On the Prevalence of Multiple-Account Cheating in Massive Open Online Learning
A replicant study

Bao, Yingying; Chen, Guanliang; Hauff, Claudia

Publication date
2017

Document Version
Final published version

Published in
Proceedings of the 10th International Conference on Educational Data Mining

On the Prevalence of Multiple-Account Cheating in Massive Open Online Learning

A replication study

Yingying Bao, Guanliang Chen* and Claudia Hauff
Web Information Systems
Delft University of Technology
Delft, the Netherlands
Y.Bao-1@student.tudelft.nl
{guanliang.chen, c.hauff}@tudelft.nl

ABSTRACT

Massive Open Online Courses (MOOCs) are a promising form of online education. However, the occurrence of academic dishonesty has been threatening MOOC certificates' effectiveness as a serious tool for recruiters and employers. Recently, a large-scale study on the log traces from more than one hundred MOOCs created by Harvard and MIT has identified a specific cheating strategy viable in MOOCs: *Copying Answers using Multiple Existences Online* (CAMEO). In essence, learners create several accounts on a MOOC platform, request assessment solutions via some of the accounts, and then submit these "harvested" solutions in their main account to receive credit. In our work, we *replicate* the CAMEO implementation and apply it to ten edX MOOCs created by the Delft University of Technology. Our results show that in those MOOCs, 1.9% of certificates were likely earned through CAMEO cheating, a number comparable to the fraction of cheating observed in Harvard and MIT MOOCs.

Keywords

MOOCs, Academic Dishonesty, Multiple-Account Cheating, Educational Data Mining

1. INTRODUCTION

Cheating is generally defined as using dishonest means to gain an undeserved reward of ability or to get rid of an embarrassing situation [3]. Academic dishonesty is a type of cheating that occurs in relation to an academic exercise. It is a widespread occurrence across different levels and forms of education [4]. There are diverse cheating strategies adopted by students to implement academic dishonesty such as impersonation, bringing notes into the exam hall, using an unauthorized digital device, and so on.

MOOCs, which are courses designed with open access for a large number of online participants, have become a vital part of scalable and large-scale education. However, the effectiveness of MOOCs has been threatened by academic dishonesty. For instance, as early as 2012, some instructors have voiced concerns about various forms of cheating in their MOOCs [7].

One of the main issues in exploring the issue of cheating in MOOCs is the general lack of ground truth data — MOOC providers may be reluctant to confront learners (as a definite proof of cheating is difficult to come by and a time-consuming endeavour) and MOOC learners are reluctant to admit their misbehaviour. Recently, Northcutt et al. [5] proposed a first approach to automatically detect a particular kind of cheating purely based on the log data that is collected in major MOOC platforms; they termed this method CAMEO or *Copying Answers using Multiple Existence Online*. In brief, this method is able to detect learners that cheat in the following way: (1) A learner registers multiple accounts on a MOOC platform and enrolls in a MOOC of interest with all these accounts; one of those registered accounts is the learner's *main* account. (2) The learner uses some of the registered accounts to randomly submit answers to assessment questions (which in MOOCs are often multiple-choice or fill-in-the-blank questions to enable automatic grading) as a way to *harvest* the correct solutions. This is made possible by a design decision of major MOOC platforms which allows learners to check their submitted solutions immediately after submission. (3) The learner then submits the harvested solutions through the main account, allowing the learner to successfully complete the course and earn a certificate. Commonly, achieving 60% (or a similar percentage) of all possible points is sufficient to receive a MOOC certificate.

Among the many potential ways of cheating in MOOCs, CAMEO is of particular concern for a number of reasons: (1) the CAMEO cheating strategy can be performed by every learner individually, it does not require learners to collaborate with others; (2) CAMEO cheating is efficient and easy to execute as it directly utilizes the solutions provided in a MOOC; and (3) CAMEO cheating can be applied across many different MOOCs, largely independent of the subject

*The author's research is supported by the *Extension School* of the Delft University of Technology.

or course level.

Northcutt et al. [5] observed CAMEO cheating in 69 Coursera MOOCs (out of 115 investigated) provided by MIT and Harvard University; among those 69, approximately 1.3% of the certificates were issued to learners identified as CAMEO users. Given that MOOCs provided by different universities usually attract varying sets of learners, in this work, we investigate the following two **R**esearch **Q**uestions:

RQ1 What is the *prevalence* of CAMEO cheating in the MOOCs provided by TU Delft?

RQ2 What are *characteristics* of learners identified to have employed the CAMEO strategy?

