Detecting Fraudulent Student Communication in a Multiple Choice Online Test Environment

Mariana Carrasco, António Rito Silva, and Rui Henriques

***Abstract*— Online evaluation systems, pervasive nowadays, are known to be susceptible to higher fraud risks. This work proposes a novel and robust method to detect potential fraud acts in online multiple-choice question (MCQ) exams. For the first time, the communication probability between the examinees is statistically assessed based on the concordance of responses and answer time against null expectations and is subsequently used to identify potential fraud behavior. The model is sensitive to the direction of communication acts, distinguishing content consumption from production, as well as multiwise communication channels. Online remote tests from engineering courses at Técnico Lisboa are used as a case study. We show that the cumulative contribution of concordant responses between students, when recurrent, offers a way of signaling fraud behavior. Separating content production from consumption reveals the underlying student role played in potential fraud acts. Collusion behavior is assessed against null models of fraud and conformity, and therefore being statistically framed and offering a solid criterion to guide tutors in ascertaining fraud and discouraging communication.**

***Index Terms*— Communication network, fraud detection, multiple choice quiz, online remote evaluation, statistical significance.**

I. Introduction

OVER the last decade, alongside the developments in technology [1], [2], came, for students, the possibility to enroll in a wide variety of online courses and, in some colleges, to choose between the traditional face-to-face classes and the computer-based classes. Online courses attained their popularity by providing students the flexibility to work in a self-paced manner and reduce attendance costs [3], [4], [5]. University administrators are motivated to present online content and assessments to ensure a broader student reach [6]. The COVID-19 pandemic converted this possibility into a necessity [7], [8], [9]. However, with the remote way of teaching comes the challenge of unsupervised online testing, shown to yield a higher possibility of fraud [9], [10], [11], challenging the fair principle of evaluation [12].

We define fraud as all forms of illegitimate activities that are aimed at increasing one's assessment performance. These activities include using unauthorized materials, copying, collusion among examinees, acquisition of test contents (also termed preknowledge), impersonation [13], or external assistance from someone who is not taking the test [9]. In this study, the focus is on collusion among examinees, the arguably most common form of online cheating in multiple choice question (MCQ) exams performed at individual homes [14].

Several statistics have been proposed to assess collusion [15], [16], [17], [18], [19], [20], ranging from item response theory to the analysis of response times. Nevertheless, the existing methods generally suffer from three major drawbacks:

1) Assume fixed question orders and reversible answering;
2) Neglect the distinguished roles and multiwise cumulative effects from inadvertent content sharing exerted in unauthorized communication platforms;
3) Do not reliably test the deviation of the acquired behavioral statistics against plausible expectations.

As the first comprehensive effort to address these limitations, this work establishes a novel method to detect potential fraud acts based on timestamped answer records from online quizzes with shuffled questions, capturing potential multiwise acts of information exchange by examinees taking the test at the same time. In the context of multiple choice online test environments, this is an increasing need as fraud can be attained by either direct in-room communication or via instant messaging applications, for instance, *Whatsapp* and *Messenger*, as electronic communication is becoming more pervasive worldwide with the spread of the Internet [21]. As such, handling collusion, irrespective of the communication method, is the pivotal requirement tackled in this study.

In this context, the following major questions arise: Is it possible to identify collusion fraud taking into consideration both the selected options and their timestamps? How can collusion candidates be statistically tested to minimize false discoveries? Can we further inquire into the nature of inadvertent communication between students, including its directionality (inadvertent content sharing and/or consumption) and cardinality (number of involved students)? This work offers a comprehensive discussion of these research questions.

To this end, we propose a disruptive methodology to assess fraud communication which starts from a preanalysis of the data to accommodate distinct patterns of fraudulent behavior. The methodology sustains itself in the following four major principles:

1) a statistical frame to assess the probability of pairwise student communication considering: a) matched answers; b) choice probability; c) response times (directionality); and d) recurrence of suspicious behavior;

Manuscript received 15 November 2022; revised 28 January 2023; accepted 27 February 2023. Date of publication 20 March 2023; date of current version 31 January 2024. This work was supported by the Fundação para a Ciência e Tecnologia (FCT) through the projects DACOMICO under Grant PTDC/CCI-COM/2156/2021, LAIfeBlood under Grant DSAIPA/AI/0033/2019, ILU under Grant DSAIPA/DS/0111/2018, and INESC-ID pluriannual under Grant UIDB/50021/2020. *(Corresponding author: Rui Henriques.)*

The authors are with the INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, 1000-029 Lisbon, Portugal (e-mail: rmch@tecnico.ulisboa.pt).

Digital Object Identifier 10.1109/TCSS.2023.3254504

2) a network representation of potential communication acts grounded on the previous probabilistic stance, allowing the assessment of directional multiwise communication acts in a multiple choice online test;
3) null models of compliance and fraud from the principled understanding of inadvertent communication to test collusion dynamics and identify them with strict guarantees of statistical significance;
4) scoring, clustering, and visualization principles to facilitate the understanding of inadvertent communication pathways and promote the actionability of recommendations, supporting the course's tutor with subsequent inquiry acts and advertence initiatives.

As a case study, we consider online remote tests performed on the *Quizzes Tutor*'s platform, developed at Técnico Lisboa. Students receive the questions in different orders (shuffling) and cannot return to a question they have already answered or skipped. There is limited monitoring capacity of the students' behavior. Periodic quizzes from software architecture (SA) course are used to validate the proposal.

The remainder work structure is organized as follows. Section II introduces essential background. Section III discusses relevant work. Section IV introduces the target fraud detection model. In Section V, null models of student behavior are specified to assess differences between compliant and fraudulent behavior. Section VI proposes a methodology to detect potential fraud acts against behavioral expectations and find multiwise communication channels. Section VII discusses the acquired results, comparing the target model against existing scores. Deployment notes are discussed in Section VIII. Concluding marks are finally provided.

II. Problem Formulation

The target notation is now presented. Consider a course edition to be described by the attendees, a set of n students, $S = \{s_1, \ldots, s_n\}$, and a set of questions $Q = \{q_1, \ldots, q_g\}$. A question set Q is lightly used to either represent a single quiz (default), the questions from a set of quizzes, or, when accommodating confounding aspects, such as optional questions and examination shifts, the set of shared questions for a given group of students.

Let Σ be the set of all the available options for the questions contained in Q and, in particular, Σ_k be the set of available options for question $q_k \in Q$. In the context of our work, one response r_{ik} is a pair that contains an ordered set of selections in Σ_k, $\mathbf{y}_{ik}$, performed by student s_i to question q_k, and an ordered set of timestamps, $\mathbf{t}_{ik}$, which are monotonically increasing given that we are targeting evaluation settings where it is prohibited to return to a previous question. The order of the first set conforms to the order of the later.

Let, in addition, $\mathbf{r}_i$ be the set of responses r_{ik} from student s_i to the question set Q. Finally, let x_{ik} be the last answer student $s_i \in S$ gave to question q_k, f_{ik} be its timestamp, $\mathbf{x}_i$ be the sequence of all final answers to Q, and $\mathbf{f}_i$ be the timestamps corresponding to the final answers.

Let $\tau : S \times Q \rightarrow \{0, 1\}$ be the scoring of a question. If a response is scored as correct, $\tau(x_{ik}) = 1$; otherwise, $\tau(x_{ik}) = 0$.

