MIT Libraries | DSpace@MIT

MIT Open Access Articles

Using Multiple Accounts for Harvesting Solutions in MOOCs

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

Citation: Ruiperez-Valiente, Jose A., Giora Alexandron, Zhongzhou Chen, and David E. Pritchard. "Using Multiple Accounts for Harvesting Solutions in MOOCs." Third Annual ACM Conference on Learning at Scale (April 2016).

As Published: http://learningatscale.acm.org/las2016/accepted-papers/

Publisher: Association for Computing Machinery (ACM)

Persistent URL: http://hdl.handle.net/1721.1/101136

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike

Massachusetts Institute of Technology

Using Multiple Accounts for Harvesting Solutions in MOOCs

Jose A. Ruiperez-Valiente[a,b,c,*], **Giora Alexandron**[a,*], **Zhongzhou Chen**[a], **David E. Pritchard**[a]
[a] Massachusetts Institute of Technology, Massachusetts Ave 77, 02139 Cambridge (MA) USA
[b] Universidad Carlos III de Madrid, Av. Universidad 30, 28911 Leganés (Madrid) Spain
[c] IMDEA Networks Institute, Av. del Mar Mediterráneo 22, 28918 Leganés (Madrid) Spain
{jruipere, giora, zchen22, dpritch}@mit.edu

ABSTRACT
The study presented in this paper deals with copying answers in MOOCs. Our findings show that a significant fraction of the certificate earners in the course that we studied have used what we call *harvesting* accounts to find correct answers that they later submitted in their main account, the account for which they earned a certificate. In total, ∼2.5% of the users who earned a certificate in the course obtained the majority of their points by using this method, and ∼10% of them used it to some extent. This paper has two main goals. The first is to define the phenomenon and demonstrate its severity. The second is characterizing key factors within the course that affect it, and suggesting possible remedies that are likely to decrease the amount of cheating. The immediate implication of this study is to MOOCs. However, we believe that the results generalize beyond MOOCs, since this strategy can be used in any learning environments that do not identify all registrants.

Author Keywords
Academic dishonesty; educational data mining; learning analytics; MOOCs

ACM Classification Keywords
H.2.8. **[Database Management]**: Database Applications - *Data mining*; J.1. **[Computer Applications]**: Administrative Data Processing - *Education*; K.3. **[Computing Milieux]**: Computers and Education - *Computer and Information Science Education: Information systems education*

INTRODUCTION
The issue of academic dishonesty in higher education has been studied for at least half a century. According to McCabe and Trevino [4], studies report that between "13 to 95 percent of college students engage in some form of academic dishonesty" (p. 523). The survey study by Davis [3] reported cheating rates of up to 88%. The specific numbers depend on how one defines and measures academic dishonesty, but the overall picture is that of a severe, widely spread phenomenon. Singhal [9] described cheating as one of the major problems in education today.

Massive Open Online Courses (MOOCs) offer free access to complete courses often containing both lectures and problems. Students who demonstrate a satisfactory level of achievement can earn a certificate of accomplishment. While these certificates currently do not have formal value, the fact that many people tend to include them in resumes and professional social networks indicates that they are perceived by some as an indication of knowledge and proficiency. There is also evidence that they have value in the job market[1]. Another use of MOOC certificates is in academic admissions, e.g. Wharton Business School announced that it will use its MOOCs as an additional tool for selecting MBA candidates [2]. This can explain, at least partially, the motivation of some students looking for easy ways to earn MOOC certificates.

In this study we detect the use of multiple accounts by the same student for copying answers. The method works as follows. The student uses one or more harvesting accounts (the *harvester/s*) to obtain the correct answer, and then submits it in the *master* account, the account for which the student intends to earn a certificate. Finding the answer in the harvester account can be done either by asking to see the correct answer after using all the attempts ('show answer' on the edX platform), or by exhaustive search (e.g. pure guessing for multiple choice questions) until the correct answer is found. This is a clear case of academic dishonesty since it explicitly violates the edX honor code (which all registrants agree to) that requires that users "Maintain only one user account", and "Not engage in any activity that would dishonestly improve my results"[2]. The findings that we present below show that this phenomenon is quite widespread. For example, about 5% of the certificate earners in our course harvested at least 10% of their correct answers; 2% used it for more than 70% of their correct answers.