To answer these questions, we implement the detection approach as described in [5] and apply it on the log traces of 10 edX MOOCs. We find that 1.9% of the certificates are earned by CAMEO learners (our answer to **RQ1**), with some types of MOOCs more prone to cheating than others. While we did not observe any CAMEO behaviour in a MOOC on political debates, we found more than 6% of certificates to be CAMEO certificates in a business and technical course respectively. With respect to **RQ2**, we observe cheating to be most prevalent mid-course and to be more prevalent in some user demographics than others.

2. RELATED WORK

There are a few works proposed to investigate the prevalence of cheating in MOOCs. Two of the earliest works were proposed by [5] and [6]. Both of these two works focused on the detection of CAMEO cheating based on learnersâĂŹ traces in MOOCs provided by MIT on edX.

In [5], 1.3% of the certificates among 69 MOOCs covering different subjects were earned by learners who adopted CAMEO cheating strategies. Learners who applied CAMEO are more likely to be young, male and international than the other certified learners. In [6], the number is 10.3% of the certificates in an introductory physics MOOC.

In both of these works, researchers set patterns of CAMEO and select learners whose behaviors satisfy the patterns. There are overlaps between the criteria adopted by the two works. Ruiperez-Valiente et al. [6] has relatively more detailed assumptions to CAMEO in different modes. Northcutt et al. [5] was conducted in more than 100 MOOCs, which helps to avoid the accidental bias in the prevalence of CAMEO caused by courses.

Compared to these works, our goal is to investigate the prevalence of this cheating behavior in the MOOCs provided by TU Delft and what the common characteristics are among the detected cheaters.

3. DETECTION METHOD

In this section we recap the main assumptions that underpin Northcutt et al. [5]'s approach. Note that these assumptions are derived from intuitions about MOOC learners' (or more generally online users') behaviours on the learning platform. Our implementation of the approach matches the original paper's algorithmic formulation as closely as possible.

- **CAMEO users hold at least two accounts.** Each CAMEO user (i.e. a learner who cheats to gain an advantage in a MOOC) should use one or more accounts to harvest solutions (so-called Harvest Account(s)) and one main account to submit the correct solutions (i.e., the Master Account) so as to earn the certificate. Initially, every possible pair of user accounts having enrolled in a particular MOOC is a candidate Master/Harvester pair.

- **CAMEO users harvest solutions before entering them into their Master Account.** In other words, for questions that learners cheat on, the candidate Harvester Account should precede the candidate Master Account in time for the gathering of solutions.

- **CAMEO users quickly pass collected solutions from Harvester Accounts to Master Account.** It is reasonable to assume that a cheater may simultaneously log in both the Harvest Account and the Master Account, and once the learner collects the correct solutions, he may immediately submit the correct solutions through the Master Account. This assumption requires the time difference between the correct submission from the candidate Master Account and the request to solutions from the candidate Harvester Account to be small.

- **Master Accounts are certified, the Harvester Accounts are not.** Given that Harvester Accounts are mainly used to gather correct solutions via randomly submitting answers, more often than not, the Harvester Accounts do not reach the passing threshold of a MOOC. At the same time, the Master Accounts should perform well in that respect and earn a certificate.

- **Master Account and Harvester Account are connected via IP addresses.** As noted before, a CAMEO user may simultaneously log into multiple accounts on one and the same or different devices in the same location; thus, it is likely that Master and Harvester account share a common logged IP address during the MOOC.

In the CAMEO approach, these intuitions are transformed into filtering rules (that filter the initially created account pairs) and only candidate Master/Harvester pairs that meet all of these criteria are considered to be CAMEO users, that is, learners who cheat through multiple account usage in a MOOC. Most of these rules contain ad-hoc parameters (e.g. the time limit between a Harvester and Master account submission); we have followed the parameter settings described in [5] in our implementation.

4. EXPERIMENT

4.1 Dataset

Our study is based on the log data generated during 10 edX MOOCs (eight different MOOCs of which two ran twice) which were provided by TU Delft between 2014 and 2016. The MOOCs cover various scientific areas including data science, programming paradigms, biotechnology, business and political science. An overview of the MOOCs, including the number of enrolled learners and the number of certificates earned is shown in Table 1.

Table 1: Overview of the ten MOOCs included in this study. **#Enrollments** shows the number of user accounts that registered for each MOOC and **#Certificates** lists the number of registered participants that achieved a certificate (the passing threshold is 50% for Frame101x and 60% for all other MOOCs). Note that FP101x and EX101x are listed twice, as they both ran in two different time periods.

