer.

[Walk et al., 2014] Walk, S., Singer, P., and Strohmaier, M. (2014). Sequential action patterns in collaborative ontology-engineering projects: A case-study in the biomedical domain. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1349–1358. ACM.

[Wilkinson and Friendly, 2009] Wilkinson, L. and Friendly, M. (2009). The history of the cluster heat map. *The American Statistician*, 63(2):179–184.

[Yang et al., 2013] Yang, D., Sinha, T., Adamson, D., and Rosé, C. P. (2013). Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop*, volume 11, page 14.

[Yuan et al., 2013] Yuan, L., Powell, S., CETIS, J., et al. (2013). Moocs and open education: Implications for higher education.