The phenomenon of using harvesting accounts was also reported by Northcutt, Ho, and Chuang, who examined multi-

* The two first authors have contributed equally to this work

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
L@S 2016, April 25-26, 2016, Edinburgh, Scotland Uk
© 2016 ACM. ISBN 978-1-4503-3726-7/16/04 $15.00.
http://dx.doi.org/10.1145/2876034.2876037

[1] http://www.pcworld.com/article/2071060/employers-receptive-to-hiring-it-job-candidates-with-mooc-educations.html
[2] https://www.edx.org/edx-terms-service

ple courses offered by MITx and HarvardX [6]. They coined the term CAMEO (Copying Answers using Multiple Existence Online) for this phenomenon, and we adopt this term. In the context of cheating in digital learning environments, another closely related work is the one by Palazzo et al. [8]. They studied the amount of copying answers from another student using the same online homework tutor, and found that the fraction of copied answers ranged between 3 and 11%. CAMEO also resembles what Baker et al. [1] defined as *gaming the system* – "Attempting to succeed in an interactive learning environment by exploiting properties of the system rather than by learning the material" – both in terms of the motivation (improving grades), and in terms of the method (exploiting technical features of the system).

The rest of the paper is organized as follows. In the next section we define the phenomenon and the detection method. In the Results section we present the results of applying the method on a specific course. In the Discussion section we discuss the findings, followed by summary and main conclusions in the last section.

HARVESTING SOLUTIONS USING MULTIPLE ACCOUNTS

This section is organized as follows. First, we define the phenomenon. Second, we state the criteria that we use for identifying an answer as having been harvested. Third, we define behavioral patterns. Fourth, we give a high-level description of the algorithm that implements the patterns.

Defining the phenomenon

CAMEO refers to an event in which a user uses one or more accounts to find the correct answer to a question, and then submits this answer in his/her main account. We refer to the account(s) used for finding the solution as the harvesting accounts (or the *harvester/s*), and to the account in which the answer is submitted as the *master* account. A CAMEO event can thus be described as a triplet of the form *<master, harvester, question>*. The criteria for identifying CAMEO events are described below.

Criteria

Each triplet must also adhere to the following criteria:

1. ***Harvester and master belong to the same IP group.*** IP group is a group of accounts that shared the same IP at least once in the course, or are connected through an account with whom both shared an IP (this criterion is applied recursively). It is defined as follows. Let *G=(U, I, E)* be a bipartite graph in which *U* represents the set of the users, *I* represents the set of the IPs, and *E* are the edges between *U* and *I*, when an edge *(u,i)* denotes that user *u* has used IP *i* at some point in the course. For each *connected component cc* within *G* (a connected component is a subgraph in which there is a path between each two nodes), then the nodes of *cc* that belong to *U* (the 'user' nodes) form an IP group. Identifying connected components in a graph is a basic problem in graph theory, and can be computed in linear time.

2. ***Harvester does not have extrinsic motivation.*** This is operationalized as harvester does not earn a certificate. The rationale is clear - earning a certificate indicates that the user seeks a reward from time invested in this account, reducing the likelihood that this is a 'service' account. We note that this requirement may ignore 'heavy' harvesting accounts, i.e., accounts that accumulate enough course credits to earn a certificate solely due to extensive harvesting.

3. ***Harvested questions are actually used in the master account.*** The rationale that underlies this criterion is that if a student establishes an account to harvest answers, then most of the questions done by this account will be used by a master account. Thus, we require that at least 85% of the solutions obtained by the harvester were actually used by a master. We picked 85%, and not 100%, to allow for some flexibility, for example in case that the user solves a question in the harvester account, but then decides to skip it in the master account because he/she realizes that it is a random question (thus the solution in the master account can be different; we elaborate on the effect of randomization below).

4. ***Master harvests at least 10 answers.*** The intention is to increase our certainty that the student used this pattern to harvest answers; therefore this filter adds a minimum count of harvested answers to consider the student as using the pattern. This threshold represents about 1% of the questions required for a certificate.

5. ***Master account is not used for harvesting solutions.*** The rationale is similar to the one in criterion 2, in the opposite direction.

Harvesting patterns

We define two CAMEO patterns: *Immediate*, and *delayed batch*. In both patterns, a harvesting event is composed of the harvester getting the correct answer in his/her account either using 'show answer' or by using exhaustive search, followed by the master submitting the correct answer in his/her account. The harvester and master should also adhere to the criteria listed above (the Criteria subsection). The patterns are described below. A graphical representation of them is given in Figure 1.

