Giora Alexandron, José A. Ruipérez-Valiente, Zhongzhou Chen Pedro J. Muñoz-Merino, David E. Pritchard (2017). Copying@Scale: Using Harvesting Accounts for Collecting Correct Answers in a MOOC. *Computers & Education*, *108*, 96-114.

http://www.sciencedirect.com/science/article/pii/S0360131517300234

DOI: 10.1016/j.compedu.2017.01.015

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Copying@Scale:
Using Harvesting Accounts for Collecting Correct Answers in a MOOC

Giora Alexandron[a1]* José A. Ruipérez-Valiente[bc]* Zhongzhou Chen[a]
Pedro J. Muñoz-Merino[b] David E. Pritchard[a]

[a] *Massachusetts Institute of Technology*
[b] *Universidad Carlos III de Madrid*
[c] *Institute IMDEA Networks*

Abstract

This paper presents a detailed study of a form of academic dishonesty that involves the use of multiple accounts for harvesting solutions in a Massive Open Online Course (MOOC). It is termed CAMEO – Copying Answers using Multiple Existence Online. A person using CAMEO sets up one or more *harvesting* accounts for collecting correct answers; these are then submitted in the user's *master* account for credit.

The study has three main goals: Determining the prevalence of CAMEO, studying its detailed characteristics, and inferring the motivation(s) for using it. For the physics course that we studied, about 10% of the certificate earners used this method to obtain more than 1% of their correct answers, and more than 3% of the certificate earners used it to obtain the majority (>50%) of their correct answers. We discuss two of the likely consequences of CAMEO: jeopardizing the value of MOOC certificates as academic credentials, and generating misleading conclusions in educational research. Based on our study, we suggest methods for reducing CAMEO. Although this study was conducted on a MOOC, CAMEO can be used in any learning environment that enables students to have multiple accounts.

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

[1] Corresponding author

* The first two authors contributed equally to this work

1. Introduction

This paper deals with a new form of academic dishonesty that involves the use of multiple accounts for copying solutions in Massive Open Online Courses (MOOCs). The method works as follows. A *harvesting* account is used for collecting correct answers; Collecting the answers is accomplished by either asking to see the answer ('show answer' on edX platform), or, by exhaustive search – successive guessing till the correct answer is found. The answer is then copied to the user's main account, termed the *master* (account), for credit. We adopt the term CAMEO – Copying Answers using Multiple Existence Online that was coined by Northcutt, Ho and Chuang (2016; hereafter denoted NHC) to describe this phenomenon. (Throughout this paper, we distinguish between *user* -- a real persona interacting with the system, and *account/s*, which are operated by users. *Master* and *harvester* are attributes of accounts, and a user who operates such accounts is referred to as a *CAMEO user*.)

CAMEO is related to two previously investigated ways of answering questions without possessing the knowledge they aim to assess: *academic dishonesty*, and *gaming the system.* We contend that CAMEO is a form of academic dishonesty that most of its practitioners would classify as cheating, because it clearly violates two provisions of the edX honor code (which all users are required to agree to): "Maintain only one user account" and avoid engaging in "any activity that would dishonestly improve my results"[2]. Baker et al. defined gaming as "Attempting to succeed in an interactive learning environment by exploiting properties of the system rather than by learning the material" (2009). Gaming the system is not generally condemned as a serious form of academic dishonesty, but more as exploiting a path that was not intended by the designer. Still, CAMEO resembles gaming in terms of *why* students are using it -- improving grades -- and in *how* they are doing it – by exploiting technical features of the system.

CAMEO is especially harmful because, when feasible, it is a very efficient way for getting answers dishonestly. First, it is self-contained, namely, it does not rely on collaboration with other users. Second, the solution is readily available, i.e., no need to spend time on searching for the solution and adopting it to the problem at hand. Considering these, it may not be surprising that we detected some users who "earned" certificates while harvesting most of their correct answers using CAMEO, including users who submitted correct answers in an 'inhuman' pace (up to 700 correct answers with average time of less than 30 seconds between opening the problem and answering it).

[2] https://www.edx.org/edx-terms-service (accessed 2016-04-01)

The most common master:harvester relation that we observed is of one master operating one harvesting account. However we also detected few cases in which a single master operated multiple harvesting accounts (probably to get more attempts when using 'exhaustive search' on questions that do not have 'show answer'), and one case in which a group of master accounts together operated several harvesting accounts, and seemed to divide the work between them.

We also distinguish between two *modes* in which CAMEO is used. One is what we call *help seeking*, and it refers to cases in which the master user turns to using the harvesting account after trying to solve the question legitimately (i.e., CAMEO serves as a fallback strategy). The other CAMEO mode, which we term *premeditated*, involves harvesting the answer prior to looking at the question in the master account. As we show in the Results section, we found that most of the CAMEO events were premeditated, and that 'heavy' CAMEO users tended to be more premeditated.

The context in which this research was conducted was the Introductory Physics MOOC 8.ReV, run on the edX.org platform. In this course, users can typically ask to see the correct answer (*show answer*) after exhausting all the attempts. When 'show answer' is not allowed (in this course it is disabled for most quiz questions), correct/incorrect feedback is mostly always available, and can be used to exhaustively search for the right answer on multiple-choice questions (and even on open response questions if the harvester is persistent).

The goal of this research is to understand the extent, the method, and the motivation for using CAMEO in MOOCs, and how it can be reduced, using our course as a case study. CAMEO is a new topic of research and to the best of our knowledge our work is the first to study it in depth. The current paper extends the initial findings reported in (<Reference removed for anonymity>, 2015; <Reference removed for anonymity>, 2016). It centers on the amount of CAMEO, its distribution over time (showing that it almost stops after the CAMEO users earn enough points for certificate), whether it is a *premeditated* or a *help seeking* behavior, and what instructional design selections, such as using randomization, are correlated with decreased CAMEO. Following this, our research is guided by the following questions:

- *How many students are practicing CAMEO in our course, and how do they use it (e.g., is it 'help seeking' behavior, or is it premeditated harvesting of answers)?*
- *What can we infer about students' motivation to practice CAMEO from their observed behavior?*
- *What instructional design features correlate with reduced CAMEO?*

Our main method is an educational data mining approach – to operationalize CAMEO as a temporal pattern that can be detected using time-series analysis of clickstream data. The cornerstone of our method is that different user accounts operated by the same person share their IP address. On top of that, we add a few criteria designed to eliminate accounts that share IP address but are operated by different people.

The findings that we present below show that CAMEO is already significant, at least in our course – more than 10% of the certificate earners used it to obtain more than 1% of their correct answers, and more than 3% of the certificated users used it to obtain the majority of their correct answer. Since CAMEO decreases the reliability of the assessment, it reduces the confidence that a certificate is a valid evidence of proficiency, and thus, poses a threat to the certificate system. Also, it interferes with learning, and as we also show, it can alter the results of educational research. We note that CAMEO is not detected by the current methods that MOOC providers like edX use to validate the certificates.

Our findings on the amount of CAMEO greatly exceed those of NHC, who examined multiple courses offered by MITx and HarvardX, including ours. They use different criteria for CAMEO than we do, and the amount of cheating that we report is much higher than what they report in our course - 10%, vs. 2.5% detected by NHC's algorithm in our course (private communication). In the Discussion section we argue that our results are a more accurate representation of the *true* amount of CAMEO (the purpose of NHC was to establish a very strict lower bound, which at least for our course is not tight). Our results are still only a lower bound on the amount of cheating, as our algorithm does not detect other forms of cheating.

This study is relevant to MOOC researchers, instructors, and administrators. More broadly, it sheds light on a relatively new form of academic dishonesty.

The rest of the paper is organized as follows. Section 2 presents the methodology. Section 3 presents the findings, which are discussed in Section 4. Section 5 surveys related work. Section 6 summarizes the paper, derives conclusions and suggests directions for further research.

2. Methodology

Discussion of our methodology starts with a short description of the edX log files, which are the input to our algorithm. Then we describe the time-series analysis procedure that we use to detect CAMEO events in the log files, and how we operationalize the notion of *help-seeking* vs. *premeditated* CAMEO. Finally, we explain how we verified the results.

2.1. EdX log files

Our CAMEO algorithm analyzes the edX log files of our course. The log files contain clickstream data about users' interaction with the platform. Particularly relevant to the algorithm described below is the information kept for submissions and 'show answer' events. Both contain information on the user's name, the ip from which the interaction was made, and the identifier of the question. For submissions, information on the correctness is also stored. For more details on the format, see edX documentation of the tracking logs [3]. We process the logs to create per-user, time-sorted log files.

2.2. Algorithm and Criteria

2.2.1. IP groups

Our algorithm searches for CAMEO only between user accounts that share the same IP address at some point during the course. Since one user account can connect via different IPs (due to switching between physical locations, using wireless connection, etc.), we define *IP group* as all accounts linked through IPs. A formal definition is given below.

IP group. 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. We now look on the *connected components* (*cc*) of G (a connected component is a subgraph in which there is a path between each two nodes (Hopcroft & Tarjan, 1973)), then for each *cc*, 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, which can be computed in linear time using standard algorithms.
We note that in the algorithm that we used in (<Reference removed for anonymity>, 2016) we removed IPs that were used by more than 10 accounts (referred to as 'public

[3] http://edx.readthedocs.org/projects/devdata/en/latest/internal_data_formats/tracking_logs.html (accessed 2016-04-01)

IPs'), and IP groups that contained more than 10 accounts (as computed after removing 'public IPs'). The rationale was to reduce the likelihood of identifying different users who share IP address, for example because they browse through a shared router, as CAMEO users. However, we discovered that this constraint causes the algorithm to overlook real CAMEO users. Thus we decided to remove it, and to compensate for that we made the filters more stringent by adding an additional filter (*filter 4*; see below) and raising the threshold in another one (*filter 3*).

