Teaching a Class to Grade Itself using Game Theory

Nick Kaashoek and William Wu

September 8, 2013

1 Introduction

Over the past few years, there has been a tremendous increase in the popularity of MOOCs (Massive Open Online Courses), as well the importance of MOOCs 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 [1]. However, the MOOC's main importance comes from its scalability. MOOCs are able to educate massive numbers of students from anywhere in the world [2]: By the end of 2012, 1.7 million students had attended a course through Coursera [3]. The sheer number of students leads to high student-professor ratios that can reach 150,000:1 in some courses.

High student-professor ratios lead to problems for professors. Professors 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. Automated grading relies on machines to grade assignments. However, machines can only check certain types of answers (i.e. multiple choice), severely limiting the depth of the questions asked [4]. Even though automated grading for written essays is an active area of research with much recent progress, a usable solution does not yet exist [8]. On the other hand, peer grading can grade any type of question, but a system utilizing it could easily be "hacked" by the students [2]. Although both of these systems are interesting concepts, they are limiting or inaccurate in practice [9].

The problem of grading large numbers of submissions is important to solve. Without a solid solution, MOOCs cannot provide efficient feedback for their students, rendering them unable to effectively evaluate their knowledge. Because this limits their learning abilities, it is imperative to create an efficient system that enables students to receive feedback to learn efficiently.

There have been many attempts at solving such an important problem [6] [7]. However, this work introduces what previous attempts lack: a rigorous mathematical analysis of the system using tools from game theory and mechanism design. This work sets a concrete set of constraints that a grading system must satisfy to mitigate student incentive to cheat, as well as a concrete benchmark to analyze the efficiency of various systems. Together, these create a concrete measure for evaluating peer grading systems in terms of efficiency (how much time the professor and students spend grading) and fairness for the students (how accurate their

final scores are). Having this measurement model is a large contribution in itself, because it allows comparison between different systems. 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. 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 [5].

Game theory allows the creation of a unified system of assumptions that simulate student behavior in a class. The grading system can be viewed as a "game" where the set of "rules" are known to everyone and the "players" — the students — "play" the game in order to get the best possible result. Using the assumptions and rules, game theory can predict how students will behave when they "play". This way, mechanism efficiency and effectiveness cam be proved, allowing the evaluation of different sets of rules.

In addition to modeling grading systems, this project develops a mechanism in which students accurately grade the assignments of other students. Students are assumed to be self-interested; they care about their own grades, not the grades of their peers. With this, the problem becomes how to find the perfect incentive for students to spend time to grade assignments of others. Incentivizing students seems simple at first: let the professor give bonus points to students who grade, for example. However, this may not return anything useful because a clever student will just take the points and assign a generic grade (i.e. they will just give every assignment a 100). The next step would be to implement simple checking: e.g. have each paper be graded by two students who both receive bonus points if their grades match. This case may work if the students are isolated, but clever students will still just give all papers 100. This way, their papers will match those of other clever students, giving them points without grading. These examples show that properly incentivizing graders can be a difficult task.

As demonstrated by the above examples, the complexity of the problem lies in the complex behavior of selfish human beings. Because human nature dictates that people will do whatever is in their best interest and may try to outsmart the system, coming up with a strict set of rules to encourage the desired behavior can be quite challenging. A mechanism should be made as simple as possible, but the above examples show that a small amount of simplicity must be sacrificed to obtain a system that can't be manipulated.

This research opportunity is made interesting by the complexity of human nature as well as its necessity for practical use. This paper will cover the thought process of models leading up to the final model of grading, as well as explain the final model and its implications. All models and related calculations will be presented. Section 2 describes the mathematical definition of the problem as well as its rules and assumptions. Section 3

2 Materials and Methods

To create the models that are used to predict the behavior of the students, a set process was used each time. To begin with, a set of assumptions is created. These assumptions are used

to explain how students will act in a given situation; for example, one assumption could be that students want good grades. This is a logical assumption that is likely true, as all the assumptions are, and can be used to classify the behavior of the students in a model. The same set of assumptions can be used for multiple models, which will allow us to evaluate different models side by side. To begin with, a simple set of assumptions was created:

1. Students all share a common happiness function, $H(g)$, where g represents the grade the students received, and the output is their happiness. w.l.o.g. $H(0) = 0$
2. Doing work costs one unit of happiness, so after grading W papers, a student's happiness can be expressed as $H(g) - W$
3. Students want to maximize their final happiness, expressed as $H(g) - W$
4. Students only care about the expectation of their final happiness, which, put in game theory terms, makes them risk neutral (i.e, if $H(x) = 10$ and $H(y) = 5$, then getting x and y each with 50 percent chance, gives a happiness of 7.5).

