Game Theory Based Peer Grading Mechanisms For MOOCs

William Wu
Acton Boxborough Regional High School
36 Charter Rd
Acton, MA 01720, USA
willy.vvu@gmail.com

Constantinos Daskalakis
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
costis@csail.mit.edu

Nicolaas Kaashoek
Lexington High School
251 Waltham Street
Lexington, MA 02421, USA
nick.kaashoek@gmail.com

Christos Tzamos
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
ctzamos@gmail.com

Matthew Weinberg
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
smweinberg@csail.mit.edu

L@S 2015, Mar 14-18, 2015, Vancouver, BC, Canada
ACM 978-1-4503-3411-2/15/03.
http://dx.doi.org/10.1145/2724660.2728676

Abstract
An efficient peer grading mechanism is proposed for grading the multitude of assignments in online courses. This novel approach is based on game theory and mechanism design. A set of assumptions and a mathematical model is ratified to simulate the dominant strategy behavior of students in a given mechanism. A benchmark function accounting for grade accuracy and workload is established to quantitatively compare effectiveness and scalability of various mechanisms. After multiple iterations of mechanisms under increasingly realistic assumptions, three are proposed: Calibration, Improved Calibration, and Deduction. The Calibration mechanism performs as predicted by game theory when tested in an online crowd-sourced experiment, but fails when students are assumed to communicate. The Improved Calibration mechanism addresses this assumption, but at the cost of more effort spent grading. The Deduction mechanism performs relatively well in the benchmark, outperforming the Calibration, Improved Calibration, traditional automated, and traditional peer grading systems. The mathematical model and benchmark opens the way for future derivative works to be performed and compared.

Author Keywords
Massive Open Online Courses; MOOC; game theory; mechanism design; peer grading; learning at scale;

ACM Classification Keywords
K.3.m. [Computers and Education]: Miscellaneous. See: http://www.acm.org/about/class/1998/ for help using the ACM Classification system.

Introduction
Over the past few years, there has been a tremendous increase in the popularity of MOOCs (Massive Open Online Courses) and their importance to education as a whole. Popular MOOC systems such as Coursera or EdX are well funded, which explains their rapid growth: 60 million dollars were invested in EdX when it started in May of 2012 [8]. The main importance of MOOCs is their ability to educate massive numbers of students worldwide [3]: by the end of 2012, 1.7 million students had attended a course through Coursera [6]. This leads to high student-professor ratios, reaching 150,000:1 in some courses.

High student-professor ratios lead to problems for professors, who are simply unable to grade hundreds of thousands of submissions. Currently, two types of solutions are used to remedy the problem: automated grading and peer grading [7]. Automated grading relies on machines, which can only check certain types of answers (i.e. multiple choice), severely limiting the depth of the questions asked [1]. Even though automated grading for written essays is an active area of research with much recent progress, the quality and accuracy of such systems is under heavy debate [4]. Students who know the machine's grading criteria can fool the system [9], yielding inconsistent grades. On the other hand, peer grading can grade any type of question. However, such systems can easily be "hacked" by the students [3]. Additionally, lack of feedback from peer grading is an area of complaint in systems such as Coursera [10]. These limitations render the students unable to effectively evaluate mastery of course material.

We propose several peer grading mechanisms based on game theory and our student model - a set of assumptions we believe students abide by. We also create a benchmark to compare between our and existing mechanisms. Although a theoretical model cannot predict exactly what will happen in practice, game theory and mechanism design have a history of generally determining mechanisms that work in practice from ones that do not [5]. Mechanisms that do not follow game theoretic constraints may work in the short-term, but they will be exploited if possible in the long term [2].

Model and Assumptions
Our student model consists of assumptions we believed students abide by, as follows:

1. Let H be a function of a student's grade, returning a student's happiness, such that a grade of zero yields zero happiness ($H(0) = 0$).
   Happiness is an arbitrary numerical unit.
2. Students want to maximize their happiness.
3. Grading an assignment costs 1 (one) happiness.
4. Happiness is not affected by external factors, such as the grades of peers.
5. Students can communicate with their peers.
6. Students are not perfect graders.