2.2.2. CAMEO Detection Algorithm

The CAMEO algorithm searches for CAMEO between all pairs of accounts in each IP group. It is composed of two main steps.

The first step collects events that adhere to the general scheme of CAMEO – one account gets the solution to a problem, another account within the same IP group submits a similar answer to this question shortly after. The output of this step is a list of master:harvester pairs, and for each pair, a list of questions in which this master is suspected to use this harvester.

The second step concentrates on filtering the false positives by applying various filters. The two steps are described in details below.

Step 1. For each account *a1*, for each correct submission made by *a1* to a question *q*, we check whether any other account *a2* within the IP group of *a1* obtained the correct answer to *q* (operationalization for 'exhaustive search'), or asked to see the answer to question *q*, in the previous 24 hours.

If a match is found, we add <*a1, a2, q*> to the list of potential CAMEO events. We allow the pair *a1,q* to appear (at most) with one harvester *a2* (a master cannot use several harvesters for the same problem).

A graphical illustration of two kinds of CAMEO are given in Figure 1. The left chart shows what we call *immediate* CAMEO – an event in which the user gets the solution in the harvester account and submits it in the master account immediately after. The right chart shows what we call *batch mode* CAMEO – a modus operandi in which the user harvests several solutions in the harvester account, and then submits them in a rapid sequence in the master account (we observed sequences of up to 40 questions, with less than 20 seconds between successive submissions at the master side).

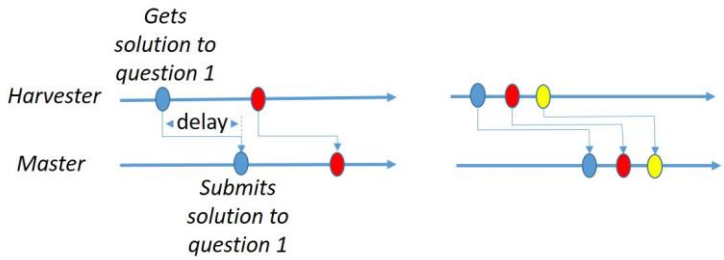

Figure 1: Immediate (left chart) and batch mode (right chart) CAMEO

Step 2. On the list of events collected at *step 1*, we apply the following filters, in the order they appear. The filters were fine-tuned on the data from our course, and for each filter, we report its filtering level – the amount of master-harvester pairs that it leaves out.

1. ***The harvester account does not earn a certificate***. The rationale behind this criterion is that if an account receives any benefit for its behavior, it is less likely that this is an account whose sole purpose is supporting another account. We note that honor certificates[4] are given automatically for any account that reaches 60% of the points in the course, so a purely harvesting account could actually receive certificate unintentionally, hence we may be excluding 'heavy' harvesting accounts in order to lower false positives. On our dataset, this filter removes 34% of the initial set of master-harvester pairs (pairs that appear in at least one of the events collected in *step 1*).

2. ***Master-harvester pair appears in at least 10 questions.*** This is a sort of a 'high-pass filter' that aims to eliminate noise (false positives) created by pairs of users who exhibit master-harvester relation on a small number of questions. The rationale behind it is that real master-harvester pairs would exhibit this behavior on a significant amount of questions. The specific value was picked by examining the log-like accumulative distribution function of the number of questions for each master-harvester pair. On our dataset, this filter removes 50.6% of the pairs that pass *filter 1*.

[4] From December 2015 edX does not issue honor certificates verified ones.

3. ***More than 5% of the master's correct submissions are potentially harvested.*** The rationale is to have a 'significance level' threshold on the amount of questions that the master is suspected to harvest. The specific value is a modification of the 1% threshold used for obtaining the results reported in (<Reference removed for anonymity>, 2016). When verifying manually a subset of 4 users who are suspected to harvest 1-5% of their correct submissions, we found that one of them was a false positive (not a master). Therefore, we decided to raise the limit used in (<Reference removed for anonymity>, 2016). Overall, on our dataset, this criterion removes 2% of the pairs that pass *filter 2*.

4. ***Evidence of 'inhumanly fast' submissions.*** A very short delay between opening the problem and submitting a correct answer to it makes the submission extremely suspicious (Palazzo, Lee, Warnakulasooriya, & Pritchard, 2010). In order to pass this filter, a potential master has to have a minimum number of inhumanly fast events. Thus this filter has two parameters:
   t = the upper bound for 'inhumanly fast' (i.e., if the submission takes *more* than t seconds, it is not considered suspicious).
   n = the minimum number of CAMEO events in which the delay between open and submit is *shorter* than t.
   We use t = 30 second (Palazzo et al., 2010), and n =6. We note that as opposed to Palazzo et al., we do not use the fast submission criterion as a mandatory criterion for an illegitimate *event* (in our case CAMEO; Palazzo et al.'s considered copying). We use it as a mandatory criterion for a CAMEO *user*. Once a user is identified as a 'CAMEO user' (after passing all the criteria), all the suspected-to-be-CAMEO events of this user are treated as CAMEO events. On our dataset, this filter removes 4.5% of the pairs that pass *filter 3*. This filter was not considered in (<Reference removed for anonymity>, 2016).

5. ***Harvester works for masters***. We require that most of the questions done by the harvester (more than 55%) were *actually used* by a master account. The rationale is that an account whose sole purpose is supporting another account should not do useless work. The specific value (55%) was picked by observing an elbow in the graph of the function (the amount of master-harvester pairs as a function of the fraction of questions that the harvester solved, or asked to see the answer for, and were used by the master). On our dataset, this filter removes 50.3% of the pairs that passed *filter 4.*

We note that there are several reasons to allow for some flexibility here (namely, allow for some fraction of questions done by the harvester that are not actually used). For example, to compensate for cases in which the user harvested the solution to a randomized question in the harvester account, and then either discovered that it is randomized and decided to skip it in the master account, or submitted a wrong answer in the master account. (Our algorithm overlooks such 'unsuccessful CAMEO' events in which the master submits an incorrect answer due to randomization; however we did observe such events.)

6. ***The harvesting account must not exhibit 'master' behavior (and vice versa)***. The rationale is that an account that is a 'service' account is not likely to use harvester accounts, and that an account that is a *master* is not likely to 'service' other accounts. We note that such a behavior would be expected of two students who collaborate, but identifying this behavior is not in the scope of this study. Technically, this criterion means that for each couple <*master*, *harvester*> that passes this criterion, the order of correct submissions made by *master* and *harvester* to in-common questions is always the same. (Such correct submissions to in-common items are identified as 'exhaustive search' CAMEO in *step 1*.). On our dataset, this filter removes 7.1% of the pairs that passed *filter 5*.

   Users whose master accounts pass these filters are termed *CAMEO users.*

2.2.3. Effect of parameters on detection rate

Using different choices for the algorithm changes the amount users and events that are detected.

In *Appendix A* we demonstrate the effect on the detection rate when using different parameters for: i) the delay between harvesting the solution and submitting it (see *step 1* of the algorithm); ii) considering only events from the same IP vs. event from IP group, and iii) considering only 'show answer' harvesting vs. considering both 'show answer' and 'exhaustive search'.

2.2.4. Verifying the results

The success of our methodology in identifying obvious cases of CAMEO was verified using a quantitative and a qualitative approach. First, we compare the statistical signature of

the CAMEO events that we detect (i.e., the CAMEO events of the master-harvester pairs that passed all the filters) to the signature of similar events between pairs of random users. This is presented in *Appendix B*, which shows that the distribution of the delay between the master-harvester events, and the distribution of the delay between such events among pairs of random users (with delay<24 hours), are significantly different.

Second, we analyze in-depth the log files of a sample of 10 master accounts that were picked at random from the 65 certified master accounts that our algorithm detects, and the harvesting accounts that these masters operated (some of these 10 master accounts operated more than one harvesting account). In our judgement, all the masters that we examined were real CAMEO users.

Figure 3 provides two examples taken from the log files of two of these ten masters, and their harvesting accounts. Each example contains a series of actions made by the master and the harvester, placed on a joint time-scale, with the time gap $\delta(t)$ between each event and the event that preceded it, to emphasize the proximity and pace. We classify these examples as unequivocal CAMEO. Both of these two master accounts were not detected by NHC's algorithm.

These examples also demonstrate some of the variety found in the behavior of the CAMEO users. The example on the left illustrates a harvesting sequence in which the user first opens the questions in the harvester account (without making any attempt to solve it in the master first), asks to see the solutions for few questions (which reside on the same html page), and then, in the master account, inserts these solutions in a row. The example on the right demonstrates a harvesting event in which the user first makes an unsuccessful attempt to solve the question legitimately in the master account, then goes to the harvester account, asks to see the answer, and returns to submit it in the master account.

In the next subsection we elaborate on these two modus operandi, which we term *premeditated* and *help-seeking* CAMEO.

Time	δ(time)	Harvester	Master
09:34:27		Goto page	
09:34:28	0:01		Goto page
09:36:13	1:45	Show answer Pivoted_bar_1	
09:36:14	0:01	Show answer Pivoted_bar_2	
09:36:16	0:02	Show answer Pivoted_bar_3	
09:36:19	0:03	Show answer Pivoted_bar_4	
09:36:38	0:19		check_problem_correct Pivoted_bar_1
09:36:59	0:21		check_problem_correct Pivoted_bar_2
09:37:23	0:24		check_problem_correct Pivoted_bar_3
09:37:45	0:22		check_problem_correct Pivoted_bar_4

Time	δ(time)	Harvester	Master	Comments
22:42:48			check_problem_correct Yo-Yo_Part_5	
22:46:29	3:41		check_problem_incorrect Yo-Yo_Part_4	The long delay before the incorrect attempt might be indicating that the user tried to solve the question legitimately
22:48:27	1:58	Log into edX		
22:49:05	0:38	Goto page of Yo-Yo_Part_4		
22:49:24	0:19	check_problem_incorrect Yo-Yo_Part_4		'show answer' is enabled after exhausting all the (two) attempts
22:49:45	0:21	check_problem_incorrect Yo-Yo_Part_4		
22:49:48	0:03	Show answer Yo-Yo_Part_4		
22:49:56	0:08		check_problem_correct Yo-Yo_Part_4	

Figure 2: Events from log files of two CAMEO users: premeditated at left, help-seeking on right.