TABLE I
Notation

symbol	description
$S = \{s_1, ..., s_n\}$	set of students
$Q = \{q_1, ..., q_g\}$	set of questions
Σ and Σ_k	set of options for questions in Q and $q_k \in Q$, respectively
$\mathbf{y}_{ik}$	ordered set of all selections in Σ_k by student s_i to question q_k
$\mathbf{t}_{ik}$	ordered set of increasing timestamps of selections in Σ_k by s_i
$r_{ik} = (\mathbf{y}_{ik}, \mathbf{t}_{ik})$	response pair for a student s_i
$\mathbf{r}_i$	set of responses r_{ik} from student s_i to the question set Q
$\mathcal{R} = \{\mathbf{r}_i \mid i{=}1..n\}$	set of all student responses
x_{ik}	last answer student $s_i \in S$ gave to question q_k
f_{ik}	timestamp of the last answer student $s_i \in S$ gave to question q_k
$\mathbf{x}_i$	sequence of all final answers to Q by s_i
$\mathbf{f}_i$	sequence of timestamps f_{ik} of all final answers to Q by s_i
τ	$S{\times}Q \rightarrow \{0, 1\}$, scoring of a question (1: correct, 0: incorrect)
$a(\mathbf{x}_i)$	grade of a student s_i
p_k	probability of correctly answering question q_k
$answer$	$S \times Q \rightarrow \Sigma$, last answer x_{ik}
$time$	$S \times Q \rightarrow \mathbb{N}$, timestamp of last answer f_{ik}
$correct\_answer$	$Q \rightarrow \Sigma$, correct answer to a particular question
$sequence$	$S \times Q \rightarrow \mathbb{N}$, order of the given question for the given student
α_{ik}	weighting factor on the communication with student s_i in q_k
$M = \{p, c\}$	communication mode between students (production p, consumption c)
$weight$	$S{\times}S{\times}M{\times}Q \rightarrow [0, 1]$ amount of communication via mode in M between two students in S
$score$	$S{\times}M{\times}\mathbb{N}^+ \rightarrow [0, 1]$, fraud score

In the context of a given question set, Q, we can obtain the grade of a student s_i using, for instance, the sum of scores

$$a(\mathbf{x}_i) = \sum_{k=1}^{g} \tau(x_{ik}) \tag{1}$$

and the probability of correctly answering a question q_k in Q as an average of scores over the set of students S

$$p_k = \frac{1}{n} \sum_{i=1}^{n} \tau(x_{ik}). \tag{2}$$

Consider the input course data to be the set of all student responses, $\mathcal{R} = \{\mathbf{r}_i \mid i = 1, \ldots, n\}$, to the undertaken online quizzes. Given $\mathcal{R}$, the targeted problem is to identify and describe fraud behavior within and across quizzes. To this end, particular care is necessary to guarantee the *statistical significance* of the found associations, the *traceability* of the undertaken fraud behavior (together with the potentially involved students), and the *actionability* of recommendations.

III. Related Work

As precedent research shows, students cheat for numerous reasons, which are not strictly associated with online testing [22], [23]. These reasons may include low grades and ineffective study strategies; poor time management skills; personal values and views which relate to achievement, fear of punishment, class attendance, and peer pressure; extrinsic versus intrinsic motivations to learn; and age [10]. Ladyshewsky [10] observed that some student profiles will attempt to cheat regardless of the mode of instruction. Although earlier studies yield no conclusive evidence that remote online assessments increase cheating likelihood [24], [25], recent results show that there is a significant increase in dishonest behaviors in remote assessments [11], [23].

Several statistics, grounded on item response theory and the analysis of response times, have been proposed to detect unexpected gain scores, collusion, preknowledge, and other

abnormal behaviors [15], [26], including l_z^* [27], H^T [17], L_s [18], S [20], erasure index [19], and cumulative distribution indices [16], each with varying degrees of success [18], [28].

Ranger et al. [13] compared several cheating indicators and were unable to find indicators that could discriminate preknowledge from test collusion. Withal, the authors found out that indicators based on response times were capable to detect preknowledge but not test collusion, and indicators based on the response revisions were capable to detect test collusion but lack the power to detect preknowledge.

Of the various cheating indicators purposed in the recent Ranger et al. study [13], we highlight three statistics based on the selected responses—U1, U3, and CS—and four which are based on the editions/revisions to a given response—N1, NC1, N2, and NC2. The indicators based on the selected responses constitute the most basic way to analyze an examinee's performance since they solely require that, for each question, the option chosen by each student is saved, while the indicators based on the response's revisions are of particular interest as capable of detecting test collusion.

In our work, each student receives the questions that compose a quiz in distinct order, with shuffled possible options, and is further prevented from going back and editing responses to previous questions. As such, collusion statistics based on response revisions are insufficient. In addition, most of the previous statistics, including those in [13], do not consider the concordance of responses between students against chance agreements, which is a normal condition for communication attempts within the class. Furthermore, some of the existing indicators neglect the rich temporal frame at which responses are provided, preventing the possibility to assess the significance and directionality of potential copy acts between students.

In the context of online courses with long-duration assessments, Ruipérez-Valiente et al. [29] found that close submitters needed a statistically significant lower amount of activity in the platform to successfully complete a course. Results show that most of the student user accounts were grouped as couples of close submitters, with some large communities also observed. In a similar context, Balderas et al. [30] considered fraudulent collaboration involving an arbitrary number of students. Given students s_1 and s_2, the targeted forms of suspicious behavior include s_2 starting examination after s_1 submission and showing a better grade/completion time ratio. The sequential rules produced under the aforementioned principles were used to produce clusters of students involved in potential fraud acts. In spite of the relevance of these studies to find multiwise collaboration patterns, they focus on a specific single form of dishonest behavior observed in long-duration assessments.

Blockchain principles to separate malicious attacks from truthful events in online systems can be arguably considered for fraud detection purposes by considering the multiplicity of fraud statistics as voters. Considering online test environments, Cai et al. [31] propose a decision schema that tackles the problems of majority voting in the presence of dishonest voters (i.e., false-positive scores of fraud) by assigning awards when a voter's report is trusted according to a peer prediction scheme. The proposed scoring scheme is incentive compatible, with a maximum attained with honest reporting [31].

Comprehensive policy assessments undertaken by Bilen and Matros [32] conclude that capturing each student's computer screen and room is pivotal to decrease fraud intentions and further recommend avoiding grading on a curve to decrease cheating behaviors motivated by peer competition. Tiong and Lee [33] developed an e-cheating intelligence agent for online assessments that is further able to access the Internet Protocol (IP) of the students, issuing alerts when students changed their device or initial location. The agent is capable of preventive behaviors as it is dynamically able to reassign questions in instances where abnormal behavior is detected.

One of the challenges of working with statistical models of fraud is the inherent difficulty of identifying the cut-off values that separate normal from atypical response profiles. Man et al. [26] placed a supervised stance on fraud detection to tackle this challenge. Using predictive learning, the authors were able to compare the discriminative power of the statistics using the collected fraud evidence and further conclude that the use of predictive models able to combine multiple sources of information can lead to a higher detection rate over traditional item response and response time methods. Alexandron et al. [34] proposed a semisupervised anomaly detection approach, trained on a known set of cheaters, to detect fraud. The approach is shown to be capable of generalizing well toward cheaters with distinct behaviors. A new time-based statistic—the fraction of items that were solved correctly in significantly lower time than the average time of correct responses on those items—is proposed to assess aberrant behaviors.

Despite the relevance of the placed (semi)supervised stances, their transfer and deployment across different courses and cultures are arguably limited as it requires the presence of expressive forms of cheating behavior from different contexts. The presence of ground truth to develop and assess academic dishonesty is a well-recognized difficulty [15]. Man et al. [26] considered a case study where students had the opportunity to illegally steal exam content before assessment. Complementary cheating behavior during the exam was further flagged via postinvestigation clearance. To validate fraud detection in mixed face-in-face and online settings, Balderas et al. [30] considered the differential analysis of grades between settings to validate findings. Understandably, these assumptions disregard the fact that changes in academic performance can be undertaken with integrity and are further restricted to mixed evaluations in the context of a course or academic path. Bilen and Matros [32] and Tiong and Lee [33] validated fraud models by predefining cheating as the ability to quickly answer difficult questions. Although useful, these labeling assumptions are arguably biased and limited to specific forms of fraud and dependent on parameterizable cut-off thresholds that assume the homogeneity of student profiles.

IV. Fraud Detection Model

A sound statistic of collusion likelihood in online quizzes, able to integrate state-of-the-art stances on answer concordance and compatible response times, is now introduced. The intuition behind the proposed model is that the underlying communication acts between students are

generally associated with concordant answers, with the direction of communication—whether content production or content consumption—being generally reflected on the time of response associated with a concordant answer. Complementarily, the lower the concordance likelihood for a given question (unlikely item selection against overall selections), the higher the collusion likelihood. Finally, collusion likelihood further increases with recurring suspicious behavior with the same set of students within a quiz or along multiple quizzes. Ground on the aforementioned assumptions, the proposed fraud detection model measures the weight of direction-sensitive communication between students.