Immediate mode. This pattern refers to a situation in which the user harvests the solution in the harvesting account, and inserts it in the master account shortly after. The threshold that we use is of up to 15 minutes between the two events. The main hypothesized modus operandi behind this pattern is that of a user who is progressing in the two accounts simultaneously, using two browsers (in order to login with two accounts at the same time). In this case we require that both actions were done by the same IP address.

Delayed batch mode. This pattern refers to a situation in which the user harvests the solution for multiple answers, and then submits them in a rapid sequence in the master account. The main hypothesized modus operandi behind this pattern is of a user who harvests multiple solutions, stores them, and later submits them in a batch, possibly in a different physical location and/or using a different machine, but one in the same

IP group as explained in the first criterion. The threshold that we use for the length of the sequence is of (at least) 10 correct submissions, and the threshold that we use for 'rapid' is less than 20 seconds between two consecutive ones. To avoid overlap with the *immediate* mode, we also require that there will be at least 15 minutes gap between the last harvesting event and the first master event in this sequence. We note that cheating events that are part of a batch mode sequence, but with less than a 15 minutes delay, will be caught by the *immediate* mode. So in such cases the event would still be detected.

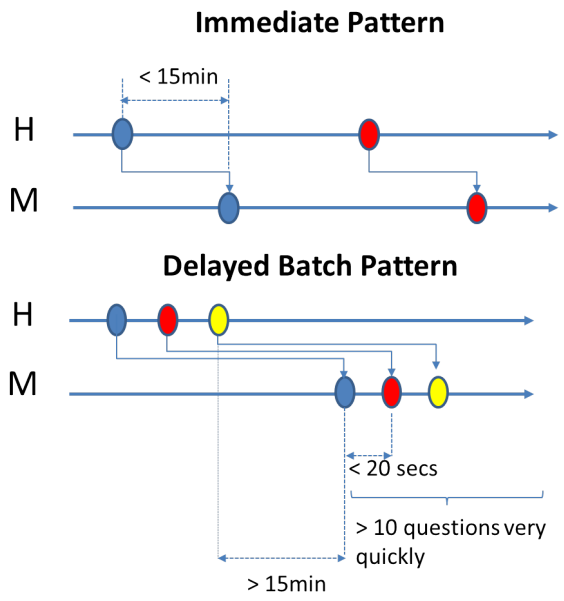

Figure 1. Schematic representation of the two harvesting patterns (H=Harvester, M=Master)

Completeness of the two patterns. We note that the two patterns do not fully cover the range of harvesting possibilities. For example, a sequence consisting of less than ten questions that are harvested and then submitted by the master more than fifteen minutes after they were harvested will not be detected. Also, we have used global thresholds that were not adjusted for specific questions. We deliberately picked very strict definitions so as to follow our general approach that in the study of academic dishonesty, it is better to decrease the false positive ratio at the expense of increasing the false negative one. The implication is that our algorithm is probably not detecting all the harvesting events. Still, the results are quite significant. Thus, we believe that at this stage it is more important to bring this issue to the attention of the community, rather than to polish the patterns in order to identify all of the harvesting events.

Algorithm

Below we give a high-level description of the algorithm. Its input is the tracking logs of the course, processed to create per-user, time-sorted log files. Each event in the tracking log represents a student action, and contains the IP from which the action that this event represents arrived. A submission to a question creates an event that contains, among other things, information about the question (i.e., question ID), the student, the answer, and whether it is correct or not (more information on edX tracking logs can be found on edx.org).

1. Process the user log files to build two dictionaries - one that maps each student to all the IPs used by this student, and one that maps an IP to all the students who used it at least once during the course.

2. For *immediate pattern*: Per user U, for each successful submission to a problem P at time t, search if any student V from that IP group has submitted a correct answer, or a request to see answer, to P in the 15 minutes before t, from ip. If so, add the triplet to the list of harvesting events.

3. For *delayed batch pattern*: Per user U, for each sequence of ten or more successful submissions to questions $Pi...Pk$, with no more than 20 seconds between each consecutive submissions, search if any students V in the IP group of U, got the answer to $Pi...Pk$ in a similar pattern to the one described in 2. If so, add the triplets $<U, V, Pi>...<U, V, Pk>$ to the list of harvesting events.

4. Test for the criteria described in the Criteria subsection. Remove entries that contain items that do not fulfill these criteria.

