Analyzing Millions of Submissions to Help MOOC Instructors Understand Problem Solving

Fang Han
Massachusetts Institute of Technology
Cambridge, MA
elaineh3@mit.edu

Kalyan Veeramachaneni
Massachusetts Institute of Technology
Cambridge, MA
kalyan@csail.mit.edu

Una-May O'Reilly
Massachusetts Institute of Technology
Cambridge, MA
unamay@csail.mit.edu

ABSTRACT

We present an analytics framework that helps analyze student behavior during problem solving in Massive open online courses. There are some important differences in how students solve problems online when compared to on campus education. However, massive amounts of information, at a very high granularity, capturing how students access the content and solve problems is currently stored. We present a structured way for assembling, aggregating, visualizing and statistical approaches to analyze this data. We are developing a comprehensive set of tools that will help and inform instructors. This paper presents our first few steps towards achieving that goal.

1. INTRODUCTION

MOOC platforms allow instructors to pose problems to their students in the context of a highly interactive but automated environment. Multiple sorts of problem designs are available to instructors so that problems can be tailored to factors such as nature of content, student progress or assessment performance. The nature of problem design and solving differs from its counterpart in standard campus pen and paper approaches, the most prominent difference being that students may be granted multiple attempts while receiving real time feedback on response correctness. As well, in the MOOC environment students more frequently answer questions under an "open book" policy, i.e. they are granted access to any of the available resources, e.g. lecture videos, e-books, the course wiki, forum and tutorials *while they problem solve.*

All educators want information on how students answer problems. It can show what problems are hard or easy. It can be used to guide content revision, adjust teaching with immediate feedback and generally help to understand the differences between problem designs or different groups of students with respect to problem solving. In the case of MOOC instruction, some of this behavorial information resides in data that is, in fact, captured and which can be knowledge mined for answers. As a MOOC student attempts to solve problems, regardless of their context (e.g homework, lecture exercise), his/her activities in interaction with the website are captured in raw data streams. These can subsequently be processed to extract multiple statistics on a problem basis. For example, a student's number of attempts, the number of resources a student visited while solving the problem, or the amount of time a student spent working on the problem can be derived from the data.

Given this data related to how students address problems, our goal is to elicit from it knowledge which is informative and concretely helpful to instructors. We are developing means to provide instructors with direct, descriptive analytics like how many attempts, on average, students needed for a problem through to helping them distinguish between problems given what combination of resources were frequently consulted, or how students who achieved different grades handled them. In this paper we:

- describe what behavior we have to date isolated in order for instructors to study problem solving behavior, see Section 2.
- explain how we have extracted quantifiable variables on a per-student or per-problem basis from the data, see Section 3.
- show what visualization and comparisons, based on these variables, we provide instructors with, to answer questions of the following nature, see Section 4:

  Problem oriented analysis: Considering two different problems, how do they differ in terms of resources consulted to answer them and the duration of time they required to answer them. For example: Do students spend more time doing labs or homeworks? In a particular course, did students spend more time problem solving in one module in comparison with the others?

  Cohort oriented analysis: Considering two student cohorts, based on grade, geography or any other basis, how did they differ in terms of how long they spent answering a problem or what resources they consulted to answer a problem. For example: Do students who earned an A in the course spend less time on the homework problems than students who earned a B? Did B students consult different resources than A students to solve the problem?

2. STUDENT BEHAVIOR

To develop our analytic methods and visualizations we are exploit our adoption of MOOCDB [4]. MOOCDB facilitates

.

the efficient re-organization of originally recorded MOOC data by offering a general purpose data model for analytics. The model, while no relevant content is lost, reduces the raw data size and structures the data efficiently for subsequent behavioral research studies. Our research data are organized according to the model and this makes the extraction of specific data for different analytics relatively convenient. We heavily reference the *Submitting Mode* and *Observing Mode* tables within MOOCDB for this paper's tools. The *Submitting Mode* tables form a structured representation of the data that records student interactions with the assessment modules of a course.[1] The *Observing Mode* tables record student interactions with resources.

