Teaching a Class to Grade Itself using Game Theory

William Wu and Nicholaas Kaashoek
Matt Weinberg and Christos Tzamos

Fourth Annual MIT PRIMES Conference
May 18, 2014
Overview
Overview

- Problem
- Model
- Benchmark
- Mechanisms
  - Calibration
  - Deduction
- Experiment
- Conclusion
Problem
Problem

MOOCs - Massive Online Open Courses
Problem

MOOCs - Massive Online Open Courses
150,000:1 Student/professor ratio
Problem

MOOCs - Massive Online Open Courses
150,000:1 Student/professor ratio

Computer grading - Limited by multiple choice
Problem

MOOCs - Massive Online Open Courses
150,000:1 Student/professor ratio

Computer grading - Limited by multiple choice
Peer grading - Hackable by clever students
Model
1) Let \( H \) be a function of a student's grade, returning a student’s happiness, such that \( H(0)=0 \). Happiness is an arbitrary numerical unit.
Model

1) Let \( H \) be a function of a student’s grade, returning a student’s happiness, such that \( H(0)=0 \). Happiness is an arbitrary numerical unit.

2) Students want to maximize their happiness.
Model

1) Let $H$ be a function of a student’s grade, returning a student’s happiness, such that $H(0)=0$. Happiness is an arbitrary numerical unit.

2) Students want to maximize their happiness.

3) Grading an assignment costs 1 happiness.
Model

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

1) Let $H$ be a function of a student’s grade, returning a student’s happiness, such that $H(0)=0$. Happiness is an arbitrary numerical unit.

2) Students want to maximize their happiness.

3) Grading an assignment costs 1 happiness.

4) Happiness is not affected by external factors, such as the grades of peers.

5) Students can communicate with their peers.
Model - New Assumptions
Model - New Assumptions

6) Students are not perfect graders.
Model - New Assumptions

6) Students are not perfect graders.
7) There is no such thing as partial-grading.
Model - New Assumptions

6) Students are not perfect graders.
7) There is no such thing as partial-grading.
8) Students can report their level of uncertainty when they grade. Let this factor be equal to $U$. 
Model - New Assumptions

6) Students are not perfect graders.
7) There is no such thing as partial-grading.
8) Students can report their level of uncertainty when they grade. Let this factor be equal to U.
9) More effort spent in grading lowers uncertainty.
Model - New Assumptions

6) Students are not perfect graders.
7) There is no such thing as partial-grading.
8) Students can report their level of uncertainty when they grade. Let this factor be equal to $U$.
9) More effort spent in grading lowers uncertainty.
10) When a student assigns a grade $G$, the chance of the grade being $N$ off from the actual grade is proportional to $U$. 
Benchmark
A *numerical score* defined by maximum work done by any person + highest possible error in grading.
Benchmark

A *numerical score* defined by maximum work done by any person + highest possible error in grading.

$$\max_{i \geq 1} \{|H(g_i) - H(o_i)|\} + \max_{i \geq 0} \{w_i\}$$
Mechanisms - Calibration

- "Calibrated" assignment pre-graded by professor
- Student receives two assignments
- Randomly assigned student assignment

Grade neither assignment
0 units of work
- Receive 0
  - Earn $H(0)$
  - Least happy
    - Total happiness $H(0)$

"Calibrated" assignment
50% chance
0.5 $H(0)$
- "Calibrated" assignment
50% chance
0.5 $H(G)$
- Slightly happy
  - Total happiness $(H(0)+H(G))/2 - 1$

Student assignment
50% chance
0.5 $H(G)$
- Student assignment
50% chance
0.5 $H(G)$
- Most happy
  - Total happiness $H(G) - 2$

Grade both assignments
2 units of work
- Receive full credit
  - Earn $H(G)$
Mechanisms - Calibration

Max work: 2
Max error: 2
Benchmark Score: 4

"Calibrated" assignment pre-graded by professor → Student receives two assignments → Randomly assigned student assignment

Grade neither assignment → 0 units of work
Grade one assignment → 1 unit of work
Grade both assignments → 2 units of work

Receive 0 → Earn $H(0)$
"Calibrated" assignment 50% chance $0.5 \cdot H(0)$
Student assignment 50% chance $0.5 \cdot H(G)$
Receive full credit → Earn $H(G)$

Least happy → Total happiness $H(0)$
Slightly happy → Total happiness $(H(0)+H(G))/2-1$
Most happy → Total happiness $H(G)-2$
Mechanisms - Calibration

"Calibrated" assignment pre-graded by professor

Grade neither assignment
- 0 units of work

Receive 0
- Earn $H(0)$

Least happy
- Total happiness $H(0)$

Max work: 2

Grade one assignment
- 1 unit of work

"Calibrated" assignment 50% chance 0.5 $H(0)$

Slightly happy
- Total happiness $(H(0)+H(G))/2$-$1$

Most happy
- Total happiness $H(G)$-$2$

Max error: 2

Grade both assignments
- 2 units of work

Student assignment 50% chance 0.5 $H(G)$

Receive full credit
- Earn $H(G)$

What if students can communicate?