While we hope that this description can be a sufficient starting point for ones who wish to implement this algorithm, we intend to make the source code of the algorithm publicly available.

RESULTS

In this section we present the results of running our algorithm on the data from edX.org MOOC 8.MReVx given in summer 2014. First, we describe the setting. Second, we present the findings.

Context - Introductory Physics MOOC 8.MReV

Population: The MOOC 8.MReVx was run on edX.org in Summer 2014. It attracted about 13500 registrants, of which 502 earned a certificate. Gender distribution was 83% males, 17% females. Education distribution was 37.7% secondary or less, 34.5% College Degree, and 24.9% Advanced Degree. Geographic distribution includes US (27% of participants), India (18%), UK (3.6%), Brazil (2.8%), and others (total of 152 countries). (All numbers are based on self-reports.)

Course structure: The course covers the standard topics of a college introductory mechanics course with an emphasis on problem solving and concept interrelation. It lasted for 14 weeks, with content divided between 12 mandatory units and two optional ones on advanced topics. The course contains 273 e-text pages, 69 videos, and about 1000 problems. The problems include checkpoints problems embedded within the e-text and videos, and homework and quiz questions which are given at the end of the units.

Tracking logs: The interaction data of the students with the learning environment is saved on edX servers and was downloaded and delivered to us through MITx, according to the

	#Master accounts	#Harvester accounts	#Harvested answers
All students	99 (6.3%)	112	19602 (3.06%)
Certificate earners	52 (10.3%)	63	12396 (3.09%)
Non-certificate earners	47 (4.35%)	49	7206 (3%)

Table 1. Amount of CAMEO in our course

usage agreements between edX, MITx and us. The interaction logs were subdivided by user and sorted by time.

Findings

This subsection presents the results of applying the script that implements the criteria described above on the tracking logs of 8.MReV. For the purpose of this study, we consider only students who completed at least 5% of the questions in the course - a total of 1581 students. Among these, 502 earned a certificate. The output of the CAMEO detection script is a list of such triplets <*master, harvester, question*> labeled as 'immediate' or 'batch'. This list could be then analyzed and combined with other information on the course to obtain various statistics. The findings are divided into 4 subsections. In the first we present an overall view of the amount of cheating in the course. In the second we compare the performance of the master and the harvester accounts with the rest of the students. In the third we give some descriptive statistics about CAMEO events. In the fourth we show findings regarding question characteristics and amount of cheating.

Amount of CAMEO

Number of master and harvester accounts. In total, we identified 99 master accounts, and 112 harvester accounts, which constitute 6.3% of the 1581 who completed at least 5% of the questions in the course. Among the 99 master accounts, 52 were certificate earners (10.3% of the 502 certificate earners) and 47 were non-certificate earners (4.35% of the non certificate earners who completed at least 5% of the questions). We note that some masters operated more than one harvester account. The rationale is probably to enable exhaustive search on questions for which the number of allowed attempts is not enough to find the solution. See Table 1.

CAMEO among submissions. Approximately 3% of all the correct submissions in the course were harvested (19602 out of 639863 correct submissions). See Table 1.

CAMEO among certificate earners. Figure 2 shows the amount of cheating among certificate earners. A point on the graph represents the amount of the certificate earners who harvested at least y% of their correct submissions. For example, 5.37% of certificate earners harvested $\geq$10% of their correct answers. About 2.5% of the certificate earners harvested more than 50% of their correct answers. We judge that to the left of the shoulder at (2, 70%) are students 'harvesting for a certificate' without any attempt to master the required knowledge.