We isolate multiple aspects of student problem solving behavior that will benefit instructors:

Problem context We isolate which of 5 different contexts within which a student encounters a problem: *homework*, *lecture quizzes*, *midterm*, *final* exam or *lab*. We include subjective or objective problems.

Problem hierarchy We break down a specific problem if it has sub-problems and consider separately each non-decomposable subproblem. We use the hierarchy encoded in MOOCDB's *Problems* table.

Resource consultation We isolate how a student references resources inbetween when s/he accesses a problem and enters each attempt to solve it. We take advantage of the *Observing Mode* tables in our database which collectively capture each exact resource consultation and, when available, the context in which the resource was accessed.

Problem attempts We isolate how many times a student attempts a problem.

Correctness information We isolate whether or not a student provided the correct answer to the problem. Attempts are tagged with whether or not the answer was correct and can be found in the *assessments* table defined by the MOOCDB data model.

3. STUDY VARIABLES

To proceed we define quantifiable variables which we measure on a per student and per problem basis. These are respectively superscripted as i and j in notation:

Response Formulation Duration, variable d: This variable expresses how long a student spent answering a question. Obviously this is a complex and best approximation measurement given browser transaction records are the sole source of information.[2] We rely upon timestamps attached to problem submission events and preprocessed fields in the MOOCDB data model's observing mode tables. To measure response formulation duration we first identify the time stamp t_s for the first submission and the last submission t_e by the student and problem. We then identify all the events in the observing events table that correspond to this student between those two time stamps. We then add the duration for each of the event we identified to add up to the total duration.

Resource consultations, variable r: To count how many resources have been consulted, we consider a resource to be a unique url within the course website. To obtain the count, we simply count the number of resource URL events between the two submission time stamps. We filter out the resources of duration less than 10s as they could could be a touch-and-go landing from a bookmarked page. We also filter out pages that correspond to *profile*, *informational* as they are likely not related to the problem.

Problem Attempts, variable a: A count of problem attempts per problem can be directly calculated from the submissions table in the MOOCDB data model. There each submission by the student is recorded as a separate event.

Correctness matrix, C of elements c_{ij}: We assemble a $n-by-m$ matrix where n is the number of students and m is the number of problems. The entry $c_{ij} \in [01]$ represents whether student i got j problem correct.

To prepare to provide visualizations and comparisons to instructors, we assemble these variables in simple tables, one per problem, that have columns for d, r and a and rows for each student. We can thus easily provide summaries per problem or on a per student basis. For example, we can calculate percentage of students (among all) who got a problem correct from the ratio of number of students who got it correct to the total number of students who attempted that problem. Similarly for each problem, we can assemble average response formulation duration, average number of problem attempts, and average count of resource consultations.

4. PROBLEM ANALYTICS

We now present a set of analytics we have developed for instructors based on this data. In Section 4.1 we provide per problem summaries and use statistical methods to provide means of comparing student behavior when problem context differs (e.g within labs vs homeworks). In Section 4.2 we show how student grade-based cohorts can be statistically compared on the bases of **response formulation duration**, number of **problem attempts**, and counts of **resource consultations**.

As demonstration data we assembled these *per-student* variables *for every problem* of an MITx offering of "6.002x: Circuits and Electronics". The course had 154,763 registrants, of which 69,221 students looked at the first homework and 26349 students earned at least one point on the homework. 7157 students completed the course successfully and earned a certificate. There were 633 problems in this course including labs (73), lecture exercises (328), midterm (26), final (47), and homeworks (154) and miscellaneous (5). There were a total of 17,812,152 submissions (including all attempts) from students across different problems. Out of these 9,985,753 were from students who earned the certificates. To assemble our variables we had to process and scan 132,286,335 events in the observed events table out of which 68,608,730 events were from certificate winners. Assembling this data involved a significant amount of computational effort.

[1] A typical MOOC student is in the submitting mode when trying homework problems, exams, quizzes, exercises in between lectures and labs (for engineering and computer science).

[2] For example, no information is available about whether a student is distracted while on a web page.