Confidence Interval. Based on this detailed analysis, there is 90% confidence that at least 52 of the master accounts receiving certificates are demonstrably using this technique.

2.3. Help-seeking vs. Premeditated CAMEO

We distinguish two modes in which CAMEO might be used that have different educational implications. One is using CAMEO as a *help seeking* strategy, i.e., for finding the solution after a student tries unsuccessfully to solve the question (and without losing credit). Alternatively, and worse both morally and educationally, the student may make a *premeditated* decision to use CAMEO to obtain correct answers before even trying the question in the master account.

Operationalization. Each CAMEO event may be classified as either *help seeking* or *premeditated* using the following operational criterion:

- *Help seeking*: The master made an incorrect attempt, or observed the question for at least 30 seconds, before the question was opened in the harvester account. This gives an indication that the student tried to solve the question legitimately before resorting to CAMEO.
- *Premeditated*: Otherwise, we consider a CAMEO event as premeditated.

A graphical illustration of help mode CAMEO is given in Figure 2.

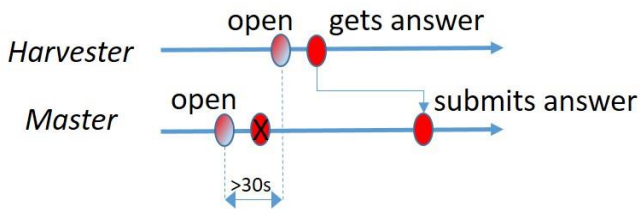

Figure 3: Help mode CAMEO

Not surprisingly, we found a strong correlation between the amount of CAMEO events performed by a CAMEO user, and the fraction of these events that were premeditated (see the Results section).

3. Results

This section is arranged as follows. In the first two subsections, we focus on CAMEO among certificate earners, and show how the total amount is distributed over the time-span of the course. In the third subsection, we investigate CAMEO found among accounts who did not earn certificates, and present findings that shed light on the motivation for non-certificatees to use CAMEO. Last, we present findings that suggest means for decreasing the amount of CAMEO. In some of the subsections, the findings are immediately followed by relatively straightforward conclusions that we want to present when the data is still fresh in the reader's mind. Interpretation that relies on a longer chain of reasoning is placed in the Discussion section.

3.1. The course

We used the algorithm described on the Methodology section to analyze the amount of CAMEO in the 2014 instance of the introductory physics MOOC <removed for anonymity> offered by <removed for anonymity> through edX. The course lasted for 14 weeks, with content divided between 12 mandatory units and two optional ones.

3.1.1. Research population

The course attracted about 13500 registrants, from which 502 earned a certificate. Gender distribution was 83% males, 17% females. Age varied (roughly) from 15 to 75, with 45% of the students under 26, 39% in the range of 26 to 40, and 16% 41 and above.

Education distribution was 37.7% secondary diploma or less, 34.5% College Degree, and 24.9% Advanced Degree. Geographic distribution included the 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 and are typical of MIT MOOCs.)

3.2. Total amount of CAMEO among certificate earners

First, we look at the total amount of CAMEO performed by the certificate earners in the course. This is demonstrated in Table 1. It shows that our algorithm detected 65 master accounts – 12.9% of the certificate earners in the course. As we explained in Subsection 2.4 (“Verifying the results”), the 90% confidence interval is of at least 52 CAMEO users. From here after we refer to final result of our algorithm (65 certified master accounts) as the reference point. These accounts operated 78 harvesting accounts (some masters used more than one harvester, probably to increase the number of tries – very useful for multiple choice questions with small number of allowed attempts). These masters harvested 17350 correct answers – 4.3% of *all* the correct answers submitted by certificate earners (including non-CAMEO users). The table also shows the distribution between *help seeking* and *premeditated* CAMEO modes. The table also shows the results for the Non-certificate earners, which are analyzed in Subsection 3.4 (among the non-certificated users, we consider only those who completed at least 5% of the assessment items in the course; total of 1079 accounts).

Table 1: Amount of CAMEO by certificate earners

	#Master accounts	#Harvester accounts	#Harvested answers	Help vs. Premeditated
Certificate earners	65 (12.9%)	78	17350 (4.3%)	22% help 78% pre'
Non-certificate earners	84 (7.7%)	74	12438 (5.1%)	15% help 85% pre'

Next, we look on how the events are distributed between the master accounts. This is shown in Figure 3a (left). The figure shows the percentage of the certificate earners (*x*-axis) who harvested at least y% of their correct answers. The point (3.7, 50) means that 3.7% of the certificate earners used CAMEO to obtain more than 50% of their correct answers

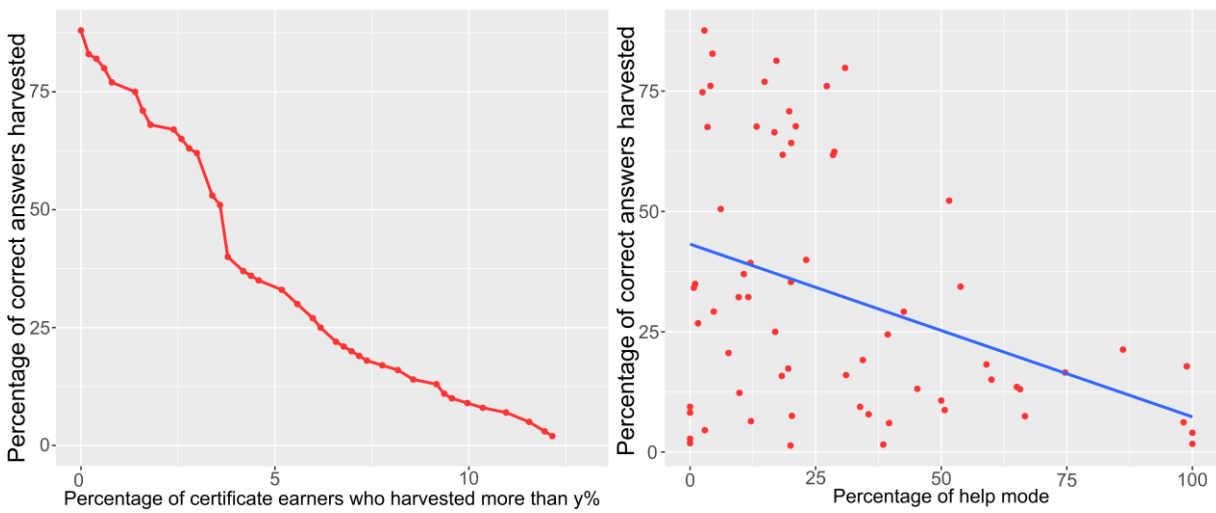

Figure 3: a) Amount of CAMEO among students. b) Amount of CAMEO vs. amount of help mode

Distribution of events between the accounts. As can be seen in the graph, the CAMEO events are distributed unevenly between the accounts. The shoulder at ~(3,60) likely bounds those must employ CAMEO to reach certification (60% of total credit).

Heavy CAMEO users are more premeditated. 'Heavy' CAMEO users tend to be more premeditated. This is illustrated in Figure 3b. There is a negative correlation of 0.359 between

the amount of CAMEO (fraction of correct answers that were obtained using CAMEO), and the fraction of it that is 'help seeking'. The obvious cluster of points above 50% CAMEO usage shows very few who used over 25% of help mode, showing that they rarely made a serious attempt to answer a question they first saw in the master account.

3.3. Distribution of CAMEO over course timeline

We now analyse the distribution of CAMEO by certificate earners over the different sections of the course. We partitioned the 13 chapters of our course into 10 sections. Some chapters were considered together because they have common quiz and homework (the combined chapters are 1 and 2, 4 and 5, and 9 and 10). For each section, we calculated the amount of CAMEO on questions that belong to the chapters that were mapped into this section. Most students follow the linear order of chapters working mostly in the days prior to the due dates in the course, hence this binning of activity by sections is a good approximation of binning by time intervals. Thus we refer to it as 'temporal' analysis. Moreover, it also reflects students' progress in terms of cumulative points earned. The rationale for binning behaviour over time is that correlating CAMEO with other temporal measures can shed light on students' behavior.

The findings of this temporal analysis are presented in Figure 4. The figure shows, the percentage of the questions in each section that were *attempted, correct,* and *harvested* (for all the lines, the '100%' baseline is the total number of questions in the section). In addition, the figure also shows, per section, that fraction of the accounts that passed the certification criterion in this section (i.e., moved from below to above 60% of total points in the course).

As can be seen in the graph, about 85% of the certificatees passed the certification point in section 7 (chapters 9+10). It is clearly seen that in this section and subsequently, both the percentage of questions tried, correct, and harvested drops significantly. We believe that this finding strongly supports the hypothesis that for most students using CAMEO, the main motivation is obtaining enough credit for a certificate.