Let M define the communication mode between students, $M = \{p, c\}$, where p denotes content production and c content consumption. Given a set of students S and questions Q, weight: $S \times S \times M \times Q \to [0, 1]$ is a function that assesses the amount of communication between two students in S through the communication mode in M, in the context of a set of questions Q, possibly spanning one or multiple quizzes.

Note that weight is a total function, which means that it takes the value 0 when there is no communication of a particular type between two students. Given two students, $s_i, s_j \in S$, $\text{weight}(s_i, s_j, c, Q)$ denotes the amount of information that s_i consumed from s_j, and $\text{weight}(s_i, s_j, p, Q)$ denotes the amount of information that s_i shared with s_j. Note that

$$\text{weight}(s_i, s_j, c, Q) = \text{weight}(s_j, s_i, p, Q). \tag{3}$$

Since weight is a total function, it respects the restriction

$$\text{weight}(s, s, c, Q) = \text{weight}(s, s, p, Q) = 1 \quad \forall s \in S. \tag{4}$$

To inquiry about the mode of communication between two students, access to responses' time is necessary. Recovering the problem formulation (see Section II), the response of a student s_i to a question q_k, r_{ik}, is a tuple of containing the answer selections $\mathbf{y}_{ik}$ and their timestamps $\mathbf{t}_{ik}$. Similarly, the time, after the start of the quiz, of s_i's final attempt x_{ik} in a particular question q_k, defined as f_{ik}, is given by time: $S \times Q \to \mathbb{N}$.

Given the set of possible selections Σ, answer: $S \times Q \to \Sigma$ denotes the final element in the sequence of answers given by a student to a question (i.e., the response which is going to be taken into consideration when evaluating the test), previously defined as x_{ik}. The function correct_answer: $Q \to \Sigma$ presents the correct answer to a particular question.

As students can receive different sequences of questions for the same quiz, it is relevant to define their order, sequence: $S \times Q \to \mathbb{N}$, specifying the permutation of questions per student.

The timestamps of the answer selections $\mathbf{y}_{ib}$ for question q_b should be greater than the timestamps $\mathbf{y}_{ia}$ for question q_a if q_b comes after q_a in the sequence of questions assigned to student s_i in Q. In this context,

$$\begin{aligned}\forall_{s_i \in S} &\min(\text{time}(s_i, q_b)) > \max(\text{time}(s_i, q_a)) \\ &\text{iff } \text{sequence}(s_i, q_b) > \text{sequence}(s_i, q_a).\end{aligned} \tag{5}$$

The communication between two students, $s_i, s_j \in S$, for a question, $q_k \in Q$, where s_i is the producer, occurs when

$$\text{answer}(s_i, q_k) = \text{answer}(s_j, q_k) \wedge \text{time}(s_i, q_k) < \text{time}(s_j, q_k). \tag{6}$$

The functions $\text{share}, \text{consume}: S \times S \times Q \to \{0, 1\}$ determine whether or not there is a sharing or consumption between two students. The values of $\text{share}(s_i, s_j, q_k)$ and $\text{consume}(s_j, s_i, q_k)$ are characterized by condition 6, which also illustrates that consumption of information between two students is related with a sharing of information between the same students.

In this context, given a set Q with g questions, the estimated consumption weight between two students, $s_i, s_j \in S$ is

$$\text{weight}(s_i, s_j, c, Q) = \sum_{k=1}^{g} \alpha_{ik} \frac{\text{consume}(s_i, s_j, q_k)}{g} \tag{7}$$

where α_{ik} is a weighting factor, possibly dependent on the student profile s_i and question q_k. Similarly, the production weight is defined by

$$\text{weight}(s_i, s_j, p, Q) = \sum_{k=1}^{g} \alpha_{ik} \frac{\text{share}(s_i, s_j, q_k)}{g}. \tag{8}$$

By default, the weighting factor α_{ik} is defined using the frequency of the response of a selected option

$$\alpha_{ik} = 1 - \sum_{j=1}^{n} \frac{\mathbf{I}(\text{answer}(s_j, q_k) = \text{answer}(s_i, q_k))}{n} \tag{9}$$

where n is the number of students taking the exam, and $\mathbf{I}$ is the identity function, which returns 1 when the condition in parentheses yields True and 0 otherwise. Values of α_{ik} closer to one indicate that the selection is uncommon, implying higher weights when both students identically select a highly infrequent item.

It can be observed that the given definition fulfills the conditions stated about the weight. Additionally, the fraud score of a student is an estimator based on the weight of the most prominent communication channels held with other students. The fraud score is represented by the function score: $S \times M \times \mathbb{N}^+ \to [0, 1]$ that is defined in terms of the communication channels held between students. Consider $\beta \in \mathbb{N}^+$ to be the number of communication channels. Given $s_i \in S$ and $m \in M$

$$\begin{aligned}&\text{score}(s_i, m, \beta) \\ &= \sum_{k=1}^{\beta} \text{top}([\text{weight}(s_i, s_j, m, Q) : s_j \in S \ \wedge s_j \neq s_i], k)\end{aligned} \tag{10}$$

where *top* denotes the kth highest value in the multiset containing the mode m weights between s_i and all other students, excluding itself, and $0 < \beta < n$.

Variations to the fraud detection model correspond to different parameterizations of the weight function and values of β applied to the *top* function. For instance, by fixing (9) and $\beta = 3$, we are considering a fraud stance that considers the contributions from the three student communication channels

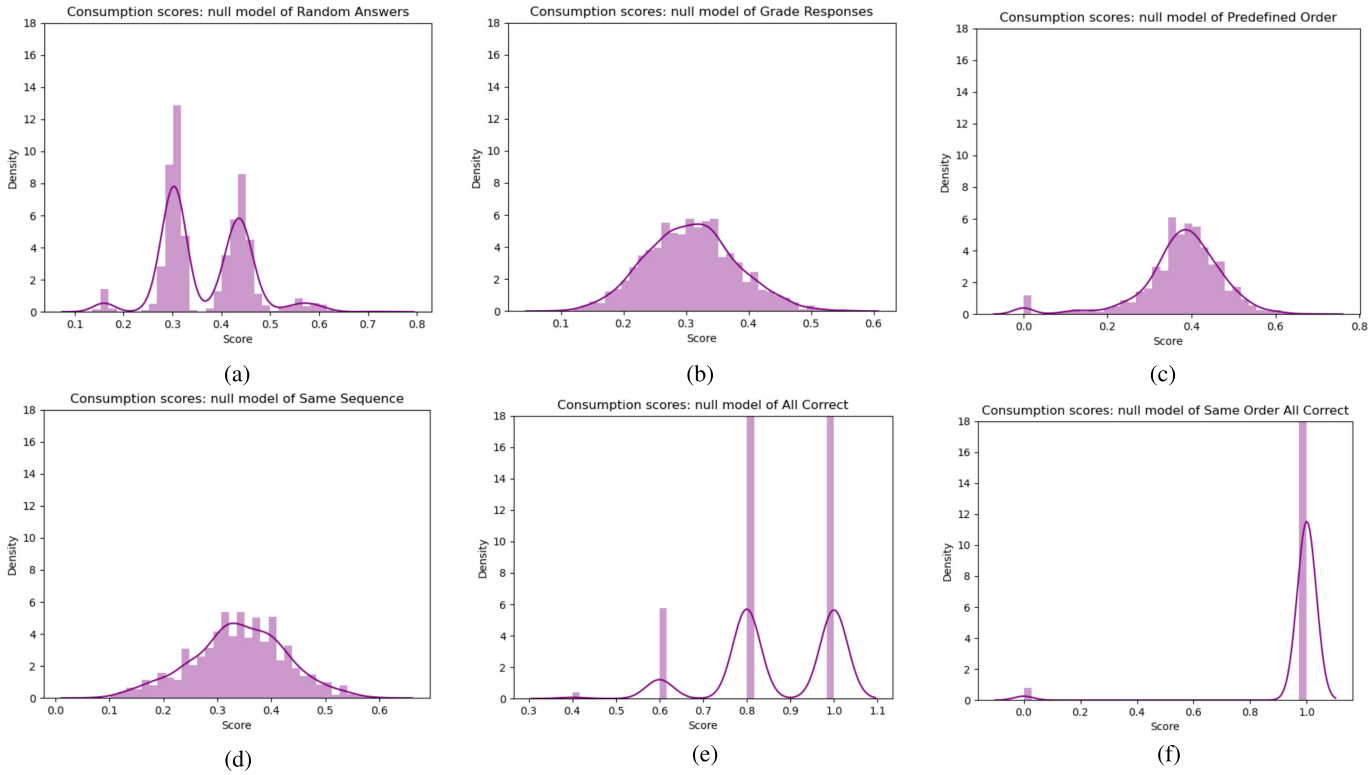


Fig. 1. Null distribution of consumption scores. (a) Students answer randomly. (b) Grade is the unique answer influence. (c) Students answer in a predefined order. (d) Students receive the same question sequence. (e) Students correctly answer all questions. (f) Students correctly answer questions in a predefined order.

with the highest production and consumption of contents. Fixing β considerably below n, $\beta \ll n$, is suggested to remove contributions from channels with residual weights produced by spuriously concordant responses.

