Detecting Multiple-Accounts Cheating in MOOCs

Master's Thesis

Yingying Bao

Detecting Multiple-Accounts Cheating in MOOCs

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE
TRACK SOFTWARE TECHNOLOGY

by

Yingying Bao
born in Anhui

Web Information Systems
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
http://wis.ewi.tudelft.nl

Detecting Multiple-Accounts Cheating in MOOCs

Author: Yingying Bao
Student id: 4341899
Email: y.bao-1@student.tudelft.nl

Abstract

Massive Open Online Course (MOOC) is a course designed for unlimited participation and can be accessed by anyone through the Web. As a promising education form, it has attracted lots of attentions from institutions, learners and employers [68]. However, the effectiveness and fairness of MOOC have been encroaching by academic dishonesty. Academic dishonesty is defined as using dishonest means to gain an undeserved reward or to get rid of an embarrassing situation in relation to an academic exercise [45]. It is a widespread occurrence in different levels of education and various education forms [45, 48, 63].

In this thesis, we focus on a cheating strategy, Copying Answers using Multiple Existence Online (CAMEO), in MOOCs. The strategy involves learners who use fake accounts for harvesting solutions that they later submit in their main accounts [51]. On the basis of user logs, we identify potential CAMEO users in 10 MOOCs provided by Delft University of Technology (TU Delft) on edX with three different detection methods. Besides, we analyze the characteristics of the detected users. Our results reveal that among the 8171 certificates issued in the 10 MOOCs, an estimated 2% of the certificates are earned by CAMEO users. We find that the CAMEO users are more likely to cheat at the midterm of a MOOC than the other periods of the course.

The research makes contributions to understanding the popularity of cheating especially CAMEO in MOOCs and getting the knowledge of cheaters' behaviors preferences in MOOCs. With the knowledge, at the end of the thesis, targeting at CAMEO, we purpose preventions to MOOC platforms and instructors to defend the effectiveness of MOOCs and the value of MOOC certificates.

Thesis Committee:

Chair:	Prof. dr. ir. G.J.P.M. Houben, Faculty EEMCS, TU Delft
University supervisor:	Dr. ir. C. Hauff, Faculty EEMCS, TU Delft
Committee Member:	Dr. ir. M. Zuniga, Faculty EEMCS, TU Delft

Preface

The thesis is one of the most challenging and significant projects of my study life. I would like to thank everyone who helped me to get through the tough year.

First and foremost, I would like to offer my sincerest gratitude to my daily supervisor, Dr. Claudia Hauff, who consistently provided critical and precise feedback to lead my research and writing in the right direction over the year. Thanks to Claudia's guidance, I learned how to be a researcher.

Next, I would like to thank Guanliang Chen for supporting me with his knowledge on processing edX data and academic writing. Without his patience and advice, I could not conduct the research successfully.

Furthermore, I would also like to acknowledge José A. Ruipérez Valiente, Dr. Giora Alexandron and Prof. David E. Pritchard of the Research in Learning, Assessing and Tutoring Effectively (RELATE) Project Group at Massachusetts Institute of Technology (MIT) for useful communications and cooperation.

Last but not least, I want to express my deepest appreciation to my parents for their unfailing support in my life. I am also very grateful to my friends who encouraged me throughout the year. Besides, I owe special thanks to all study colleagues from Web Information Systems Group. Thank you for the interesting and meaningful talks.

Now, I provide the thesis about detecting multiple-accounts cheating in MOOCs to you. Hope you enjoy reading the thesis. Cheating does not help you to learn but reading a research about detecting cheating does.

Yingying Bao
Delft, the Netherlands
February 9, 2017

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cheating is generally defined as using dishonest means to gain an undeserved reward of ability or to get rid of an embarrassing situation [23]. Academic dishonesty is a type of cheating that occurs in relation to an academic exercise. It can be implemented with diverse strategies including plagiarism, impersonation, bringing a cheat sheet into the exam hall, using an unauthorized digital device, etc. Academic dishonesty is a widespread occurrence in different levels of education [45]. As early as 1941, researchers [21] started to investigate the prevalence of academic dishonesty in higher education. In the experiment conducted in 1941, 126 students in a women's college were required to take six weekly tests on an unnamed subject. The previously-scored test papers without marks on papers were returned to the students for self-grading with correct solutions which were clearly announced by instructors in the classroom. By comparing students' self-graded scores to the scores instructors previously graded, 23.8% students were found to cheat on one or more questions by marking correct for their wrong answers. In another survey [3], in 1980, researchers defined academic dishonesty as 33 specific cheating behaviors including copying others assignments, obtaining test information from other students, concealing professor's errors, taking a test for someone else, bribery or blackmail, etc. and surveyed 200 students in Bloomsburg State College in the U.S. whether they had the kind of cheating behaviors, 75.5% of surveyed undergraduates admitted to cheating over their college work. Ten years later, in 1990, ABC's Nightline reported a Miami University study indicating that at Miami University, 91% of college students had cheated during their schooling [25]. The result was consistent with another study conducted in 1993. In that research [60], 91% of surveyed 60 MBA students in the U.S. admitted they had engaged in academic dishonesty including some relatively minor cases such as telling instructor a false reason for missing a class. On the basis of the proportions of students who admitted to cheating reported in prior researches [3, 21, 25, 29, 60], the prevalence of academic dishonesty had been continuing to rise over time from 1941 to 1990 [14]. The undesirable situation about academic dishonesty happened not only in higher education but also in secondary education among adolescents. In a survey of 24,000 students at 70 high schools in the U.S in 2007, 95% of the respondents stated that they had participated in academic dishonesty, whether it was on a test or for an assignment, with plagiarism or copying homework [64]. Although the disparate samples and the different definitions of academic dishonesty in the investigations mentioned above make it difficult to directly compare the percentages of students who involved in academic dishonesty, generally, academic dishonesty extensively exists in education.

The negative effect of academic dishonesty is reflected not only in the classroom but also in society at large. One of the most direct results of academic dishonesty is the destruction of fair competitions in the classroom. Meanwhile, researchers found that a successful academic dishonesty might become an example driving other students to cheat [46, 47]. Besides, the relationship between academic dishonesty and unethical business practices were revealed [16, 35, 60]. It was found that people's tolerance for unethical business behaviors such as employee theft is positively related to the attitude and frequency of academic dishonesty during their school life. To prevent the negative effect in classroom and business in future, detections and preventions for academic cheating should be proposed.

As time passes and technology advances, the teaching forms of education gradually become diversified. Campus-based learning, which is one of the most traditional education forms, is still widely adopted today. In campus-based learning courses, students can communicate with instructors face-to-face, participate in social interactions with peers and access the facilities provided on campus. However, mainly because of the geographic restrictions, many people cannot access the educational resources located on campus. Therefore, distance learning developed as a new form of education breaking through the geographic constraints. In 1728, a teacher named Caleb Phillips in Boston taught students who lived in the country shorthand by posting printed weekly lessons, which is regarded as the first attempt to distance education [30]. The media of distance education evolved from letter to radio, to television over time [7, 11, 55]. In the 1960s, the creation of the Internet brought a revolution to distance education. Illich [31] proposed his ideas of delivering the content of lessons with the Internet in *Deschooling Society* in 1971, which inspired the birth of online education. Online education is a type of distance learning that students do not have to physically attend a class and can access the study materials and instructors through the Web. Compared with other media of distance learning, the Web facilitates the instant-interactions between learners and instructors. According to a report released by the U.S. Department of Education in 2008, between 2000 and 2008, enrollment in online courses increased rapidly in almost every country [56].

The early online courses are courses with static content in a closed environment, which in most cases, learners had to pay for their enrollments to access the study material. In 2008, a variant of online courses, Massive Open Online Course (MOOC), was born, which broke the limitation of access for learners. MOOC was defined by Bryan Alexander and Dave Cormier in 2008 [2, 4]. It is a course designed for unlimited participation and can be accessed by anyone through the Web. Connectivism and Connective Knowledge (CCK2008) is recognized as the first MOOC, which was organized by George Siemens and Stephen Downes in 2008. It attracted more than 2200 participants at that time. In early MOOCs, instructors were mainly responsible for creating a framework as the backbone of the MOOC, while learners would use blogs, wikis, and other social media to learn, to share and to enrich the course. However, these MOOCs had not received extensive attentions from participants and mainstream media until the appearance of Introduction to Artificial Intelligence (CS271) in 2011. Compared with early MOOCs, CS271 has a settled course structure and fixed teaching materials, including reading texts, short video lectures, assignments and a shared discussion space, created by instructors. MOOCs represented by CS271, which have a relatively fixed course structure and content, are called xMOOCs. xMOOCs rapidly became the

mainstream of MOOCs and took over the market of online learning. According to a statistic[1] reported in December 2015, there have been more than 500 universities providing more than 4200 xMOOCs which have attracted more than 35 million learners by 2015.

Most MOOCs rely on specially designed platforms that allow the registrations of learners, provide facilities for storing and streaming digital study materials and automate the assignment assessment procedures [6]. According to a summary[2] on the network, up till December 2015, there have been more than 80 MOOC platforms. Among these platforms, based on the number of MOOCs provided, by 2015, Coursera[3] and edX[4] are two leading MOOC platforms which respectively offered 35.6% and 18.1% MOOCs throughout the network. Millions of learners have registered on the two platforms. As of April 2016, Coursera had more than 18 million registered learners[5] and edX has attracted over 7 million participants[6] worldwide.

The rise of Coursera and edX not only motivates learners to study online but also grasps the attentions of employers. When learners study a MOOC, their grades are calculated based on their performances on the course assignments. If the grade passes a cutoff instructors set, a certificate will be issued to the learner to show his/her completeness of the MOOC. The value of MOOC certificates has been confirmed by an investigation [57] involving human resources staff in 10 organizations from various fields including business, technology, health, etc. in North Caroline, the U.S. In each of the 8 investigated career fields, more than half surveyed HR staff stated that MOOC certificates gave a positive influence on making their hiring decisions. Besides, over 350 industry leading companies including Google, Amazon and Twitter have paid Coursera to match high-ranked learners with their open job profiles [69].

The benefit of the good marks and MOOC certificates in labor market may be a motivation that drives learners to participate academic dishonesty to earn an undeserved MOOC certificate or to fake an incommensurate high score in MOOCs. At the early days of the establishment of Coursera, instructors had realized and complained of plagiarism assignments by students [66]. However, the strategy to implement academic dishonesty in MOOCs is abundant and diverse[7], which is far more than plagiarism. To develop effective measures detecting and preventing academic dishonesty in MOOCs, in 2015, we conducted a small-scale survey investigating learners' attitudes towards cheating in MOOCs and potential cheating strategies they can imagine.

[1] http://www.class-central.com/report/moocs-2015-stats/
[2] http://www.knowledgelover.com/best-mooc-massive-open-online-course-providers-list/
[3] http://www.coursera.org
[4] http://www.edx.org
[5] https://blog.coursera.org/post/142363925112
[6] http://blog.edx.org/edx-celebrates-4-years
[7] nation.time.com/2012/11/19/mooc-brigade-can-online-courses-keep-students-from-cheating/

1.1 Survey of Academic Dishonesty in MOOCs

The questionnaire[8] consists of three questions, two multiple-choice questions asking respondents their familiarity to MOOCs and the difficulty of cheating in MOOCs according to their thinking, and an open-end question surveying the respondents what kind of cheating behaviors they can imagine in MOOCs. The questionnaire sheet was shared online via social media websites such as facebook.com.

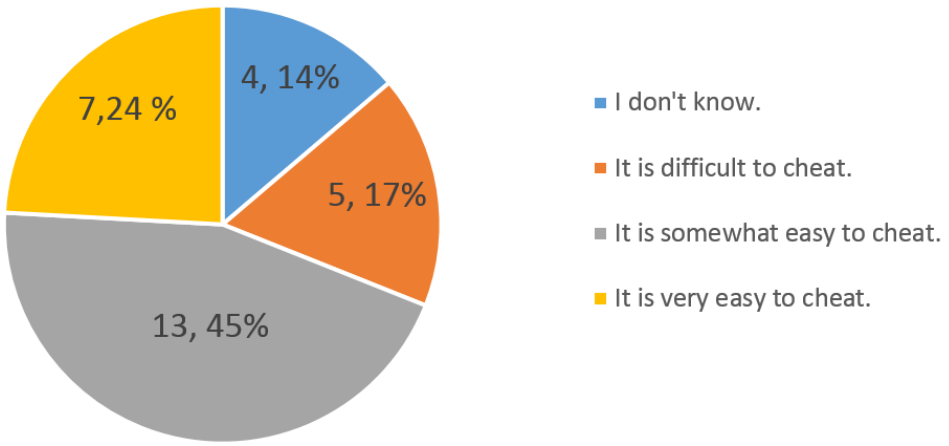

Figure 1.1: Level of Difficulty of Cheating in MOOCs

We received 35 valid responses. Among them, 37% (13/35) of the respondents had successfully completed one or more MOOCs; 26% (9/35) of the respondents had started at least one MOOC, but had not completed any successfully; 14% (5/35) of the respondents knew what MOOCs are, but never participated in one; 6% (2/35) of the respondents only had a vague idea about what MOOCs are; and the rest 17% (6/35) of the respondents declared that they had no idea about MOOCs.

Excluding the six respondents who had no idea about MOOCs, respondents' attitude towards cheating in MOOCs can be seen in Figure 1.1. Around 70% (20/29) of the respondents stated it is very easy or somewhat easy to cheat in MOOCs.

Besides, respondents nominated several possible cheating strategies they can imagine according to their understanding of MOOCs. There are four most mentioned cheating strategies in the responses as follows.

- Search Google for answers shared by other participants.
- Hire surrogate exam-taker who is an expert in the subject.
- Probe correct answers by registering multiple accounts.
- Ask or Search in subject-related Question & Answer forums

[8] http://goo.gl/forms/T5kwjzsyEX

1.2 Multiple-Accounts Cheating Strategy

Among the four most mentioned cheating strategies, we focus on probing correct answers by registering multiple accounts. The mechanism of the cheating strategy is as follows: The openness of MOOCs enables a learner to register multiple accounts with different email addresses on a MOOC platform. Cheaters who use the multiple-accounts cheating strategy utilize the openness to create multiple accounts on the MOOC platform. The learner who cheats with the multiple-accounts cheating strategy utilizes the openness to create multiple accounts on the MOOC platform. Then the cheater enrolls a same MOOC with the multiple accounts. For most course questions, MOOC platforms provide learners a chance to check solutions right after the learners complete the questions. The cheater takes the opportunity to submit a randomly-selected answer and access the solution to a question with an account. Then he/she submits the harvested solution via another account to fake a high performance for the account in the MOOC. There are different patterns in which cheaters can apply the cheating strategy[9]. Researchers [51] called the multiple-accounts cheating strategy, Copying Answers using Multiple Existence Online (CAMEO). Learners employ the strategy are referred as CAMEO users. We adopt the name in the thesis.

Compared with other cheating strategies, CAMEO attracts our attention for its three peculiarities. Firstly, CAMEO users are independent. They are able to complete the entire cheating process without any help from others. For instance, asking Q&A forums, which is also one of the most mentioned cheating strategies in our survey, can only be effective when there is someone replying correct answers to the question posts. However, CAMEO users can implement their cheating just by their interactions with the learning management systems underlying the MOOC platforms. The independence makes CAMEO more feasible, compared with other cheating strategies. Secondly, since the solutions CAMEO users harvested are prepared by instructors, the absolute correctness of solutions makes CAMEO quite efficient. Thirdly, cheaters can apply CAMEO to most MOOCs with no limitation on MOOC subjects. For instance, the effect of hiring surrogate exam-takers relies on the level of professionalism of the surrogates, and for MOOCs with different subjects, several experts from various fields should be hired. However, CAMEO is not limited by the content of the MOOCs. The independence, high efficiency and scalability of CAMEO make it stand out from other cheating strategies. Except for the three peculiarities, CAMEO is also featured by its detectability. Massive user logs recorded on the servers of MOOC platforms enable us to detect the activities of CAMEO users.

1.3 Research Questions

The ultimate purpose of preventing academic dishonesty is to defend the value of education and the fairness of competition. MOOC, as a promising online education form, should guarantee the integrity of education while establishing its credibility and proving its effectiveness[10].

[9] We detail the patterns of CAMEO in Section 4.1.

[10] http://nation.time.com/2012/11/19/mooc-brigade-can-online-courses-keep-students-from-cheating

For CAMEO, researchers [51, 59] have designed and implemented some detection methods on their experimental courses. However, different MOOCs provided by diverse universities involves various learners who may behave completely differently. In other words, although with a same detection method, the results of detecting CAMEO users can differ greatly on different MOOCs. For the MOOCs created by our home university, Delft University of Technology (referred as TU Delft below), up to 2017, there is no research conducted about detecting and preventing CAMEO. This leads to our first research question as follows:

Research Question 1: What is the prevalence of CAMEO in MOOCs created by TU Delft?

To our knowledge, Northcutt, Ho and Chuang published the first scientific paper [51] about detecting and preventing CAMEO in MOOCs in 2015. In this thesis, we first replicate the detection method in our MOOCs dataset based on [51]. During the replication, we find there is a limitation of the detection method (referred as Singleton Detection Method below) that the authors utilized a unitary set of criteria to detect different patterns of CAMEO. To deal with this, on the foundation of the Singleton Detection Method, we design a new detection method (referred as Hybrid Detection Method below) to identify CAMEO users. During the course of the thesis, another research [59] about detecting CAMEO in MOOCs were published. Compared with the other detection methods, the detection designed by Ruiperez-Valiente et al. [59] is featured by its requirement to the number of questions a CAMEO user continuously cheat on. Thus we refer the method as Long Batch Detection Method below and we also implement the method in our MOOCs dataset.