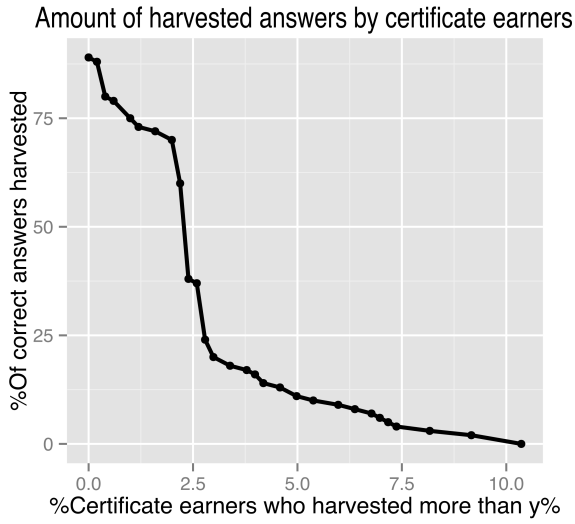

Figure 2. Percentage of cheating by certificate earners

Cheaters compared with other students

Success rate and response time among certificate earners. The findings show that within the certificate earners, masters tend to have a very high average success rate on individual problems, and very fast response time (elapsed time between opening a problem and answering it correctly), compared with the rest of the students. (Here we focus only on certificate earners in order to compare groups of students that answered similar numbers of questions). This is illustrated in Figure 3. The *x*-axis shows the percentage of correct submissions at the first attempt, and the *y*-axis shows the average time needed for a correct submission. Master accounts are marked in red, with the size of the circle proportional to the fraction of answers that they harvested. As can be seen, there is a cluster of red points on the bottom right part of the figure (fast and correct). In general the trend is that the bigger the red circle (more cheating), the higher the success rate and the faster the submission. Also, the top three performers of the course in terms of minimum time and performance at first attempt are cheaters. Additionally, we also found that it was statistically significant that cheaters have better performance in first attempt submissions and require less time for correct submissions.

Distribution of performance among masters, harvesters, and the rest of the students. The distribution of performance among masters, harvesters, and the rest of the students (ones who attempted more than 5% of the questions) is presented in Figure 4. In terms of mean values, masters have a success rate of 82.5%, harvesters have a success rate of 43.9%, and the rest of the students have a success rate of 61.6%. The results of an ANOVA test confirm that the success rate (masters) >success rate (rest of students) >success rate (harvesters). ($F=72.43$, $p=10^{-16}$).

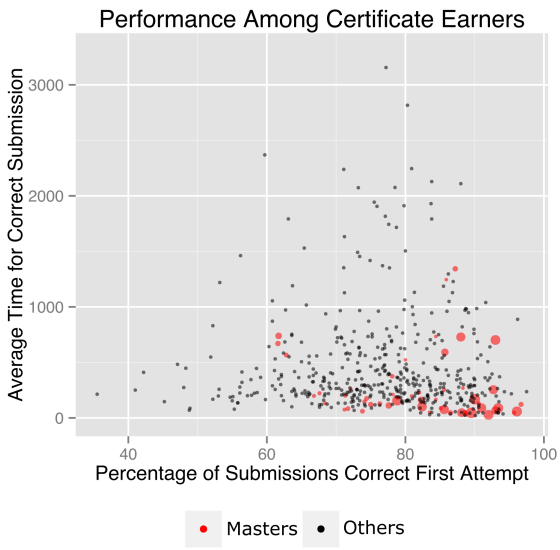

Figure 3. Response time vs. percentage correct for certificate earners; size of red dot indicates amount of cheating

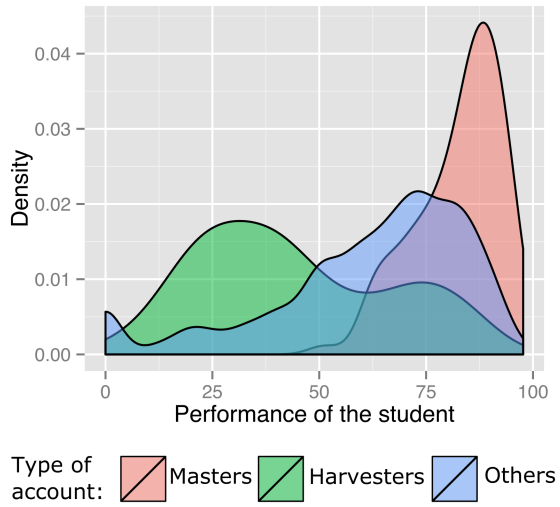

Figure 4. Distribution of performance

Characteristics of CAMEO events

Distribution of harvesting between patterns. Approximately 90% of the harvesting events followed the immediate pattern, and 10% followed the delayed batch pattern.

Harvesting technique. 53.5% of the cheating events were harvested by asking to see the answer ('show answer' button). 46.5% were harvested using exhaustive search.

Harvesting precedes first master answer. 91% of the harvested answers preceded the masters first attempt (which was therefore correct).

Delay of the cheating event. Figure 5 shows the distribution of the elapsed time between the harvesting event and the submission in the master account for events in which this delay was less than 5 minutes (90% of the CAMEO event detected). The median is 27 seconds with mode=5 seconds; 75% of the *immediate* mode submissions were done within 72 seconds after the harvesting event.