The next step is to create a system that will allow multiple different models to be compared and evaluated side by side; in other words, a benchmark needs to be created. To do this, two things need to be addressed: how accurate the grades the mechanism produces are, and how happy people are. Because happiness can be directly expressed as a function of work, the two things that need to be evaluated are how accurate the grades are, and how much work is being done. As such, the benchmark becomes a sum of the amount of work done by the person doing the most work and the maximum deviation from the correct grades; the lower the sum, the more efficient the mechanism.

Being able to evaluate two mechanisms side by side is key to the success of the problem, because it allows us to definitively determine whether one mechanism is better than the other. In the benchmark, the two important factors are both necessary to consider, as without one or the other, it would become impossible to create an efficient mechanism. Without caring about the workload, a model could be considered efficient if it simply asked the professor to grade every paper, because the grades would be very accurate, but a mechanism like this is clearly inefficient. Also, without caring about accuracy, not grading the papers becomes an efficient design.

Once a simple set of assumptions and a benchmark were created, the assumptions need to be further developed into a true model that will become the "rules" of the game. To do this, a model was created that will simulate the environment of a classroom, based on the observations of students inside real classes. This turned the list of assumptions from the simple into:

1. The objective score of an assignment is from 0 to 100
2. The objective score assigned by the professor is the correct objective score
3. Grading is binary, a grader can either grade or not grade, *i.e* there is no way to use 0.5 units of effort
4. Not everyone has the ability to get the correct objective score

5. Every assignment costs 1 unit of effort, no matter who grades or how difficult the assignment
6. Students want good grades
7. People don't like doing work
8. Students are selfish and don't care about their friends
9. Students all share a common happiness function, $H(g)$, where g represents the grade the students received, and the output is their happiness. w.l.o.g. $H(0) = 0$
10. Doing work costs one unit of happiness, so after grading a paper, a student's happiness can be expressed as $H(g) - W$,W is number of work units used
11. Everyone is risk neutral with respect to happiness
12. All students can communicate with all other students

These assumptions provide the necessary environment to begin creating a mechanism, as they state everything that is necessary to understand how students behave. Simply put, every assignment has a grade from 1-100, and whatever the professor says goes. Every student has a different capacity for grading, and as such not everyone will be able to correctly grade very assignment, but they can only choose to either grade or not grade. All students are capable of communicating with each other, and will do so, and the rest is from the simple set of assumptions.

Because it would be illogical to start creating a mechanism to encompass all of these assumptions, every mechanism grows in complexity, with the first one being the most simple. Throughout the design process, the assumptions were changed to allow for simplicity, which allows for the creation of simpler models. The complex set of assumptions were always kept in mind, however.

The next step in creating a mechanism

3 The Problem

In a given class of n students and 1 professor, students must submit assignments, each of which have grades on a scale of 1 to 100. All of these assignments must be graded by someone, and no one can grade their own assignment. Our goal is to come up with a scheme that minimizes the maximum amount of work done by a single person, as well as the maximum deviation between any student's given grade and the grade their assignment deserves.

[Note from Matt and Christos: Add some discussion about the benchmark: why do you care about workload? Why do you care about accurate grades? If you just care about workload, give everyone 100. If you just care about grades, have the professor grade everything. Need to consider both to properly model the problem.]

3.1 Rules

Grading one of the assignments costs 1 unit of effort, and will always find the correct score, *i.e* the *objective score*, and grading is the only way to find this objective score.

3.2 Assumptions

These are the assumptions that were made in creating this model.

1) The objective score is from 0 to 100

2) Grading is binary, a grader can either grade or not grade, *i.e* there is no way to use 0.5 units of effort

3) Everyone, including the professor, has the ability to get the correct objective score

4) Every assignment costs 1 unit of effort, no matter who grades or how difficult the assignment

5) Students want good grades

6) People don't like doing work

7) Students are selfish and don't care about their friends

8) Students all share a common happiness function, $H(g)$, where g represents the grade the students received, and the output is their happiness. w.l.o.g. $H(0) = 0$

9) Doing work costs one unit of happiness, so after grading a paper, a student's happiness can be expressed as $H(g) - W$,W is number of work units used

10) Everyone is risk neutral with respect to happiness

11) *Students cannot communicate with each other*

4 The Calibration Model

This model is quite unrealistic, but not by such a degree that it is unimaginable. Because of this, it was possible to create a perfect solution if the problem took place in this situation.