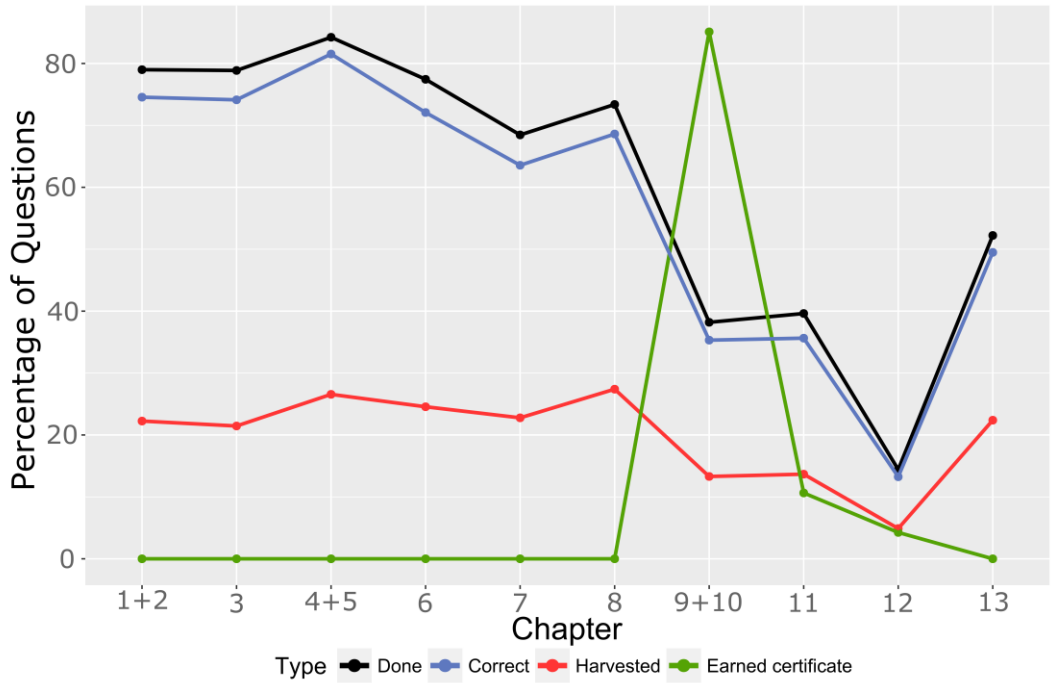

Figure 4: Amount of activity by chapter

3.4. CAMEO by non-certificate earners

In Subsection 3.2 and 3.3 we presented findings on the amount of CAMEO performed by certificate earners. Now we turn our attention to CAMEO among those non-certificated accounts (who completed at least 5% of the assessment items in the course).

The overall amount of CAEMO among this group is shown in Table 1. The table shows that 84 (7.7%) of the non-certificate earners were master accounts, and that these accounts operated 74 harvesting accounts. The fact that in this group we detect fewer harvester than masters might indicate that some of the non-certificated master accounts are also fake accounts. For example, we found that 15 of the harvester accounts operated by the non-certificated masters were also used by *certificated* master accounts. In total, 5.1% of the correct answers submitted by non-certificate earners were harvested. Among the CAMEO events in this group, 15% were help seeking.

Since it appears that the main motivation for CAMEO is improving grades for earning a certificate, finding CAMEO among non-certificatees was somewhat surprising. We now examine factors that we believe can shed light on the motivation of uncertificated users to

perform CAMEO, by comparing them with the certificated CAMEO users. The first factor is the percentage of questions done that were harvested. Second is the distribution of events between *help-seeking* and *premeditated* CAMEO. Third is the level of CAMEO over time (again, operationalize as by-section).

Table 2 shows the fraction of correct answers that were harvested, and the fraction of the harvested events that were 'help seeking' (i.e., the user tried to solve the question before going to the harvester account) among the two groups. The table shows *mean values* for certificate/non-certificate CAMEO *users* (as opposed to table 1, which aggregates over all the events).

Table 2: Percentage and purpose of harvesting among certificated/non-certificated users

	Certificated	Non-certificated	p.value (cert. <non_cert)
% Correct submissions that were harvested	32.5%	51.0%	*p.value* < 0.0001
Fraction of harvesting that is premeditated	70.4%	82.9%	*p.value* = 0.001

The table shows that:

- Non-certificated CAMEO users harvested a much higher fraction of their correct submissions than certificated CAMEO users.
- Non-certificated CAMEO users are more premeditated than certified CAMEO users.

Behavior over time. Examining the behavior of the *non-certificated* masters over time (figure is omitted) reveals that the non-certificated masters were very active on the beginning of the course – attempted about 75% of the question of Chapters 1 and 2 (with average cheating of more than 50%), but then decreased their amount of activity rapidly and steadily, attempting less than 10% of the questions on Chapter 8. We interpret this behavior as stopping out from the course (since they did not formally signed-out, we don't consider them "drop-outs").

Conclusion on motivation of non-certificated masters. In our judgement, the most reasonable explanation for the use of CAMEO by those who did not receive a certificate is that they entered the course with the intention of earning a certificate using this method, but stopped-out. Quite probably they found our course 'less friendly' for CAMEO, since it contains a large number (~1000) of questions, many of them randomized (which makes CAMEO more difficult). This is furthered discussed in the Discussion section.

3.5. *Question parameters associated with reduced CAMEO*

As educators, it is very important for us to find ways to reduce cheating in general and CAMEO in particular. Two practices that instructors can take to make CAMEO more difficult to perform are using randomized questions, and delaying exposing the solution until after the due date. We use these practices on some of the questions in our course. The dependence of CAMEO on these variables is shown below.

3.5.1. Randomization

Problems in the edX platform can be set to use randomized numerical parameters so that, different user accounts would have different angles of slope in a question dealing with a ball rolling on a ramp, with (therefore) different values for the answer. Randomization is motivated by reducing answer transfer between different students, and this strategy frustrates CAMEO since the answer is different in the harvester and master accounts. Table 3 presents the percentage of correct submissions that were harvested for randomized and non-randomized questions.

Table 3: Harvesting on random and non-random questions

	Questions with random parameters (N=52744)	Non-random questions (N=501442)
Percentage of harvesting out of total submissions	4.06% (1956 out of 48127)	6.07%. (27832 out of 458223)

A *t-test* confirms that a submission to a random question is less likely to be harvested, with *p-value* < 0.001. We believe that this finding indicates a causal relation, and demonstrates the effectiveness of randomization against CAMEO. This is further discussed in Section 4 (Discussion) and 5 (Implications).

3.5.2. Delaying 'show answer' feedback

CAMEO is based on getting the correct answer in the harvester account either by 'show answer', or by exhaustive search. Thus, limiting the feedback, or the number of allowed attempts, is expected to make CAMEO harder. Given that CAMEO depends on the weight of the question (<Reference removed for anonymity>, 2016) the effect of 'show answer' feedback must be evaluated on questions of the same weight. Fortunately, in Quiz_9_10 (the

quiz given after chapters 9 and 10) show answer was accidentally enabled due to mistake in settings. For this quiz question, 6.4% of the submissions were harvested. For the rest of the chapter quizzes the average number of harvested submissions was 3.83%. Similarly, we found that in the midterm exam, with show answer disabled, the fraction of submissions that was harvested was 3.43%, whilst in the final exam, in which show answer was enabled (again, by mistake), it went up to 6.35%. Note that by the final exam, many students had earned their certificate, lowering the incentive to CAMEO. Otherwise we would expect even higher levels of CAMEO.

Clearly removing 'show answer' feedback substantially reduces CAMEO. 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 (on edX). We discuss the pedagogical downsides of limiting feedback in Sections 4 and 5.

3.6. Summary of findings

To summarize the findings, we see that:

- 12.9% of the certificate earners obtained more than 1% of their correct answers using CAMEO.
- 3.7% of the certificate earners obtained more than 50% of their correct answers using CAMEO.
- Most of the CAMEO users significantly reduced their harvesting upon qualifying for a certificate.
- The majority of the CAMEO events (78% among the certificate earners) are *premeditated*, namely, there is no evidence that the master tried to solve the question legitimately before opening it in the harvester account. Also, 'heavy' CAMEO users are more premeditated (correlation of 0.359 between fraction of correct answers that were harvested, and fraction of harvested that are premeditated)
- Non-certificated master accounts tended to be more cynical users than the certificated masters (used CAMEO for a larger fraction of their correct answers and were more premeditated) in the beginning of the course, and then stopped-out.
- Randomization and omitting 'show answer' feedback are independently correlated with a reduction of ~ 2x in the amount of CAMEO

4. Discussion

In the Introduction we have formulated the following questions:

1. *How many students are practicing CAMEO in our course, and how do they use it?*
2. *What can we infer about students' motivations to practice CAMEO from their observed behavior?*
3. *What is found to reduce CAMEO?*

The Discussion is arranged as follows. First we discuss the amount of CAMEO in our course, and discuss its internal validity and likely presence in other courses. Second, we discuss the motivation for CAMEO. Third, we explain why CAMEO is a special form of academic dishonesty. Forth, we analyze the threat that it poses to the value of MOOC certificates. Last, we discuss ways to reduce CAMEO.

4.1. Amount of CAMEO in our Course and Implications for Other Courses

In Subsection 3.2 we presented findings regarding the amount of CAMEO found in our course. According to our algorithm, 65 accounts, representing 12.9% of the certificated users, used CAMEO to obtain more than 1% of their correct answers (and at least 10 questions). We verify the results of the algorithm in two ways. First, by comparing the distribution of the delay between master:harvester events identified as CAMEO, to the delay between similar events between pairs of random users. As we show in *Appendix B*, the two distributions differ significantly. Second, we analysed in-depth a random sample of ten users, and to our judgement, all were obviously real CAMEO users. As argued in Subsection 2.4, based on the sample, the 90% confidence interval is of [52…65] master accounts among the certificated earners.