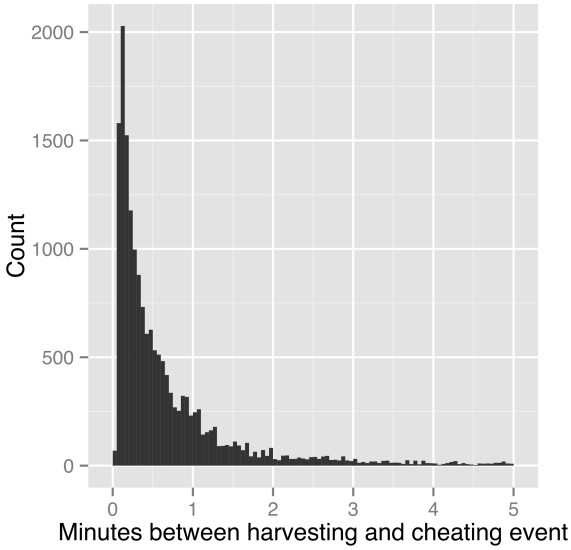

Figure 5. Delay between harvesting and submission: *Immediate* mode

Factors that correlate with cheating

This subsection presents results on the relation between characteristics of problems and the amount of cheating on these problems. The main objective is to focus on aspects that we believe can help instructors design their course in a way that is less vulnerable to CAMEO.

More harvesting on high-stake questions. Our course contains three types of problems - quizzes, homework and checkpoints. Additionally, we also have a midterm and a final exam, not included in the analysis for this subsection. In terms of contribution to the final grade, quiz questions have the highest weight, then homework questions, then checkpoints. The information about the weight of the questions is given in the syllabus of the course, so we can assume that students are aware of it.

We calculated the average percentage of harvested submissions per type of assignment. On average, the fraction of harvested submissions for quizzes, homework and checkpoints is 1.2%, 0.98% and 0.77%, respectively (of the total submissions; we note that the 3% in Table 1 refers to the percentage of *correct* submissions). An ANOVA test confirms that this difference is statistically significant ($F = 3.03$, $p = 0.046$).

Limiting feedback reduces harvesting. We have several places where questions that are similar in terms of weight are accompanied by different levels of feedback regarding the correct answer. The 'show answer' feature was not available before the deadline for quizzes except for one quiz (due to

a mistake in settings). On this specific quiz, harvesting accounted for 3.77% of correct answers, vs. 1.25% on those without show answer.

Similar findings were obtained when looking at the midterm and final exams. On the midterm, with the 'show answer' disabled, the average amount of cheating was 1.2%, while on the final exam, on which 'show answer' was enabled, the average amount of cheating was 2.6%. It might have been higher except that students who harvest a majority of their answers tend to stop out when they reach the passing grade and make little of no effort on the final exam. These results clearly show that limiting the feedback reduces harvesting. That it does not reduce it further reflects that students can still harvest solutions using exhaustive search utilizing the correct/wrong feedback that is always given.

Randomization reduces harvesting. In most of the questions in our course requiring a numerical (vs. algebraic) answer, the question parameters are randomized (for example, the angle of an inclined plane), so different accounts see numerically different variations of the same problem.

We found that there is less cheating on randomized questions, both globally – the total amount of cheating events, and after normalizing by the total number of submissions to the problem.

For randomized questions, the percentage of cheating is 5.84%, while for non-randomized questions, it is 9.23%. In terms of absolute numbers, the average number of submissions that were harvested per problem with randomized variables was 10, whilst for the rest of the questions, it was 20. The normalized results are statistically significant, with p-value under 0.01 for both the normalized and the absolute comparisons. Although this was not a randomized trial, we believe that this indicates a causal relationship, namely, that randomization decreases cheating by making the harvesting more difficult. This is further discussed in the next section.

DISCUSSION

Significant use of CAMEO

The findings that we have presented above reveal that a significant fraction of the certificate earners in our course - almost 10%, used CAMEO to obtain at least $\sim$1% of their correct answers. Moreover, they reveal that $\sim$2.5% of the certificated users obtained most of their points by using this method, typically without even inspecting the questions to see if they might answer them legitimately. Our study is consistent with the hypothesis that the main motivation for this kind of cheating is to earn a certificate. Although the observation that a significant amount of cheating was by non-certificate earners seems to contradict this, we note that the non-certificate earners that used CAMEO tended to drop out from the course early, but till then, actually harvested a higher fraction of their answers than the certificated CAMEO users. Thus our conclusion is that they started the course with the intention of earning a certificate dishonestly, but decided to quit for various reasons (for example, our many questions, or the many randomized questions that thwart CAMEO), maybe even moving their CAMEO efforts to somewhere else. This is also consistent with the finding of Northcut et al. that CAMEO is more prevalent among users who have many certificates [6].