V. Understanding Fraud With Null Models

The proposed fraud detection model in Section IV offers the possibility to quantify potential communication acts, weight them according to the likelihood of response selections, and further separate modes of communication considering response times. In the absence of fraud, communication scores can differ from zero due to the presence of spurious concordant responses. In this context, assessing expectations on the communication scores in the absence of fraud is of paramount importance to identify student-specific deviations that are potentially associated with fraudulent behavior.

To this end, this section introduces null models to understand how scores vary in the absence and presence of fraud. Accordingly, we propose null models of compliance (see Section V-A2) where students' answers are placed in the absence of communication acts, and null models of fraudulent behavior (see Section V-B) assuming explicit communication between the students within a group. The reasoning behind each null model is presented. For simplicity's sake, this analysis is pursued taking into consideration the final response each student gives to each question. For all the proposed null models, condition (5) introduced in Section IV holds.

Fig. 1(a)–(f) presents the distribution of consumption scores for each of the nonfraudulent null models. All results presented were obtained using 30 simulations and considering the regularities found in the set of questions of the quiz with id 14225, performed by SA students in *Quizzes Tutor*'s platform (details in Section VII-A). Here, the consumption score of a student is defined as the highest weight found in the set of consumption edges for that student.

The distribution of the production scores is omitted since it is similar to the distribution of consumption scores, verifying property (3).

A. Null Models of Nonfraudulent Behavior

1) Students Answer Randomly: To gather expectations on the weight of communication channels between nonfraudulent students, it is relevant to analyze a model where all students answer in a random fashion. The amount of information that exists in this situation helps us to define a threshold that establishes when there is explicit communication between students. For this model, there are no restrictions, every output of the functions *answer* and *time* is viable.

The generation of this model is trivial, a sequence of questions is randomly (uniform) generated for each student, with incremental timestamps, and then the students' selections to each question are also randomly (uniform) picked.

Fig. 1(a) describes this null model, showing density peaks and communication weights likely contained in [0.3, 0.45].

2) Students' Grades Influence Their Answers: In hopes of analyzing the weight of communication channels between nonfraudulent students whose performance on the test is dictated by their knowledge (grade), a null model in these conditions

was created. To generate this model, for each question, each student's performance (correct/incorrect answer) is determined by the course's mark. If the answer is incorrect, the selected option is randomly chosen given the probability of picking each incorrect option obtained from real data. However, if for some question in the real data, no student picked an incorrect answer, the chosen incorrect option in the null model is uniformly selected. The timestamps are randomly generated.

Fig. 1(b) shows that the distribution of the consumption scores in this null model follows a Gaussian with a mean communication weight of 0.3.

3) Students Answer in a Predefined Order: To study the impact of different assumptions, a third null model of nonfraud is considered where students answer questions in the same order, that is,

$$\forall_{s_i,s_j\in S}\forall_{q_k\in Q}\ \text{time}(s_i,q_k) < \text{time}(s_j,q_k) \quad \text{for } i<j \tag{11}$$

but the answer given is determined by each student's course mark, to assess the maximum weight of spurious communication between two students answering in predefined order.

To generate this model, a permutation of students is computed, timestamps are predetermined by this order, and the remaining parameters are according to the second null model.

In Fig. 1(c), one can see two peaks: at 0 (student which is always the first to answer) and around 0.4 (mean spurious communication weight).

4) Students Receive the Same Sequence of Questions: It is of particular interest to study a model where students receive the same sequence of questions, which can be formulated as

$$\forall_{s_i,s_j\in S}\forall_{q_k\in Q}\ \text{sequence}(s_i,q_k) = \text{sequence}(s_j,q_k). \tag{12}$$

The purpose lies in identifying cases where the weight of sharing from one student to another is more evident, not because one shared information with the other, but because one answered before the other and, by chance, their options coincided. For a model generation, the sequence of questions is primarily settled. As in the previous model, each student's performance is determined by their course's mark. The timestamps for each question are randomly selected.

In Fig. 1(d), showing the distribution of scores for this null model, one peak is evident, around 0.35, similar to what happens in Fig. 1(b), yet tails are now heavier.

5) Students Correctly Answer All Questions: A null model where students correctly answer all questions is also relevant to study weights under conformity and can be formulated as

$$\forall_{s_i\in S}\ \text{answer}(s_i,q_k) = \text{correct\_answer}(q_k). \tag{13}$$

The timestamps dictate the weight of communication between two students. With this approach, it is possible to highlight the maximum values of weight communication between students who answer in fluctuated orders. To generate the model, it is only necessary to assign the order of questions and the corresponding timestamps. For this approach, the α_{ik} factor, responsible to adjust a contribution in accordance with the probability of selecting a given option (9), is zero to prevent null consumption and production weights.

The distribution of scores in Fig. 1(e) identifies four peaks (0.4, 0.6, 0.8, and 1). Since every answer is correct, students agree on the chosen option in all questions. Hence, there is nonzero communication between every two students.

6) Students Answer in a Predefined Order and Correctly: This model combines the principles of the previous two conditions: conditions (11) and (13). Fig. 1(f) depicts the distribution of the acquired scores. Understandably, density peaks are associated with a 0 score (first one to answer) and 1 score (remaining, fully concordant case).

7) Final Remarks: Various null models representing nonfraudulent behavior were tested in pursuance of obtaining insights into our scoring methodology. Distinct patterns, supported by occasional arrangements of answers, promoted different score distributions. As expected, the null model in which the unique influence on students' answers is their grades [see Fig. 1(b)] is the one better resembling real dynamics of quiz answering in academic integrity. This distribution follows a Gaussian (Shapiro–Wilk at $\alpha = 1\text{E}{-3}$), yielding statistical properties of interest.

B. Null Model of Fraudulent Behavior

The existence of collusion implies that there is at least one student sharing information and one student receiving it. Therefore, in a fraudulent scenario, it is expected that students organize themselves in groups and can communicate with each other. Following this logic, we can define two roles, *leader* and *copycat*, which are not necessarily disjoint.

Collusion may occur in the context of pairwise communication between two students, as well as within larger student groups (multiwise communication channels) where the shared contents are accessible by a community.

In this context, a *leader* is someone who answers a question independently and shares that information with the group. The elements of the group which are not leaders, the *copycats*, may be in one of the two following situations when answering a question: the question has already been answered by a leader, so they can use the shared option, or the question has not yet been answered by a leader and they can decide on whether to wait until they receive the answer from a leader. This strategy is described by Krueger as picking "a 'sacrificial lamb' to take the online test first and bring back the questions to the group" [35].

Under different pressure conditions, collusion can be observed among knowledgeable peers [32]. In alternative forms of collusion, students with low performance can divide efforts in accessing and sharing external information within a single communication channel. In both the aforementioned scenarios, several leaders should be considered. In this context, we assume that if one of the leaders is about to answer a question that has already been answered by another leader, it chooses the option disclosed by the first leader.