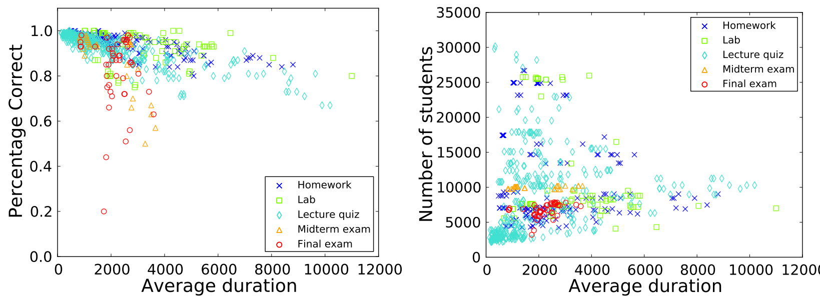

Figure 1: (Left) Average response formulation duration vs. percentage of students who both attempted a problem and eventually answered it correctly. (Right) Average response formulation duration vs. number of students that attempted the problem.

	p-value	*Confidence interval*		*Sig?*
Average:		l	u	
Response duration	1 e-28	-1.19e3	-0.83e3	Yes
Response consultation	8.85e-26	-6.70	-4.59	Yes
Attempts	8.86e-14	-0.33	-0.19	Yes

Table 1: Differences in student behaviors: lab vs. homework problems.

4.1 Problem summaries

Our first set of analytics are simple visuals and bivariate plots which help an instructor examine the variation within a problem context, e.g. lab, homework. The analytics calculate for each problem: average response formulation duration, average number of problem attempts, average resource consultations and percentage of students who both attempted a problem and eventually answered it correctly. Figures 1, 2, 3 show the bivariate plots. Each point is a different problem. We color and mark different problems types using different legend. We are currently working to make these plots interactive (using $D3.js$) so the instructor can browse through the plot and mouse-over or click on a data point to get detailed information about the problem.

Is student behavior different in problem context? Our second set of analytic allow contexts of problems to be compared. Instructors can learn whether there is a significant difference in these variables: average response formulation duration, average problem attempts and average resource consultations for labs *vs.* homeworks. An instructor could potentially use this information to consider how students respond to the different degrees of interaction or visualization use in labs compared to homeworks.

To facilitate comparison, we first assemble aggregate statistics over each problem. We then apply a *two-sample t-test* which tests null hypothesis that the data in both vectors comes from independent random samples from normal distributions with equal means and equal but unknown variances. Table 1 presents results from the two-sample t-test where we compare labs to homeworks. It informs the instructor that, for all three statistics, the difference was significant. It indicates to the instructor that, for almost every statistic, the ranges for the confidence interval are negative. Given the ordering for our test was homeworks and labs, this would be interpreted as students spent less time doing homework, consulted fewer resources and attempted less number of times in the context of homeworks. Both homeworks and labs were equally weighted towards the grade in this course. Without automated confounding variable analysis, we would expect an instructor to interpret this information in light of informal confounding information s/he has.

4.2 Cohort analysis

Next we attempt to help instructors identify whether the behavioral patterns within the context of different problem types are different for different student cohorts. Our demonstration focuses on students who got A's *vs.* B's vs. C's and students from the top 5 countries in terms of certificate winners which are IN (India), US (USA), ES (Spain), GB (Great Britain), CO (Columbia). We analyze each cohort in the contexts of homeworks and labs. This involves a different aggregation of the data. We assemble the average problem attempts, average response formulation duration, and average resource consultations for every student across all homework and all lab problems s/he attempted. We then perform 1-way ANOVA and pass the results of the ANOVA [1] to Tukey's Honestly Significant Difference Test [2]. This test provides statistic evidence regarding a pairwise comparison between cohorts. Demonstration results for the homework problems are shown in Table 2, 3 and demonstration results for the lab problems are shown in Table 4, 5. This sort of analysis and results is common in education research literature [1, 3]

Each row in these tables compares two cohorts. The range in the brackets corresponding to an aggregate, presents 95% confidence interval for the true difference of the means for that aggregate when conditioned on the two cohorts under consideration. If 0 is not present in the range the difference is considered significant. The differences that are considered significant are highlighted.