We were surprised by the relatively large number of students who used CAMEO in our course. One would expect that people would be interested in earning a certificate without actually learning something in courses that can be used for gaining external benefits, such as an advantage in the labor market. This is more naturally associated with programming and engineering courses, and less with introductory physics course as ours. So there is no specific reason to believe that the amount of cheating in our course is particularly high.

The high level of CAMEO may well result from the fact that MOOCs share features identified by previous work on academic dishonesty as associated with more cheating. For example, in [5] it was reported that cheating is more likely to happen in large and public institutions (vs. small private ones). According to the analysis of around 80 studies done in [7], Classroom Environment also has a considerable effect, with cheating more associated with bigger classrooms in which there is less individualized attention. Also, according to [4], cheating is restricted by what is perceived to be the social norm.

Altogether, due to the characteristics of MOOCs, of cheating being a phenomenon that tends to spread, and of MOOC certificates becoming more valuable, we can expect to see an increase in the amount of CAMEO in MOOCs, making it a more significant issue.

Implications

The most important finding of our work is that CAMEO is quite widespread. Furthermore, CAMEO is only one form of academic dishonesty: we are not detecting when answers obtained by one student are given to another student, which has been observed to occur in $\sim$ 11% of all answers [8] (vs the 3% CAMEO found here). Nor have we investigated the availability of plagiarizing answers from solutions found on the Internet. So cheating is undoubtedly highly prevalent in MOOCs.

The main threat is that CAMEO (and other forms of cheating) will decrease the confidence that the certificates provide a reliable evidence of knowledge and competency. Thus we believe that it is important to address CAMEO and other methods of academic dishonesty that are possible in online learning environments before they jeopardize the value of the MOOC certificates. We note that the current 'verified certificates' take steps to assure that the name on the certificate belongs to the individual who entered the answers in the master account, but is not a good defense against CAMEO and other forms of academic dishonesty.

Besides being a threat to the integrity of MOOC certificates, CAMEO also interferes with educational research. Our findings show that the users with the highest skill accounts in our course are masters, and that the lowest skill accounts are harvesters. This means that any research that tries to study the behavior of successful students vs. that of unsuccessful ones, might be heavily influenced by these two outlying

groups. For example, if we try to identify the variables that most strongly correlate with student's skill, this subset of harvester and master accounts would have a significant effect on the results.

CAMEO almost certainly impacts learning. Users employing batch mode are foregoing the struggle of trying to answer questions on their own, and it seems obvious that this will have a negative effect on their learning. However, it is not clear that students who use a harvester account to find the answer to a few percent of the questions that they have struggled with is hurting their learning - indeed it might even be beneficial. Certainly this is a topic for future research.

Remedies

We now discuss steps that our research shows should reduce, or at least frustrate, CAMEO.

Our findings clearly show that using randomization of the numbers in questions reduces CAMEO. Randomization of question parameters is already supported by the edX platform, and we recommend that instructors use it where possible, certainly in preference to multiple choice (which can be harvested using exhaustive search). Randomizing does not completely defeat CAMEO. A more general form of randomization would be using question pools, i.e., group of questions that require very similar skills at similar level of difficulty, but have significantly different surface features (i.e. wordings or symbols chosen for various quantities).

A much simpler-to-implement prevention method is to avoid giving any feedback on important questions such as exams - including even the usual true/false feedback. Even allowing 'show answer' after the due date enables students to harvest the answers for use the next time the course is given. The obvious disadvantage of omitting feedback is that it compromises the learning experience, since immediate feedback is very important for learning. Individual instructors will have to trade off whether they want to help the great majority of students that wants to learn vs assuring the security of the edX.org certificate.

In the future, CAMEO can be addressed on the platform level, by adding cheating detection, and by supporting more general classes of randomization and question pooling along with algorithms for fairly grading students who do not do exactly the same questions.