More formally, let us define the partitioning P of the set of students S, such that $\cup_{P_i\in P}\ P_i = S \wedge P_i \cap P_j = \emptyset$ for $P_i, P_j \in P$. Each P_i is then further partitioned in two groups: P_{il} (set of *leaders* in group i) and P_{ic} (set of *copycats* in group i), yielding $P_{il} \cup P_{ic} = P_i \wedge P_{il} \cap P_{ic} = \emptyset$. Here, the groups of leaders and copycats are disjoint. Given a question $q_k \in Q$ and group P_i, let us assume that every element of the

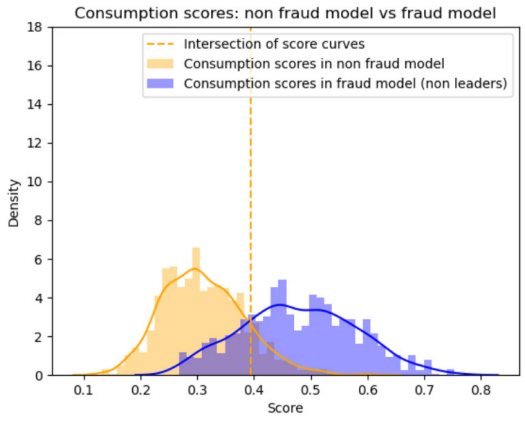


Fig. 2. Distribution of consumption scores: the null model of nonfraud versus null model of fraud with groups of 3 and one leader.

group will answer the option of the first leader, that is,

$$\text{first\_leader} = \text{argmin}_{s \in P_{il}} \text{time}(s, q_k). \tag{14}$$

As a result, restrictions are placed to define this model

$$\begin{aligned} &\forall_{s \in P_i} \text{answer}(s, q_k) \\ &\quad = \text{answer}(\text{first\_leader}, q_k) \\ &\qquad \wedge \text{time}(s, q_k) > \text{time}(\text{first\_leader}, q_k)). \end{aligned} \tag{15}$$

1) Collusion With w Leaders per Group: This model is essential to study the communication weights between the leader students, and the remaining fellow students in a group considering that there is an explicit transmission of information between them.

The model is generated, first, by doing a partition on the set of students. Each resulting group is partitioned into two, one set describing the leaders and the other one representing the copycats. In each group, the sequence of the quiz questions is randomly generated for each student. The leaders' performance is defined as explained in previous models and the timestamps for the leaders are randomly generated. The first leader to answer each question is determined and the chosen option chosen is the selected for every element of the group. The timestamps for the copycats are randomly generated, with the restriction that each should be greater than the timestamp of the first leader to answer that question.

To assess the range of communication weights associated with likely fraudulent behavior, Fig. 2 shows the intersection point of *consumption* scores between the distributions of the nonfraud and fraud models (where only the copycats in the group are considered). The intersection occurs at 0.39. Scores above this value have higher density in the fraud model, while above 0.6 are only present in the fraud model.

2) Every Element of the Group Is Leader: This model is a particular case of the previous one. Communication weights highlight the explicit broadcast nature of sharing information. Students in the group access shared content, yet they also actively broadcast content.

VI. Collusion Fraud Detection

The next step is to identify, from the group of students in the analysis, the ones whose score is indicative of fraud. These are signaled as possible cheaters and the course's professor may request an oral examination or other clearance initiatives. The end-to-end pipeline describing the proposed fraud detection methodology is provided in Fig. 3.

A. Assessing Individual Fraudulent Behavior

To understand the distribution of weights and scores, computed according to (10), in theoretical fraudulent and nonfraudulent environments, the null models presented in Section V are analyzed. The identification of potential fraudulent students is done by three methods. First, unilateral Wilcoxon signed-rank testing of each student's score against the baseline scores produced under the fraudulent null model.

More formally, let Y be the random variable describing the students' scores in a fraudulent null model, which is generated by a significant number of simulations, and X be the random variable describing the students' scores in the real model, for the set of questions Q, where $x_i \in X$ is the score of s_i, that is, $x_i = \text{score}(s_i, m, \beta)$, where $m \in \{c, p\}$, $\beta \in \mathbb{N}^+$. Given Q, when analyzing student s_i, we are interested in the variable $Z_i = Y - x_i$, which is obtained by subtracting the score of s_i in Q, x_i, to the observations drawn from Y.

Consider the random samples $\mathbf{y}$ and $\mathbf{z}_i$ obtained from the described random variables. Since we wish to identify scores with upward deviation, the null hypothesis is defined to state that the median of $\mathbf{z}_i$ is positive (the score of the student in the analysis is smaller than the scores of the students in the null model) against the alternative (the score of the student in the analysis is greater than the scores of the fellow students in the null model)

$$H_0 : \text{median}(\mathbf{z}_i) \geq 0, H_1 : \text{median}(\mathbf{z}_i) < 0. \tag{16}$$

If the score of a particular student is above the median of scores obtained in a scenario where fraud is present, we may inquiry, with some confidence, that they may be dishonest. Otherwise, we can hypothesize, under some confidence, that they may have not committed fraud. The output p-value is then assessed at a significance level (1%). Students with p-values below this threshold should be signaled for postanalysis.

An alternative method is to use interquartile range (IQR), or an alternative outlier statistic, inferred from the interval of scores computed from the real data. Students with outlier scores above the higher bound of the interval are noted as possible cheaters.

A third alternative is to rely on the intersection point of the score curves given by the null model of the nonfraud and null model of fraud, identifying as devious students those whose scores are above the threshold defined by this point.

In the end, students are identified as *fraudulent* (against the reference null models) if H_0 on (16) is rejected for the chosen significance level; the student's score is above the higher bound of the IQR interval; and the score is above the threshold defined by the intersection point between null models in the absence and presence of collusion; and as *nonfraudulent* otherwise.

B. Fraudulent Group Identification

Collusion fraud often involves more than two students who establish a channel of communication to inadvertently share

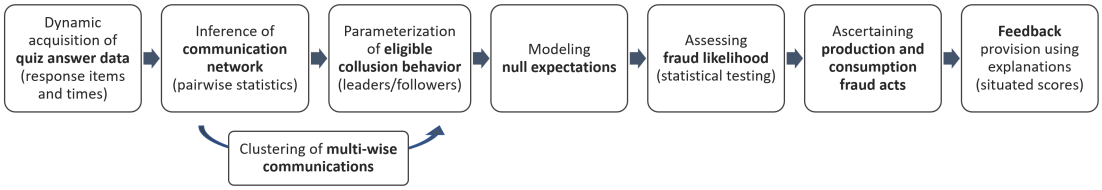


Fig. 3. Major steps of the proposed fraud detection methodology.

and access information [30]. In online assessments, although collusion can occur via physical channels when students choose to occupy the same space, messaging platforms are believed to be the most common communication channel. Parks et al. [36] examined how social media can promote a collective movement toward cyber-cheating, identifying motivations, and channels for group collusion.

The seek for fraudulent groups may be interpreted as an unsupervised machine-learning problem since it is unknown a priori who they are but their characteristics may reveal some evidence. Clustering is here suggested to identify groups of users with dense connections. This is particularly relevant since students may organize in groups to facilitate collusion behaviors.

With this analysis, it is possible to study the number of students per cluster. This information can guide the parameterization of the null models in Section V-B. For simplicity, let $\text{weight}(s_i, s_j, p, Q) = \text{weight}(s_j, s_i, c, Q)$ (by definition) be denoted as $w_{s_i s_j Q}$ and $\text{weight}(s_i, s_j, c, Q) = \text{weight}(s_j, s_i, p, Q)$ (by definition) be denoted as $w_{s_j s_i Q}$.

To apply clustering algorithms to the data, a similarity measure should be defined. Here, we use

$$\text{sim}(s_i, s_j, Q) = w_{s_i s_j Q} + w_{s_j s_i Q} \tag{17}$$

where s_i and s_j are students in S, and Q is a set of questions. It computes the amount of communication between these students by summing the weights of consumption and production.

As a result, we pretend to group students with a potentially high transmission of contents, either corresponding to high values of production, consumption, or, in accordance with the introduced similarity measure, their sum. On the other hand, cheating is frequently unidirectional, given that a student helps another one. In this case, it is relevant to consider the maximum communication weight between them. In this context, the similarity could be alternatively defined as

$$\text{sim}(s_i, s_j, Q) = \max(w_{s_i s_j Q}, w_{s_j s_i Q}). \tag{18}$$

This way, the focus resorts to the one-way transmission of information: production or consumption.

Either way, strong connections between candidate students or clusters of students are indicative of fraud predisposition and therefore students can be grouped together.

The dissimilarity between two students can be defined as

$$d(s_i, s_j, Q) = \text{sim\_max}(Q) - \text{sim}(s_i, s_j, Q) \tag{19}$$

where $\text{sim\_max}(Q)$ is the maximum value in the similarity matrix. The proofs that (17) and (19) measures are valid (dis)similarities can be found in the Appendix.