Course Code	Course Title	Session	#Enrollments	#Certificates
FP101x	Functional Programming	2014 Fall	37,940	1,356
CTB3365DWx	Drinking Water Treatment	2014 Fall	10,458	246
EX101x	Data Analysis	2015 Spring	33,515	2,190
Frame101x	Framing: How Politicians Debate	2015 Spring	34,017	919
Calc001x	Pre-university Calculus	2015 Summer	27,857	358
EX101x	Data Analysis	2015 Fall	21,041	1,156
IB01x	Industrial Biotechnology	2015 Fall	8,143	329
FP101x	Functional Programming	2015 Fall	20,936	1,143
RI101x	Responsible Innovation	2016 Spring	2,741	113
CTB3365sTx	Urban Sewage Treatment	2016 Spring	9,566	361

Table 2: Overview of the detected CAMEO users and the percentage of certificates gained by CAMEO users. The last row shows the numbers across all ten MOOCs.

Course Code	#CAMEO Users	% CAMEO Certificates
FP101x (2014)	13	0.96%
CTB3365DWx	4	1.63%
EX101x (2015S)	27	1.23%
Frame101x	0	0
Calc001x	13	3.63%
EX101x (2015F)	20	1.73%
IB01x	12	3.65%
FP101x (2015)	16	1.40%
RI101x	7	6.19%
CTB3365sTx	25	6.93%
Total	137	1.89%

4.2 CAMEO Detection Results

For each of the MOOCs, we present the number of detected CAMEO users (and subsequently the percentage of certificates gained through CAMEO) in Table 2. CAMEO users are detected in 9 out of the 10 MOOCs and overall account for 137 (or 1.89%) of all certificates. This percentage is slightly higher than Northcutt et al. [5]'s (1.3%). The percentages vary across courses, with *Urban Sewage Treatment* being the MOOC with the largest percentage of CAMEO learners, nearly 7%. On the other hand, our only MOOC without CAMEO cheating detected is *Framing: How Politicians Debate*. In future work we will investigate this variance in CAMEO between courses; we hypothesize that for participants in Frame101x a certificate has less intrinsic value (the self-development aspect is more important) and thus cheating is less likely to occur.

4.3 Verification of CAMEO Users

To explore how plausible the detection results are — i.e., are the detected account pairs actually belonging to the same learner and did the learner indeed cheat — we manually verified key account characteristics. It is sensible for instance to assume that at least some CAMEO users register with the same/similar name across the Harvester and Master Account. Indeed, among our 137 detected CAMEO users, 20% have similar or even same registered full names attached to their Harvester and Master Accounts[1]. To provide the reader with some intuition on the similarities, we now describe for a randomly picked CAMEO user in our dataset the similarities between the detected Master and Harvester Account:

- The Harvester & Master Account have the same registered full name.
- The registered email addresses of the Harvester & Master Account contain a common long character sequence (eight characters).
- The Harvester & Master Account utilize the same IP address to answer every question.
- The Harvester & Master Account submit answers within 60 seconds for every harvested question and the Harvester Account always submits before the Master Account.
- The Harvester Account submits answers for all questions in the course, but the correctness is only 11.5%.

Based on these observations, we are highly confident that the learner is indeed a CAMEO user.

4.4 Characteristics of CAMEO Users

To gain a better understanding of the detected CAMEO users, we analyze their characteristics and patterns. With respect to the nationality of the certified learners, we find them to come mainly from the US, the Netherlands and the UK. However, the detected CAMEO users are mainly from India (27), the US (12) and Germany (7).

We are also interested in the motivation of CAMEO cheaters, i.e., what drives them to cheat in MOOCs. Intuitively, we believe that most CAMEO users to be strongly goal-oriented with the goal being the certificate (instead of the goal being related to knowledge gains). To verify this intuition, we compute how many detected CAMEO users would be

[1]We compute the similarity between two account names according to the Ratcliff/Obershelp sequence match method [1].

Table 3: Overview of the identified CAMEO learners and their certificate status (pass or fail) if the assessments points they gained through CAMEO were removed.

Course Code	Pass w/o CAMEO	Fail w/o CAMEO
FP101x (2014)	2	11
CTB3365DWx	0	4
EX101x (2015S)	3	24
Frame101x	0	0
Calc001x	0	13
EX101x (2015F)	4	16
IB01x	0	12
FP101x (2015)	2	14
RI101x	1	6
CTB3365sTx	0	25
Total	12	125

able to earn a certificate without CAMEO cheating. Specifically, we calculate the grades of CAMEO users on the condition that they only receive credits for questions they did not cheat on and evaluate whether the scores are sufficient to pass the course. As shown in Table 3, nearly 90% of the CAMEO users cannot pass the MOOCs without cheating, which implies that most of the CAMEO users are purely certificate-driven.