Research Question 2: What are the characteristics of the potential CAMEO users in MOOCs created by TU Delft?

Except for detecting potential CAMEO users with different detection methods in MOOCs created by TU Delft, understanding features of these CAMEO suspects is also necessary for amending our assumptions of cheating behaviors of CAMEO users and customizing pertinent preventions for CAMEO in MOOCs. The features of the detected CAMEO suspects are analyzed in the aspects of their certificate mode, geographical location, certificate mode, potential motivation of cheating and the timing when they are most likely to cheat during a MOOC.

1.4 Research Approach

The research is divided into three phases in Figure 1.2. We start the research by introducing the operating environment of CAMEO. To be specific, in the first phase, we illustrate the structure of MOOCs on edX, which is the only MOOC platform that TU Delft cooperates with, and emphasize the problem component of MOOCs on which cheaters apply CAMEO. Besides, we describe the certificate mode on edX, which may be one of the main motivations that drive learners to cheat due to its positive effects on the labor market [57].

The second phase is to detail the three different detection methods we implemented in our MOOCs dataset. On the basis of the previous publications, we replicate the Singleton Detection Method and the Long Batch Detection Method in 10 MOOCs

created by TU Delft. Besides, according to our analysis of the Singleton Detection Method, we point out some omissions of the method, then design and implement the Hybrid Detection Method with the target to remedy the indicated disadvantages.

With the three detection methods, we get discrepant detection results. In the third phase, we report the number of detected potential CAMEO users by the different detection methods. Besides, we do a verification for detected potential CAMEO users to illustrate the effectiveness of the methods. To provide pertinent suggestions of preventing CAMEO to the institutions which create MOOCs and MOOC platforms, the characteristics of the detected potential CAMEO users are discussed.

1.5 Thesis Outline

In the second chapter, we provide a review of literature about academic dishonesty. The chapter starts with a summary of various previous investigations about the prevalence of academic dishonesty in secondary and higher education. Then we list potential motivations of academic dishonesty researchers proposed in their previous research. After that, we report some detections and preventions that have been adopted in offline or online learning to deter academic dishonesty.

The third chapter corresponds to the second phase of the research. In this chapter, we introduce the operating environment of CAMEO, edX, including the structure of MOOCs on edX, the problem component of the MOOCs, and different types of the certificates issued to learners who complete the MOOCs.

The fourth chapter contains a detailed description of the three detection methods we implemented in our MOOCs dataset. It starts with an illustration of the Singleton Detection Method. Then some drawbacks of the method are pointed out and the Hybrid Detection Method is introduced aiming at remedying the disadvantages. Afterward, another detection method, Long Batch Detection Method, is explained.

We start the fifth chapter with a detailed illustration to our MOOCs dataset. Then we report the detection results of applying the three detection methods described in the fourth chapter in the 10 MOOCs to answer our first research question about the prevalence of CAMEO in MOOCs created by TU Delft. Afterward, to evidence the detection results can possibly be true, we build a case for verification. We select one of the detected potential CAMEO user and do a manual check to his/her activities on the MOOC platform. We also analyze some characteristics of the detected potential CAMEO users including their certificate mode and their geographical location, etc.

In the last chapter, we answer our two research questions on the basis of the analysis of detection results and summarize the contributions of the research we make. According to the features of CAMEO and detected potential CAMEO users, possible measures to deter CAMEO in MOOCs are purposed.

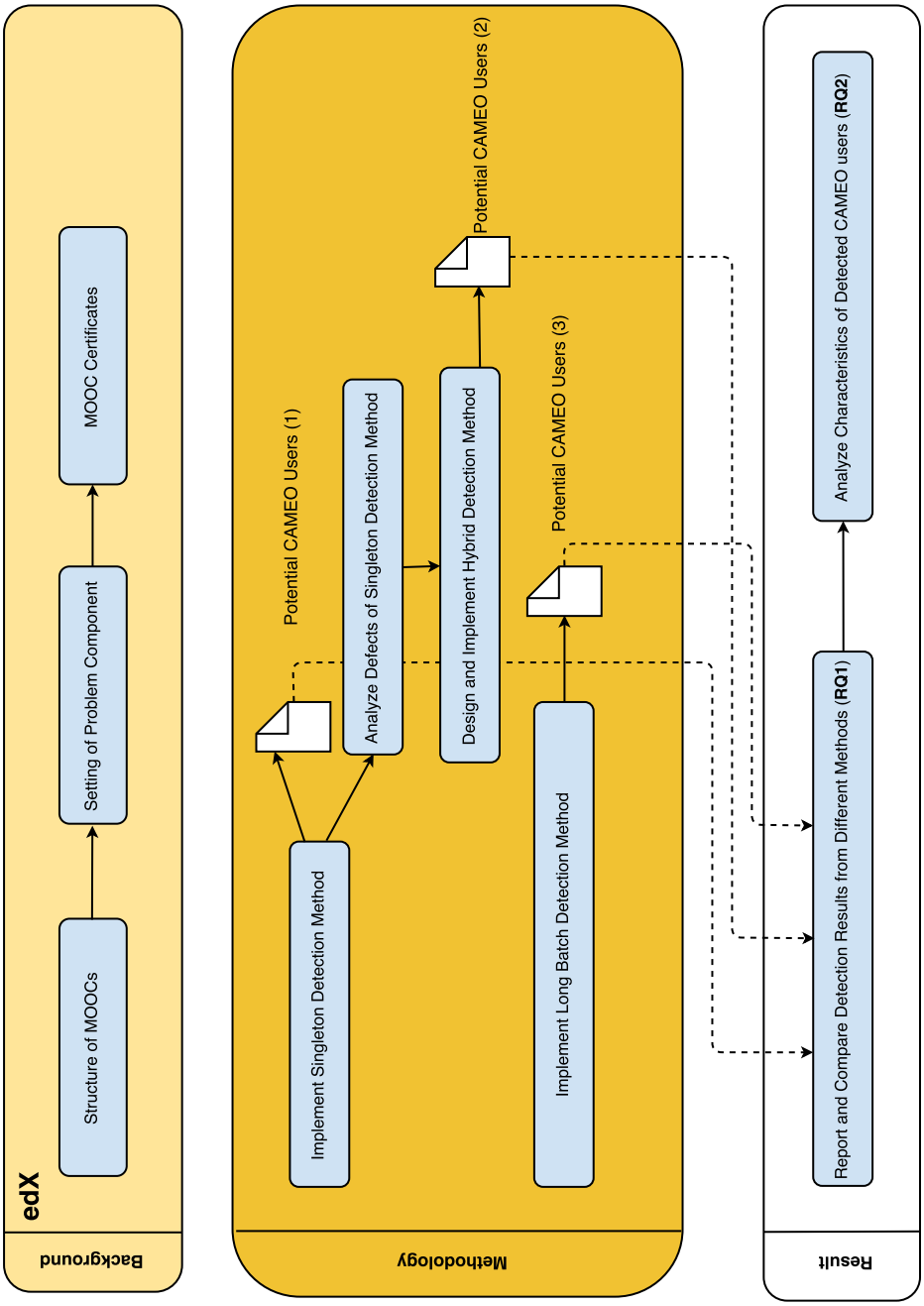

Figure 1.2: Research Approach Diagram

Chapter 2

Related Work

In the pursuit of knowledge, learners are expected to adhere to honesty to defend the value of education and the fairness of competition. However, with the diversification of education forms and the advancement of technology, academic dishonesty has become easier for learners to implement and harder for instructors to detect [33]. Researchers have realized the severity of the problem and conducted many investigations with disparate samples to survey the prevalence of academic dishonesty in various subjects, in different regions and in different levels of education. In Section 2.1, we summarize parts of the studies which indicate the prevalence of academic dishonesty, especially in higher education. Contextual factors such as grades are suggested as one of the main factors that contribute to the popularity of academic dishonesty among students [10, 19, 47]. In Section 2.2, we summarize the factors researchers analyzed which influence students cheating behaviors. Concerning the popularity of academic dishonesty, faculties have adopted some preventions such as honor codes, plagiarism detection software to combat cheating [42, 46]. We review these preventions instructors used in Section 2.3. In the same section, targeting at MOOCs, we emphasize the measurements mainstream MOOCs platforms utilized to prevent academic dishonesty.

2.1 Prevalence of Academic Dishonesty

In the past decades, researchers [4, 13, 28, 34, 43, 47] have been aware of the popularity of academic dishonesty and the rate at which it is increasing. Most of them concluded that the advancement of technology greatly pushes forward the prevalence of academic dishonesty by making it easier for students to cheat [9, 32, 52]. Except for some well-known traditional cheating strategies such as sneaking notes into the exam hall, whispering to others during the exam and copying from others without their knowledge, in recent years, students start to implement their academic dishonesty with high-tech means such as utilizing cell phones to communicate with experts outside the exam room, searching answers on the web etc. [33]. Besides, researchers also have noticed the existence of some relatively extreme cheating strategies such as breaking the offices or hacking learning management systems to access the test paper or answer keys before the exam [53]. Here we list the most common cheating behaviors researchers defined in previous research [12, 27, 44, 47].

List of Common Academic Dishonesty Behaviors Defined in Previous Research [12, 27, 44, 47]

1. Telling instructors a false excuse to delay/miss an exam.
2. Asking others to complete an exam/assignment with your name.
3. Using an unauthorized cheat sheet/note during an exam.
4. Collaborating with others on an individual assignment/exam.
5. Copying from others during an exam with/without their knowledge.
6. Learning questions for an exam from an unauthorized source prior to taking it.
7. Using a not-allowed digital device as an extra aid during an exam.
8. Turning in work done by others without proper references for an assignment.
9. Helping others to implement academic dishonesty behaviors listed above.

Students implement their cheating with diverse means in various academic contexts from elementary to post-graduate education [45]. Among different levels of educations, researchers said that secondary and college education are the hardest-hit areas of academic dishonesty [45]. Researchers stated that students are exposed to significant cheating cultures during their secondary education and their cheating behaviors are likely to intensify once they reach college [45].

Previous research has revealed that academic dishonesty has always been popular [12, 28, 43]. For instance, for U.S. undergraduate students, the prevalence of academic dishonesty from the 1960s to the 2000s can be seen in Table 2.1. To be specific, in 1964, William J. Bowers surveyed 5,000 undergraduate students from 99 colleges and universities in the U.S. by sending them a questionnaire asking whether they had taken some specific cheating behaviors such as bringing notes into exam halls and the frequency [12]. Three-fourths of the respondents in the investigation admitted they had participated in at least one incident of academic dishonesty. Twenty years later, in 1984, Haines et al. interviewed 380 undergraduate students in various subjects at a small state university in the Southwest, the U.S. with a 49-item questionnaire about academic dishonesty [28]. More than half of the surveyed students self-reported their cheating during the academic year in one or more areas including exam, quiz and homework assignment. In a recent study, Donald L. McCabe investigated 50,000 undergraduates from more than 60 campuses in the U.S. with a web-based questionnaire from 2002 to 2005 and found that on most campuses, 70% of the respondents admitted to their cheating [43].

Although many investigations about academic dishonesty were conducted in the U.S. [4, 12, 13, 28, 43, 48, 63], researchers have noted that the popularity of academic dishonesty is a global education problem which is not confined to the U.S. [22, 27, 38, 40]. As seen in Table 2.2, in 2004, Grimes surveyed 1,810 undergraduate business students from eight eastern European countries with a questionnaire, 74% of

the respondents admitted they had cheated on an exam or course assignment in college [27]; in the same year, another questionnaire about attitudes to academic dishonesty was distributed to another 600 undergraduate business students in Botswana and Swaziland in Africa [22], among 460 valid responses in the experiment, 56% agreed that students must pass examinations by all means including cheating strategies . Except for the U.S, Europe and Africa, academic dishonesty is also popular among students in Australia and Asia [38, 40]. In 2005, Marsden et al. surveyed 954 students in diverse majors from four Australian universities with a self-report questionnaire and found that 81% of the respondents admitted their cheating behaviors [40]. Two years later, in Asia, Lin et al. investigated 2,068 college students throughout Taiwan on four domains of academic dishonesty including cheating on exams, on assignments, plagiarism and falsifying documents [38]. On the basis of the experiments, the prevalence of academic dishonesty among college students in Taiwan was estimated to be 61.72%.

Table 2.1: Prevalence of Academic Dishonesty at Different Times

Year	**Samples**	**Subject of Samples**	**Detection Method**	**Percentage of Cheating**	**Region**
1964	5,000 Undergraduates	Unspecified	Questionnaire (Self-Report)	75%	U.S.
1984	380 Undergraduates	Diverse	Questionnaire (Self-Report)	54%	U.S.
2005	50,000 Undergraduates	Unspecified	Questionnaire (Self-Report)	70%	U.S.

Table 2.2: Prevalence of Academic Dishonesty at Different Regions

Year	**Samples**	**Subject of Samples**	**Detection Method**	**Percentage of Cheating**	**Region**
2004	1,810 Undergraduates	Business	Questionnaire (Self-Report)	74%	Europe
2004	460 Undergraduates	Business	Questionnaire (Self-Report)	56%	Africa
2005	50,000 Undergraduates	Unspecified	Questionnaire (Self-Report)	70%	U.S.
2005	954 Multi-Grades	Diverse	Questionnaire (Self-Report)	81%	Australia
2007	2,068 Undergraduates	Unspefiied	Questionnaire (Self-Report)	62%	Asia

The popularity of academic dishonesty is neither influenced by time, regions nor swayed by the majors of students [48]. In 1992, Meade sent a survey to 15,000 students at 31-top ranked universities in the U.S. asking respondents whether they had cheated during their college career [48]. Among the 6,000 valid responses, 87% business students, 74% engineering students, 67% science students and 63% humanities students recognized their cheating behaviors.

Except for the investigations to the prevalence of academic dishonesty in campus-based courses as we mention above, with the diversification of the media of education, researchers started to pay attentions to the academic dishonesty on the premise that students are allowed to take exams and assignments at a remote end without the supervision of instructors and their peers [63]. In 2010, Watson et al. surveyed 635 students studying various subjects at a medium sized university in Appalachia, the U.S, with 44 multiple-choice questions to compare the prevalence of academic dishonesty offline and online [63]. 32.1% respondents admitted they had cheated on assignments or tests in campus-based courses, while there were 32.7% surveyed students admitted their cheating in online courses, which indicates that academic dishonesty not only extensively exists offline but also can be found in online courses.

We have noticed that the means of investigating the prevalence of academic dishonesty is quite simple. Almost all experiments [12, 22, 27, 28, 38, 40, 43, 50, 63], to our knowledge, are based on a self-reported questionnaire/scale. We suppose the questionnaire was utilized by researchers since it is able to collect and compile data from many people in a short period of time with a relatively lower cost [1, 54]. However, there are also some disadvantages of questionnaires. Firstly, the return rate of questionnaire usually is low. The low response rate has been shown in some investigations we described above [22, 28, 40]. Secondly, the research result of a questionnaire is very likely to be influenced by the design of the questionnaire. For instance, compared with McCabe's [45], the sheet designed by Grimes [27] had a broader definition of academic dishonesty and it included some extra cheating behaviors such as doing less work than you shared in a group project which was not covered in McCabe's questionnaire. These differences of content may result in the respondents feeling different about whether they have involved in academic dishonesty.

To sum up, many investigations conducted by researchers in the past fifty years [12, 22, 27, 28, 38, 40, 43, 50, 63] have indicated the popularity of academic dishonesty regardless of time, region, discipline, and education form.

2.2 Factors Influencing Academic Dishonesty

After the investigations to the prevalence of academic dishonesty, researchers went deeper to analyze the factors influencing students cheating behaviors [10, 12, 15, 19, 20, 24, 47, 58, 61]. The factors can be roughly divided into two groups, individual difference factors including student's age, gender, parents' education and contextual factors such as the levels of academic dishonesty among peers. Here we list some factors researchers assumed that influence students cheating behaviors.

List of Common Factors Influencing Students Cheating Behaviors Assumed in Previous Research [10, 12, 15, 19, 20, 24, 47, 58, 61]

- Individual Difference Factors
  - Age
  - Gender
  - Source of Funding
  - Academic Achievement
  - Parents' Education
- Contextual Factors
  - Fraternity Membership
  - Peer Cheating Behaviors
  - Peer Disapproval of Cheating
  - Quality of Courses
  - Severity of Penalties

Compared with individual difference factors, researchers stated that contextual factors have more influence on academic dishonesty [47]. Among the contextual factors, the effect of peers was repeatedly mentioned by researchers [12, 20, 47, 58, 61]. According to a multi-campuses investigation conducted by McCabe and Trevino in the U.S. in 1997, among 12 factors researchers assumed including age, gender, perceived severity, etc, the behavior of peers was the strongest factor influencing academic dishonesty [47]. Supported by social learning theory and differential association theory [5], researchers concluded that a successful cheating by peers would increase the tendency of the observer to behave in a similar way [47]. The influence of peers reflects in two aspects. On the one hand, the cheating behaviors by peers establish a climate where non-cheaters feel left at a disadvantage [20]. In the climate, students are more likely to adjust their attitudes to academic dishonesty and tend to cheat voluntarily to keep up with peers who benefited from dishonest means [61]. On the other hand, if students are in a cheating environment, they are highly likely to be involved in academic dishonesty passively to retain acceptance within the large peer group [58].