With the aim of clustering fraudulent communities of students, similarity and dissimilarity matrices are produced using the described formulas. Then, agglomerative hierarchical clustering methods with Single and Average linkage are suggested to identify groups of students yielding either local or spread interactions with peers in a communication channel.

VII. Results

A. Case Study

We consider online remote tests performed on the *Quizzes Tutor*'s platform, developed at Técnico Lisboa, as a study case. In particular, we analyze quizzes from two courses: SA, lectured in the first Semester of 2020/2021, and software engineer (SE) lectured in the second Semester of 2020/2021. Due to space limits, SA results are primarily discussed. The design of the quizzes ensures that students receive the questions in distinct order, with shuffled possible options, and further prevents students to go back and edit responses to previous questions. The quizzes are designed to be a part of the course's continuous evaluation, so their final contribution to the course grade is low, known to be associated with a lower tendency to cheat [10]. All the quizzes are performed at the end of each lecture, twice per week, with five MCQs and four options each. The order in which the questions and options appear to each student is randomly chosen. SA students had 6 min to complete each exam, whilst SE students had 5 min.

B. Fraud Detection Experiments

The principles placed along Sections IV–VI form a methodology to assist tutors detecting collusion events (see Fig. 3). Once the network with directional communication weights is computed, a natural subsequent step is the identification of collusion groups, that is, to detect if students are organized in groups with the aim of sharing or accessing real-time information about an exam when it is taking place. To this end, the introduced clustering stance (see Section VI-B) is pursued. Major results on SA are presented.

Considering S to be the set of students taking the SA course and Q to be the set of questions of the quiz with id 14225. Fig. 4 illustrates the clustering case with average linkage. A color threshold of 0.2 is set to better separate clusters whose dissimilarity between their elements is below this value. It is possible to identify a cluster of students 810 and 19979, potentially involving student 19988 (dark green); and a cluster of students 13089 and 19867 (orange). Although fraud appears to occur in compact groups of two or three students, larger communication channels should not be excluded at this stage to study the different possibilities of collusion.

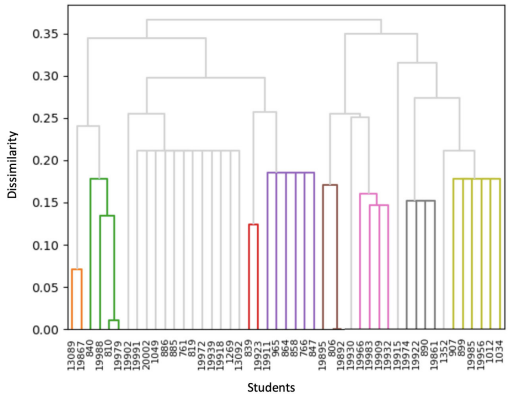


Fig. 4. Dendrogram (average linkage) on real data from quiz 14225.

We now assess whether these associations deviate from expectations. This is a subsequent step in the methodology. To this end, let us assume students organize themselves in groups of three where two consume content from the remaining student that acts as content producer. In this context, we can fix a null model of fraud (expectations of fraudulent behavior) assuming the size of communication channels and the number of content producers. Results on the presented null models were obtained using 30 simulations and considering the regularities found in Q, the set of questions of the SA quiz with id 14225.

Fig. 5(a) presents the distribution of the weight of associations in the respective null model of fraud, where the weight corresponds to the likelihood of fraudulent behavior. Lower values of weight (around 0.1) have higher density, which is expected since these weights correspond to interactions between students outside collusion groups. As such, their communication is reduced when compared to the communication between students of the same group, yielding higher weights (approximately 0.35) and lower density. Fig. 5(b) depicts the distribution of association weights in the null model where the unique influence on the answers is the grade. Finally, Fig. 5(c) presents the distribution of association weights on the real data. The computed weights combine both channels of communication: production and consumption. Peaks are observed in lower weights, around 0.05 and 0.15, which appear to indicate that the communication is, in general, artificial and no fraud has been committed, or else we would observe higher density around 0.35 values in accordance with the null model of fraud. Fig. 6 provides the distribution of the weight function within a fraudulent group against association weights outside a group, revealing significant differences between the weights of edges connecting students in the same group and edges connecting students of different groups, as expected given the pursued null definition of fraud.

In previous examples, where we consider the presence of groups of three students with one leader, the score of fraud of a student in this quiz is computed according to (10) with $\beta = 1$. If there is evidence of access to multiple content producers (leaders), β can be increased in accordance.

In the presence of expectation levels on what is likely a fraudulent behavior (e.g., scores above the intersection point between null distributions in Fig. 6), we can now move to the comprehensive network-based view of associations to assess collusion between pairs of students. Fig. 7(a) shows a representation of the communication between students in the null model of fraud with groups of three and one leader. Here, the leaders and copycats are easily distinguished as the former are represented by big purple circles (as they produce more than consume) and the latter by big pink circles (higher consumption). In Fig. 7(b), representing the communication between students in the null model of nonfraud where grades are the unique influence to the responses, circles are smaller than the ones presented in the previous graph as the existent communication between students is spurious. Understandably, the difference between producers and consumers is less evident.

Fig. 8 provides the network model from the real answers to quiz 14225, SA. Generally, nodes are generally smaller than in previous networks, and the differences between consumption and production are subtle, indicating that, if existent, the occurrences of fraud in quiz 14225 are scarce.

Fig. 2 (in Section V-B) showed the presence of statistically significant differences between the scores in the null models of fraud and nonfraud, as theoretically expected. Complementarily, we now assess how fraud intersection thresholds vary for communication channels with a higher number of consumers and producers (leaders). Tables II–V show that for a fixed group size, the intersection point is lower when the number of leaders is higher. The mean and standard deviation of the curve correspondent to the null model of nonfraud for consumption edges is 0.3129 and 0.0696, respectively. For a fixed number of students in a group, the greater the number of leaders, the closer the scores to the null model of fraud, hampering the separation of behaviors, as further illustrated in Fig. 9(a) and (b).

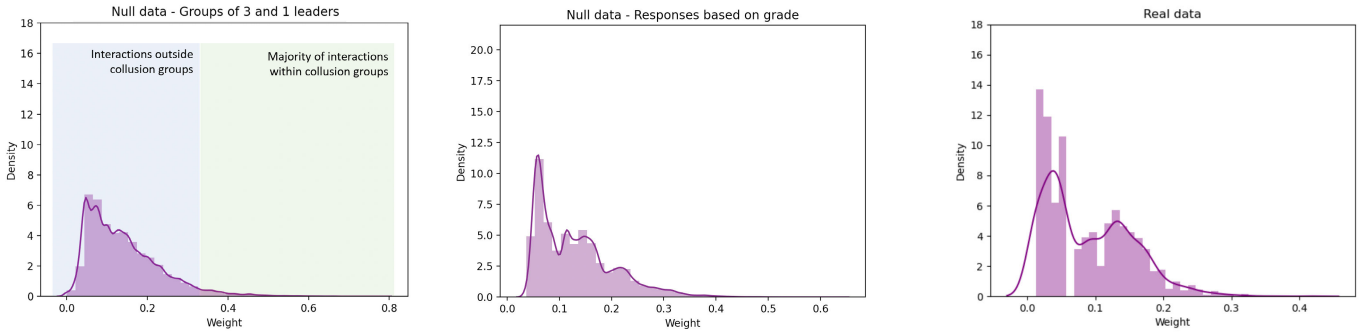
Fraud detection is the final step of the proposed methodology. Decisions under $\alpha = 0.1$ significance levels are illustrated

TABLE II
INTERSECTION POINT OF SCORE CURVES OF NONFRAUD (GRADES CONSIDERED) AND FRAUD FOR CONSUMPTION EDGES

group size	number of leaders 1	2	3	4	5	6
2	0.3995	-	-	-	-	-
3	0.3951	0.3813	-	-	-	-
4	0.3904	0.3810	0.3787	-	-	-
5	0.3878	0.3786	0.3760	0.3703	-	-
6	0.3862	0.3796	0.3726	0.3755	0.3748	-
7	0.3889	0.3820	0.3757	0.3700	0.3751	0.3733

TABLE III
MEAN AND DEVIATION OF CONSUMPTION SCORES (NULL MODEL OF FRAUD)

group size	number of leaders 1	2	3	4	5
2	0.4909±0.09	-	-	-	-
3	0.4767±0.10	0.4376±0.09	-	-	-
4	0.4637±0.10	0.4284±0.10	0.3946±0.11	-	-
5	0.4486±0.11	0.4132±0.10	0.3952±0.10	0.3731±0.10	-
6	0.4527±0.09	0.4116±0.11	0.3803±0.10	0.3738±0.10	0.3579±0.10
7	0.4356±0.12	0.3927±0.11	0.3859±0.10	0.3680±0.10	0.3129±0.07



(a) (b) (c)

Fig. 5. Distribution of association weights (potential acts of fraud) considering the regularities of AS quiz 14225. (a) Null model of fraud with groups of three students and one leader. (b) Null model where grades are the unique response influence (no fraud). (c) Real student data.