Comparison to NHC Our algorithm detects about 4-5x more master accounts than reported by NHC for our course (private communication) – 12.9% vs. 2.4% of the certificated earners respectively (our 90% confidence interval of 52...65, vs. the 12 accounts detected by NHC's algorithm). This raises the question of which result is more representative. We take for granted that their result is a lower bound (both because of their rigorous method, and because all their master accounts are also identified by our algorithm), and also note that obtaining a rigorous lower bound was the purpose of their study. We basically claim that our algorithm

represents more accurately the *true* amount of CAMEO users in our course, while still being conservative (fewer false positives than false negatives).

From analysing NHC's algorithm, we see several reasons that might explain the difference in the amount of detection. First, our algorithm detects more events, principally due to two facts:

i. We consider two harvesting methods – 'show answer' and 'exhaustive search', while NHC's algorithm considers only harvesting using 'show answer'. *Appendix B* show that including exhaustive search increases the number detected events by nearly a factor of 2.
ii. We consider a time limit of 24 hours, while NHC's algorithm requires that 90% of the harvesting events fall within 5 minutes.

Since NHC's algorithm requires a relatively long sequence of events in order to identify someone as 'master' (explicitly stated in p. 4 of their paper), ignoring events might lead this algorithm to overlook 'true' masters'.

Their algorithm requires that 90% of the in-common events between a master:harvester pair will be within 5 minutes (master after harvester). This constraint is very strict and can cause the algorithm to overlook CAMEO users that some of their events are of longer delay. Longer delays can be caused for example when the user performs what we call *batch mode* CAMEO (<Reference removed for anonymity>, 2016). This refers to a very efficient modus operandi in which the user harvests a sequence of solutions in the harvester account, and then submits them in a row in the master account (we saw sequences of up to 40 questions, with less than 30 seconds between successive submissions). In such cases the delay between the event of harvesting the solution to a specific question, and the event of submitting this solution in the master account, can be longer than five minutes even if the 'submission sequence' is performed immediately after the 'harvesting sequence'. We believe that this is one of the reasons why NHC's algorithm did not detect seven out of the ten most 'heavy' master accounts that we found in our course.

Additionally, NHC's algorithm can exclude real masters on questions that have several sub-questions where the 'show answer' button is common to all subsections. If the master account submits a correct answer to one of these subsections (even by guessing), then the user goes to the harvester account and asks to see the solution. This will be interpreted by NHC's algorithm as a mis-ordered pair (since the sub-problem was answered correctly before asking to see the answer for this sub-problem in the harvester account). Again, because of

their *90% criterion*, even a few such mis-ordered pairs can cause the algorithm to exclude a true master-harvester pair.

This does not mean that our algorithm is less strict. We apply filters that are not considered, or are stricter, than the ones considered by NHC:

i. Master does not exhibit harvester behaviour, and vice versa (not included in their criteria).
ii. Most items done by the harvester are *actually used* by the master. NHC require that among the items that are *in-common* between the harvester and the master, at least 90% are actually used. However they do not examine the items done by the harvester that are *not* in-common with the master. To our judgment, it is not reasonable to have many such items, as they might indicate that the harvester is working for a purpose different than serving the master account. Thus we constrain the amount of items done by the harvester that are not done by the master to less than 0.45% of the items done by the harvester (the threshold is determined by observing an elbow in the relevant curve; see filter 5).

We have also observed unsuccessful CAMEO events. These occur on randomized questions – users who tried to plug-in the answer from the harvester account in the master account, probably because they did not notice that the parameters of the question in both accounts are different; A detailed example is provided in Subsection 4.5.

To conclude this analysis of the difference in the results between the two algorithms, NHC were intentionally conservative, and our algorithm is more inclusive, yet still conservative. We have argued that our results are a much more accurate estimation of the *true* amount of CAMEO in our course. Whether such a big difference exists in other courses is a question that we intend to study.

Generalizability of the results. We believe that the results of our research are generalizable to other edX MOOCs that are similar to our course in terms of demographics, and that enable to receive feedback on a significant fraction of the questions in the course. These characteristics represent many of the MITx courses in science and engineering. For such courses, and based on our findings, we can anticipate that 5 to 10% of the certified users are CAMEO users.

One reason to believe that the amount of CAMEO in our course is on the lower bar is the fact that we teach Introductory Physics, which carries little value for the labor market relative to computer programming, business skills, etc.
On the other hand, there are reasons that can lead to a relatively high amount of CAMEO in our course. For example, the fact that it contains many questions (more than 1000), make it very time consuming, thus maybe pushing some students to look for shortcuts. This is inline with the fact that we saw a fair number of CAMEO users who stop out (and quite possibly move their CAMEO efforts to another MOOC).

4.2. Motivation: Certificates

We argue that the main motivation for harvesting solutions is earning points for a certificate. This is supported by our findings that:

1. Certificated CAMEO users reduce their level of activity dramatically after earning the certificate (we note that certificated users who did not use CAMEO, showed a less dramatic decline after certification (Subsection 3.3).
2. High-stakes questions are more likely to be harvested (see <Reference removed for anonymity>, 2016)
3. CAMEO users preferentially harvest questions that are either for high credit or quick for CAMEO, suggesting that quickly accumulating enough points to pass is a priority.
4. An alternative explanation that CAMEO is used on questions that the student perceives as difficult contradicts with the finding that in most cases, CAMEO is premeditated. Also, the ‘help seeking’ mode decreases in favor of the more cynical premeditated mode until they qualify for certification.
5. Findings on CAMEO performed by non-certificate earners, which could potentially weaken the hypothesis that CAMEO is for certificate, actually show that these individuals are probably experienced CAMEO users who are looking for easy course to pass using this method.

4.3. CAMEO is a form of academic dishonesty

We refer to academic dishonesty as a “transgression against academic integrity which entails taking an unfair advantage that results in a misrepresentation of a student’s ability and grasp of knowledge” (King, Guyette, & Piotrowski, 2009, p. 4), or as “any fraudulent action

or attempt to use unauthorized or unacceptable means in any academic work" (Lambert, Hogan, & Barton, 2003; Palazzo, 2006). Following these, CAMEO is obviously a form of academic dishonesty (or cheating, the term used by King et al.). As we pointed out in the introduction, CAMEO is unauthorized because it clearly violates the user agreement of edX.org.

In the previous subsection we argued that the main motivation for CAMEO is earning a certificate. Certificate is an acknowledgement of proficiency. Undoubtedly, CAMEO leads to misrepresentation of ability. We find it unlikely for a reasonable student to assume that CAMEO is a legitimate strategy for earning a certificate, as the purpose of CAMEO is answering correctly without the need to possess knowledge.

There is a question of what is the threshold for identifying someone as a CAMEO user (namely, a cheater). For example, our threshold of 10 correct answers (which is only one of the filtering criteria) is motivated by prediction accuracy, and it does not mean that less than that is acceptable. Obviously, using CAMEO to obtain the 70% of one's correct answers is much more severe than using it to obtain 1% of them, because it is a higher level of 'misrepresentation of ability'. Additionally, we judge that using CAMEO in a premeditated mode is more severe than using it in help-seeking mode. But discussing these moral issues, and their translation to actions (for example, what level of CAMEO is severe enough to cancel one's certificate) are outside the scope of this paper.

4.4. Threat to the Value of MOOC Certificates

More than 10% of the certificated earners in our course are already using CAMEO to a significant extent, and about 3.7% of them used it to obtain the majority of their correct answers. Moreover, unsupervised students have other ways to obtain answers dishonestly, that we do not detect yet, for example getting answers from other students (Palazzo et al., 2010). CAMEO in particular, and cheating in general, decrease the evidentiary value of the certificate as evidence of proficiency. Thus, CAMEO and other cheating methods pose a threat to the professional and academic value of MOOC certificates.

MOOC providers are already sensitive to cheating issues, and have addressed concerns about identity and impersonation (e.g. getting an expert to earn a certificate in an account bearing the cheater's name). Methods currently used (monitoring active webcams and analyzing keystroke patterns) assure the identity of each student, but offer no obvious defense against CAMEO.

4.5. Instructional Design that Reduces CAMEO

Following the findings that show that CAMEO is already significant, and likely to increase, we recommend ways to reduce it. Our focus is on instructional design methods, i.e., means that are at the hand of the instructors and course designers. The findings that we present in Subsection 3.5 shows that on randomized questions, and on questions with delayed feedback, there is about 2x less CAMEO, respectively. Examining the effect of randomization using a broader brush, NHC also found that courses that use randomization has 2x less CAMEO.

Randomization. EdX allows randomization of some of the parameters of the question, so different accounts get a question with different parameters that (are designed to) lead to different correct answers. We have seen this frustrate students who found that the correct answer in the harvester account was graded "wrong" in the master. For example, one of the students sent an email to the teaching staff, claiming that the answer to one of the questions in the course was changed during the last week, so his answer that was graded as 'correct' a week ago is now grade as 'incorrect'. The student also attached two screenshots, claimed to be taken 'a week ago' (with the answer graded as 'correct'), and 'yesterday' (with the answer graded as 'incorrect'). None of this screenshots included the user name on them. Upon checking the settings of this question, the instructors found that it is a random question, and realized that the user is trying to submit in one account a solution that is correct in the version of the question that the other account sees (in fact, the screenshots did presented two slightly different questions, but the student did not realized that the questions are different, or that this difference affects the result). Later, we found that the user who sent this email appeared as one of the master accounts detected by our algorithm.

Though this kind of randomization makes CAMEO harder, it does not eliminate it entirely. Students can check a symbolic expression, or infer the scaling of the solution with parameters if they use several harvester accounts. The randomization is typically limited to several options (because randomization is not fully automated and requires manual work per alternative), so by using multiple harvesting accounts masters can increase the likelihood of seeing the same variation in their master account and in one of their harvesting accounts.