Except for the behaviors of peers, the quality of courses is regarded as another contextual factor that impacts the prevalence of academic dishonesty [10, 15, 20, 24]. Cole et al. in 2000 conducted a survey among high school and college students confirming the view. According to the survey, students are more likely to cheat when they think their assignments are meaningless, while they are less likely to cheat when they are interested in what they learn [15]. The existence of preventions and the severity of penalties also impact the prevalence of academic dishonesty [10, 20, 47]. The Center for Academic Integrity conducted a survey involving 48 different college campuses in the U.S. in 2002 and found that existence of academic preventions was able to deter students cheating behaviors, with the prevalence of academic dishonesty being 30% to 50% lower at classes with preventions vs. the classes without [20]. Besides, the instructor-student relationship is also considered to have an impact on academic dishonesty [24, 47]. Students tend to be more honest when they admire and respect their teachers, while they are more likely to cheat if they think their teachers are unfair [24].

Another contextual factor influencing students cheating behaviors is the focus on grades from the outside including parents and instructors [10, 19]. Upon most occasions, grade is used as the main measure of the value of a student and it is also closely related to the future success of the student including the admission of higher education and position in the workplace [10]. "The 'achievement ethic' throws academic honesty and integrity out the window in favor of personal gain", a respondent in the survey

conducted by Greene et al. in 1992 said [26]. Many students in the survey blamed the pressure from their parents and teachers for their cheating behaviors [26].

As mentioned, students practice academic dishonesty mainly to maximum their personal rewards [10, 19, 26]. The execution of their cheating behaviors is influenced by multiple factors including peers' attitudes towards cheating, quality of the courses, and punishments for caught cheating.

2.3 Detections and Preventions

In this section, we introduce four common means adopted by educational institutions to detect and prevent academic dishonesty. They are proctored examination, plagiarism detection software, academic honor code and biometric verification. We respectively detail the theories of these means, their applications in both campus-based learning and distance learning and their effectiveness in detection and prevention.

Proctored Examination

For written tests of campus-based courses, arranging a face-to-face proctored examination is often regarded as the most common mean to detect and prevent academic dishonesty. However, for distance learning, there is no campus and it is not easy to gather massive students all around the world to organize a face-to-face proctored exam. With the advancement of technology, online proctoring service was introduced, which allows invigilators to proctor students through web cameras. Besides, there are some business examination agencies offering global testing services for specific tests, which enables students around the world to find a nearby test center to complete their face-to-face proctored exams. Currently, for MOOCs, edX provide learners the option of proctored exams at test centers provided by Pearson VUE[1], and Coursera utilizes online proctoring services provided by ProctorU[2].

Plagiarism Detection Software

With the digitization of paper documents, the form of assignment and exam has shifted from paper to electronic documents. For the digital assignments and exams, various computer-assisted plagiarism detection systems[3] have emerged since 2000. The main function of these plagiarism detection systems is to highlight the text of the checked documents which has appeared in other documents in the system database. The database of a mature business plagiarism detection software usually contains billions of web pages, millions of student papers and millions of journal articles, periodicals and books. The recall of a plagiarism detection software partly depends on the coverage of its database. It also relies on the detection approach of the software. Some plagiarism detection software complete the detection process based on the similarity between word strings in two documents [62]. In allusion to the mechanism, there are students trying to rewrite, translate or redraft sentences to conceal the evidence of plagiarism [50]. Although researchers have improved the performance of plagiarism detection software by utilizing multiple detection approaches, the plagiarism detection

[1] https://home.pearsonvue.com/

[2] https://www.proctoru.com/

[3] https://en.wikipedia.org/wiki/Comparison_of_anti-plagiarism_software

is beginning to look more like a cat-and-mouse game that students try to get the idea of detection software and then evade the detection according to the mechanism of the software [67]. Besides, false positive is a limitation to the precision of a plagiarism detection system. The high frequency of some common phrasings may result in a coincidental and unfair match detected by the plagiarism detection software and in the case, manual detection should be involved to distinguish the coincidence with the real academic dishonesty [65]. Nevertheless, there are many universities adopting plagiarism detection software as a means to deter academic dishonesty. Turnitin[4], which is one of the most popular plagiarism detection software, claims that it has been trusted by 15,000 institutions in 140 countries. For MOOCs, Coursera has reported dozens of plagiarism especially in humanities MOOCs and considered adding a plagiarism detection software [66]. However, maybe because of the high costs in time and money [70], up to November 2016, to our knowledge, there is no MOOC platforms deploying plagiarism detection software.

Academic Honor Code

(Academic) Honor code, which is a set of ethical principles defining and illustrating honorable behaviors within an academic community, is one of the most popular preventions adopted by educational institutions. Learners within the academic community are required to follow the honor code. The effectiveness of honor code depends on an assumption that student can accept the significance of the ethical principles and can be trusted to act honestly. Researchers have affirmed the efficiency of honor code for campus-based learning in previous research [12, 42, 46]. To be specific, in 1964, Bowers surveyed 5,000 undergraduate students from 99 colleges and universities in the U.S. and found that students in the institutions with honor code were less likely to cheat [12]. McCabe et al. in 1993 [46] and Mazar et al. in 2008 [42] also confirmed that honor code is to the benefit of reducing the prevalence of academic dishonesty in campus-based learning among undergraduate students in the U.S.

However, the utility of honor code for distance learning is in dispute. Mastin et al. [41] conducted an experiment investigating 439 undergraduate students who respectively studies an introductory psychology online course at three separate times in 2005 and found that the honor code in the online course had no significant effect on students cheating behaviors. LoSchiavo et al. in 2011 [39] surveyed 84 undergraduate students who participated in a 5-credit online psychology course and drawn a similar conclusion with Mastin et al. that the effectiveness of honor code might be missing from an online environment. Meanwhile, there are some researchers confirming the efficiency of honor code for online learning [18]. In 2014, Corrigan-Gibbs et al. [18] designed three different exam environments including online exam without any additional instructions, with an honor code, and with both an honor code and a warning illustrating the severity for violating the honor code for 409 students who studied an online algorithm course. For detecting students who completed the exam with unauthorized resources from the Web, researchers designed a honeypot website containing all exam questions and a "Click to show answer" button. The fully accessed honeypot website was the first hit returned if students searched the exact text of the exam with Google [18]. When students clicked the show answer button on the honeypot website, no an-

[4] http://turnitin.com/

swer would be displayed but information about the user who asked for answers would be recorded on the server for identifying cheaters. In the experiment, 34.4% students who took the online exam without any external instruction were detected as cheaters, while 25.5% students who completed the exam with an honor code and 15.5% students with both an honor code and a warning were detected [18], which indicates that honor code is able to deter academic dishonesty in online courses. Both Coursera[5] and edX[6] have utilized honor code to restrain students and prevent academic dishonesty in MOOCs.

Biometric Verification

Biometric verification stops surrogate exam-takers with the aid of computer science. It is a system to authenticate user's identity for secure access to electronic systems by evaluating one or more distinguishing biological traits. There are mainly two kinds of biometric verification systems adopted for detecting and preventing academic dishonesty in MOOCs. One is facial recognition, another one is keystroke recognition. The mechanism of facial recognition system on a MOOC platform is that the platform asks learners to update their government-issued photo IDs at the beginning of a MOOC, and during the assignments/exams of the MOOC, the platform photographs the person who is answering the questions via web camera, then maps the picture with photo on the corresponding uploaded government-issued ID with facial recognition algorithm to confirm the identity of the respondent. The principle of keystroke recognition system on a MOOC platform is that the platform requires learners to transcribe a string of words at the beginning of a MOOC, and during the transcription, the time of each key-down and key-up event is recorded in milliseconds. If the learner's keystroke pattern during the MOOC is different from the record, he/she is suspected as a potential cheater who recruits others to help him/her complete the MOOC. On the basis of OhKBIC dataset, keystroke recognition has shown a promising performance with a correct learners verification over 90% using only 100 keystrokes [49], which indicates it can be used an effective measure to detect academic dishonesty in digital assignments/exams. Currently, edX has deployed facial recognition system, while Coursera adopted a combination of the facial and keystroke recognition system to prevent academic dishonesty with surrogate exam-takers.

[5] https://learner.coursera.help/hc/en-us/articles/209818863-Coursera-Honor-Code
[6] https://www.edx.org/edx-terms-service

Chapter 3

edX

edX, a leading MOOC platform which is both non-profit and open source[1], was established by Massachusetts Institute of Technology and Harvard University in May 2012. Up to November 2016, edX has offered more than 950 MOOCs, collaborated with more than 2,300 faculties and staff, released more than 840,000 certificates to edX learners[2]. TU Delft is one of the edX charter members and has created 43 MOOCs[3] on edX (the courses are referred as DelftX[4] below) covering various fields including art, business, computer science etc. In this chapter, to understand the operating environment of CAMEO, we respectively introduce the structure of MOOCs in Section 3.1, the problem component of MOOCs in Section 3.2, and the certificate for learners who complete the MOOCs on edX in Section 3.3.

3.1 MOOCs

edX has defined some certain guidelines[5] for the structure of MOOCs. Every MOOC on edX has its basic information including unique course number, course name, and organization which creates the course, to identify itself from other courses. Besides, each MOOC has its start date and time which specify when learners can access the content of the course. Here we show a common four-layer structure of MOOCs on edX in Figure 3.1.

A MOOC on edX is composed of several sections. edX defines that section[6] is the topmost category of a MOOC. A section can represent a time period (study week), a chapter or another organizing principle of the MOOC. It consists of multiple subsections. Each subsection may represent a topic and it is made up of one or more units. Learners view a unit as a single web page[6]. A unit contains at least one component. Component is the smallest element of the structure of MOOCs on edX. It contains the actual course content. There are four types of components on edX, HTML, discussion, problem and video. The HTML component is to arrange the content and location of text, images and hyperlinks on the page. The discussion component provides a space

[1] http://edx.org/about-us

[2] https://www.edx.org/schools-partners

[3] https://online-learning.tudelft.nl/courses/?course_types=M

[4] The detailed information about DelftX will be given in Chapter 5.

[5] http://edx.readthedocs.io/projects/edx-partner-course-staff/en/latest/

[6] http://edx.readthedocs.io/projects/edx-partner-course-staff/en/latest/developing_course/index.html

on the page for learners to post and exchange their opinions on specific topics. The video component offers a short film that instructors post on the page as teaching material. The problem component includes a question to assess learners' understandings to the course material. The assembly of these components is very flexible. For instance, for the problem component, it can not only be deployed with other components in a teaching section as a formative assessment but also exclusively constitute a test section for a mid or final exam. With correct submitted answers to the problems, learners can earn enough credits to get a certificate showing their completeness of the course.

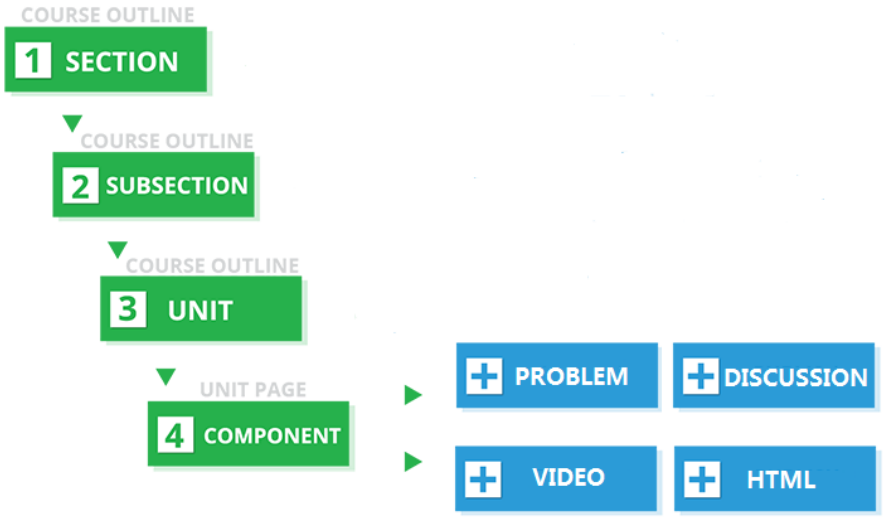

Figure 3.1: Structure of MOOCs on edX
(Reproduced on Creation Workflow from edX Documentation[7])

Except for the start time of the course, course pacing also decides the timing when learners can access the content of the MOOC. There are two different course pacing styles on edX, instructor-paced and self-paced. For instructor-paced courses, instructors set the release and due time for each section. Learners cannot access the section before its release date and they are required to complete questions in the section before the deadline. For self-paced courses, there are no time constraints except the end date of the MOOC. Each MOOC on edX has its end date and time. After the end date of a self-paced MOOC or the deadline of a section in an instructor-paced MOOC, enrollments are still able to access the course content and answer questions, however, the questions in the expired course or section can no longer make contributions to the learners' credits and certificates for the MOOC.

3.2 Problem Component

Assessment is an important part of education which is to measure and document the knowledge and skills of students [17]. edX allows instructors to implement assessments to learners with the problem component in MOOCs. One problem component[8]

[8] http://edx.readthedocs.io/projects/edx-partner-course-staff/en/latest/exercises_tools/index.html

contains one question. Currently, edX fully supports five types of questions including checkbox question, dropdown question, multiple choice question, numerical input question and text input question. All of these questions have the following common elements as shown in Figure 3.2.

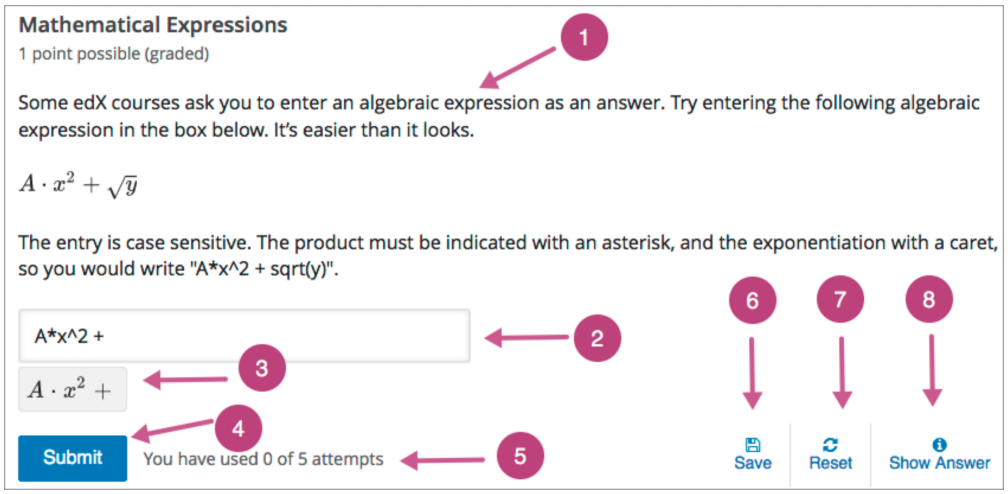

Figure 3.2: An Example of Question on edX
(Copied from Building and Running an edX Course[9])

1. **Problem Text.**

2. **Response Field.**

   It is the space where learners enter their answers. The form of the response field depends on the type of the question.

3. **Rendered Answer.**

   For some questions especially math expression questions, there is a rendered answer for learners to pretty print their answers.

4. **Submit.**

   When learners click the "Submit" button, the answer on the response filed will be submitted to the learning management system of edX. An immediate feedback will be given based on the correctness of the submitted answers. If the answers are correct, a green check mark will appear, otherwise, a red x will occur.

5. **Attempts.**

   It specifies the maximum times that learners are allowed to submit their answers with the "Submit" button and their current left chances.

6. **Save.**

   The "Save" button helps learners to keep their answers displaying on the response field. To be noticed, the answers have not been submitted and graded until learners click the "Submit" button.

7. **Reset.**

   The "Reset" button enables learners to clear any input that has not been submitted for the question.

8. **Show Answer.**

   By clicking the "Show Answer" button, an instructor-prepared answer and explanation for the question will show up. The button is closely related to CAMEO we focus on. For the button, there are eight alternative timings as follows that instructors can choose and decide when the button is visible to learners.

   - *Always.* The "Show Answer" button always appears.
   - *Answered.* The "Show Answer" button appears when the learner has submitted a correct answer to the question.
   - *Attempted.* The "Show Answer" button appears when the learner has submitted answers to the question.
   - *Closed.* The "Show Answer" button appears when the learner has run out of all attempts for the question or the due date of the question has passed.
   - *Finished.* The "Show Answer" button appears when the learner has submitted a correct answer or run out of all attempts or the due date of the question has passed.
   - *Correct or Past Due.* The "Show Answer" button appears when the learner has submitted a correct answer or the deadline of the question has passed.
   - *Past Due.* The "Show Answer" button appears when the due date of the question has passed.
   - *Never.* Never show the "Show Answer" button.

Except for the five types of questions, edX provisionally supports[10] some other advanced problem types such as random questions, open-response assessment questions etc. The randomization can happen within a question, which means that different learners in the MOOC get the same question but with different numeric values. Meanwhile, there is problem randomization, which means that different learners in the MOOC get different questions. Unlike most questions that are automatically graded by the learning management system of edX, open-response assessment question is manually checked by the learners themselves or their peers or instructors of the MOOCs.