SUMMARY, CONCLUSIONS AND FUTURE WORK

In this paper we have presented research on academic dishonesty in MOOCs. The specific method that we have studied, termed CAMEO [6], is based on using multiple accounts to harvest solutions. Our results show that a significant fraction of the certificate earners in our course - about 10%, have used this method to some extent, and that 2.5% of them obtained the majority of their correct answers by using it. We also showed that students who used CAMEO tended to have high success rate and fast response time compared with the other students. We then found that a question's characteristics correlate strongly with the amount of cheating on it. This led to our suggestions for instructional design practices that are less vulnerable to CAMEO.

Our main conclusions are as follows:

1. CAMEO is already significant among MOOC students, our study represents a lower bound on CAMEO, and does not include other forms of academic dishonesty like copying from other students. Our results suggest that cheating can jeopardize the validity of the MOOC certificate system.

2. The main motivation for CAMEO appears to be earning a certificate, consistent with our observation that questions that weigh heavily towards the overall grade are most likely to be cheated upon.

3. We show that delaying feedback and using randomization of problems can reduce CAMEO. Instructors can employ these with trade-off between pedagogy, prevention, and the amount of extra work required.

Though the research was conducted in MOOCs, the results and conclusions are also relevant to other learning environments that allow users to register additional accounts under a different user name. This research is only a first step into studying this phenomenon. Our plans for future research include:

Generalizing the results. In order to get a better estimation of the severity of CAMEO, our plan is to extend our research, and analyze a larger sample of courses. We also intend to make the source code publicly available.

Run-time detection. In addition to using methods for preventing CAMEO, it is important to detect it while it happens, so timely intervention can be made. Thus, we plan to develop a run-time version of the algorithm. Since the algorithm is quite scalable, and with small optimizations can be made linear in the amount of actions in the course, the cost in performance should be minimal. Also in the context of the algorithm, we are interested in developing a detector that does not rely on the IP, as sophisticated users can use various method to hide their IP address.

Understanding CAMEO. Pedagogy-wise, we are interested to understand the specific purposes for using CAMEO (for example, is it a *help seeking* strategy?). Among other things, this will extend our understanding of the motivation of learners in MOOCs, and eventually, can help to improve them.

ACKNOWLEDGMENTS

The first author wants to thank the projects 'eMadrid' (Regional Government of Madrid) under grant S2013/ICE-2715 and 'RESET' (Ministry of Economy and Competiveness) under grant RESET TIN2014-53199-C3-1-R for partially supporting this work. The authors would like to thank Christopher Chudzicki for his help in carrying out this study, and to Curtis Northcutt, Isaac Chuang, and Andrew Ho for useful discussions. We also thank the anonymous reviewers for their useful comments.

REFERENCES

1. Baker, R., Corbett, A., Koedinger, K., Evenson, S., Roll, I., Wagner, A., Naim, M., Raspat, J., Baker, D., and Beck, J. Adapting to when students game an intelligent tutoring system. In *Intelligent Tutoring Systems*, M. Ikeda, K. Ashley, and T.-W. Chan, Eds., vol. 4053 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2006, 392–401.
2. Byrne, J. A. Wharton Using MOOCs To Recruit MBAs. http://poetsandquants.com/2015/02/11/wharton-using-moocs-to-recruit-mbas/, 2015. Accessed: 2015-10-30.
3. Davis, S. Academic dishonesty in the 1990s. *The Public Perspective* (1993).
4. Donald L. McCabe, L. K. T. Academic dishonesty: Honor codes and other contextual influences. *The Journal of Higher Education 64*, 5 (1993), 522–538.
5. McCabe, D. L., Trevino, L. K., and Butterfield, K. D. Cheating in Academic Institutions: A Decade of Research. *Ethics & Behavior 11*, 3 (July 2001), 219–232.
6. Northcutt, C., Ho, A. D., and Chuang, I. L. Detecting and preventing "multiple-account" cheating in massive open online courses. *CoRR abs/1508.05699* (2015).
7. Palazzo, D. J. Detection, patterns, consequences, and remediation of electronic homework copying, 2006. Masters Thesis.
8. Palazzo, D. J., Lee, Y.-J., Warnakulasooriya, R., and Pritchard, D. E. Patterns, correlates, and reduction of homework copying. *Phys. Rev. ST Phys. Educ. Res. 6* (Mar 2010), 010104.
9. Singhal, A. C. Factors in students' dishonesty. *Psychological Reports 51*, 3 (2015/10/29 1982), 775–780.