A more general solution is having question pools. By that we mean that per topic, there is a pool of questions with different levels of complexity. Developing comprehensive question pools is an effort that requires considerable resources. Thus practically, it is not a solution that can be implemented by a single instructor or even a small course team. It requires a more systematic pipeline that also needs to be supported by appropriate technology

to develop the questions, assess level of difficulty, and share them between course developers. For fairness to students, it would require a grading scheme that accounted for the measured differences among the questions such as Item Response Theory.

Delayed feedback. Another option to decrease CAMEO is to delay feedback. Our findings show that on questions for which the 'show answer' is available only after the deadline, there is about 2x less CAMEO, than when show answer was available also before the deadline (see Subsection 3.5). Obviously, 'show answer' is the most convenient way to harvest solutions. Without it, users are left to harvest the solutions using the correct/incorrect feedback that is always given – an approach that is much less efficient, especially on questions that are not multiple choice.

The main disadvantage with 'delayed feedback' is that this instructional design pattern is counter-pedagogic. Feedback, especially instant one, is very important for learning. Also, on self-paced courses that do not have rigid deadlines, designers who want to use this method are more or less left with the option of not giving feedback at all. Altogether, using 'delayed feedback' as a cheating prevention method means favoring security considerations over pedagogic ones. Since this choice means reducing the learning experience of the all the users because of the dishonest behavior of some of them, it a sort of 'collective punishment'. Thus we advise using it only on high-stake questions.

Recommendation. Bottom line, randomization and delayed feedback are both effective means against CAMEO that are at the hand of the instructors, with a trade-off between pedagogy, prevention and the amount of time needed to set-up. Our recommendation is to use randomization as the first choice, and on high-stake questions, to delay the feedback and change the exams from year to year so that students who fail one year cannot use it the second year.

For recommendations that are not specific to MOOCs and CAMEO, such as the use of *honor code,* please refer to Section 5 – Related Work. This section surveys (among other things) factors that were found to reduce academic dishonesty in conventional educational settings, and might be effective also in MOOCs.

4.6. Implications for Educational Research

Besides being a threat to the value of the certificates, CAMEO also has the potential of seriously interfering with educational research in MOOCs. We find that those accounts with

the highest ability in our course are masters, and that those with the lowest ability are harvesters (See <Reference removed for anonymity>, 2016), and the figure in *Appendix C*: Distribution of success among masters, harvesters, and the rest of the students). Thus in any research that tries to identify the variables that most strongly correlate with students skills, this subset of harvester and master accounts would have a disproportionate weight in the results.

For example, consider a study that tries to quantify the effectiveness of various kinds of learning activities. Masters tend to have a very high success rate, achieved in a way that does not require interacting with the course materials (videos, e-text pages). Thus, if considering only certificated students (a common approach in MOOC research) and therefore including master accounts but not the harvesting ones, one might observe a relation that is stronger than it ought to be between doing problems and success, whereas the relation between using the instructional materials and success would be weakened. Or, since masters tend to have very fast submissions, it can bias the results towards negative correlation between time on task and success.

Thus, being able to detect and remove master and harvester accounts from the data seems essential to reaching reliable results about education in MOOCs.

4.7. Limitations of this study

The main risk to the internal validity of this research lies in the lack of external evidence that specific users are cheating. Thus we rely on an unsupervised learning approach, i.e., detect CAMEO by analyzing patterns in the data. To limit the likelihood of false positive identification (namely, identifying 'innocent' users as using CAMEO), we use very strict criteria, and verify the results in various ways (see Subsection 2.2.4).

The main risk to the generalizability of the results to other MOOCs lies in the fact that we focus on one course. Our claim for generalizability is based on analyzing various characteristics of our course with respect to other MOOCs, and on the fact that our algorithm detects 4X more CAMEO users than the number of CAMEO users detected in our course by the algorithm of Northcutt et al (2016). However, it is in our roadmap to analyze courses in other domains and from other universities. Such a research will naturally be a more broad-brush kind of research, and will complete the in-depth analysis that is the focus of the current study.

5. Related Work

Academic dishonesty in MOOCs is a new research topic that was pioneered by (<Reference removed for anonymity>, 2015; Northcutt et al., 2016; <Reference removed for anonymity>, 2016). We find two lines of research particularly relevant to this topic – *gaming the system*, and *academic dishonesty* in general. Below we survey these topics and place our work in this context.

5.1. Academic Dishonesty

Our study extends the body of work on academic dishonesty with a detailed study of a new form of cheating in MOOCs, termed CAMEO, which is also relevant to other online learning platforms that enable the users to register with multiple accounts, and in which students can receive feedback on the correctness of their answers.

CAMEO can be classified either as 'General Cheating', or maybe a subcategory of 'General Plagiarism from Exterior Sources', using the categorization suggested by (Lambert et al., 2003).

Amount of cheating. Comparing our findings on the amount of cheating to numbers reported by previous studies on academic dishonesty shows that the numbers that we report fall on the lower scale. In the context of cheating in an online learning system, Palazzo et al. (2010) reported that overall, between 3 and 11% of the submissions were copied. We found that ~4% of the correct submissions made by certificate earners were harvested. Regarding cheating in more traditional settings, McCabe and Trevino (1993) surveyed studies reporting that "anywhere from 13 to 95 percent of college students engage in some form of academic dishonest" (p. 3). Our algorithm detected that 12.9% of the certificate earners used CAMEO to some extent – also in the lower scale on the amount of students. Examples of other studies include the work of Witherspoon, Maldonado, and Lacey (2012) who reported that most students cheat occasionally, but that only small number are a flagrant cheaters; Balbuena and Lamela (2015), who reported that 67% of the students cheated on more than one exam, and more. A methodological shortcoming that is common to most of the previous work on academic dishonesty is reliance on students' self-report (Palazzo, 2006). A key advantage of detection algorithms such as ours, is that they do not rely on self-report of the research subjects.

Academic dishonesty is affected by many factors. Below we review some of them.

Demographic factors. Such factors that were studied in the context of academic dishonesty are *gender* (Anderman & Midgley, 2004; Bogle, 2015; Harding, Mayhew, Finelli, &

Carpenter, 2007; Witmer & Johansson, 2015), *age* (Anderman & Midgley, 2004) and *educational level* (Harding et al., 2007).

Effect of Personality. Another aspect that was considered is one's personality, and its relation to academically dishonest behavior (Anderman, Cupp, & Lane, 2009; De Bruin & Rudnick, 2007; Giluk & Postlethwaite, 2015; Harding et al., 2007; Jensen, Arnett, Feldman, & Cauffman, 2002; Jordan, 2001; Sanecka & Baran, 2015).

Learning Environment. Regarding the effect of the learning environment, MOOCs bear characteristics that have been found to correlate with dishonest behavior. An analysis of ~ 80 studies (Palazzo, 2006) showed that academic dishonesty is significantly increased for large and public institutions, vs. small private ones – certainly MOOCs seem more like the former. Furthermore, this study cited several papers showing the effects of Classroom Environment, concluding that "smaller classes with more individualized attention and increased student professor interaction" can reduce cheating – MOOCs have quite opposite characteristics.

Peers. Regarding the effect of peers – McCabe, Trevino and Butterfield (2001) argued that dishonest behavior is strongly affected by students' perception of peers' behavior. Peers provide not only methods, but also a kind of normative support (Payan, Reardon, & McCorkle, 2010). At some point, non-cheaters can feel that they are left at a disadvantage, pressing them to adopt dishonest behaviors even if they initially perceived them as illegitimate (McCabe & Trevino, 1993) . Once being involved in cheating, one might change his/her attitude towards it, and see it less as in conflict with moral rules, as was shown for example by Shu and Gino (2012). This view of cheating as a slippery slope is in line with our findings that over time, CAMEO tended to be used more in the *premeditated* form (which we interpret as more severe), and less as a *help-seeking* behavior.

Other environmental factors that was studied in the context of academic dishonesty include the role of the teachers and their attitudes (Anderman et al., 2009; Broeckelman-Post, 2008), the learning objectives (Kauffman & Young, 2015), and even features of the software that students use, such has the existence of copy/paste (Kauffman & Young, 2015), and more.

Prevention. Preventing academic dishonesty can be done through education (what is considered as cheating, why it is bad, etc.), thwarting (e.g., by making it harder to perform), and deterrence. In the context of education, several studies reported that when honor codes were clearly presented in class, the amount of cheating decreased significantly (LoSchiavo & Shatz, 2011).

Thwarting is of course tightly connected to the specific form of cheating it comes to prevent, such as human proctors to prevent cheating in exams, or individualizing the assignments to the class.

Regarding deterrence, according to reports, most college students believe that cheater do not get caught (Kleiner & Lord, 1999). Such perception makes it hard to diminish cheating (ref Scanlan, "Strategies to Promote a Climate of Academic Integrity and Minimize Student Cheating and Plagiarism"). Publishing the existence of technological detection methods such as ours, the results of applying them, and the enforcement acts that were taken, can help to change the perception that "crime pays".

Overall, cheating decreases the reliability of the assessment, and eventually can reduce the confidence that certificates of accomplishment that are based on this assessment carry meaningful information on one's abilities. They can alter the results of educational research (<Reference removed for anonymity>, 2016), and thus affect also policy decisions. Also, cheating most likely interferes with learning. Thus, preventing cheating is a challenge of significant importance for higher education. It is also true for MOOCs (Daradoumis, Bassi, Xhafa, & Caballé, 2013; Gupta & Sambyal, 2013; Siemens, 2013). Our study shows that cheating is indeed a serious issue in MOOCs, by providing an in-depth study of a new form of cheating that was reported in previous studies (<Reference removed for anonymity>, 2015; Northcutt et al., 2016s; <Reference removed for anonymity>, 2016).