For homework problems an instructor will note that stu-

	Duration		*# of attempts*		*# of resources*	
	l	*u*	*l*	*u*	*l*	*u*
A *vs.* B	-728.0771	291.7774	-0.6141	-0.3467	-4.8766	-2.3188
A *vs.* C	-190.2763	436.8347	-0.2356	0.1488	-2.6303	1.0462
B *vs.* C	295.2343	971.1786	0.2299	0.6441	0.8242	4.7870

Table 2: Homework problems

	Duration		*# of attempts*		*# of resources*	
	l	*u*	*l*	*u*	*l*	*u*
IN *vs.* US	-872.3882	3.5861	-0.6428	-0.1051	-2.4724	2.6720
IN *vs.* GB	-1224.7	-131.1148	-0.9644	-0.2930	-4.6386	1.7840
IN *vs.* ES	-546.6704	558.1158	-0.7726	-0.0943	-2.3314	4.1568
IN *vs.* CO	-886.3729	272.7918	-0.8243	-0.1127	-4.1661	2.6415
US *vs.* GB	-748.3169	261.2665	-0.5646	0.0551	-4.4917	1.4375
US *vs.*ES	-70.7088	950.9564	-0.3731	0.2541	-2.1871	3.8129
US *vs.*CO	-412.5076	667.7286	-0.4261	0.2371	-4.0342	2.3099
GB *vs.*ES	76.9200	1290.4	-0.1772	0.5677	-1.2232	5.9032
GB *vs.*CO	-260.4477	1002.7	-0.2275	0.5480	-3.0442	4.3742
ES *vs.* CO	-948.9352	323.9087	-0.4257	0.3557	-5.4126	2.0626

Table 3: Homework problems

	Duration		*# of attempts*		*# of resources*	
	l	*u*	*l*	*u*	*l*	*u*
A *vs.* B	-1117.8	-257.4382	-0.5284	-0.2817	-6.3102	-1.2575
A *vs.* C	-1133.1	103.5755	-0.2432	0.1113	-8.7658	-1.5034
B *vs.* C	-493.6336	839.3696	0.1480	0.5301	-5.2647	2.5633

Table 4: Lab problems

	Duration		*# of attempts*		*# of resources*	
	l	*u*	*l*	*u*	*l*	*u*
IN *vs.*US	-922.6375	803.7899	-0.5299	-0.0345	-2.5717	7.5698
IN *vs.*GB	-1810.2	345.2250	-0.5842	0.0342	-6.6015	6.0598
IN *vs.*ES	-977.8678	1199.5	-0.8449	-0.2201	-4.8257	7.9648
IN *vs.*CO	-2034.1	250.4387	-0.8462	-0.1906	-10.6366	2.7834
US *vs.*GB	-1667.9	321.8334	-0.2783	0.2926	-8.6140	3.0743
US *vs.*ES	-836.5339	1177.0	-0.5392	0.0385	-6.8436	4.9846
US *vs.*CO	-1896.9	232.0838	-0.5417	0.0692	-12.6788	-0.1725
GB *vs.*ES	-352.4896	2039.1	-0.6007	0.0856	-5.1839	8.8647
GB *vs.*CO	-1404.1	1085.4	-0.6006	0.1138	-10.9678	3.6563
ES *vs.*CO	-2257.0	251.6374	-0.3458	0.3741	-12.8643	1.8719

