Using Multiple Accounts for Harvesting Solutions in MOOCs

José A. Ruipérez Valiente\textsuperscript{b,c,*} @JoseARuiperez
Giora Alexandron\textsuperscript{a,*}
Zhongzhou Chen\textsuperscript{a}
David E. Pritchard\textsuperscript{a}
Contact: \{jruipere, giora, zchen22, dpritch\}@mit.edu

\textsuperscript{a} Massachusetts Institute of Technology
\textsuperscript{b} Universidad Carlos III de Madrid
\textsuperscript{c} IMDEA Networks Institute
* Equal contribution to this work
Overview

- Identify students who use fake accounts for harvesting solutions that they later submit in their main account.
  - **CAMEO**: Copying Answers using Multiple Existence Online (Northcutt, Ho, and Chuang, 2015)

- Main goals:
  - Study the amount of CAMEO, the motivation for using it, and how it can be reduced

- Course studied: MITx 8.0MReV Introductory Physics on edX

- Some of the results:
  - ~3% of the certificated students used this method in more than 50% of their correct submissions
  - More cheating on high stake questions
  - Less cheating on randomized questions and when feedback is limited

- Importance of the research:
  - Threat to the value of the certificates
  - Interfere with educational research

- Closely related work:
  - Related to academic dishonesty (students break edX’s code of Honor) and gaming the system (exploit properties of system)
  - CAMEO in MITx/HarvardX courses (Northcutt, Ho, and Chuang, 2015)
  - Online homework tutor system (Palazzo, Lee, Warnakulasooriya, and Pritchard, 2010)
Methodology

- **What we detect:** Cheating using multiple accounts
  - *Harvesting* account is used to collect correct solutions (using ‘show answer’ or exhaustive search)
  - *Master* account submits these solutions
- **Methodology:** Educational data mining on tracking logs
- **Algorithm:**
  1. Collect all events *(harvester, master, q)* with the properties:
     - *Harvester* gets the solution for *q*, then *master* submits it
     - *Harvester* and *master* share IP group
     - Delay between events < 24h
  2. Apply criteria:
     - *Master* never behaves as *harvester*, and vice versa
     - *Harvester* works for others:
       - Most of submissions are actually used by the master
       - Does not earn a certificate
     - *Master* harvests at least 10 questions
  3. Remove events of users who don’t adhere to these criteria

→ We made a manual verification of a sample
Findings: Amount of CAMEO

- Population of the analysis:
  - Students who completed at least 5% of the questions (1581 students)
  - 502 certificate earners

<table>
<thead>
<tr>
<th></th>
<th># master accounts</th>
<th># harvester accounts</th>
<th># harvested answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>All students</td>
<td>99 (6.3%)</td>
<td>112</td>
<td>19602 (3%)</td>
</tr>
<tr>
<td>Certificate earners</td>
<td>52 (10.3%)</td>
<td>63</td>
<td>12396 (3%)</td>
</tr>
<tr>
<td>Non-Certificatees</td>
<td>47 (4.35%)</td>
<td>49</td>
<td>7206 (3%)</td>
</tr>
</tbody>
</table>

2% of the c"e obtained 70% of correct answers using CAMEO
Findings: Characteristics of CAMEO events

- Harvesting technique: 53.5% using ‘show answer’ and 46.5% using exhaustive search
- Harvesting precedes first master answer in 91% of the cases
- Delay between the harvesting event and the submission in the master account:
  - 90% events below 5 min
  - Median 27 seconds
  - Mode 5 seconds
Findings: Cheaters vs. other students

Density distribution. Performance per type of account.

ANOVA (F=72.43, p < 0.001)

Scatterplot. Response time vs. performance
Findings: Certificate vs. non-certificate earners

- Comparison of cheating over time
- Non-certificate earners tend to cheat more \(\rightarrow\) Cheat for a certificate
- Drop in chapter 9+10 due to certification peak
- Different behaviors depending the student

85% of certificates were earned within this chunk!
Findings: Reduction methods

- Limiting feedback reduces harvesting
  - Our findings suggest that questions with delayed feedback are cheated 3x less than the rest

- Randomization reduces harvesting
  - Decreases harvested answers about 2x
  - Statistically significant per student and per problem
Discussion and conclusions

- **Motivation? Certificate!**
  - Master’s do not try to solve the problem first (no learning?)
  - High-stake questions more likely to be cheated
  - Change behavior after receiving certificate
  - Non-certificate earners start with the intention of ‘cheat for certificate’

- **Problematic issue**
  - 10% students harvested around 1% of their solutions
  - Relative large number of students cheating in introductory physics
  - It might increase if MOOC certificates become significant for industry

- **Implications**
  - Decrease confidence in the assessment → Affect the perceived value certificates
  - Can interfere with MOOC research
  - Can apply to other online learning environments as well
Discussion and conclusions

 Remedies
  ▪ Problems using randomized variables
  ▪ Delaying (or removing) feedback on high-stake questions

 Limitations
  ▪ Internal validity: No definite evidence that what we detect is CAMEO, or a tagged sample (‘training set’).
    • Too high? (4x more CAMEO users than Northcutt et al. found in our course)
    • Too low? (very strict criteria; findings are low comparing to general literature)
  ▪ Generalizability: We studied only one course. How typical is it?

 Future research
  ▪ More courses broadening the understanding of the phenomenon
  ▪ Detection without relying on IP
  ▪ Run-time detection [by the platform itself]
  ▪ Other ways of copying
Thank You!

José A. Ruipérez Valiente\textsuperscript{b,c,*} @JoseARuiperez
Giora Alexandron\textsuperscript{a,*}
Zhongzhou Chen\textsuperscript{a}
David E. Pritchard\textsuperscript{a}
Contact: \{jruipere, giora, zchen22, dpritch\}@mit.edu

\textsuperscript{a} Massachusetts Institute of Technology
\textsuperscript{b} Universidad Carlos III de Madrid
\textsuperscript{c} IMDEA Networks Institute
* Equal contribution to this work
Appendix 1: Example pattern

Harvester

Master
Appendix 2: Example of individual profiles

Archetypal Patterns of Students' Behaviour across Chapters

Student A - Certificate earner

Student B - Certificate earner

Student C - Certificate earner

Student D - Non-Certificate earner

Submission type:
- All Correct
- Harvested