Computers & Education 100 (2016) 71–80

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Detecting and preventing "multiple-account" cheating in massive open online courses☆

Curtis G. Northcutt [a, *], Andrew D. Ho [b], Isaac L. Chuang [a]

[a] Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[b] Harvard Graduate School of Education, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:
Received 23 August 2015
Received in revised form 13 April 2016
Accepted 19 April 2016
Available online 3 May 2016

Keywords:
Massive Open Online Courses (MOOCs)
Cheating detection
Educational certification
Educational Data Mining (EDM)
Security
Architecture for educational technology system
Learning communities
Lifelong learning
Pedagogical issues
Teaching/learning strategies

ABSTRACT

We describe a cheating strategy enabled by the features of massive open online courses (MOOCs) and detectable by virtue of the sophisticated data systems that MOOCs provide. The strategy, Copying Answers using Multiple Existences Online (CAMEO), involves a user who gathers solutions to assessment questions using a "harvester" account and then submits correct answers using a separate "master" account. We use a small-scale experiment to verify CAMEO and estimate a "lower bound" for its prevalence among 1.9 million course participants in 115 MOOCs from two universities. Using conservative thresholds, we estimate CAMEO prevalence at 1237 certificates, accounting for 1.3% of the certificates in the 69 MOOCs with CAMEO users. Among earners of 20 or more certificates, 25% have used the CAMEO strategy. CAMEO users are more likely to be young, male, and international than other MOOC certificate earners. We identify preventive strategies that can decrease CAMEO rates and show evidence of their effectiveness in science courses.

1. Introduction and motivation

Massive Open Online Courses (MOOCs) began receiving significant media coverage in 2012 (McNutt, 2013; Pappano, 2012), coincident with the widespread commitment by established universities to providing free courses online (Christensen et al., 2013; Ho et al., 2014; Stanford Online, 2013). These MOOCs distinguished themselves from predecessors like MIT's Open Courseware (Smith, 2009; d'Oliveira, Carson, James, & Lazarus, 2010) by providing not only free content but a course-like structure, including enrollment, synchronous participation, periodic graded assessments, online discussion forums, interactive simulations, and of greatest relevance for our purposes, certification of successful completion (DeBoer, Ho, Stump, & Breslow, 2014; Linn, Gerard, Ryoo, McElhanney, Liu, & Rafferty; 2014). One theory of MOOC proliferation holds that free certification of proficiency in college courses can reduce inefficiencies in higher education by replacing high-cost residential courses with low-cost online certification (Hoxby, 2014).

☆ This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (#1122374). The work presented is the sole work of the authors. The authors assume full responsibility for the contents of this manuscript.

* Corresponding author.
E-mail addresses: cgn@mit.edu (C.G. Northcutt), andrew_ho@gse.harvard.edu (A.D. Ho), ichuang@mit.edu (I.L. Chuang).

http://dx.doi.org/10.1016/j.compedu.2016.04.008
0360-1315/

In this paper, we reveal a particular cheating strategy that is detectable across the 115 MOOCs in our sample and currently presents a serious threat to the trustworthiness of their certifications. We call the strategy, Copying Answers using Multiple Existences Online [1](CAMEO). A user employing this strategy, whom we refer to as a CAMEO user, earns a certificate by creating at least two MOOC accounts: (1) one or more "harvester" accounts used to acquire correct answers by guessing at test answers and then accessing instructor-provided solutions via a "Show Answer" button, and (2) one or more "master" accounts used to submit these solutions as correct test answers.

The CAMEO strategy lies at the intersection of a number of other copying techniques and contexts. We distinguish between 1) what is copied, 2) why it is copied, 3) how it is copied, and 4) how copying is detected. The CAMEO strategy occurs in similar contexts as community collaboration in online courses (Yang, Wen, Kumar, Xing, & Rose, 2014), and detection of both involves analyzing the interactions of multiple accounts. However, prior efforts have focused on how communities of different users affect learning outcomes (Kumar, Rose, Wang, Joshi, & Robinson, 2007), in contrast with CAMEO behavior, where a single user exploits multiple accounts, potentially circumventing the learning process entirely. CAMEO is most similar to "multiple account" sharing strategies in online games (e.g., Kafai & Fields, 2009), where a single user can increase scores or other in-game outcomes by creating multiple accounts and interacting them strategically. However, CAMEO behavior distinguishes itself from online game strategies due to what is copied (correct answers to tests) and why it is copied (to fake or expedite certification of proficiency). As we show, the specificity of these differences enables targeted detection, quantification, and prevention of CAMEO use in these MOOCs.