Table 5: Lab problems

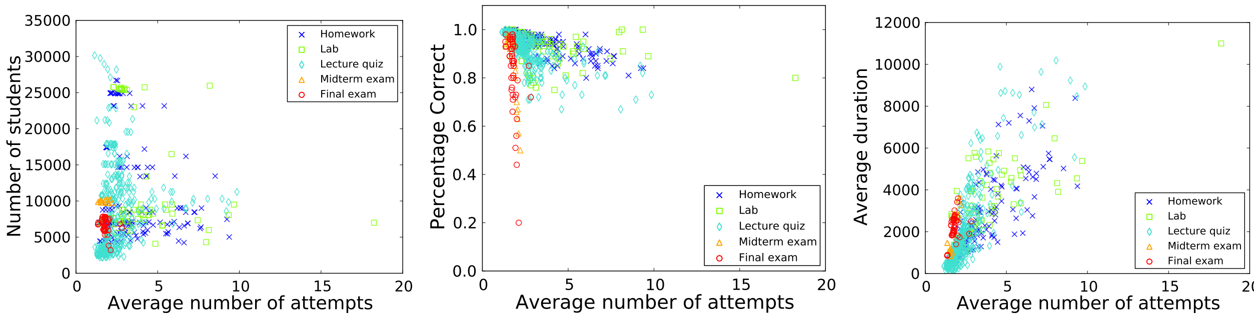

Figure 2: Left: Average problem attempts vs. number of students who attempted the problem. Middle: Average problem attempts vs. percentage of students who both attempted a problem and eventually answered it correctly. Right: Average problem attempts vs. average response formulation duration.

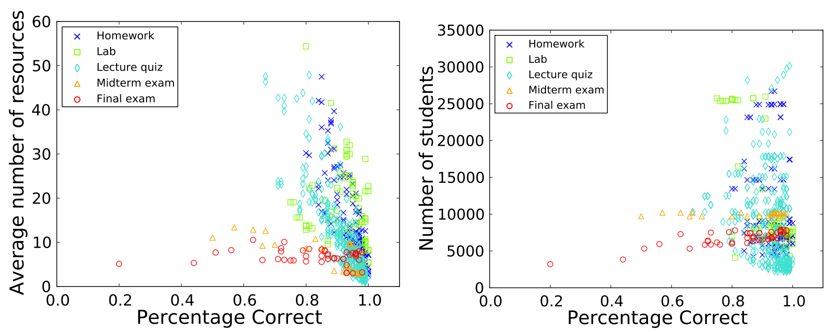

Figure 3: Left: percentage of students who both attempted a problem and eventually answered it correctly vs. average resource consultations. Right: percentage of students who both attempted a problem and eventually answered it correctly vs. number of students who attempted the problem.

dents who got A did not spend as much time, or use as many attempts or view as many resources as B students. As well, there is no significant difference between A and C students in these respects. In B students spending more time on an average on each problem, they made more attempts and viewed more resources. An instructor could find this pattern to be of interest and take other measures to validate additional intuitions around "A students can solve problems fastest" or "students who got C are not making enough effort".

For country based analysis of cohorts, we note that there really not many differences that are significant. The most prominent difference we see is for students from India who on an average have less number of attempts in both homework problems and lab problems. Other than this we do not see many differences between students from different countries.

5. CONCLUSIONS

We have begun to help MOOC instructors analyze how students interact with the learning platform, problem designs and content when solving problems. We are building a very generalizable analytics framework that will enable analyses on the bases of student cohorts or problem context. We are also providing them with straight forward plots of descriptive statistics. This current work is part of the larger framework in which we are developing a number of ways in which we can give extensive feedback to the instructor on how students are solving the problems and where there is the most student engagement. This information has never been available to instructors at this scale before.

Acknowledgements

This work was supported by Quanta Research.

6. REFERENCES

[1] G. V. Glass and K. D. Hopkins. *Statistical methods in education and psychology.* Prentice-Hall Englewood Cliffs, NJ, 1970.

[2] Y. Hochberg and A. C. Tamhane. *Multiple comparison procedures.* John Wiley & Sons, Inc., 1987.

[3] R. F. Kizilcec, C. Piech, and E. Schneider. 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, 2013.

[4] K. Veeramachaneni, F. Dernoncourt, C. Taylor, Z. Pardos, and U.-M. O Reilly. Moocdb: Developing data standards for mooc data science. In *AIED 2013 Workshops Proceedings Volume*, page 17, 2013.