4.1 Mechanism

In the model, the professor grades one paper, called X. The professor then distributes all the student's papers randomly among the students, while making sure that no student receives his or her own paper. Every student is also given a copy of X.

The professor then tells the students that "Out of the two papers you were given, one of them has been graded by me, and the other hasn't. If you fail to give the correct grade for the one I graded, then you will receive a 0." The students then go about their business grading the papers, and individually report the grades they gave out to the professor. Then, just like the professor promised, if the students graded the pre-graded, or *calibrated* paper incorrectly, then that grader gets a 0 on the assignment. If they graded correctly, their assigned grade is guaranteed to be at least some minimum value.

4.2 Why it Works

G is the maximum of { some minimum value, the assigned grade }.

The happiness of one student grading one paper is equivalent to

$$\frac{H(G) + H(0)}{2} - 1$$

Because the students have a 0.5 chance of getting a 0, and a one-half chance of getting G.

If they grade both papers, their happiness will be equal to

$$H(G) - 2$$

Because they spent two units of energy, and are guaranteed to get G.

If they grade neither then the happiness will be

$$H(0)$$

Because they automatically get a 0, but spend no energy.

This means that if $H(G) - 2 > H(0)$, then student will grade both papers. So, for the professor to guarantee that the students use effort and obtain the correct grades, he or she would choose a minimum value such that $H(G) > H(0) + 2$.

4.3 Results

The teacher expends one unit of effort, and the students each spend 2, as long as the teacher set the minimum value high enough. Everything above $H^{-1}(2)$ will be graded correctly, while everything below it will receive a score of $H^{-1}(2)$. This means that the maximum deviation is $H^{-1}(2)$, because this is the score that might be given to someone who deserves a 0. Expressed in happiness units, the value is 2. The means that the value of the objective function in this case is 4, a very low value, which shows how effective this solution is.

4.4 Problems

The goal for this model was to make it more realistic while still making it simple to understand, and have a simply solution that would be easy for students to understand. There is now a model that works very well in an unrealistic scenario, so we moved on to a new model.

5 Improved Verification Model

The major change made in this model was one to assumption number 11. Now assumption 11 reads as follows: *All students are able to communicate with each other*

This provides a very different scenario than the original model, because now students will be able to figure out which papers are calibrated. However, it would be better to try and generalize the previous solution and apply it to the new model, which leads to the following scheme.

5.1 Mechanism

The first attempt to apply a solution to this model was a near carbon copy of the previous scheme. The class is arranged in a virtual circle, where every student grades the papers of the k students to the right of them. The professor also grades every *kth* paper starting from some random point in the circle. If a student mis-grades the calibrated paper, they get a 0, otherwise they get the maximum of their objective score and the minimum value from the previous mechanism.

5.2 Why it Works

In this scheme, because of the overlapping papers, the students have no better way of figuring out which papers are calibrated than random guessing, even though they can all communicate with each other. As such, there are 3 options for the students.

1) They can grade 0 papers, giving them a happiness of $H(0) = 0$

2) They can grade all the papers they are assigned, giving them a happiness of $H(g) - k$

3) They can grade i papers giving them a happiness of $(k - i) \div k \cdot H(G) - i$. This means that it is always worse to grade i, because happiness from grading i is equal to i times happiness for grading 0 plus i-k times happiness for grading k.

5.3 Results

Even though the students can communicate with one another, they will have no way to determine which paper is calibrated, and are thus best off grading all the papers.

5.4 Problems

The major problem with this method is the amount of work. The professor has to grade $\frac{n}{k}$ papers, which can be massive in a class of 1000+ students. For this reason a different mechanism was necessary.

6 The Deduction Model

In the new mechanism, a different strategy was created; one that would ideally work better than the previous ones.

6.1 Mechanism

For each assignment, a new assignment is created to evaluate the performance of the graders. Every student receives 2 papers from different people, so that each paper overlaps with one other person, and every paper is distributed twice. The student then chooses how many points to subtract, and writes a justification for each one of these points. Graders are then given a contribution score from 1 to -1. From each of the 2 assignments they grade, they get their points deducted / total points deducted-0.5. Then, if the original writers of the papers can appeal their grade if they are unhappy with it. The professor will grade the paper, and if any of the graders were wrong, they get a 0. To compute the final score, the grader's final grade on the assignment is calculated as follows: 4 points times their contribution score are added or subtracted to H(the student's assigned grade). (Then the student's grade is converted from Happiness to an actual score).

6.2 Why it Works

The grader's have two choices, they can either grade the papers, or not grade them.

If the grader does not grade the paper, what score should he give it?