[10] edX categorized the level of support as full, provisional or no support.

3.3 Certificates

Every question on edX has its corresponding credits and weight specified by instructors. The product of the credits and weight of a question is the contribution that learners can make to their final grades by correctly answering the question. If a learner's grade successfully passes a cutoff instructors set, he/she will get a certificate showing his/her achievement on the MOOC. So far, edX has provided three different types of certificates, *Honor Code*, *Verified* and *XSeries*. The content of these certificates depends on the design of the instructors of the MOOC. Generally, the learner's registered full name, the MOOC's course code, the institution which creates the MOOC and the signatures of the course instructors are displayed on the certificate. Here we show an example of edX Verified Certificate in Figure 3.3.

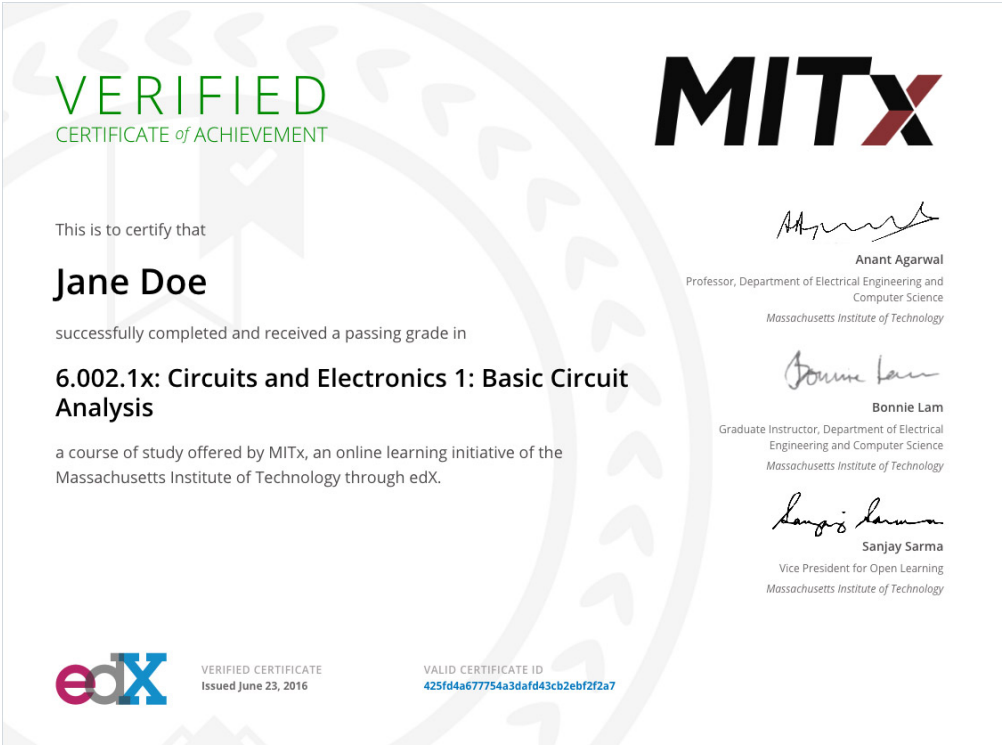

Figure 3.3: An Example of edX Verified Certificate

Honor Code Certificate

An Honor Code Certificate shows the learner has earned a passing grade in the MOOC with the acceptance of edX honor code but has not completed extra identity verification steps. It is free of charge. The learner is asked to confirm an honor code and a warning about the results of violation at the beginning of the course. edX conveys the following requirements to learners with the honor code and warning.

- Learners should have exclusive authority to their unique account.
- Learners should verify the answers are their own work.
- Learners should not make solutions they get available to anyone else.

If a violation of the honor code is found, the lightest penalization is to cancel the credits for the question on which the learner cheats, while the most severe punishment is to ban the cheating account. However, up till November 2016, edX has not designed or utilized any mechanism to detect or report the violations of honor code.

Verified Certificate

A Verified Certificate shows the learner not only has successfully passed the MOOC with a confirmation to the edX honor code but also has verified his/her identity with the face recognition system deployed on edX. Since the learners are verified, the Verified Certificate is supposed to be more formal than the Honor Code Certificate. Meanwhile, it carries a fee that varies by course.

XSeries Certificate

In September 2013, edX introduced XSeries as a series of multiple MOOCs in a specific subject. Simultaneously, with the launch of XSeries, a new certificate, XSeries Certificate was released on edX. An XSeries Certificate shows that the learner has earned passing grades in all of the courses that make up the series. It also requires learners to accept edX honor code and verify their identities. The certificate charges a fee that varies by series.

Chapter 4

Detection Methods

For investigating the prevalence of CAMEO in MOOCs, researchers have designed and implemented some detection methods [51, 59]. At the beginning of the chapter, we first introduce two different CAMEO patterns in Section 4.1. Then we start our detection experiments with a replication of the existing Singleton Detection Method [51]. In Section 4.2, we firstly introduce the method. After that we analyze some limitations of the method at the beginning of Section 4.3. To obtain a theoretically better detection performance, targeting at the limitations we pointed out, we design and illustrate the Hybrid Detection Method in the same section. During the process of the thesis, the Long Batch Detection Method [59] was published. In Section 4.4, we detail the method based on the publication and scripts respectively. Finally, in Section 4.5, we briefly introduce the detection results of the Singleton Detection Method and the Long Batch Detection Method in previous research. Then we make a table to compare the differences among the three detection methods we adopt in the experiment.

4.1 Patterns of CAMEO

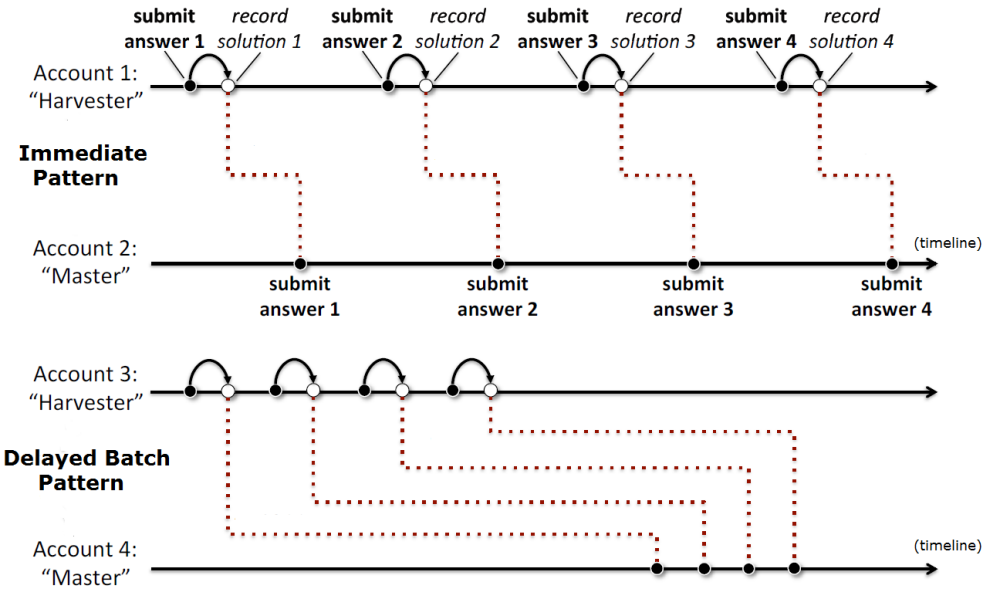

Figure 4.1: Two Types of Prototypical CAMEO Patterns
(Reproduced from Northcutt et al. 2015 [51])

There are two patterns in which cheaters apply CAMEO, *Immediate Pattern* and *Delayed Batch Pattern*. With Figure 4.1, we describe the patterns. In *Immediate Pattern*, CAMEO users probe the solution to a question with an account and submits the harvested solution to the question via another account immediately. In *Delayed Batch Pattern*, CAMEO users probe a batch of solutions to several questions and then switch to another account to submit the batch of harvested solutions. The accounts used for probing answers are called *Harvester Accounts*, while the accounts used for submitting harvested solutions and defrauding undeserved MOOC certificates or high grades are called *Master Accounts*.

4.2 Singleton Detection Method

The research about detecting and preventing CAMEO cheating in MOOCs was started in 2015. The earliest published experiment, to our knowledge, was conducted by Northcutt et al. [51]. For the two separate CAMEO patterns, Immediate Pattern and Delayed Batch Pattern, Northcutt et al. assumed a unitary CAMEO user behavior model and utilized a single detection method with constant thresholds to detect the model, thus we refer the detection method designed by Northcutt et al. as Singleton Detection Method. In this section, we illustrate the assumptions of CAMEO user behavior model, the criteria to detect the model and the operations in Singleton Detection Method.

4.2.1 Assumptions of CAMEO User Behaviors

Northcutt et al. made five assumptions of CAMEO user behaviors.

1. CAMEO user should hold at least two accounts.

The first and the most basic assumption is that a CAMEO user should hold at least two accounts: (1) one or more "Harvester" accounts used to acquire correct answers by submitting randomly-chosen answers and then accessing instructor-provided solutions via the "Show Answer" button and (2) one or more "Master" accounts used to submit the harvested solutions as their correct answers.

2. CAMEO user should harvest solutions first, then submit the correct answers.

For a question that CAMEO user want to cheat on, the CAMEO user should harvest the solution with his/her Harvester Account at first. Then it will become possible for him/her to submit the correct answer via his/her Master Account.

3. CAMEO user should transfer solutions from Harvester to Master Account quickly.

The close synchronicity between two accounts increases the possibility that there is a CAMEO user logging in simultaneously to both Harvester Account and Master Account on different browsers or computers.

4. CAMEO user's Master Account should be certified, and the Harvester should not.

The goal of a CAMEO user is presumably to earn a certification or even a higher score for Master Account. As for Harvester Account, it is mainly used to harvest solutions by submitting random guesses. It should not be interested in certifications, and the quality of answers it submitted is supposed to be too low to earn a certificate.

5. The Harvest and Master Account should be connected via IP addresses.

The relationship in the IP addresses two accounts used increases the probability that the two accounts indeed belong to a same person (the CAMEO user).

On the basis of the five assumptions, there are five corresponding data filters constituting the Singleton Detection Method to identify potential CAMEO users who satisfy all the assumptions above.

4.2.2 Detection of Potential CAMEO Users

Firstly, Singleton Detection Method begins by considering every account in the MOOC as a Candidate Master Account (referred as CM below). For a CM account, with the exception of the account itself, all the other accounts in the MOOC are considered as its Candidate Harvester Accounts (referred as CH below). Then, pair the CM with its corresponding CH respectively. A CM - CH pair represents a candidate CAMEO user.

For CM - CH pairs in a MOOC, a variable $\Delta t_{\mathrm{m,h,c,i}}$ is introduced in Figure 4.2.

$$\Delta t_{\mathrm{m,h,c,i}} = t_{\mathrm{m,c,i}} - t_{\mathrm{h,c,i}} \tag{4.1}$$

The variable $\Delta t_{\mathrm{m,h,c,i}}$ stands for the time difference between the candidate CAMEO user submitting the correct answer for question i in course c with CM m and the candidate CAMEO user harvesting solution for the same question with CH h.

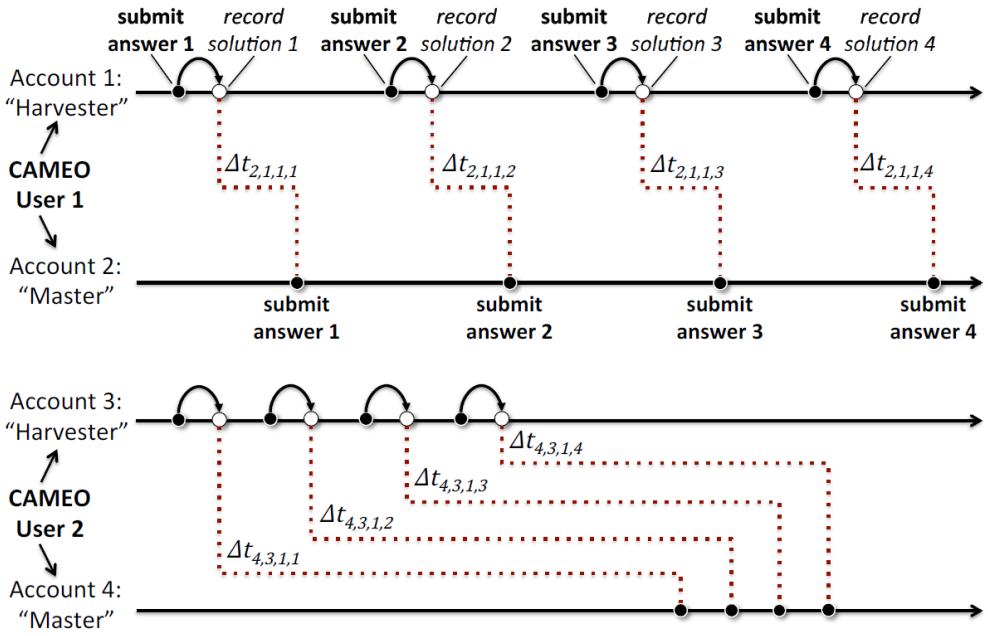

Figure 4.2: Samples of $\Delta t_{m,h,c,i}$ in CM - CH Pairs
(Copied from Northcutt et al. 2015 [51])

Most of the data filters in Singleton Detection Method rely on the value of $\Delta t_{m,h,c,i}$. For a CM (m)- CH (h) pair in a course (c), for different questions (i, j, k,...) in the course, there are several Δt ($\Delta t_{m,h,c,i}$, $\Delta t_{m,h,c,j}$, $\Delta t_{m,h,c,k}$,...). Here we use $\Delta t_{m,h,c}$ below to refer to the set of all Δt for the CM (m) - CH (h) pair in course (c).

4.2.2.1 Filter 1: Bayesian Filter

The Bayesian Filter is based on the assumption that CAMEO user should harvest solutions first, then submit the correct answers. According to the assumption, the $\Delta t_{m,h,c,i+1}$ is expected to be positive. The Bayesian Filter predicts a proportion π, which indicates the posterior probability that $\Delta t_{m,h,c,i+1}$ is positive, on the basis of the existing $\Delta t_{m,h,c}$. If the proportion π is larger than 90%, the CM - CH pair is kept as a candidate awaiting the process of next data filter, while the other pairs are discarded. Compared with other prediction, Bayesian inference is more robust when data is limited [37].

The detailed prediction process is as follows. There are several parameters needed. First of all, not for all questions in the course, exists the variable $\Delta t_{m,h,c,i}$. The variable $\Delta t_{m,h,c,i}$ exists only when the CM have submitted the correct answers and the CH have clicked the "Show Answer" button for the same question i. The first parameter n is the number of existing $\Delta t_{m,h,c,i\,/\,j\,/\,k\,/...}$ for the CM (m) - CH (h) pair in the course (c). The second parameter x, is the number of positive $\Delta t_{m,h,c,i\,/\,j\,/\,k\,/...}$. Since $\Delta t_{m,h,c}$ can be represented as a series of positiveness and non-positiveness, the parameter x follows the binomial distribution in Equation 4.2.

$$x|n,\pi \sim Binomial(\pi,n) \tag{4.2}$$

On the basis of Bayesian inference, Beta distribution is conjugate prior to the binomial distribution. The conjugate prior distribution is for representing a probabilistic distribution of the probability (π in this case) before some evidence (parameters n and x in this case) is taken into account. The probability π in Equation 4.2 follows a Beta distribution.

$$\pi|\alpha,\beta \sim Beta(\alpha,\beta) \tag{4.3}$$

For a Beta distribution, there are two parameters, α and β. The initial values of the two parameters are both 0.5. The shape of a Beta distribution with $\alpha = \beta = 0.5$ is shown in Figure 4.3. The distribution of proportion π with $\alpha = \beta = 0.5$ is a gentle U-shape, which is consistent with an assumption that most CM and CH are irrelevant and most $t_{\mathrm{m,c,i}}$ should deviate from $t_{\mathrm{h,c,i}}$ in one direction, due to the asynchronous nature of MOOCs.

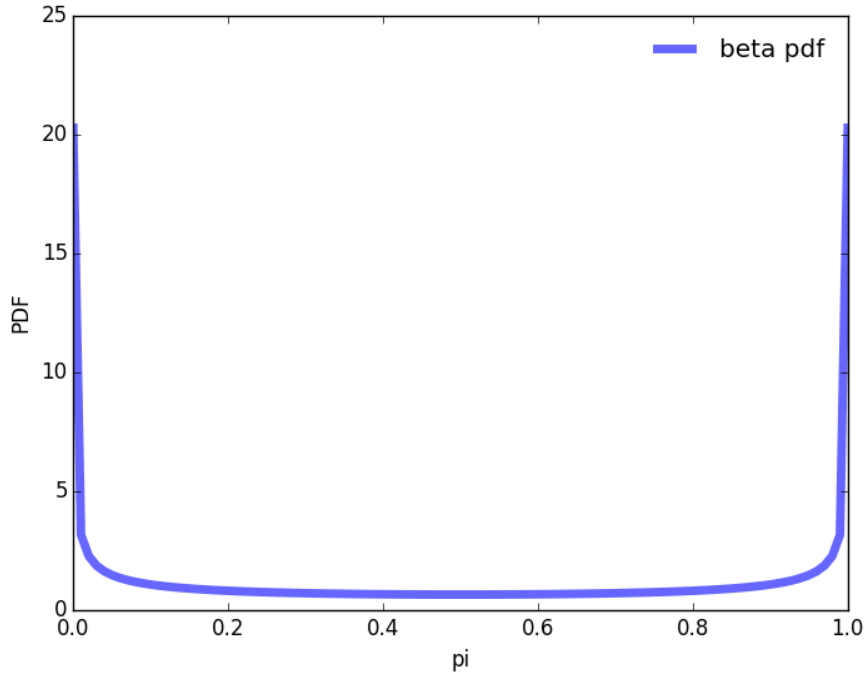

Figure 4.3: Probability Density Function of Prior Beta Distribution ($\alpha = 0.5, \beta = 0.5$). The probability density function (PDF) is a function that describes the relative likelihood for this random variable to take on a given value. The U-shape depicts the value of π (the proportion of positive $\Delta t_{\mathrm{m,h,c,i}}$ is more likely to extremely close to 0 or 1.