7. There is no such thing as partial-grading. Students either grade or do not grade.
8. Students can report their level of uncertainty (U) when they grade.
9. More effort spent in grading lowers uncertainty.
10. The chance of a student-assigned grade G being N off from the actual grade is proportional to U.

With a student model in place, it is now possible to simulate student behavior with game theory.

Benchmark

In order to eventually determine the effectiveness of various mechanisms, as well as to compare mechanisms, we created a numerical benchmark (objective function) where a lower score is better. The score is computed by adding the highest possible error in student grading to the most work done by any person. Mathematically:

$$\max_{i\geq 1}\{|H(g_i) - H(o_i)|\} + \max_{i\geq 1}\{w_i\}$$

where w_i is the work done (assignments graded) by the ith person, g_i is the grade given by student grader on the ith assignment, and o_i is the accurate "objective" grade that would have been given by the professor on the ith assignment. H is the happiness function defined in the *Models and Assumptions* section.

Mechanisms

Calibration Mechanism

The Calibration mechanism, described visually in Figure 1,achieves a low benchmark score of 4, consisting of a 2 in max work done and 2 in max error in grade.

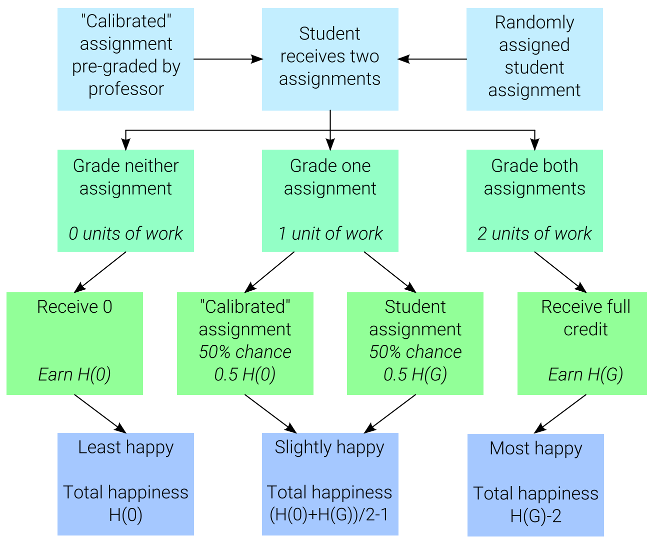

Figure 1: A flowchart of the Calibration Mechanism from the student perspective.

We tested and verified the Calibration mechanism through an anonymous crowd-sourced experiment. We wrote a program that presented participants with two sets of randomly generated orange and blue colored objects. Each set represented an assignment, and "grading" an assignment involved counting the number of orange objects in each set. After observing the set for a randomly chosen amount of time, participants were asked to input the number of orange objects they thought they saw in each set. Initially unknown to the participant, one set is "calibrated" and will be used to reward the participant based on the accuracy of the grading. This reward, awarded for accurately grading the Calibrated set, is synonymous to the punishment administered by the professor upon improperly grading the Calibrated set.