Cheating by CAMEO shares similarity in purpose with copying in online and conventional courses (Baker, Corbett, & Koedinger, 2004, January; Kauffman & Young, 2015; McCabe, Butterfield, & Treviño, 2012; Palazzo, Lee, Warnakulasooriya, & Pritchard, 2010). However, three features of CAMEO make it a unique threat as a cheating strategy in online education. First, it is internally sufficient. Whereas most users copy from other students or external resources, CAMEO users employ multiple accounts to copy from themselves, making the cheating strategy highly accessible by removing dependence on outside resources. As a result, the strategy is extremely effective. Second, in asynchronous MOOCs, where students can access course materials and assessments at their own pace, a CAMEO user can employ the CAMEO strategy for every question they attempt, allowing certification for full course completion in a single sitting. Third, it is unrestricted, employable in a nonselective, open admission setting. Degrees from selective institutions assert, at the very least, that users have been prescreened, but MOOC certificates do not. Because MOOC users, unlike most postsecondary students, are not selected by any merit-based process or criteria, the considerable accessibility of CAMEO in these MOOCs holds the potential to render their certificates valueless as an academic credential.

The key contributions of this paper are a detection algorithm for the CAMEO-based cheating that allows for a lower bound estimate of prevalence and a small-scale experiment confirming CAMEO behavior. This latter experiment is an extension of "honey pot" cheating detection (Corrigan-Gibbs, Gupta, Northcutt, Cutrell, & Thies, 2015), where copied answers can be confirmed directly. These contributions complement the considerable literature that estimates cheating prevalence through surveys, where survey responses may be influenced by social desirability, interpretation of item prompts, concerns about anonymity, and inflation in self-reported performance (Mastin, Peszka, & Lilly, 2009). This paper investigates a specific cheating strategy using an algorithm customized to big datasets that contain detailed user interactions with online course content, including activity timestamps. With 115 courses, this is also the largest analysis of cheating in online courses of which we are aware.

CAMEO also represents an example of a more general tendency for open online learning systems to enable both new strategies for cheating and new strategies for detection (Horton, Rand, & Zeckhauser, 2011; Li, Chang, Yuan, & Hauptmann, 2015; Raines et al., 2011). Although CAMEO is technically a copying strategy, we argue that its use in MOOCs constitutes "cheating." At a minimum, employing CAMEO is a violation of policy, because MOOC honor codes forbid the creation of multiple accounts (Coursera, 2012; edX, 2014; Udacity, 2014). The CAMEO strategy also threatens perceptions of the value of MOOC certification. Any reasonable interpretation of standard MOOC certificates, which refer to "successful completion" (edX, 2015), includes proven student proficiency with course content. Yet, the prevalence of the CAMEO strategy justifies a starkly contrasting interpretation of MOOC certification—that a user merely copied answers from a "dummy" harvester account. Combined with growing evidence that the reputation and usefulness of MOOC certification are predictors of MOOC persistence (e.g., Alraimi, Zo, & Ciganek, 2015), we anticipate that widespread awareness of MOOC susceptibility to the CAMEO strategy could depress MOOC popularity and persistence among general users.

2. Methodology

We begin by describing a CAMEO detection algorithm that relies on the distribution of differences in time between particular user actions across particular user pairs. The CAMEO detection algorithm is comprised of five filters with highly conservative cutoffs intended to reduce false positives, including a Bayesian criterion for the timestamp difference distributions. After we present these filters, we describe a small-scale experiment that confirms CAMEO cases, and we show that the CAMEO algorithm detects these cases as expected.

[1] CAMEO is an abbreviation for Copying Answers using Multiple Existences Online.