Given the prior probability distribution $P(\pi|\alpha,\beta) = Beta(\alpha,\beta)$ and a CM - CH pair data $D = (x, n-x)$, the posterior Beta distribution is

$$\begin{aligned} P(\pi|\alpha,\beta,x,n-x) &\propto P(x,n-x|\pi)P(\pi|\alpha,\beta) \\ &\propto \pi^x(1-\pi)^{n-x}\pi^{\alpha-1}(1-\pi)^{\beta-1} \\ &= \pi^{x+\alpha-1}(1-\pi)^{n-x+\beta-1} \\ &= Beta(\alpha+x,\beta+n-x) \end{aligned} \tag{4.4}$$

$$\pi|x,n,\alpha,\beta \sim Beta(\alpha+x,\beta+n-x) \tag{4.5}$$

On the basis of Equation 4.4, we draw a conclusion (Equation 4.5) that the posterior proportion π follow a Beta distribution ($Beta(0.5+x,0.5+n-x)$).

According to the criterion of the Bayesian Filter, only when the Cumulative Distribution Function $F_\pi(0.9) \leq 0.1$, the CM - CH pair is kept. It is a quite stringent criterion that requires considerable data before concluding that for a CM - CH pair, the CH precedes the CM. To illustrate the strictness of the criterion, we plot several curves with different $D = (x, n-x)$ in Figure 4.4.

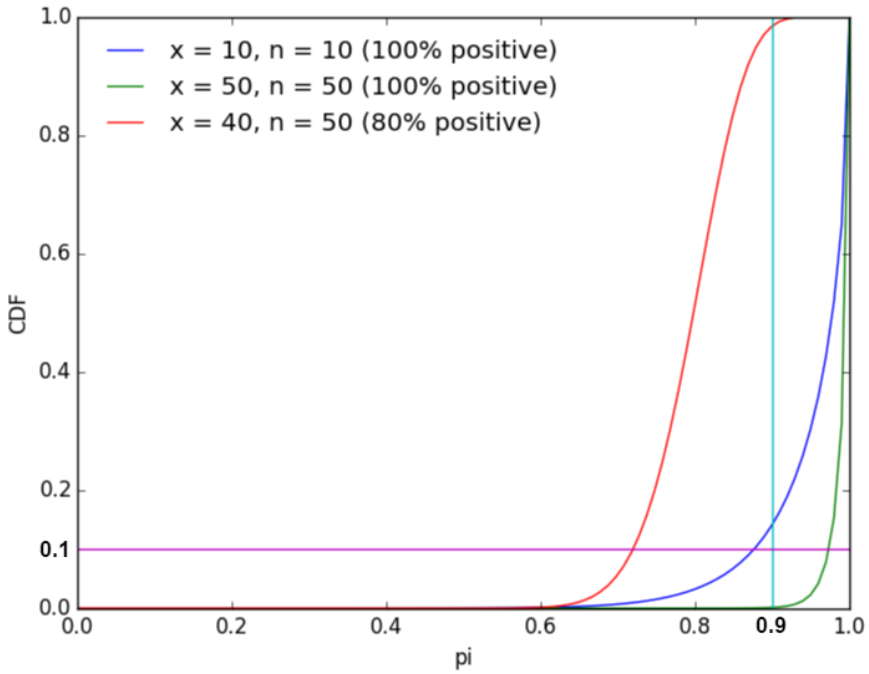

Figure 4.4: Cumulative Distribution Function of π.
(The Cumulative Distribution Function (CDF) of a variable X evaluated at y, $F_X(y)$, equals to the probability that X will take a value less than or equal to y.)

With Figure 4.4, the criterion of Bayesian Filter can be intuitively rephrased as evaluating the curve at $\pi = 0.9$, if the value is smaller than 0.1, the CM - CH is kept, otherwise, the pair is discarded. We plot three different curves in Figure 4.4. For the red curve, CH clicked the "Show Answer" button and CM submitted solutions for 50 common questions ($n = 50$). On 40 ($x = 40$) out of the 50 questions, CH clicked the "Show Answer" button before CM's correct submissions. 80% of $\Delta t_{\text{m,h,c}}$ are positive, however, the positive rate is still not large enough to satisfy the criterion. For another case, the blue curve in Figure 4.4, 100% $\Delta t_{\text{m,h,c}}$ are positive ($x = 10, n = 10$), however, the number of common questions on which both CH asked solutions and CM correctly answered ($n = 10$) is so limited that the CM - CH pair cannot meet the criterion.

CM - CH pairs left after the Bayesian Filter have satisfied two assumptions, (1) CAMEO user should hold at least two accounts, Harvester and Master Account; (2) CAMEO user should harvest solutions first, then submit the correct answers.

4.2.3 Filter 2: Time Difference Filter

The Time Difference Filter is based on the assumption that CAMEO user should transfer solutions from Harvester Account to Master Account quickly. According to the assumption, the magnitude of $\Delta t_{\mathrm{m,h,c,i}}$ should be very small. The criterion of the filter is that 90% of the $\Delta t_{\mathrm{m,h,c}}$ must be less than 5 minutes. CM - CH pairs which pass the threshold are kept.

The explanation for the threshold (5 minutes) is that the cutoff changes the number of survived CM - CH pairs dramatically when shifting between 0 to 5 minutes, and subsequent shifts past 5 minutes do not. The elbow shape curve in Figure 4.5 decides the cutoff threshold used in the Time Difference Filter.

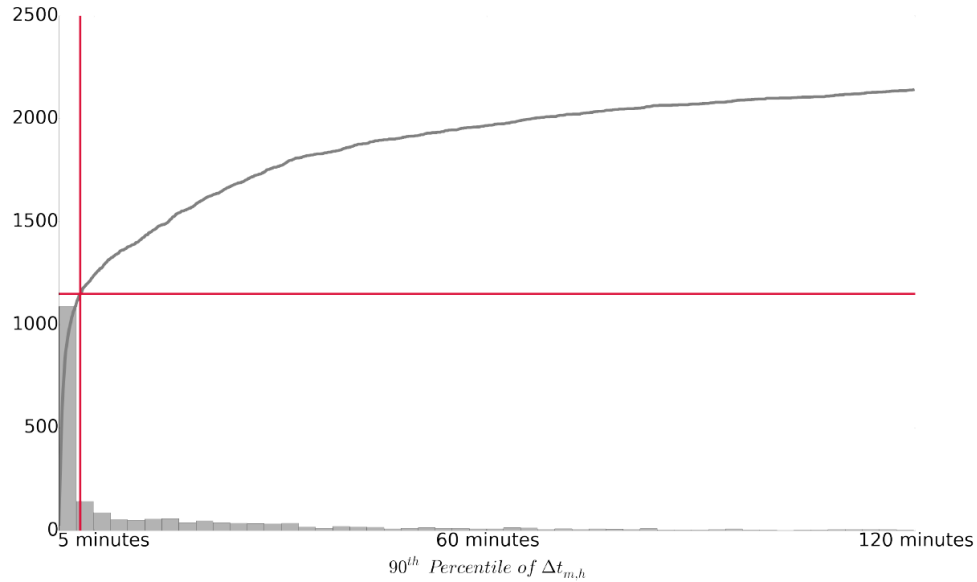

Figure 4.5: Number of CM - CH left versus the 90^{th} percentile cutoff value of $\Delta t_{\mathrm{m,h}}$. The vertical red line indicates the cutoff threshold, and the horizontal red line is the corresponding survival CM - CH pairs.

(Copied from Northcutt et al. 2015 [51])

CM - CH pairs left after the Bayesian Filter and the Time Different Filter have satisfied three assumptions of CAMEO users behaviors, (1) CAMEO user should hold at least two accounts, Harvester and Master Account; (2) CAMEO user should harvest solutions first, then submit the correct answers; (3) CAMEO user should transfer solutions from Harvester to Master Account quickly.

4.2.4 Filter 3: Certificate Filter

The Certificate Filter is based on the assumption that Master Account should earn a certificate and the Harvester should not. CM - CH pairs meeting the criterion are kept.

4.2.5 Filter 4: IP Filter

The IP Filter is based on the assumption that a CAMEO user's Harvester and Master Account should be connected via IP addresses. To be specific, in IP Filter, multiple Accounts Closures are constructed. An account closure contains multiple accounts which shared a same IP address directly or were (recursively) transitively connected via other accounts with who shared an IP address. For a CM - CH pair, if the CM and CH are in a common Account Closure, the CM - CH pair is kept.

To construct Account Closures, firstly, massive tuples (Account_ID, IP) are created recording the identification number of an account and a corresponding IP address the account has used in the course. To initialize account groups, these tuples are grouped by IP addresses. It means the tuples recording same IP addresses are assigned to a same account group. Then if there is an Account_ID appearing in two or more account groups, these groups are merged into one. Similarly, the newly-merged account groups are connected by IP addresses in the same way. The merge is repeated until the size of each account group no longer changes. The fixed-sized account groups are called Account Closures.

CM - CH pairs survived after the Bayesian Filter, the Time Difference Filter, the Certificate Filter and the IP Filter have satisfied all assumptions of CAMEO user behaviors in the Singleton Detection Method.

4.2.6 Filter 5: Router Filter

To reduce false positives, the Router Filter is used to discard CM - CH pairs which CM and CH are in an Account Closure that contains more than 10 accounts. It is supposed that IP addresses shared by these accounts are highly likely to be assigned in public space such as classroom or cafes.

4.3 Hybrid Detection Method

During the replication of the Singleton Detection Method, we find that there are some limitations existing on the method. In this section, we firstly analyze the limitations of Singleton Detection Method. Targeting at these limitations, we design a Hybrid Detection Method on the foundation of the original Singleton Detection Method. The new method refines the assumptions of CAMEO user behaviors and replaced portions of data filters.

4.3.1 Limitations of Singleton Detection Method

We suppose there are two relatively obvious limitations of Singleton Detection Method.

The first limitation is the definition of Harvester Account. In the first assumption of CAMEO user behaviors in Singleton Detection Method, the Harvester is defined as an account used to acquire correct answers by submitting randomly-chosen answers and accessing instructor-provided solutions via the "Show Answer" button. However, the "Show Answer" button is not the only pathway for CAMEO user to get access to the solutions. Learners are able to harvest correct answers by exhausting submission

attempts. For all questions on edX, after learners click the "Submit" button, an immediate feedback will be given based on the correctness of the submission. Besides, according to our observation, most of the questions on edX have at least two submission attempts. With the immediate feedback and the multiple submission attempts, it is highly possible that learners can get solutions without clicking the "Show Answer" button, especially for multiple-choice questions. Therefore, in Hybrid Detection Method, we remove the restriction of Harvester Accounts only being able to access correct answers via the "Show Answer" button.

The second limitation we find is the requirement that CAMEO user should transfer solutions from Harvester Account to Master Account quickly (within 5 minutes in this case). However, for CAMEO users who utilize the Delayed Batch Pattern, they will stay on their Harvester Accounts for a while to harvest solutions for multiple questions. The harvesting process may last longer than several minutes, which results in the time difference between the Harvester's harvesting and the Master's correct submission ($\Delta t_{\mathrm{m,h,c,i}}$) too large to pass the cutoff of Time Difference Filter. Therefore, we modify the third assumption of the Singleton Detection Method and add some new assumptions of CAMEO user behaviors in Hybrid Detection Method.

4.3.2 Assumptions of CAMEO User Behaviors

There are seven assumptions of CAMEO user behaviors in Hybrid Detection Method.

We modify the 1st assumption of Singleton Detection Method.

1. CAMEO user should hold at least two accounts.

A Harvester Accounts is used to access correct answers via show answer button or exhaustive attempts. A Master Accounts is used to submit the harvested solutions.

We inherit the 2nd, 4th and 5th assumptions from the Singleton Detection Method.

2. CAMEO user should harvest solutions first, then submit the correct answers.

3. CAMEO user's Master Account should be certified, and the Harvester should not.

4. The Harvest and Master Account should be connected via IP addresses.

We add three new assumptions of CAMEO user behaviors.

5. For CAMEO user who utilizes Immediate Pattern:

CAMEO user should transfer solutions from Harvester to Master quickly.

6. For CAMEO user who utilizes Delayed Batch Pattern:

The Master should correctly and rapidly answer several consecutive questions.

In Delayed Batch Pattern, CAMEO user has collected a set of solutions in advance. Intuitively, if a learner has already known solutions to multiple questions, the time he/she spent on these questions will be significantly reduced.

7. CAMEO user should use harvested solutions in Master Account.

The exclusive function of Harvester Accounts is supposed to be harvesting solutions. Only when these harvested answers used in a Master Account do the Harvester Accounts play their role, and the cheating process is accomplished.

On the basis of the first four assumptions, we inherit four data filters from the Singleton Detection Method. As for the last three assumptions, we design an Immediate Time Difference Filter, a Batch Time Difference Filter and a Proportion Filter to detect potential CAMEO users who satisfy all these assumptions.

4.3.3 Detection of Potential CAMEO Users

The workflow of the Hybrid Detection Method is described in Figure 4.6.

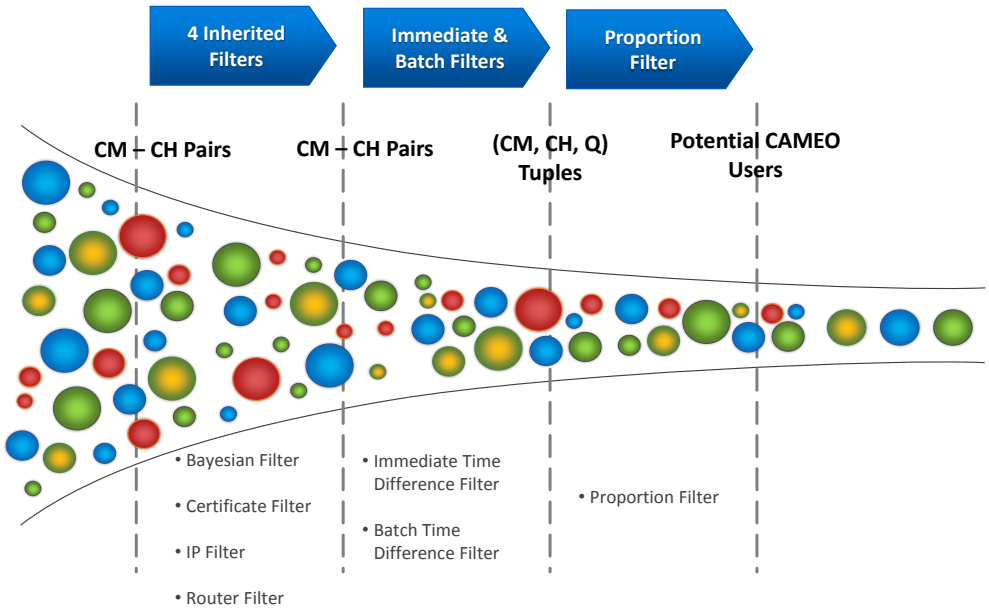

Figure 4.6: Three Phases of Hybrid Detection Method.

Firstly, similar to the Singleton Detection Method, we construct massive CM - CH pairs and calculate $\Delta t_{\mathrm{m,h,c}}$ for these pairs. On the basis of the values of $\Delta t_{\mathrm{m,h,c}}$, these pairs are selected by Bayesian Filter, Certificate Filter, IP and Router Filter. The pairs survived after the four filters have satisfied the first four assumptions of CAMEO user behaviors. After the filters inherited from Singleton Detection Method, Immediate Time Difference Filter and Batch Time Difference Filter are introduced.

4.3.3.1 Immediate Time Difference Filter

The Immediate Time Difference Filter is targeting at detecting Immediate Pattern, in which it is supposed that CAMEO users do the harvest with Harvester Accounts and the submission with Master Accounts alternately. The filter is based on the assumption

that for CAMEO user who utilizes Immediate Pattern, CAMEO user should transfer solutions from Harvester Account to Master Account quickly.

Similar with the Time Difference Filter in Singleton Detection Method, the criterion of the filter is also based on the value of $\Delta t_{\mathrm{m,h,c,i}}$. To be specific, for a CM (m) - CH (h) pair, for a question i in the course c, if $\Delta t_{\mathrm{m,h,c,i}}$ is smaller than 5 minutes, a tuple (m, h, i) is created and recorded as a potential Immediate CAMEO Event.

4.3.3.2 Batch Time Difference Filter

The Batch Time Difference Filter is targeting at detecting Delayed Batch Pattern. It is based on the assumption that for CAMEO user who utilizes Delayed Batch Pattern, the Master should correctly and rapidly answer several consecutive questions.