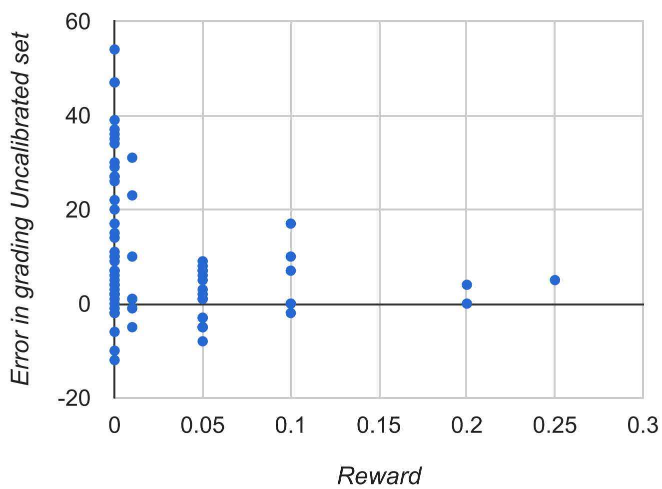

Figure 2: Uncalibrated Error vs Reward

The results of this experiment can be seen in Figure 2, where grading error is plotted against relative magnitude of reward. This shows that a higher reward correlates to lower error. As the reward is administered based on the grader's performance on the Calibrated set, the correlation implies that the graders who perform well on the Calibrated set also perform well on the non-calibrated set. This verifies that the Calibration mechanism indeed works.

Improved Calibration Mechanism
Originally designed with the assumption that students cannot communicate, the Calibration mechanism quickly breaks when student conspire to reveal the calibrated assignment to circumvent grading. The Improved Calibration mechanism mitigates this issue by introducing multiple calibrated papers at the expense of more work, raising the objective score. However, since the work created by this mechanism does not scale well with class size, the Deduction mechanism was developed.

Deduction Mechanism
The Deduction mechanism (Figure 3) achieves a very low benchmark score of 2, with 2 in max work done and a 0 in max error in grade. Incapable graders will raise the benchmark score, as they issue refutations that add work to the professor.

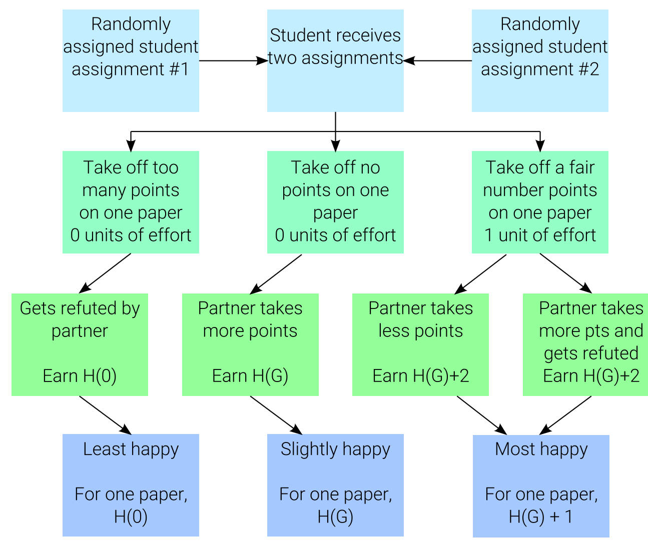

Figure 3: A flowchart of the Deduction Mechanism from the student perspective.

Results

The comparison of existing solutions and those proposed in this work can be seen in Figure 4. Each mechanism will be explained below.

Traditional Professor Grading involves one professor grading all assignments. In an online class of 1000 students, this method is extremely inefficient.

In Traditional Peer Grading, each student grades another's assignment without any supervision. However, as in this mechanism, the objective score is high due to the potential error caused by lack of motivation to grade properly.

Although without requiring effort from either professor or student, Traditional Automated Grading of open-ended responses are still under heavy research. Current solutions are quite preliminary, though can arrive at a grade within approximately 25 percent [4].

The Calibration Mechanism requires one calibrated paper from the professor and two papers graded by each student. The objective score is raised to 4 instead of 2 because students are incentivized by increasing their grade, thus sacrificing accuracy.

The Improved Calibration Mechanism requires each student to grade a subset of the other student's assignments, and the teacher to grade another subset. This mechanism addresses the flaw in the Calibration Mechanism that occurs when students can communicate, at the expense of more work. Thus, leading to poor scalability.

The Deduction Mechanism rewards graders for grading more harshly than their peers, and relies on a voting system to reject grades that are below the expected grade to be reviewed by the professor. In the dominant strategy behavior of the system, no grades should be rejected, leading to no work for the professor. Again, incentive given to the students comes at the cost of accuracy, raising the objective score by two points.