If they take off any points, the points will surely be refuted by the professor during the regrade. So if they take off any points, they will get a 0 on the grading assignment. If the grader's give it 100, they might get 50 if the other grader also gives it 100. If he gives it anything else, the original grader will get 0.

If the grader does grade the paper, what score should he give it?

If he gives it a lower score, they throw away free points, and if they give it a lower score there is a risk of them getting a 0 because the scores could be refuted. So, giving it the correct score is the best because it guarantees at least a 50.

So assuming that the student's partner gives the paper a 100, which is the best thing they could do if they don't grade, then If they don't grade the paper, they don't use any effort. So the student's happiness will H(the student's grade on the original assignment). Contribution score = 0 If they do grade, they use 1 unit of effort by default. But, the student's happiness will be H(the student's grade on the original assignment) + 3. Contribution score = 1

If the grader's partner gives the paper its true grade: If the grader doesn't grade the assignment, they will give it a 100 and so the grader's contribution score is -1. So their happiness is H(their original grade) - 4. If the grader does grade the assignment, they and their partner gave the same grades, so they get a contribution score of 0. This means that their happiness is H(their original grade).

Therefore, it is in everyone's best interest to honestly grade the paper.

6.3 Results

Assuming that everyone acts in the dominant strategy behavior, the maximum work for teacher is 0, the maximum work done for students is 2. Because everyone honestly grades the papers, the maximum deviation is 0. So, the objective function is 2, which is even better than the first mechanism.

6.4 Problems

Although the result that was produced for the previous model is rather good, it encourages very harsh competition between students, and would make the classroom environment feel very negative. This means that finding a different solution to the same problem was necessary, in which students are placed in positive competition, instead of the negative one formed by trying to find the most mistakes

7 The Flag Model

For the next mechanism, all of the assumptions were kept the same, and simply tried to find a way to better the feeling of negative competition.

7.1 Mechanism

Start by having each paper get graded by two people. Then, the original writer of this paper chooses which of the two grades he / she would like to have, and looks over the justifications for losing any points on both papers. If they see that any of these justifications are incorrect, they flag the paper and send it to the professor. Following this, the paper they chose to have is sent to a third party, who grades the paper, and checks if their grade is close to the one that was given to the paper. If it isn't, the paper is sent to the professor. After everything that needs to be flagged has been, the professor grades all the flagged papers, and rewards people who were correct with a 2 point bonus. If no papers from any group were flagged, every grader in that group gets two points. Final grades are whatever the person chose if there was no flagging, or whatever the professor gave them if there was.

7.2 Why it Works

To find out how efficient this mechanism is, two things needed to be analyzed: 1) Is everyone an honest flagger?

2) Do the graders grade honestly

7.2.1 Flagging

There are two cases for this:

1) They guess a random score and try to flag everything

-In this case, the chance of gaining points is so low that it is negligible

2) They take the time to flag correctly

-In this case they could get 2 points back, so it is the most beneficial

7.2.2 Grading

Once again, there are two cases here: People grading, and people not grading

If people don't bother grading anything, what score should they give the paper?

-100 because anything lower than the actual score will be flagged by the writer, and because of useless justifications, they will lose points. 100 is the only score guaranteed to be greater than or equal to the correct score

Their happiness in this case will be 0, because they use no effort, but also have no chance of gaining points since their paper will definitely get flagged

The happiness of people who honestly grade, on the other hand, will be a +1 gain, because they use 1 unit of effort to grade and they get 2 units back guaranteed. This higher gain means that it is beneficial to grade honestly.

7.3 Results

Because everyone is honest, we can now analyze the score of the mechanism:

Maximum work done by one person: 3 (student who flags one paper and grades 2 others)

Maximum deviation: 1 (1 point gain for being a good grader)

Final score: 4

Although this is not as good as the previous mechanism, it is still an improvement because there is now positive competition instead of negative.

References

[1] Ruth, S. (2012). *Can MOOCs and existing e-learning paradigms help reduce college costs*? International Journal of Technology in Teaching and Learning, 8(1), 21-32.

[2] Author, *Title*. Source, Year.

[3] Author, *Title*. Source, Year.

[4] Author, *Title*. Source, Year.

[5] Author, *Title*. Source, Year.

[6] Author, *Title*. Source, Year.

[7] Author, *Title*. Source, Year.

[8] Herrington, A., & Moran, C. (2012). *Writing to a machine is not writing at all*. In N. Elliot & L. Perelman (Eds.), Writing assessment in the 21st century: Essays in honor of Edward M. White (pp. 219-232). New York, NY: Hampton Press.

[9] Author, *Title*. Source, Year.