In Batch Time Difference Filter, a new kind of $\Delta t_{\mathrm{m,h,c,i}}$ is defined. To distinguish it from the $\Delta t_{\mathrm{m,h,c,i}}$ we introduced before, we use $\Delta_I t_{\mathrm{m,h,c,i}}$ to refer to the former one (time difference between the candidate CAMEO user submitting correct answers for question i in course c with CM m and the candidate CAMEO user harvesting solutions for the same question with CH h), and $\Delta_B t_{\mathrm{m,h,c,i}}$ to refer to the one we are going to describe in Figure 4.7.

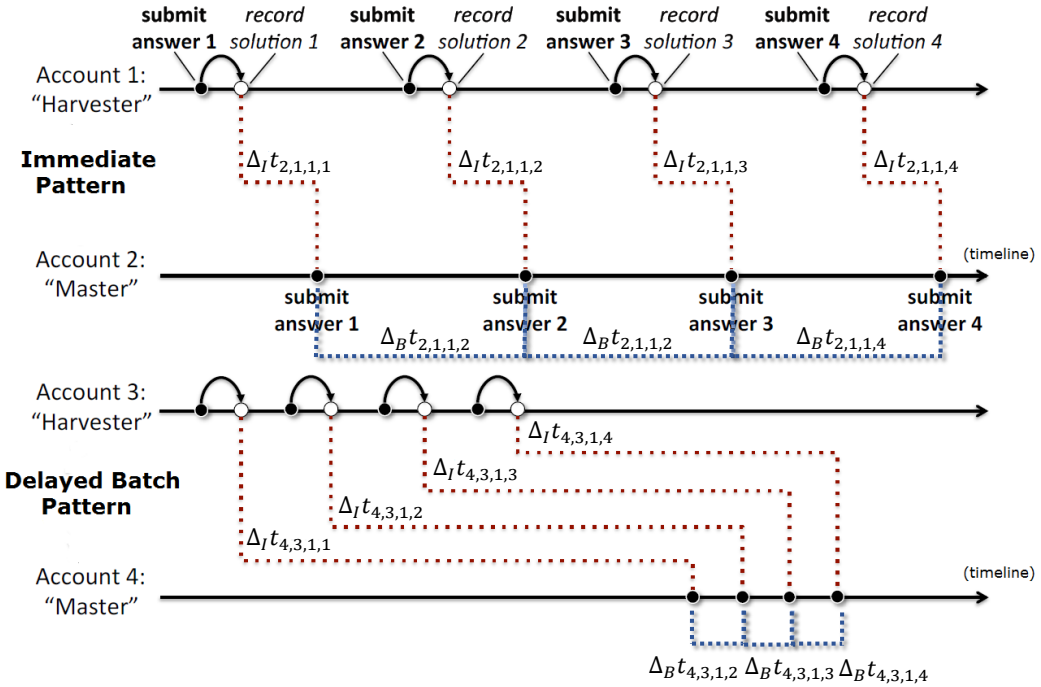

Figure 4.7: Samples of $\Delta_B t_{\mathrm{m,h,c,i}}$ in CM - CH Pairs
Reproduced on Northcutt et al. [51]

For a CM (m) - CH (h) pair in course c, for a question i for which the CH has harvested solutions before the CM submits correct answers, the time difference between the CM's correct submission and its last adjacent correct submission is recorded as $\Delta_B t_{\mathrm{m,h,c,i}}$. The last adjacent correct submission can be not only a submission which

has its corresponding harvest event in Harvester Account, but also a single correct attempt without homologous harvest. The $\Delta_B t_{\mathrm{m,h,c,i}}$ implies how much time the CM account m cost on the question i in course c.

On the basis of the sixth assumption, the response should be "unreasonable fast" when a CAMEO user has already known the solutions to the question. Therefore, for a CM (m) - CH (h) pair, for a question i in the course c, if the $\Delta_B t_{\mathrm{m,h,c,i}}$ is less than 99% time other certified accounts cost on the question, a tuple (m, h, i) is created and recorded as a potential Batch CAMEO Pattern.

4.3.3.3 Proportion Filter

The Proportion Filter is based on the last assumption that CAMEO user should use harvested solutions in Master Account. We collect all suspicious Immediate CAMEO Events and Batch CAMEO Events, and group them by CM - CH pair. For a CM - CH pair, there is a group of tuples. If more than 90% of the questions for which the CH harvests solutions can find their corresponding tuples in the group, which indicates these correct answers are used by the CM with either Immediate Pattern or Delayed Batch Pattern, the CM - CH pair is kept and regarded as a potential CAMEO user.

4.4 Long Batch Detection Method

During the process of the master thesis, another work focusing on detecting multiple accounts cheating in MOOCs was published on the 3rd ACM Learning at Scale Conference. Compared with the other two detection methods, the method is featured by its requirement to the length of the batch in Delayed Batch Pattern, thus we refer it as Long Batch Detection Method. For the method, we also summary the assumptions of CAMEO user behaviors and detection process based on the publication[59]. Besides, we get contact with the authors of Long Batch Detection Method and get access to the newest scripts. Researchers [59] have made several changes to the original method on scripts and these updates will be illustrated at the end of this section.

4.4.1 Assumptions of CAMEO User Behaviors

There are six assumptions of CAMEO user behaviors.

1. CAMEO user should hold at least two accounts.

2. CAMEO user's Harvester Accounts should not be certified.

3. CAMEO user should use harvested solutions in Master Account.

4. The Master Account should use at least 10 harvested answers.

5. For CAMEO user who utilizes Immediate Pattern:

1) CAMEO user should transfer solutions from Harvester to Master quickly.

2) For a question the CAMEO user cheats on, the Harvester and Master Account should harvest and submit with a common IP address.

There is a subtle difference between the criterion and the other assumption about CAMEO users' IP addresses we mentioned before. It is a more strict assumption that requires the Harvester and Master Account must share a same IP address directly at specific events.

6. For CAMEO user who utilizes Delayed Batch Pattern:

1) CAMEO user should correctly and rapidly answer at least 10 consecutive questions in Master Account.

2) The Harvest and Master Account should be connected via IP addresses.

On the basis of these assumptions, four procedures are set in the Long Batch Detection Method to conduct the detection of potential CAMEO users.

4.4.2 Detection of Potential CAMEO Users

Four detection procedures are as follows.

Firstly, multiple Account Closures are created. An Account Closure contains several accounts which have shared IP address directly or connected via other accounts.

Secondly, for Immediate Pattern, for each account M, for every correct submission of the account, search if there is any account H with same IP address has submitted correct answers or clicked the "Show Answer" button in 15 minutes before for the same question. If so, then create a tuple recording the accounts, account M as Candidate Master Account while account H as Candidate Harvester Account, and the question, then add the tuple to the list of candidate CAMEO Events.

Thirdly, for Delayed Batch Pattern, for each account M, for each batch of consecutive 10 or more than 10 correct submissions with no more than 20 seconds between each two adjacent submissions of the account, search if there is an account H in the same Account Closure has submitted correct answers or required solutions for the questions in the batch. If there is, and the time difference between the first correct submission account M makes in the batch and the last harvest account H makes for the questions in the batch is larger than 15 minutes, then create multiple tuples recording the accounts and every question in the batch, then push the tuples into the list of candidate CAMEO Events.

Finally, the tuples in the candidate CAMEO Events list are filtered to check whether they satisfy the second, third and the fourth assumptions of CAMEO user behaviors described in the last subsection above.

The tuples survived after the four detection procedures are regarded as suspicious CAMEO Events. The Harvester and the Master Account in a suspicious tuple are supposed to be a potential CAMEO user detected by Long Batch Detection Method.

4.4.3 Variant

During the replication of Long Batch Detection Method based on the publication [59], we contact the authors and get the latest version of the scripts for the method. Several changes have been made to the original algorithm on the scripts. According to our understanding of the scripts, the first phase of the Long Batch Detection Method (Code Version) still is to construct Accounts Closures which contains several accounts connected by IP addresses. However, two extra limitations are added in the phrase. The first limitation is to remove IP address which has been shared by more than 100 learners, and the second limitation is to remove Account Closure which contains more than 100 accounts. Then in the second phase, the distinction between the Immediate Pattern and the Delayed Batch Pattern is abolished. Instead, for each account M, for every correct submission of the account, search if there is any account H in the same Account Closure has submitted the correct answers or required solutions for the same question in 24 hours before. If there is, create tuple recording the accounts and the question. Besides, more detailed information including whether the account M has tried to solve the question before copying harvested answers and which kind of harvesting means the account H uses is also recorded on the tuple. The third phase is to filter the tuples. Parts of the criteria in the original Long Batch Detection Method including the Harvester should not be certified, the harvested solutions should be used in Master Account and the Master Account should use at least 10 harvested answers are kept. In addition, two new criteria are added. The first criterion is for a Master Account, more than 5% correct submissions should be done with the help of CAMEO. The second criterion is also for a Master Account, it should have at least 5 very fast submissions that a fast submission should be done within 30 seconds. The tuples passed the three phases are detected as suspicious CAMEO Events. The accounts in a survived tuple are regarded as the Master and Harvester Account of a potential CAMEO user.

4.5 Comparison

The Singleton Detection Method was first designed and implemented on 115 MOOCs created by Massachusetts Institute of Technology and Harvard University on edX [51]. Among these courses, researchers detected highly suspicious CAMEO users in 9 MOOCs. An estimated 1237 certificates were defrauded by the CAMEO users accounting for 1.3% of the certificates in the involved 69 MOOCs. For another existing detection method, in 2016, Ruiperez-Valiente et al. designed and applied the Long Batch Detection Method on a MOOC called Mechanics Review (8.MReVx) on edX [59]. 52 certified CAMEO master accounts were detected which constitutes 10.3% of the 502 accounts which earned certificates in the MOOC.

So far we have described three different detection methods for CAMEO, two existing methods, Singleton and Long Batch Detection Method and a new one, Hybrid Detection Method. All of these methods set assumptions of CAMEO user behaviors, and then identify CAMEO users by detecting learners who behave these assumed specific behavioral characteristics. The criteria (assumptions) used by the three detection methods overlap each other partly. Meanwhile, there are also some meaningful differences among the three methods. In Table 4.1, we make a comparison among the criteria for identifying CAMEO users in different detection methods.

Table 4.1: Comparison of Different Detection Methods

Category	Criterion	Detection Methods		
		Singleton	Hybrid	Long Batch
Harvest Approaches	CAMEO user can access solutions via **"Show Answer" button**.	✓	✓	✓
	CAMEO user can access solutions Via **Exhaustive Attempts**.	✗	✓	✓
Certified Status	**Master** Account should be **certified**.	✓	✓	✗
	Harvester Account should **not be certified**.	✓	✓	✓
IP Addresses	For a CAMEO user, Harvester and Master should **directly share** an IP address.	✗	✗	✓ (Immediate Pattern)
	For a CAMEO user, Harvester and Master should be **connected** via IP addresses.	✓	✓	✓
Immediate Pattern	CAMEO user should transfer solutions from Harvester to Master **quickly**.	✓	✓	✓
Delayed Batch Pattern	Master Account should **rapidly** submit **correct** answers to **consecutive** questions.	✗	✓	✓
	The **number** of questions in a CAMEO Batch should be **larger than 10**.	✗	✗	✓
Others	**Most harvested solutions** should be **used** in Master Account.	✗	✓	✓

*Although some criteria are shared, the thresholds may be diverse in different detection methods..

With the help of Table 4.1, we find that for Singleton Detection Method, (1) the detection criteria are the simplest among the three detection methods, (2) it cannot detect CAMEO users who harvest solutions via exhaustive attempts, (3) there is no detection criteria targeting at detecting CAMEO users who utilize Delayed Batch Pattern; for Hybrid Detection Method, (1) compared with the Singleton Detection Method, it is able to detect not only CAMEO user who cheats with "Show Answer" button but also CAMEO user who harvests answers via exhaustive attempts, (2) CAMEO users with Delayed Batch Pattern can be detected; and for Long Batch Detection Method, (1) it has the most strict detection criteria among the three detection methods, (2) compared with Hybrid Detection Method, some extra conditions about the length of a CAMEO Batch are added to enhance the suspicion of a CAMEO user and reduce false positives.

Chapter 5

Results

In the previous chapter, we detailed the three detection methods. To investigate the prevalence of CAMEO in MOOCs, these methods are implemented on 10 MOOCs created by TU Delft on edX. Before we detail the information about the 10 DelftX courses, we want to firstly illustrate an important premise of the experiment.

The premise is there is no provable data for the results of detecting CAMEO users in MOOCs. In other words, there are no labeled test samples to validate the assumptions of CAMEO user behaviors and assess the precision, recall of the detections. The absence of ground truth originates from the fact that learners of MOOCs are scattered around the world and it is impossible to supervise the answering questions process of these learners. We have considered solving the problem with questionnaires. However, because of the "sensitive" of cheating, making a questionnaire asking whether learners have utilized CAMEO to cheat seems unavailable. On the one hand, the sensitivity may make the respondents feel their privacy are invaded. On the other hand, we are afraid that learners are not willing to response truthfully towards a sensitive question about using cheating strategy.

With the issue that there is no ground truth for the detections, we review the previous research about detecting CAMEO users in MOOCs [51, 59] again. We find that for Singleton Detection Method, Northcutt et al. [51] conducted a small-scale, targeted verification experiment in a single, small course to evidence the effectiveness of detection method. In the course, there were 3 pairs of users, consisting of 3 CM and 3 CH Accounts who required and submitted answers unusually synchronous, identified by instructors. For the 3 user pairs, Northcutt et al. [51] adapted instructor-prepared answers to 7 test questions by appending a unique randomly numerical string to the answer displayed to each user. To be specific, the original answer to a question might be 3.13, while the adapted displayed answer was 3.13556 to one user, and 3.13447 to another [51]. Theoretically, the answers with unique superfluous strings should not be submitted unless the user pair indeed belongs to a CAMEO user who copies answers from an account to another account. One out of the 3 user pairs never required any adapted answers. For the other two, direct copying of at least one unique answer was detected, which indicates they are two CAMEO users. At the same time, the two user pairs were detected by the Singleton Detection Method, which partially evidences the effectiveness of Singleton Detection Method. For logistical and pedagogical reasons, Northcutt et al. [51] restricted the verification experiment to the three user pairs. For Long Batch Detection Method [59], there is no discussion about verification.

Although the lack of ground truth makes it difficult to evaluate the performance of detection methods, the research about detecting CAMEO users in MOOCs still makes sense. All of the detection methods we described in the last chapter set rational assumptions of CAMEO user behaviors and utilize relatively strict thresholds to predict the lower bound of the prevalence of CAMEO in MOOCs.

In the following sections, we detail the information about the 10 MOOCs in our dataset in Section 5.1. Then we report the numbers of suspicious CAMEO users detected by each detection methods in each MOOC in Section 5.2, which indicates the existence and popularity of CAMEO in DelftX. Since we do not have a ground truth to evaluate the detection methods, we do a "sanity check" in Section 5.3 to illustrate we indeed detect a CAMEO user and the detection results can be true. Besides, to understand CAMEO uses preferences in cheating and then to provide appropriate and suitable suggestions for preventing CAMEO in MOOCs, we analyze and illustrate characteristics of the detected suspicious CAMEO users including the certificate mode, potential motivation of cheating, distribution of region and phase when they are most likely to cheat in Section 5.4.

5.1 DelftX Dataset

TU Delft started to offer MOOCs on edX from September 2013. Up till November 2016, TU Delft has created 45 MOOCs and XSeries programs covering various scientific and engineering fields including management, data analysis, environmental studies etc. For the sake of the experiment, we temporally sample 8 courses with diverse subjects out of the 45 MOOCs. Among the 8 MOOCs, there are 2 courses have been run twice at two separate times, thus 10 courses data in total. Here we provide a table, Table 5.1, following to describe each of these courses in detail.

Table 5.1: Information about 10 DelftX MOOCs

Course Code	Course Name	Session	Number of Total Questions	Enrollment	Number of Users who has Submitted Answers	Number of Certified Users
FP101x (2014)	Introduction to Functional Programming	2014 Fall	288	37940	9327	1356 268 Verified / 1088 Honor
CTB3365DWx	Drinking Water Treatment	2014 Fall	254	10458	2132	246 71 Verified / 175 Honor
EX101x (2015S)	Data Analysis: Take It to the Max()	2015 Spring	136	33515	10205	2190 432 Verified / 1756 Honor
Frame101x	Framing: How Politicians Debate	2015 Spring	26	34017	5847	919 107 Verified / 812 Honor
Calc001x	Pre-University Calculus	2015 Summer	565	27857	6260	358 45 Verified / 313 Honor
EX101x (2015F)	Data Analysis: Take It to the Max()	2015 Fall	146	21041	5876	1156 214 Verified / 942 Honor
IB01x	Industrial Biotechnology	2015 Fall	551	8143	1977	329 103 Verified / 226 Honor
FP101x (2015)	Introduction to Functional Programming	2015 Fall	288	20936	4902	1143 276 Verified / 867 Honor
RI101x	Responsible Innovation: Ethics, Safety and Technology	2016 Spring	77	2741	380	113 34 Verified / 79 Honor
CTB3365sTx	Urban Sewage Treatment	2016 Spring	272	9566	1442	361 136 Verified / 225 Honor

5.2 Detected Potential CAMEO Users

We apply the three detection methods described in Chapter 4 on the 10 DelftX MOOCs and detect multiple potential CAMEO users. The table below, Table 5.2, shows the number of detected potential CAMEO user in each of the 10 DelftX MOOCs and the proportion of the detected potential CAMEO users in certified users in the course. From Table 5.2 we can see that in all of the 10 DelftX MOOCs, the trace of the existence of CAMEO users is detected with different detection methods. To be specific, the maximum number of potential CAMEO users in a MOOC is 65, which is detected by Long Batch Detection Method (Code Version) in course EX101x conducted in 2015 Spring. Meanwhile, there is no CAMEO user detected by the Singleton Detection Method in Calc001x.