5.2. Relationship with Gaming the System

We consider "gaming the system" as defined in (Desmarais & d Baker, 2012) -"attempting to succeed in an educational task by systematically taking advantage of properties and regularities in the system used to complete that task, rather than by thinking through the material". "Gaming the system" is considered as a tactic and strategy in tutoring systems (du Boulay & Luckin, 2015) and can be incorporated as a user feature into the user modeling (Desmarais & d Baker, 2012).

CAMEO may be thought of as "gaming the system" – a phrase familiar from work on other interactive tutors different from MOOCs. "Gaming" refers to exploiting some feature of the system to obtain the requested answer in an expedient manner that generally circumvents the intended process of learning designed into the system. For example if the designer gave a series of hints to help a student work through the problem, gaming might consist of rapidly clicking through the hints in the expectation that the last hint will reveal the answer. Similarly, trying the numbers 1-20 one after the other as the answer to an addition problem for two digits would be gaming. Different types of "gaming" have been enumerated in previous works such as help abuse, systematic guessing and checking or copying hints

(Muldner, Burleson, de Sande, & VanLehn, 2011; Wood & Wood, 1999). Thus, harvesting answers can be thought as a new specific case of "gaming" as it defeats the intended educational objectives and learning outcomes of the system. Although "gaming" might not be cheating because these behaviors might not be contrary to academic rules and it is not necessarily associated against a signed code of honor. Although other types of "gaming" have been studied in the literature of intelligent tutors, harvesting answers using multiple accounts has almost not been researched so far.

Perhaps the key issue is the extent to which gaming reduces learning, as was shown in several studies (R. S. Baker, Corbett, Koedinger, & Wagner, 2004; Fancsali, 2013; Walonoski & Heffernan, 2006). Given the analogy between "gaming the system" and CAMEO, we can assume that the latter will interfere with *real* learning, especially in the case of premeditated CAMEO. However it was shown that 'gaming' does not always interfere with learning (Aleven, Roll, McLaren, & Koedinger, 2016). In our context, it worth studying whether the use of CAMEO in help-seeking mode can have positive impact on student's learning.

Different detectors of "gaming" have been implemented, based on predefined rules (e.g., Muldner et al., 2011; Muñoz-Merino, Valiente, & Kloos, 2013) or on machine learning techniques such as decision trees, Bayesian techniques, neural networks or logistic regression (R. S. Baker, Corbett, & Koedinger, 2004; Walonoski & Heffernan, 2006). Some of the parameters that were considered by these algorithms were Response time, number of attempts, flow of the sequence, number of times the student asked for help, and problem difficulty. Several learning environments have already incorporated such "gaming" detectors, including Assisstment (Walonoski, & Heffernan, 2006), Andes (Muldner et. al. 2011), Wayang Outpost (Beal, Qu, & Lee, 2006) or a learning analytics extension of the Khan Academy platform (Ruipérez-Valiente, Muñoz-Merino, Leony, & Kloos, 2015).

To prevent students from gaming the system, several methods were suggested, including the use of specific interfaces or delaying the help, and using intervention techniques (Baker et. al, 2004, Baker et. al, 2004a).

6. Conclusions and Future Work

In this research we have presented an in-depth study of a new form of cheating in MOOCs, termed CAMEO (Northcutt et al., 2016). CAMEO is based on using 'harvesting' accounts for collecting correct answers that are then submitted in the user's main account for credit. It exploits the fact that users can set-up multiple edX accounts, and that feedback

(either full answer or correct/incorrect) is available for many questions. CAMEO was studied in (<Reference removed for anonymity>, 2015; Northcutt et al., 2016; <Reference removed for anonymity>, 2016). Our current study provides new depth of understanding of this behavior, and ways to reduce it.

Summary of main findings. Our findings show that:

- Our algorithm detects 65 master accounts (90% confidence interval of [52…65]), which are 12.9% of the users, who have used CAMEO to obtain the correct answer to more than 10 questions (which is more than 1% of their correct answers), and that 3.7% of the certificated earners have acquired *most* of their correct answers by using this method.
- The main motivation for CAMEO is most likely earning a MOOC certificate.
- CAMEO can be significantly reduced using randomized questions and delayed feedback.

The results show that CAMEO is already a significant issue that can threaten the value of the MOOCs, as it reduces the confidence that a MOOC certificate is a valid representation of student's proficiency and introduces systematic distortions into educational research. Due to the growing value of MOOCs, and previous research on dishonest behavior, we believe that if not addressed properly, CAMEO and other forms of cheating in MOOCs will become even more prevalent.

Recommendations. Thus our conclusion is that this issue should be addressed on various level. We recommend to:

- Instructors: Increase the use of randomization when possible; delay feedback, especially on high-stake questions.
- Institutes: Acknowledge the significance of this issue, and allocate resources (such as time) to enable instructors to design their courses in a way that is less vulnerable to CAMEO.
- Platform: Include in the platform tools for detecting CAMEO on run-time, and devise detection methods to guarantee that certificates indicate knowledge and skill.

Future research. Current research has made only a first step in studying CAMEO. Directions for future research include:

- Investigate the amount of CAMEO on many courses, in various domains and from different institutes to find out how serious this problem is across the board.

- Develop ways to detect CAMEO without relying on IP address, which would also thwart sophisticated users who divide their accounts between different IPs, or use systems that hide the real IP.
- Develop ways to detect other kinds of cheating, such as students obtaining answers from other students.
- Pedagogy-wise, study the actual impact of CAMEO on the learning achieved by students who practice it to varying degrees, considering that previous research generally shows that cheating is associated with poor learning.

Acknowledgements

<Names removed for anonymity>

References

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help Helps, But Only So Much: Research on Help Seeking with Intelligent Tutoring Systems. *International Journal of Artificial Intelligence in Education*, *26*(1), 205–223. http://doi.org/10.1007/s40593-015-0089-1

Anderman, E. M., Cupp, P. K., & Lane, D. (2009). Impulsivity and academic cheating. *The Journal of Experimental Education*, *78*(1), 135–150.

Anderman, E. M., & Midgley, C. (2004). Changes in self-reported academic cheating across the transition from middle school to high school. *Contemporary Educational Psychology*, *29*(4), 499–517.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. In *Intelligent tutoring systems* (pp. 531–540).

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: when students game the system. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 383–390).

Baker, R. Sj., de Carvalho, A., Raspat, J., Aleven, V., Corbett, A. T., & Koedinger, K. R. (2009). Educational software features that encourage and discourage "gaming the system". In *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 475–482).

Balbuena, S. E., & Lamela, R. A. (2015). Prevalence, Motives, and Views of Academic Dishonesty in Higher Education. *Asia Pacific Journal of Multidisciplinary Research*, *3*(2).

Beal, C. R., Qu, L., & Lee, H. (2006). Classifying learner engagement through integration of multiple data sources. In *Proceedings of the National Conference on Artificial*

Intelligence (Vol. 21, p. 151).
Bogle, K. D. (2015). Effect of perspective, type of student, and gender on the attribution of cheating. In *Proceedings of the Oklahoma Academy of Science* (Vol. 80, pp. 91–97).
Broeckelman-Post, M. A. (2008). Faculty and student classroom influences on academic dishonesty. *Education, IEEE Transactions on*, *51*(2), 206–211.
Daradoumis, T., Bassi, R., Xhafa, F., & Caballé, S. (2013). A review on massive e-learning (MOOC) design, delivery and assessment. In *P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2013 Eighth International Conference on* (pp. 208–213).
De Bruin, G. P., & Rudnick, H. (2007). Examining the cheats: The role of conscientiousness and excitement seeking in academic dishonesty. *South African Journal of Psychology*, *37*(1), 153–164.
Desmarais, M. C., & d Baker, R. S. J. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*(1-2), 9–38.
du Boulay, B., & Luckin, R. (2015). Modelling Human Teaching Tactics and Strategies for Tutoring Systems: 14 Years On. *International Journal of Artificial Intelligence in Education*, 1–12.
Fancsali, S. E. (2013). Data-driven causal modeling of "gaming the system" and off-task behavior in Cognitive Tutor Algebra. In *NIPS Workshop on Data Driven Education*.
Giluk, T. L., & Postlethwaite, B. E. (2015). Big Five personality and academic dishonesty: A meta-analytic review. *Personality and Individual Differences*, *72*, 59–67.
Gupta, R., & Sambyal, N. (2013). An understanding approach towards MOOCs. *International Journal of Emerging Technology and Advanced Engineering*, *3*(6), 312–315.
Harding, T. S., Mayhew, M. J., Finelli, C. J., & Carpenter, D. D. (2007). The theory of planned behavior as a model of academic dishonesty in engineering and humanities undergraduates. *Ethics & Behavior*, *17*(3), 255–279.
Hopcroft, J., & Tarjan, R. (1973). Algorithm 447: Efficient Algorithms for Graph Manipulation. *Commun. ACM*, *16*(6), 372–378. http://doi.org/10.1145/362248.362272
Jensen, L. A., Arnett, J. J., Feldman, S. S., & Cauffman, E. (2002). It's wrong, but everybody does it: Academic dishonesty among high school and college students. *Contemporary Educational Psychology*, *27*(2), 209–228.
Jordan, A. E. (2001). College student cheating: The role of motivation, perceived norms, attitudes, and knowledge of institutional policy. *Ethics & Behavior*, *11*(3), 233–247.
Kauffman, Y., & Young, M. F. (2015). Digital plagiarism: An experimental study of the effect of instructional goals and copy-and-paste affordance. *Computers & Education*, *83*, 44–56.
Kleiner Carolyn, & Lord, M. (1999). The Cheating Game: Everyone's Doing It, From Grade School to Graduate School. U.S. News & World Report.
Lambert, E. G., Hogan, N. L., & Barton, S. M. (2003). Collegiate academic dishonesty

revisited: What have they done, how often have they done it, who does it, and why did they do it. *Electronic Journal of Sociology*, *7*(4), 1–27.