Overall, our Calibration and Deduction mechanisms vastly outperform existing solutions with the exception of Improved Calibration.

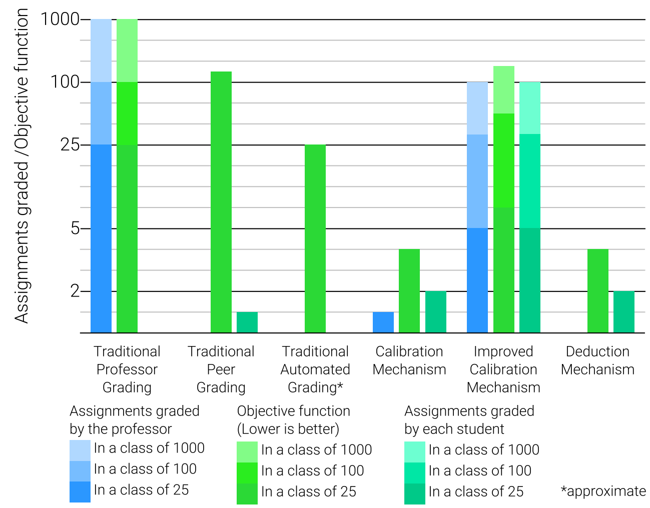

Figure 4: A comparison of mechanisms in terms of effectiveness and scalability.

Conclusion

In this paper, a student model was first created - a set of assumptions that approximate the realistic behavior of students. Based on this model, various grading mechanisms were developed: Calibration, Improved Calibration, and Deduction. These mechanisms incentivize students to grade accurately and efficiently as proven by game theory. The Calibration mechanism was tested with a crowd-sourced experiment, showing that it could work in practice. The student model can easily be reused and improved upon by future researchers who wish to develop more efficient solutions to more realistic scenarios. Mechanism efficiency can be measured in terms of benchmark defined in this paper. The benchmark is a numerical score encompassing both the accuracy of grades and the effort spent by any one person. The inclusion of effort spent by any one person in the benchmark enables the grading system scalability to be taken into account. To the best of our knowledge, these are the first game-theory-based peer-grading mechanisms.

Future Work

The further improvement and development of mechanisms may involve adding more realistic assumptions to the student model, which in turn may require more complex mechanisms. For example, a more complex mechanism may be required to generate accurate grades from incompetent graders. Of course, more testing and validation of mechanisms with crowd-sourced experiments is necessary. Eventually, new mechanisms based off the student model or existing mechanisms could be implemented in commercial MOOCs such as EdX or Coursera.

Acknowledgments

We would like to thank MIT and the MIT PRIMES program for providing the research opportunity, as well as our parents for their support. Of course, none of this is possible without the original idea proposed by our advisor, Constantinos Daskalakis.

References

[1] Bates, T. What's right and what's wrong about coursera-style moocs. *Online Learning and Distance Eductation Resources* (2012).

[2] Budryk, Z. Dangerous curves. *Inside Higher Ed* (2013).

[3] Daniel, S. J. Making sense of moocs: Musings in a maze of myth, paradox and possibility. *Journal of Interactive Media in Education* (2012).

[4] Herrington, A., and Moran, C. Writing to a machine is not writing at all. *National Writing Project* (2012).

[5] Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V., Eds. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[6] Odom, L. A swot analysis of the potential impact of moocs. *EdITLib* (2013).

[7] Randall, E. Edx now has software to grade your essays. *Boston Magazine* (2013).

[8] S, R. Can moocs and existing e-learning paradigms help reduce college costs? *International Journal of Technology in Teaching and Learning* (2012), 21–32.

[9] Theophrastus. The problems with moocs 1: Robo essay-grading. *BLT - Bible*Literature*Translation* (2013).

[10] Watters, A. The problems with peer grading in coursera. *Inside Higher Ed* (2012).