Table 5.2: Number and Proportion of Detected Potential CAMEO Users

Method / MOOC	Singleton	Hybrid	Long Batch	Long Batch (Variant)
FP101x (2014)	13 (0.96%)	38 (2.80%)	26 (1.92%)	48 (3.54%)
CTB3365DWx	4 (1.63%)	7 (2.88%)	9 (3.66%)	15 (5.69%)
EX101x (2015S)	26 (1.19%)	42 (1.92%)	5 (0.23%)	**65** (2.97%)
Frame101x	**0 (0)**	3 (0.33%)	4 (0.44%)	4 (0.44%)
Calc001x	13 (3.63%)	19 (5.31%)	17 (4.75%)	23 (6.42%)
EX101x (2015F)	20 (1.73%)	28 (2.42%)	23 (1.99%)	32 (2.77%)
IB01x	10 (3.04%)	14 (4.26%)	12 (3.65%)	22 (6.69%)
FP101x (2015)	16 (1.40%)	20 (1.75%)	13 (1.14%)	29 (2.54%)
RI101x	7 (6.19%)	8 (7.08%)	7 (6.19%)	13 **(11.50%)**
CTB3365sTx	25 (6.93%)	30 (8.31%)	25 (6.93%)	39 (10.80%)
Total	137 (1.68%)	209 (2.56%)	141 (1.73%)	289 (3.54%)

Comparing the number of potential CAMEO users detected by different methods, generally, the Singleton Detection Method identifies the minimal number of potential CAMEO users among the four methods, then the Long Batch Detection Method (Publication Version), then the Hybrid Detection Method and the Long Batch Detection Method (Code Version) can detect the most suspicious CAMEO users.

Comparing the number of detected potential CAMEO users in different MOOCs, generally, the percent of potential CAMEO users in computer science MOOCs including FP101x and EX101x (around 2%) is quite lower than the percentage of potential CAMEO users in courses with other subjects (from 3% to 11%).

Comparing the number of detected potential CAMEO users in MOOCs conducted in different sessions, the proportions of potential CAMEO users in courses conducted in 2016 including RI101x and CTB3365sTx (all above 6%) are much higher than the percentages of detected potential CAMEO users in MOOCs conducted before 2016.

Compare the number of detected potential CAMEO users with previous research [51]. Northcutt et al. implemented Singleton Detection Method on 115 MOOCs created by MIT and Harvard University. Among the 115 courses, an estimated 1237 certificates in 69 MOOCs were defrauded by CAMEO users accounting for 1.3% of the certificates. On the basis of the publication, we duplicate the Singleton Detection Method on the 10 MOOCs created by TU Delft. Among the 10 courses, 137 potential CAMEO users are detected, which accounts for 1.89% of the certified users in the 9 courses. The percentage of potential CAMEO users in MOOCs created by TU Delft is slightly higher than the proportion in MOOCs created by MIT and Harvard University.

Although there are some differences in the number of detected potential CAMEO users by different detection methods, because of the overlap in the assumptions of CAMEO users utilized by the detection methods, there is a high similarity among the potential CAMEO users identified by the methods.

5.3 Verification of A Detected CAMEO User

A sanity check is a test to quickly evaluate whether a claim or the result of an experiment can possibly be true. Since there is no ground truth to comprehensively evaluate the performance of the different detection methods, we do a sanity check instead to indicate the detection results can possibly be true.

For a CAMEO user, we have a hypothesis that his/her Harvester and Master Account are likely to have similar account profiles. There are five mandatory fields including Email, Full name (legal name used for certificates), Public username, Password and Country in an account profile on edX. We suppose the similarity between Harvester and Master account profiles maybe reflect in registered Full name and Public username. For instance, the Public username of Harvester Account is "JaneDoe", while the Public username of Master Account may be "JaneDoe_1". Thus, for every detected potential CAMEO users, we calculate two similarities across account profiles, including the similarity between the Full name of Harvester and corresponding Master Account and the similarity between Public username.

The similarity is calculated by Ratcliff/Obershelp pattern recognition. The mechanism of the algorithm is as follows: matching characters (case-sensitive) are those in the longest common subsequence plus, recursively, matching characters in the unmatched region on either side of the longest common subsequence. In Figure 5.1, we show a sample for two spellings of our affiliation, Delft University of Technology and Technology University of Delft. The similarity is 0.63. Previous research [8] shows that a similarity over 0.6 means the sequences are close matches.

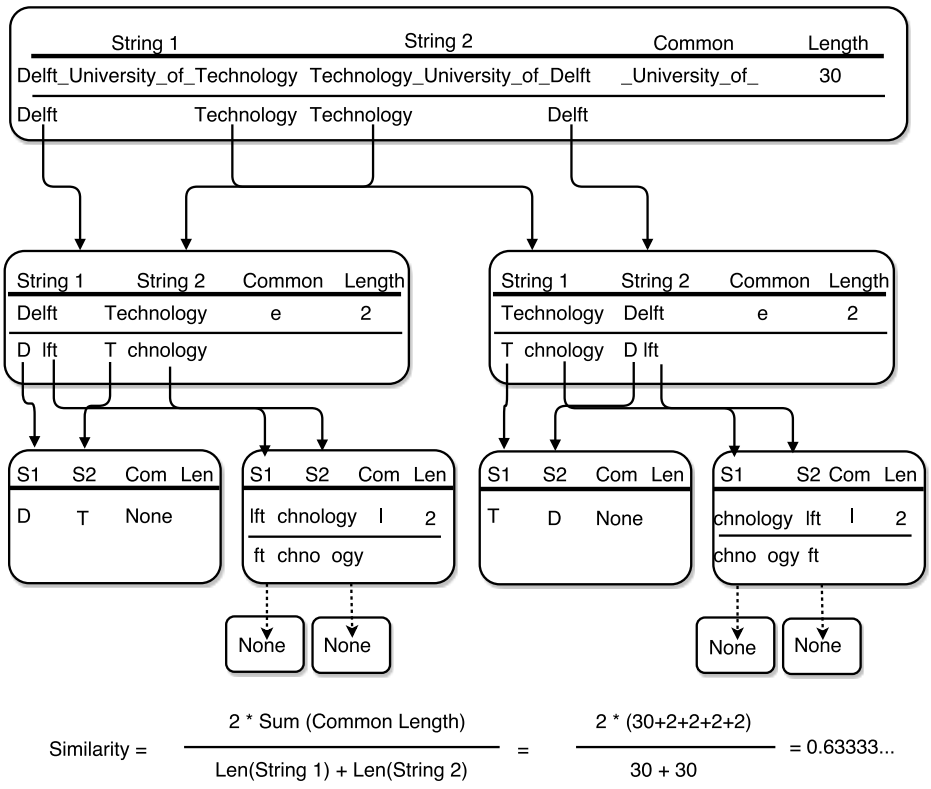

Figure 5.1: An Example of Ratcliff/Obershelp Sequence Match

For each of the three detection methods, around 20% of potential CAMEO users identified by it hit the threshold, which means the detected Harvester Accounts of these potential users have high similar registered Full Names or Public Names with their corresponding Master Accounts. Besides, we find around 10% detected potential CAMEO users registered completely same Full names for the detected Harvester and Master Accounts. (Since Public username is used to distinguish learners in courses, it cannot be completely same for any two accounts on edX.)

We select one from the 10% detected potential CAMEO users who registered same Full name for Harvester and Master Account, and go deeper into the account profiles and submission logs of his/her Harvester and Master Account. On the basis of the information, we find the following evidence.

The potential CAMEO user is detected by all detection methods in FP101x (2014).

- The Harvester and the Master Account have **same registered Full Name**.
- The **registered Email addresses** of the Harvester and Master Account contain **common long meaningless sequence (8 characters)**.
- The Harvester Account and Master Account utilize **same IP address** for answering **every** question in the course.
- The Harvester and Master Account submit answers within **a same minute** for **every** question, and the Harvester **always** submits **before** the Master Account.
- The Harvester Account submits answers for **all** question in the course, but the correctness of the answers is **only 11.5%**.

According to this evidence and our common sense, we can assert with confidence that the "potential" CAMEO user is indeed a CAMEO cheater. The certain CAMEO user is identified by all of the three detection methods in the experiment, which illustrates our detection results can possibly be true.

5.4 Characteristics of Detected CAMEO Users

To understand potential CAMEO users cheating behaviors and to then to provide suitable measures to prevent CAMEO, we analyze the characteristics of detected potential CAMEO users. The analysis in this section respectively answers (1) what is the proportion of potential CAMEO users among verified/honor learners; (2) what is the maximum/minimum number of questions a potential CAMEO user cheats on; (3) whether potential CAMEO users can get a certificate without CAMEO; and (4) where are the most potential CAMEO users from.

For the question about what is the proportion of potential CAMEO users among verified/honor learners, we have a hypothesis that paying learners (verified learners) should study harder than honor learners, which indicates that the proportion of CAMEO users should be less among verified learners than honor learners. According to the detection results, we plot Figure 5.2 showing the proportion of detected potential CAMEO users among verified/honor learners.

According to Figure 5.2, all of the results of different detection methods indicate that the percent of suspicious CAMEO users among honor learners is slightly higher the percent among verified learners, which is consistent with our hypothesis that compared with honor learners, paying learners are more honest in studying MOOCs.

We also do some statistics for the regions of the detected potential CAMEO users based on their profiles. The distribution of regions from which these users come partly reveals the prevalence of CAMEO in different countries. There are 9169 certified learners in the 10 DelftX MOOCs. They are mainly from United States (1426), Netherlands (646), and United Kingdom (488). However, according to the statistics in Table 5.3, all detection methods show that India has the largest number of potential CAMEO users in MOOCs created by TU Delft.

Table 5.3: Countries with the Most Detected CAMEO Users

Detection Method	Singleton	Hybrid	Long Batch	Long Batch (Variant)
1st	India (27)	India (42)	India (29)	India (70)
2nd	United States (12)	United States (19)	United States (11)	United States (22)
3rd	Germany (7)	Germany (10)	Colombia (7)	Netherlands (11)

Figure 5.2: Proportion of CAMEO Users Among Verified/Honor Learners

We discuss the question about how many questions a detected potential CAMEO user cheats on to check how frequently does a suspicious CAMEO user utilize CAMEO. Table 5.4 shows the minimum and the maximum number of questions a detected user applies CAMEO on and the proportion of these questions to the number of total questions in the MOOC.

As we can see from Table 5.4, the frequency of the use of CAMEO is diverse from user to user. The most serious case is that there is a potential CAMEO user identified by Long Batch Detection Method (Code Version) in MOOC RI101x. Among the 77 questions in the course, the detected potential CAMEO user is supposed to answer almost all (76) questions by utilizing CAMEO. Meanwhile, there are some other suspicious CAMEO users such as a user identified by Hybrid Detection Method in MOOC Calc001x cheat on only 3% questions in the course.

We also do an analysis about how many detected potential CAMEO users can earn a certificate without CAMEO. For the question, we have a hypothesis that most CAMEO users do not have the knowledge and ability to pass the MOOC, thus they seek the help of CAMEO. In the analysis, whether a suspicious CAMEO user without CAMEO can pass or not depends on if he/she can get a passing grade on the condition that all the credits for questions he/she cheats on are abolished. We illustrate the analysis based on the detection result of Hybrid Detection Method in Figure 5.3.

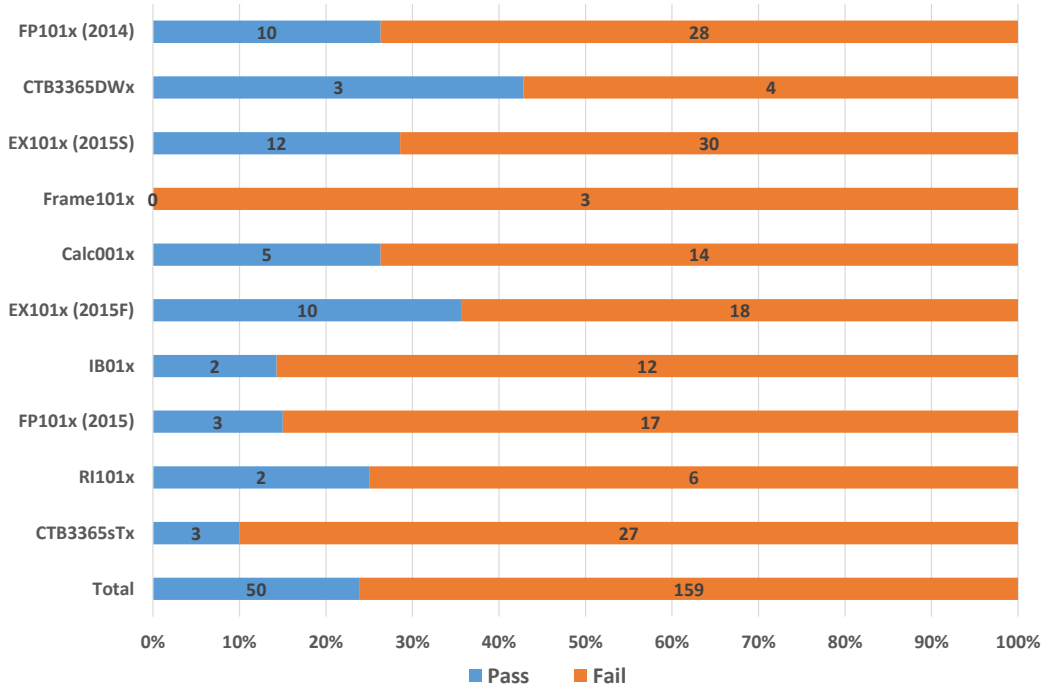

Figure 5.3: Number of CAMEO Users Can/Cannot Pass Without CAMEO (Based on the Detection Result of Hybrid Detection Method)

As shown in Figure 5.3, consistent with our hypothesis, nearly three-fourths of detected potential CAMEO users cheat to pass the MOOC, while a few of potential CAMEO users have the ability to earn a certificate and cheat maybe for a higher grade.

Table 5.4: Minimum/Maximum Number of Questions A CAMEO User Cheat On and Proportion of the Cheated Questions to Total Questions in the MOOC

Method / MOOC	Singleton	Hybrid	Long Batch	Long Batch (Variant)
FP101x (2014)	70 / 269 (24.31% / 93.40%)	13 /239 (4.51% / 82.99%)	14 /255 (4.86% / 88.54%)	12 /282 (4.17% / 97.92%)
CTB3365DWx	23 / 41 (9.06% /16.14%)	15 / 41 (5.91% / 16.14%)	17 / 41 (6.69% / 16.14%)	13 / 125 (5.12% / 49.21%)
EX101x (2015S)	15 / 130 (11.03% / 95.59%)	13 / 131 (9.56% / 96.32%)	13 / 52 (9.56% / 38.24%)	12 / 130 (8.82% / 95.59%)
Frame101x	0 / 0 (0 / 0)	14 / 19 (53.85% / 73.08%)	14 / 19 (53.85% / 73.08%)	14 / 25 (53.85% / 96.15%)
Calc001x	85 / 297 (15.04% / 52.57%)	19 / 437 (**3.36%** / 77.35%)	21 / 439 (3.72% / 77.70%)	21 / 467 (3.72% / 82.65%)
EX101x (2015F)	13 / 106 (8.90% / 72.60%)	15 / 106 (10.27% / 72.60%)	15 / 106 (10.27% / 72.60%)	11 / 107 (7.53% / 73.29%)
IB01x	108 / 325 (19.60% / 58.98%)	111 / 390 (20.15% / 70.78%)	72 / 324 (13.07% / 58.80%)	62 / 325 (11.25% / 58.98%)
FP101x (2015)	27 / 244 (9.38% / 84.72%)	27 / 275 (9.38% / 95.49%)	27 / 274 (9.38% / 95.14%)	16 / 274 (5.56% / 95.14%)
RI101x	18 / 53 (23.38% / 68.83%)	24 / 53 (31.17% / 68.83%)	24 / 53 (31.17% / 68.83%)	10 / 76 (12.99% / **98.70%**)
CTB3365sTx	25 / 156 (9.19% / 57.35%)	13 /159 (4.78% / 58.46%)	11 / 159 (4.04% / 58.46%)	14 / 159 (5.15% / 58.46%)

Another analysis we make is about the timing when CAMEO users are most likely to cheat during a MOOC. For the question, we have a hypothesis that CAMEO users tend to cheat at the midterm of a course.