LoSchiavo, F. M., & Shatz, M. A. (2011). The impact of an honor code on cheating in online courses. *Journal of Online Learning and Teaching*, *7*(2), 179.

McCabe, D. L., & Trevino, L. K. (1993). Academic Dishonesty: Honor Codes and Other Contextual Influences. *The Journal of Higher Education*, *64*(5), 522–538.

McCabe, D. L., Trevino, L. K., & Butterfield, K. D. (2001). Cheating in Academic Institutions: A Decade of Research. *Ethics & Behavior*, *11*(3), 219–232. http://doi.org/10.1207/S15327019EB1103_2

Muldner, K., Burleson, W., de Sande, B., & VanLehn, K. (2011). An analysis of students' gaming behaviors in an intelligent tutoring system: predictors and impacts. *User Modeling and User-Adapted Interaction*, *21*(1-2), 99–135.

Muñoz-Merino, P. J., Valiente, J. A. R., & Kloos, C. D. (2013). Inferring higher level learning information from low level data for the Khan Academy platform. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 112–116).

Northcutt, C. G., Ho, A. D., & Chuang, I. L. (2016). Detecting and Preventing "Multiple-Account" Cheating in Massive Open Online Courses. *Computers & Education*, Volume 100, 71-80.

Palazzo, D. J. (2006). *Detection, patterns, consequences, and remediation of electronic homework copying*. Massachusetts Institute of Technology.

Palazzo, D. J., Lee, Y.-J., Warnakulasooriya, R., & Pritchard, D. E. (2010). Patterns, correlates, and reduction of homework copying. *Phys. Rev. ST Phys. Educ. Res.*, *6*(1), 10104. http://doi.org/10.1103/PhysRevSTPER.6.010104

Payan, J., Reardon, J., & McCorkle, D. E. (2010). The Effect of Culture on the Academic Honesty of Marketing and Business Students. *Journal of Marketing Education*, *32*(3), 275–291.

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., Leony, D., & Kloos, C. D. (2015). ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform. *Computers in Human Behavior*, *47*, 139–148.

Sanecka, E., & Baran, L. (2015). Explicit and implicit attitudes toward academic cheating and its frequency among university students. *Polish Journal of Applied Psychology*, *13*(2), 69–92.

Shu, L. L., & Gino, F. (2012). Sweeping dishonesty under the rug: how unethical actions lead to forgetting of moral rules. *Journal of Personality and Social Psychology*, *102*(6), 1164.

Siemens, G. (2013). Massive open online courses: Innovation in education. *Open Educational Resources: Innovation, Research and Practice*, *5*.

Walonoski, J. A., & Heffernan, N. T. (2006). Detection and analysis of off-task gaming behavior in intelligent tutoring systems. In *Intelligent Tutoring Systems* (pp. 382–391).

Witherspoon, M., Maldonado, N., & Lacey, C. H. (2012). Undergraduates and academic dishonesty. *International Journal of Business and Social Science*, *3*(1).

Witmer, H., & Johansson, J. (2015). Disciplinary action for academic dishonesty: does the student's gender matter? *International Journal for Educational Integrity*, *11*(1), 1–10.

Wood, H., & Wood, D. (1999). Help seeking, learning and contingent tutoring. *Computers & Education*, *33*(2), 153–169.

Appendix A – The Effect of Different Parameters on Amount of CAMEO Detected

The table below illustrates the effect of the parameters on the detection rate. Each row shows the effect of changing one of the parameters, comparing to the previous row. The changed parameter in each row is bolded.

Same IP[5] Vs. IP group [6]	***Delay*[7]**	**Harvesting Method**	**% of submissions[8] % certificatees[9]**	**Comments**
Same IP	<5 minutes	Show answer	9769 (1.5%) 23 (4.5%)	
***Group_limit*=10, *IP_limit*=10**	<5 minutes	Show answer	9939 (1.5%) 25 (5.0%)	Same parameters as used by Northcutt et al.*
Group_limit=10, *IP_limit*=10	<5 minutes	Show answer + **Exhaustive search**	14691 (2.3%) 27 (5.3%)	
Group_limit=10, *IP_limit*=10	**<24 hours**	Show answer + Exhaustive search	21952 (3.4%) 41 (8.1%)	
***Group_limit*=100, *IP_limit*=100**	<24 hours	Show answer + Exhaustive search	29788 (4.6%) 65 (12.9)	- 76% of the events are from exactly the same IP - 50% of the events use 'show answer'[10]

* The algorithm of Northcutt et al. detects 12 master accounts in our course, 11 of them also identified by our algorithm with similar parameters. When running with Group/IP limit = 100, our algorithm detects all these 12 accounts (as a subset of the 65 detected master accounts).

[5] Same IP means that both the harvesting at the harvester account and the submission at the master account are done from exactly the same IP

[6] IP addresses that serve more than *IP_limit* accounts are removed. Then IP groups that are larger than *group_limit* are also removed.

[7] Between harvesting the answer at the harvester account and submitting it in the master account.

[8] From all the correct submissions in the course, the percentage that were harvested.

[9] Percentage of the certificate earner who used CAMEO to obtain at least 10 correct answers

[10] We consider as 'show answer' only events in which the harvester used show answer *without* solving the question correctly. If the question was solved correctly by the harvester, it is considered exhaustive search, even if the user asked to see the answer. This observation is especially relevant on cases where several sub-questions share the same 'show answer' button.

Appendix B – Delay between events

Figure 5 illustrates the distribution of two types of events:

- For CAMEO events (red curve): The delay between the time of getting the solution at the harvester account, and submitting it in the master account. The median value is 70 seconds, and 75% of the events are below 65 minutes.
- For non-CAMEO events: The delay between submitting the correct answer to the same question by random pairs of accounts, for a sample of 50000 submissions (only delays that are less than 24 hours are considered, as this is the delay we consider for CAMEO).

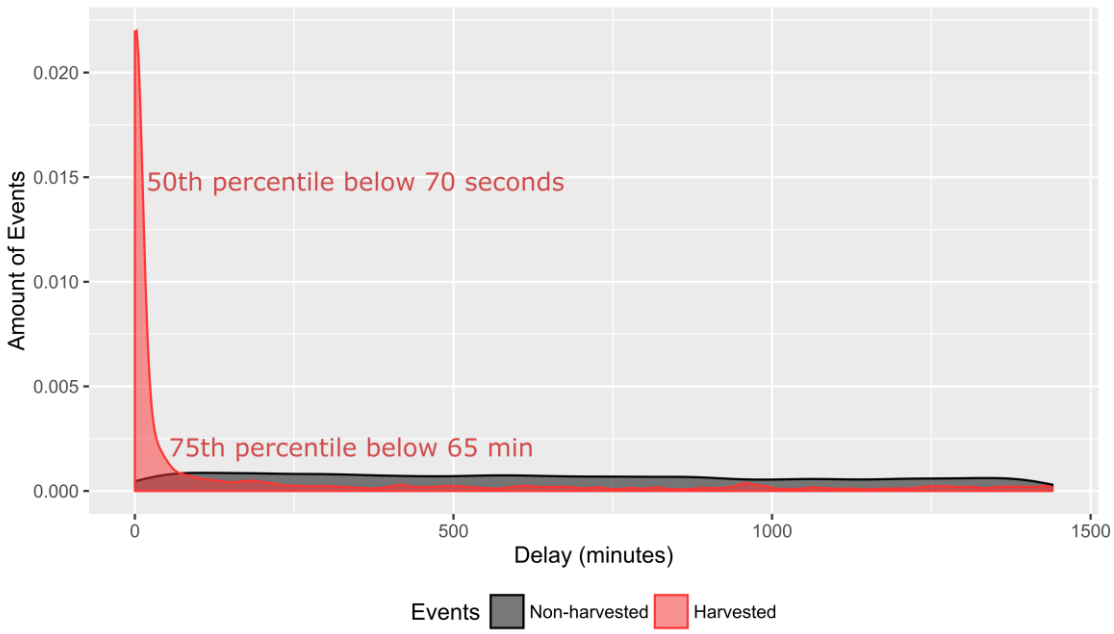

Figure 5: Distribution of delay

Time to answer. This is computed as the gap between the time in which the *master* entered into the page in which the question resides (time of 'seeing the question'), and the time in which the master submitted the correct answer to this question.

In 50% of the CAMEO events, the time to answer was less than 29 seconds. We note that in *help seeking* mode, the 'time to answer' is typically larger than in the *premeditated* mode – 56 seconds vs. 25 (median values). This is because in help mode the user opens the question in the master account, tries to solve the question, goes to the harvester account to find the answer, and then returns to the master account and submit the answer. On the premeditated

mode, the user (by definition) opens the question in the master account after the solution was found in the harvester account.

Appendix C – Success rate

The figure below shows the distribution of success on first attempt among mater accounts (certificated and non-certificated), their harvesting accounts, and the rest of the students (for each curve, the area under the curve sums to one, and does not represent the size of the group). As can be seen, the best performers on first attempt are master accounts, and the worst performers are the harvesting accounts. Interestingly, some of the harvesting accounts have a success rate that is relatively high. This might related to a pattern of use that we call 'a learning harvester' – a user who actually spend most of his/her time in the harvester account, and actually learns there, and uses the master account only for bookkeeping and for getting a certificate.