We also investigate *when* CAMEO users are most likely to cheat during the course of a MOOC. To this end, we select FP101x (2014 and 2015) and EX101x (2015 Spring and 2015 Fall) for analysis as the grading strategies adopted across the four MOOCs are very similar: almost all questions (more than 100 per course) are worth a single point and the final grade is simply based on the fraction of questions the learner answered correctly (with 60% of correct answers being the passing threshold). Figures 1 (FP101x) and 2 (EX101x) show the number of identified CAMEO users that resort to the CAMEO strategy across the different course weeks.

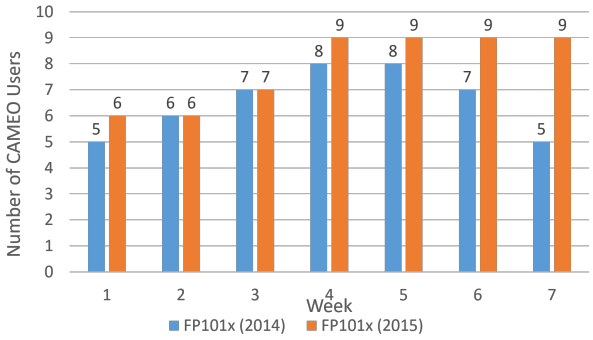

Figure 1: Average Number of CAMEO Cheater Cheating on per Question in Different Weeks in FP101x.

Few learners resort to CAMEO in the first two weeks of the course, while course weeks 3, 4, 5 and 6 attract the most cheating. This is not overly surprising considering the fact that the questions in later weeks are usually more difficult than those in early weeks. The trend of decreased CAMEO in the final week(s) can be explained by the fact that the

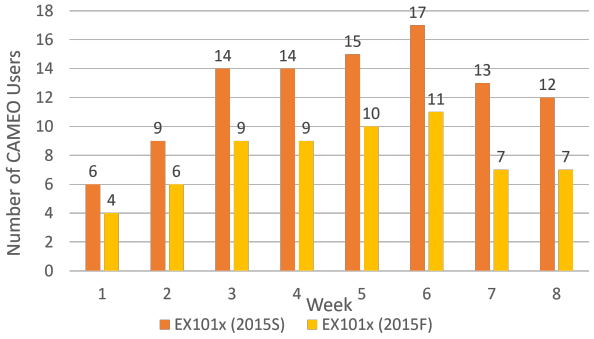

Figure 2: Average Number of CAMEO Cheater Cheating on per Question in Different Weeks in EX101x.

edX platform provides a *Progress* page where each learner can check his progress towards the passing threshold. For a learner whose main goal is the certificate, the realization of that goal (which can occur already as early as week 5 as the passing threshold is 60%) is likely to reduce or stop his CAMEO behaviour.

5. CONCLUSION

We successfully replicated the CAMEO strategy formalized in [5] and applied it to a novel set of MOOCs. Overall, we found similar percentages of CAMEO cheating in TU Delft MOOCs (1.9% vs. 1.3%), albeit with the limitation that we only explored 10 MOOCs (vs. 115 by MIT/Harvard). We are currently enlarging the study to include all 50 MOOCs that are provided by TU Delft. Our future work will place a greater emphasis on the demographic analysis of CAMEO users and on ways to reduce and prevent such cheating — either through technological means or ethical appeals and moral reminders [2].

References

[1] Paul E Black. Ratcliff/obershelp pattern recognition. *Dictionary of Algorithms and Data Structures*, 17, 2004.

[2] Henry Corrigan-Gibbs, Nakull Gupta, Curtis Northcutt, Edward Cutrell, and William Thies. Measuring and maximizing the effectiveness of honor codes in online courses. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 223–228. ACM, 2015.

[3] Melanie Ghoul, Ashleigh S Griffin, and Stuart A West. Toward an evolutionary definition of cheating. *Evolution*, 68(2):318–331, 2014.

[4] Donald L McCabe, Kenneth D Butterfield, and Linda K Trevino. *Cheating in college: Why students do it and what educators can do about it.* JHU Press, 2012.

[5] Curtis G Northcutt, Andrew D Ho, and Isaac L Chuang. Detecting and preventing "multiple-account" cheating in massive open online courses. *Computers & Education*, 100:71–80, 2016.

[6] Jose A Ruiperez-Valiente, Giora Alexandron, Zhongzhou Chen, and David E Pritchard. Using multiple accounts for harvesting solutions in moocs. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pages 63–70. ACM, 2016.

[7] K Webley. Mooc brigade: Can online courses keep students from cheating? *Time*, 2012. Retrieved May 2017, from nation.time.com/2012/11/19/mooc-brigade-can-online-courses-keep-students-from-cheating/.