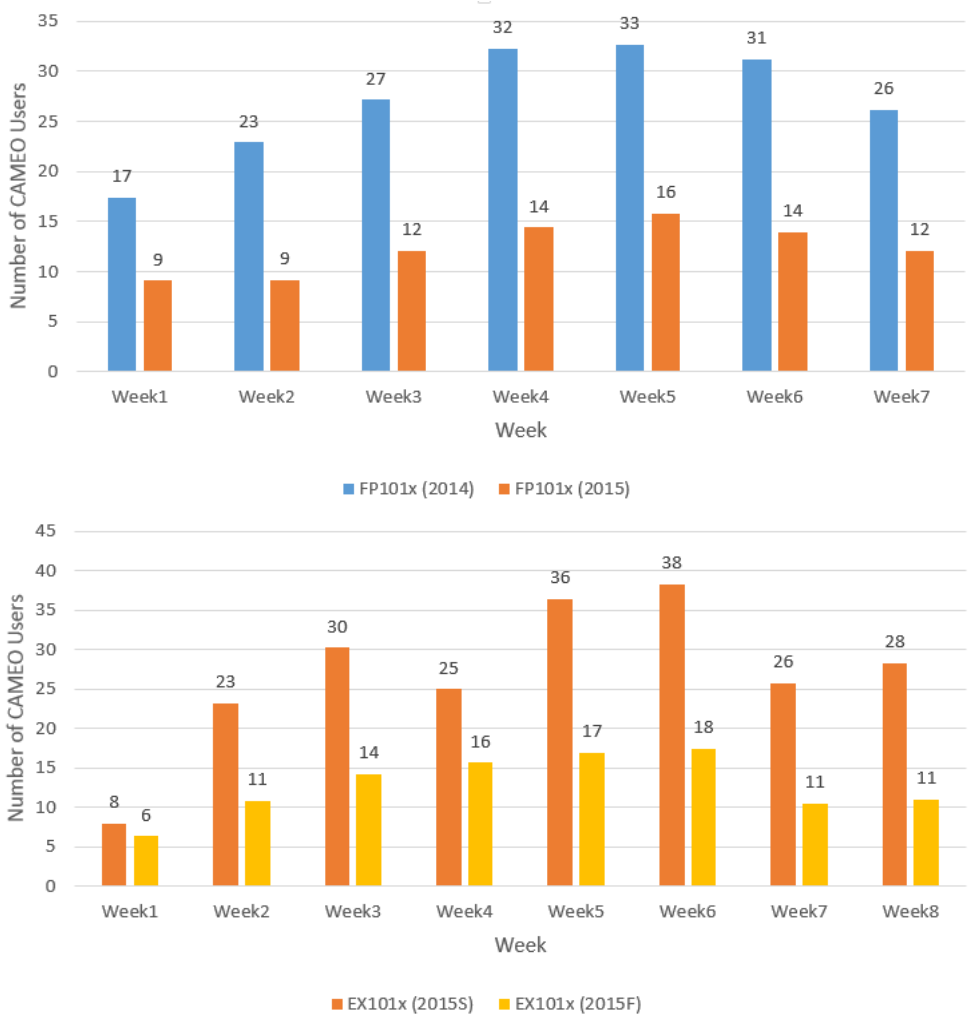

Figure 5.4: Average Number of CAMEO Users Cheating on per Question in Each Week (Based on the Detection Result of Hybrid Detection Method)

We have two reasons for the hypothesis. Firstly, as we described in Chapter 3, a MOOC on edX is composed of several sections (study weeks). During a course, the questions in latter study weeks are usually more difficult compared with the questions in the early weeks [36]. CAMEO users are supposed to cheat on more difficult questions. Secondly, learners on edX are able to check their study progress in a MOOC at any time they want. In other words, learners can instantly know if they earn satisfied credits to pass a MOOC. We suppose that CAMEO users tend to stop answering questions after they realize they could pass the MOOC. Considering the low-difficulty of questions in early weeks and the possible poor answering rate in the last few weeks, we suppose that CAMEO users are most likely to cheat at the midterm of a MOOC.

We calculate how many detected potential CAMEO users cheated on a question in average in every study week for 4 out of the 10 MOOCs and illustrate the analysis based in Figure 5.4. (The analysis is based on the detection result of Hybrid Detection Method.) According to Figure 5.4, we find that compared with other weeks (Week 1, 2, 3, 7, 8), more potential CAMEO users cheat in Week 4, 5, 6, which is the midterm of the MOOCs, which confirms our hypothesis.

Chapter 6

Conclusion

Massive Open Online Course (MOOC) as a promising education form has attracted lots of attentions from institutions, learners[1] and employers [57, 68]. Similar to other education forms such as campus-based learning, MOOC is also plagued by the situation that academic cheating is prevalent[2]. The influence of MOOC is undermined by several cheating strategies such as plagiarism[2]. Among the multiple cheating strategies, we focus on Copying Answers using Multiple Existence Online (CAMEO) in the thesis because of its three peculiarities, independence, high-efficiency and scalability. Targeting at the specified cheating strategies, we want to investigate the prevalence of CAMEO in MOOCs created by our home university, Delft University of Technology (TU Delft), and analysis the features of CAMEO users.

By replicating existing detection methods and implementing a new method in 10 MOOCs created by TU Delft, we answer our first research question about the prevalence of CAMEO. Since there is no ground truth in the experiment to prove the precision of the detection results, and all detection methods adopt quite strict filter criteria, the results are used to estimate the lower bound prevalence for CAMEO in the 10 MOOCs. Firstly, in each of the 10 courses, there are CAMEO users detected, which indicates that CAMEO really exists in the MOOCs created by TU Delft. Secondly, though the number of detected potential CAMEO users varies from methods to methods, generally, an estimated 2% of certified learners in the 10 MOOCs earned their MOOC certificates with the help of CAMEO cheating strategy. The percentage is slightly higher than the estimated proportion of certified CAMEO users (1.3%) in 115 MOOCs created by MIT and Harvard University.

To understand the cheating behaviors of these CAMEO users, and to provide suitable preventions to institutions and MOOC platforms, we analyze the features of the detected potential CAMEO users and answer our second research question. According to the analysis, we find that 1) compared with paying (verified) learners, honor learners are more inclined to cheat with CAMEO; 2) the number of questions potential CAMEO users cheat on diverse from user to user; 3) most potential CAMEO users cheat on multiple questions for getting their MOOC certificates; 4) in the 10 MOOCs created by TU Delft, India has the largest number of detected potential CAMEO users.

[1] http://www.class-central.com/report/moocs-2015-stats/

[2] nation.time.com/2012/11/19/mooc-brigade-can-online-courses-keep-students-from-cheating/

In conclusion, in this thesis, we complete a study of CAMEO in MOOCs created by TU Delft by replicating all the existing detection methods (to our knowledge), designing and implementing a new detection method in 10 DelftX MOOCs. The research makes contributions to understanding the popularity of cheating especially CAMEO in MOOCs and getting the knowledge of cheaters' behaviors preferences in MOOCs.

Besides, targeting at CAMEO, on the basis of the thesis, we want to give some suggestions to MOOC platforms and instructors who create MOOCs. Firstly, we would like to recommend problem randomization/randomization. They are two different functions which have been fully supported by edX. Problem randomization enables instructors to provide different learners with different questions, while randomization allows the learning management systems underlying edX to generate random numeric values within a question. Both of these two functions make it difficult to share solutions across accounts in a course. Secondly, we would like to ask instructors to adjust the timing about when the "Show Answer" button is visible to learners. For some instructor-paced MOOCs, we would like to suggest instructors let the "Show Answer" button appears after the past due of the question to cut off the path for CAMEO users to access instructors-prepared answers.

Bibliography

[1] Stephen Ackroyd. *Data collection in context*. Longman Group United Kingdom, 1992.

[2] Bryan Alexander. Connectivism course draws night, or behold the mooc. *Infocult: Uncanny Informatics*, 2008.

[3] John S Baird. Current trends in college cheating. *Psychology in the Schools*, 17(4):515–522, 1980.

[4] DC Baldwin Jr, Steven R Daugherty, Beverley D Rowley, and MD Schwarz. Cheating in medical school: a survey of second-year students at 31 schools. *Academic Medicine*, 71(3):267–73, 1996.

[5] Albert Bandura. Social foundations of thought and action. *The health psychology reader*, pages 94–106, 2002.

[6] T Bates. Comparing xmoocs and cmoocs. philosophy and practice. *Online Learning and Distance Education Resources*, 2014.

[7] Harriet F Bergmann. " the silent university": The society to encourage studies at home, 1873-1897. *The New England Quarterly*, 74(3):447–477, 2001.

[8] Paul E Black. *Dictionary of algorithms and data structures*. National Institute of Standards and Technology, 2004.

[9] Apiwan D Born. How to reduce plagiarism. *Journal of Information Systems Education*, 14(3):223, 2003.

[10] Mathieu Bouville. Why is cheating wrong? *Studies in Philosophy and Education*, 29(1):67–76, 2010.

[11] Beverly L Bower and Kimberly P Hardy. From correspondence to cyberspace: Changes and challenges in distance education. *New Directions for Community Colleges*, 2004(128):5–12, 2004.

[12] William J Bowers. Student dishonesty and its control in college. 1964.

[13] Kenneth J Chapman, Richard Davis, Daniel Toy, and Lauren Wright. Academic integrity in the business school environment: I'll get by with a little help from my friends. *Journal of Marketing Education*, 26(3):236–249, 2004.

[14] Gregory J Cizek. *Cheating on tests: How to do it, detect it, and prevent it*. Routledge, 1999.

[15] Sally Cole and Elizabeth Kiss. What can we do about student cheating. *About Campus*, 5(2):5–12, 2000.

[16] Sally Cole and Donald L McCabe. Issues in academic integrity. *New Directions for Student Services*, 1996(73):67–77, 1996.

[17] Harris Cooper. Synthesis of research on homework. *Educational leadership*, 47(3):85–91, 1989.

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

[19] Stephen F Davis, Patrick F Drinan, and Tricia Bertram Gallant. *Cheating in school: What we know and what we can do*. John Wiley & Sons, 2011.

[20] John Dichtl. Teaching integrity. *The History Teacher*, 36(3):367–373, 2003.

[21] Charles A Drake. Why students cheat. *The Journal of Higher Education*, 12(8):418–420, 1941.

[22] Gbolahan Gbadamosi. Academic ethics: What has morality, culture and administration got to do with its measurement? *Management Decision*, 42(9):1145–1161, 2004.

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

[24] Melody A Graham et al. Cheating at small colleges: An examination of student and faculty attitudes and behaviors. *Journal of College Student Development*, 35(4):255–60, 1994.

[25] Sharron M Graves. Student cheating habits: A predictor of workplace deviance. *Journal of Diversity Management (JDM)*, 3(1):15–22, 2011.

[26] Aleza Spalter Greene and Leonard Saxe. Everybody (else) does it: Academic cheating. 1992.

[27] Paul W Grimes. Dishonesty in academics and business: A cross-cultural evaluation of student attitudes. *Journal of Business Ethics*, 49(3):273–290, 2004.

[28] Valerie J Haines, George M Diekhoff, Emily E LaBeff, and Robert E Clark. College cheating: Immaturity, lack of commitment, and the neutralizing attitude. *Research in Higher education*, 25(4):342–354, 1986.

[29] John Harp and Philip Taietz. Academic integrity and social structure: A study of cheating among college students. *Social Problems*, 13(4):365–373, 1966.

[30] Börje Holmberg, Hrsg. Bernath, and Friedrich W Busch. *The evolution, principles and practices of distance education*, volume 11. Bis, 2005.

[31] Ivan Illich. Deschooling society. *New York*, 56, 1971.

[32] Rajesh Iyer and Jacqueline K Eastman. Academic dishonesty: Are business students different from other college students? *Journal of Education for Business*, 82(2):101–110, 2006.

[33] Sussan Johnson, Melissa Martin, et al. Academic dishonesty: A new twist to an old problem. *Athletic Therapy Today*, 10(4):48–50, 2005.

[34] Dorothy LR Jones. Academic dishonesty: Are more students cheating? *Business Communication Quarterly*, 74(2):141, 2011.

[35] Eleonora Karassavidou and Niki Glaveli. Towards the ethical or the unethical side? an explorative research of greek business students' attitudes. *International Journal of Educational Management*, 20(5):348–364, 2006.

[36] Lee Kern and Nathan H Clemens. Antecedent strategies to promote appropriate classroom behavior. *Psychology in the Schools*, 44(1):65–75, 2007.

[37] Erich Leo Lehmann and George Casella. Theory of point estimation (springer texts in statistics). 1998.

[38] Chun-Hua Susan Lin and Ling-Yu Melody Wen. Academic dishonesty in higher education - a nationwide study in taiwan. *Higher Education*, 54(1):85–97, 2007.

[39] Frank M LoSchiavo and Mark A Shatz. The impact of an honor code on cheating in online courses. *Journal of Online Learning and Teaching*, 7(2):179, 2011.

[40] Helen Marsden, Marie Carroll, and James T Neill. Who cheats at university? a self-report study of dishonest academic behaviours in a sample of australian university students. *Australian Journal of Psychology*, 57(1):1–10, 2005.

[41] David F Mastin, Jennifer Peszka, and Deborah R Lilly. Online academic integrity. *Teaching of Psychology*, 36(3):174–178, 2009.

[42] Nina Mazar, On Amir, and Dan Ariely. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of marketing research*, 45(6):633–644, 2008.

[43] D McCabe. Levels of cheating and plagiarism remain high. *Center for Academic Integrity, Duke University*, 2005.

[44] Donald L McCabe. Cheating among college and university students: A north american perspective. *International Journal for Educational Integrity*, 1(1), 2005.

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

[46] Donald L McCabe and Linda Klebe Trevino. Academic dishonesty: Honor codes and other contextual influences. *Journal of higher education*, pages 522–538, 1993.

[47] Donald L McCabe and Linda Klebe Trevino. Individual and contextual influences on academic dishonesty: A multicampus investigation. *Research in higher education*, 38(3):379–396, 1997.

[48] Jeff Meade. Cheating: Is academic dishonesty par for the course?. *Prism*, 1(7):30–32, 1992.

[49] Aythami Morales and Julian Fierrez. Keystroke biometrics for student authentication: A case study. In *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education*, pages 337–337. ACM, 2015.

[50] Maxim Mozgovoy, Vitaly Tusov, and Vitaly Klyuev. The use of machine semantic analysis in plagiarism detection. In *Proceedings of the 9th International Conference on Humans and Computers*, pages 72–77. Citeseer, 2006.

[51] Curtis G Northcutt, Andrew D Ho, and Isaac L Chuang. Detecting and preventing" multiple-account" cheating in massive open online courses. *arXiv preprint arXiv:1508.05699*, 2015.

[52] Chris Park. In other (people's) words: Plagiarism by university students–literature and lessons. *Assessment & evaluation in higher education*, 28(5):471–488, 2003.

[53] Kenneth C Petress. Academic dishonesty: A plague on our profession. *Education*, 123(3):624, 2003.

[54] Karl Popper. *The logic of scientific discovery*. Routledge, 2005.

[55] Patrick S Portway and Carla Lane. *Technical guide to teleconferencing and distance learning*. Applied Business Telecommunications, 1992.

[56] Alexandria Walton Radford. Learning at a distance: Undergraduate enrollment in distance education courses and degree programs. stats in brief. nces 2012-154. *National Center for Education Statistics*, 2011.

[57] Alexandria Walton Radford, Jessica Robles, Stacey Cataylo, Laura Horn, Jessica Thornton, and Keith E Whitfield. The employer potential of moocs: A mixed-methods study of human resource professionals' thinking on moocs. *The International Review of Research in Open and Distributed Learning*, 15(5), 2014.

[58] Erin Robinson, Rita Amburgey, Eric Swank, and Cynthia Faulkner. Test cheating in a rural college: Studying the importance of individual and situational factors. *College Student Journal*, 38(3):380, 2004.

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

[60] Randi L Sims. The relationship between academic dishonesty and unethical business practices. *Journal of Education for Business*, 68(4):207–211, 1993.

[61] Jason M Stephens and Hunter Gehlbach. Under pressure and underengaged: Motivational profiles and academic cheating in high school. *Psychology of academic cheating*, pages 107–139, 2007.

[62] Kristina L Verco and Michael J Wise. Software for detecting suspected plagiarism: Comparing structure and attribute-counting systems. *ACSE*, 96, 1996.

[63] George Watson and James Sottile. Cheating in the digital age: Do students cheat more in online courses? *Online Journal of Distance Learning Administration*, 13(1), 2010.

[64] Dan Wueste. Unintended consequences and responsibility. *Teaching Ethics*, 9(1):13–24, 2008.

[65] Robert J Youmans. Does the adoption of plagiarism-detection software in higher education reduce plagiarism? *Studies in Higher Education*, 36(7):749–761, 2011.

[66] Jeffrey Young. Dozens of plagiarism incidents are reported in coursera's free online courses. *The Chronicle of Higher Education*, 16, 2012.

[67] Jeffrey R Young. The cat-and-mouse game of plagiarism detection. *Chronicle of Higher Education*, 47(43), 2001.

[68] Jeffrey R Young. Coursera adds honor-code prompt in response to reports of plagiarism. *The Chronicle of Higher Education*, 24, 2012.

[69] Jeffrey R Young. Providers of free mooc's now charge employers for access to student data. *Chronicle of Higher Education*, 2012.

[70] Susan Zvacek, Maria Teresa Restivo, James Uhomoibhi, and Markus Helfert. *Computer Supported Education: 6th International Conference, CSEDU 2014, Barcelona, Spain, April 1-3, 2014, Revised Selected Papers*, volume 510. Springer, 2015.