Benchmark Score: 4
Mechanisms - Improved Calibration

"Calibrated" assignment pre-graded by professor

Student receives two assignments

Randomly assigned student assignment

Grade neither assignment
- 0 units of work
  - Receive 0
    - Earn $H(0)$
      - Least happy
        - Total happiness $H(0)$

Grade one assignment
- 1 unit of work
  - "Calibrated" assignment
    - 50% chance $H(0)$
    - 50% chance $0.5H(G)$
      - Slightly happy
        - Total happiness $(H(0) + 0.5H(G))/2 - 1$

Grade both assignments
- 2 units of work
  - Student assignment
    - 50% chance $H(G)$
      - Most happy
        - Total happiness $H(G) - 2$
  - Receive full credit
    - Earn $H(G)$
Mechanisms - Improved Calibration

Assumption added:
5) Students can communicate
Assumption added:
5) Students can communicate “Improved” with multiple calibrated assignments
Mechanisms - Deduction

Randomly assigned student assignment #1 → Student receives two assignments → Randomly assigned student assignment #2

- Take off too many points on one paper 0 units of effort
  - Gets refuted by partner
    - Earn H(0)
    - Least happy For one paper, H(0)

- Take off no points on one paper 0 units of effort
  - Partner takes more points
    - Earn H(G)
    - Slightly happy For one paper, H(G)

- Take off a fair number points on one paper 1 unit of effort
  - Partner takes less points
    - Earn H(G)+2
    - Most happy For one paper, H(G) + 1
  - Partner takes more pts and gets refuted
    - Earn H(G)+2
Mechanisms - Deduction

Max work: 2
Max error: 0
Benchmark Score: 2
Mechanisms - Deduction

Max work: 2
Max error: 0
Benchmark Score: 2

Unfriendly competition
Mechanisms - Comparison

- Traditional Professor Grading
- Traditional Peer Grading
- Traditional Automated Grading
- Calibration Mechanism
- Improved Calibration Mechanism
- Deduction Mechanism

Assignments graded / Objective function:

- Assignments graded by the professor
  - In a class of 1000
  - In a class of 100
  - In a class of 25

- Objective function (Lower is better)
  - In a class of 1000
  - In a class of 100
  - In a class of 25

- Assignments graded by each student
  - In a class of 1000
  - In a class of 100
  - In a class of 25

*approximate
Mechanisms - Comparison

Calibration and Deduction outperform existing mechanisms
Experiment

Online, crowdsourced, and anonymous
Experiment

Online, crowdsourced, and anonymous
Designed to validate Calibration Mechanism:
Experiment

Online, crowdsourced, and anonymous

Designed to validate Calibration Mechanism:

Presented two assignments to grade,
Rewarded on one assignment
Online, crowdsourced, and anonymous
Designed to validate Calibration Mechanism:

* Presented two assignments to grade,
* Rewarded on one assignment

Assignment - A set of “marbles”
Grading - Counting the orange “marbles”
Experiment - Screenshot
## Experiment - Reward

<table>
<thead>
<tr>
<th>Confidence</th>
<th>Within</th>
<th>Reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 marble</td>
<td>$0.25</td>
</tr>
<tr>
<td>2</td>
<td>2 marbles</td>
<td>$0.20</td>
</tr>
<tr>
<td>5</td>
<td>5 marbles</td>
<td>$0.10</td>
</tr>
<tr>
<td>10</td>
<td>10 marbles</td>
<td>$0.05</td>
</tr>
<tr>
<td>20</td>
<td>20 marbles</td>
<td>$0.01</td>
</tr>
<tr>
<td>Confidence</td>
<td>Within</td>
<td>Reward</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>--------</td>
</tr>
<tr>
<td>1</td>
<td>1 marble</td>
<td>$0.25</td>
</tr>
<tr>
<td>2</td>
<td>2 marbles</td>
<td>$0.20</td>
</tr>
<tr>
<td>5</td>
<td>5 marbles</td>
<td>$0.10</td>
</tr>
<tr>
<td>10</td>
<td>10 marbles</td>
<td>$0.05</td>
</tr>
<tr>
<td>20</td>
<td>20 marbles</td>
<td>$0.01</td>
</tr>
</tbody>
</table>

Reward is based on the reported confidence and the accuracy of the reported guess.
Experiment - Data

- Error in grading Uncalibrated set vs Reward
- Max absolute error in grading Calibrated set vs Absolute error in grading
Experiment - Data

1. Greater reward $\rightarrow$ lower uncalibrated error
1. Greater reward $\rightarrow$ lower uncalibrated error
2. Calibrated set indicates grading proficiency
Conclusion
Conclusion

- Student model - approximations for student behavior
Conclusion

- Student model - approximations for student behavior
- Benchmark - score measuring efficiency and workload of various mechanisms
Conclusion

● Student model - approximations for student behavior
● Benchmark - score measuring efficiency and workload of various mechanisms
● Calibration, Improved Calibration, and Deduction mechanisms developed
Conclusion

- Student model - approximations for student behavior
- Benchmark - score measuring efficiency and workload of various mechanisms
- Calibration, Improved Calibration, and Deduction mechanisms developed
- Calibration validated by a crowdsourced experiment
Conclusion

- Student model - approximations for student behavior
- Benchmark - score measuring efficiency and workload of various mechanisms
- Calibration, Improved Calibration, and Deduction mechanisms developed
- Calibration validated by a crowdsourced experiment
- Calibration and Deduction mechanisms outperform existing grading solutions
Conclusion - Next Steps
Conclusion - Next Steps

- Improving realism - producing accurate grades from incompetent graders
Conclusion - Next Steps

● Improving realism - producing accurate grades from incompetent graders
  ○ Proficiency test
  ○ Using multiple graders to reduce error
Conclusion - Next Steps

● Improving realism - producing accurate grades from incompetent graders
  ○ Proficiency test
  ○ Using multiple graders to reduce error

● Implementation
Conclusion - Next Steps

● Improving realism - producing accurate grades from incompetent graders
   ○ Proficiency test
   ○ Using multiple graders to reduce error

● Implementation
   ○ User testing with Mechanical Turk
   ○ Eventually in Coursera / EdX
Acknowledgements

MIT
MIT PRIMES program,
Slava Gerovitch, Tanya Khovanova, Srini Devadas
Mentors, Matt Weinberg and Christos Tzamos
Professor, Costis Daskalakis
Parents