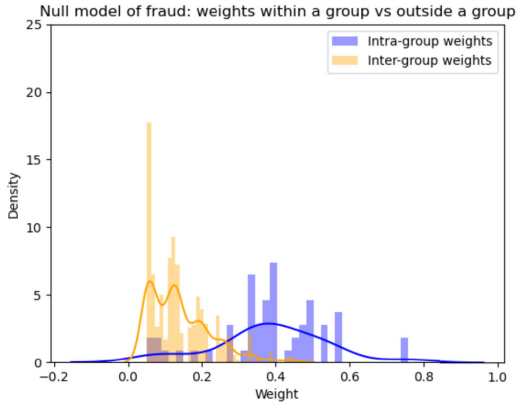
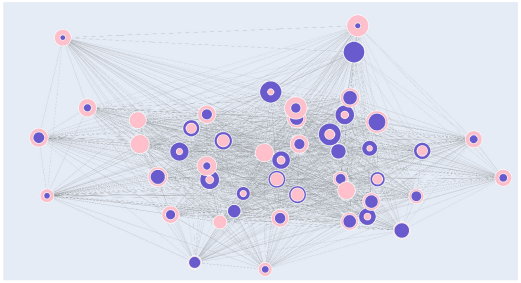
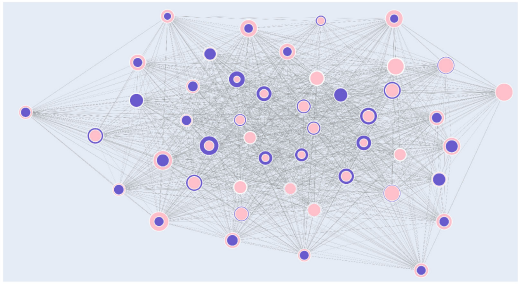


Fig. 6. Distribution of weights in the null model of fraud with groups of three students and one leader, within and outside a group.



(a)



(b)

Fig. 7. Communication graphs. (a) Students in the null model of fraud (groups of three and one leader). (b) Students in the null model where grades are the unique influence on the responses (no fraud).

in Fig. 10 for the SA quiz 14225. Exploring consumption scores [see Fig. 10(a)], we observe the presence of one student with scores higher than the median score in the reference null model of fraud; and one student with a score above the

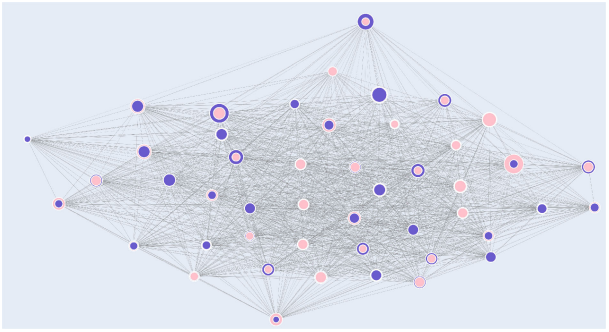


Fig. 8. Student communication graph on real data for quiz 14225.

TABLE IV
INTERSECTION POINT OF SCORE CURVES OF NONFRAUD (GRADES CONSIDERED) AND FRAUD FOR PRODUCTION EDGES

	number of leaders					
group size	1	2	3	4	5	6
2	0.3973	-	-	-	-	-
3	0.3951	0.3715	-	-	-	-
4	0.3903	0.3760	0.3792	-	-	-
5	0.3893	0.3796	0.3784	0.3805	-	-
6	0.3864	0.3804	0.3766	0.3820	0.3840	-
7	0.3891	0.3840	0.3824	0.3840	0.3842	0.3809

TABLE V
MEAN AND DEVIATION OF PRODUCTION SCORES (NULL MODEL OF FRAUD)

	number of leaders				
group size	1	2	3	4	5
2	0.4850±0.10	-	-	-	-
3	0.4769±0.10	0.3940±0.10	-	-	-
4	0.4637±0.10	0.4050±0.10	0.3459±0.10	-	-
5	0.4529±0.11	0.3953±0.11	0.3556±0.11	0.3387±0.10	-
6	0.4532±0.10	0.4033±0.11	0.3542±0.10	0.3469±0.11	0.3338±0.10
7	0.4371±0.12	0.3857±0.12	0.3580±0.11	0.3360±0.11	0.3128±0.07

higher bound of the IQR interval and the intersection of score curves. The analysis of production scores further indicates the presence of a student potentially involved in content sharing.

Consider now Q to be a set of ten quizzes in the SA course. The sets S and S' are left unchanged. In Table VI, we present the number of students in each category of fraud for each quiz. The first column refers to the result of applying the statistical test in (16), using the null model of fraud with groups of three students and one leader; the second to the

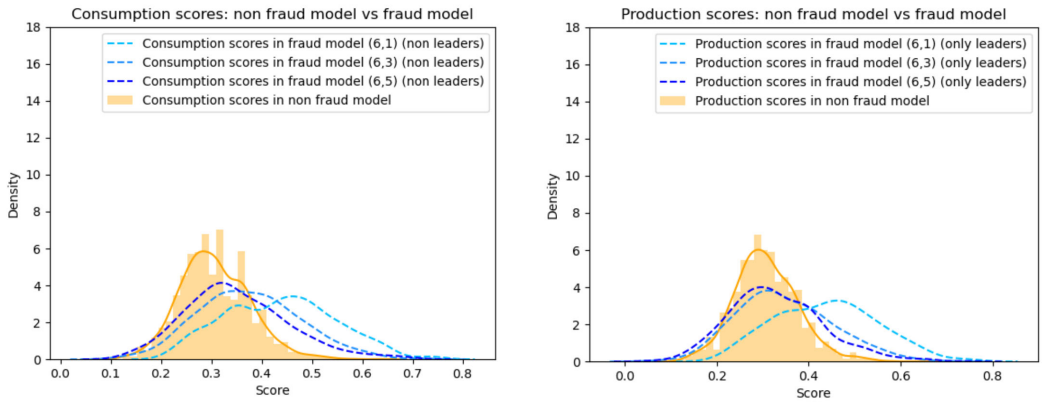


Fig. 9. Distribution of scores in the null models of nonfraud and fraud with groups of 6 and distinct number of leaders. (a) Consumption scores. (b) Production scores.

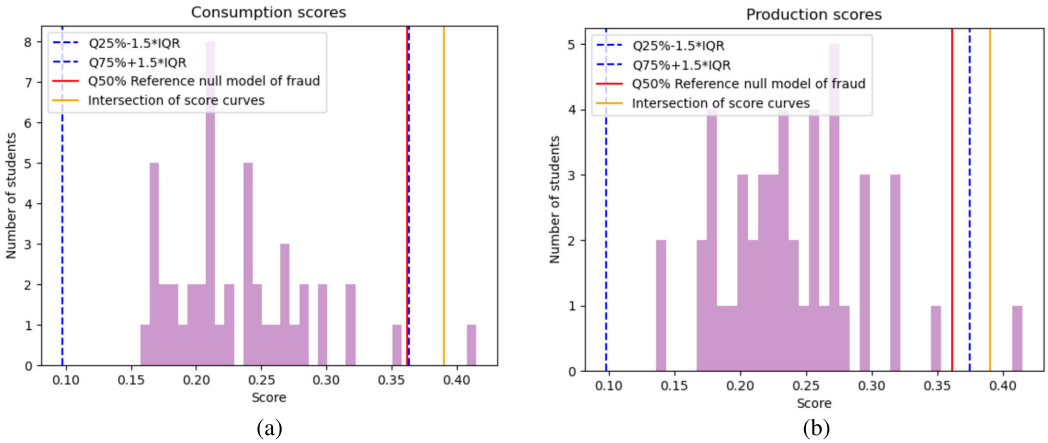


Fig. 10. Distribution of scores from quiz 14225 (SA). (a) Consumption scores. (b) Production scores.

TABLE VI
DETECTED FRAUDS PER QUIZ (CONSUMPTION MODE), SA

quiz	fraud. (null model)	fraud (int. point)	fraud. (IQR)	non fraud.
14225	1	1	1	46
14235	2	3	1	47
14236	0	0	0	48
14240	0	0	1	46
14264	0	0	0	47
14265	2	6	2	44
14331	2	7	2	44
14333	2	1	5	41
14406	0	0	0	46
14409	4	2	4	43

TABLE VII
DETECTED FRAUDS PER RANDOMLY SELECTED STUDENTS (CONSUMPTION MODE), SA

student	fraud. (null model)	fraud (int. point)	fraud. (IQR)	non fraud.
1012	1	1	1	9
1034	0	0	0	10
1049	0	0	0	10
1269	0	0	0	10
13089	0	0	0	10
13092	0	0	0	10
1352	0	0	0	10
19861	0	0	0	10
19930	1	1	1	9
19939	0	1	0	10

intersection point between the curves representing the null models of fraud and nonfraud; and the third to the IQR metric computed over the scores obtained using the real data. The fourth column contains the number of incidences which did not verify any previous criteria. The acquired results reveal quiz 14409 to be associated with the highest potential fraud acts against the null model of fraudulent behaviors (four occurrences). Quiz 14331 had the highest number of students, 7, with scores above the intersection point between the score curves of the fraud and nonfraud null models, followed by 14265. Taking into consideration outlier IQR statistics, the highest number of fraudulent students was identified in quiz 14333 (five occurrences). For all quizzes, the majority of the students were designated as nonfraudulent.

Table VII depicts, for each student, the number of quizzes with a fraud occurrence per criterion. A random sample of ten students is considered. Tables VI and VII refer to copy acts (consumption scores). Students 1012, 19930, and 19939 were the only ones signaled as fraudulent with respect to some fraud categories. In particular, students 1012 and 19930 were determined as fraudulent by the three introduced metrics. In the majority of the quizzes, fraudulent students were not encountered.

TABLE VIII
PERFORMANCE MEASUREMENT FOR FIVE QUESTION QUIZZES

Course	Quizzes	Quiz Answers	Average (ms)	Median (ms)	95% Line (ms)
Software Architecture	24	45.42	1063	1024	1364
Software Engineering	78	81.36	1762	1742	2460

C. Computational Complexity

Given n students and g questions, the time complexity to compute the item selection probability for all questions is $O(ng)$, the answer precedence between two students from their timestamps is $O(g)$, the posterior weight calculus [see (7) and (8)) is $O(g)$, the network inference is then $O(np + n^2p) = O(n^2p)$, and the subsequent scoring of all students in the network according to (10) is $O(n^2\beta)$. Accordingly, the principled generation of quiz answers and subsequent description of null models is $O(kn^2(\beta + g))$, where k is the number of simulations. The fraud detection step against the precomputed null model thresholds is linear on the number of students and null models, hence the overall time complexity is $O(kn^2(\beta+g))$, with k=1 for precomputed null models, and the memory complexity is $O(n^2)$.

To measure performance, we performed load tests using the deployed fraud detection system at the *Quizzes Tutor* platform (see Section VIII). Two tests were done on a server running Ubuntu 18.04.3 LTS with four cores and 16 GB of RAM. The data necessary to assess fraud was obtained for each one of the quizzes of the two courses, SA and SE.

Table VIII presents the results. The latency has an average between 1 and 1.7 s, where the difference is due to the number of students per quiz (average 45 and 81, respectively, for each one of the courses). We consider the values acceptable for the teacher to wait until she can start analyzing the results. Although the total number of quizzes significantly differs between the two courses (24–78), the analysis of the different values consistently shows that latency correlates with the number of quiz answers (number of students answering the quiz), in conformity with the aforementioned time complexity.

VIII. FRAUD DETECTION SYSTEM: DEPLOYMENT AND VALIDATION

The proposed fraud detection methodology, implemented in Python, is made available as an analytical module in the *Quizzes Tutor* platform.[1] This platform is frequently used for online quiz assessments by several courses at Instituto Superior Técnico, Universidade de Lisboa. The provided fraud detection facilities have undertaken successful deployment and validation stages, being available to the academic community with the necessary disclaimers for the adequate use and limits of actionability.

The deployed instance generates the communication network considering all answers, both correct and incorrect, applying the real model. Production and consumption scores are calculated using pairwise communication channels by default, that is, $\beta = 1$.

[1]https://quizzes-tutor.tecnico.ulisboa.pt/

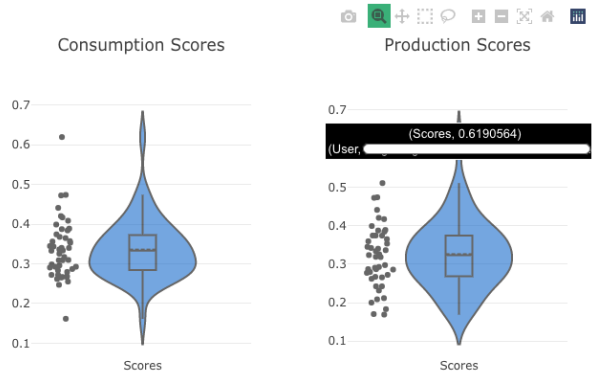


Fig. 11. Interactive violin charts for inspecting students with deviant consumption and production scores, available at *Quizzes Tutor*.

Visualizations with strict usability guarantees are provided to aid the analysis of critical cases. Fig. 11 shows the consumption and production scores violin charts for a five-question quiz. The teacher can interact with the graphs to obtain information about a particular student, for instance, an outlier student with deviant scores. In the given example in Fig. 11, the student name is anonymized.

IX. CONCLUSION

This work introduced a novel methodology to assess likely fraud communication acts in remote online MCQ exams based on the concordance of responses and answer times. Null models are produced to understand regular versus fraud dynamics and to identify collusion with strict guarantees of statistical significance. Complementarily, clustering algorithms are applied to unravel communication channels between students. Considering matched answers, choice probability, response times (directionality), and recurrence, we show that is possible to create a network of potential communication acts between students. Having constructed the network for null models representing fraudulent and honest behavior, we obtain insights into how to separate spurious communication from the actual interchange of information. Finally, employing these insights on the real data, and making use of scoring techniques, we are able to categorize each student with respect to their fraud likelihood and thus understand inadvertent communication pathways and promote the actionability of recommendations, supporting the course's tutor with the subsequent inquiry or advertence initiatives.

The application of the proposed principles in the context of the SA course reveals students with a higher fraud likelihood, already showing to be a solid criterion to guide tutors in ascertaining collusion and discouraging communication.

In this work, fraudulent behavior analysis was primarily pursued in the context of a single quiz. However, if deviant behavior is detected in more than one quiz, the chances of fraudulent behavior considerably increase. In this context, binomial testing can be straightforwardly applied to identify the probability of observing a given number of potential fraud acts.

The reported findings open guidelines to establish both preventive and reactive policies for fraud control. The disclosure of the proposed fraud detection model prevently demotivates collusion acts. The acquired results further support the role of assigning distinct orders of questions, shuffling item options, and preventing reverse editions. Complementary strategies should be considered, including continuous authentication to prevent impersonating, online proctoring (whether human or automated) to promote academic integrity [22], stratified exam contents (e.g., pools of alternative questions), attitude formation (e.g., emphasis on learning, formative assessments), and cheat-resistant software facilities, including browser tab lockers [37], IP change detectors [33], and wireless jammers [38].

ACKNOWLEDGMENT

The authors further acknowledge the support of Pedro Caldeira in the deployment of the proposed fraud detection system.

REFERENCES

[1] A. M. Duhaim, S. O. Al-Mamory, and M. S. Mahdi, "Cheating detection in online exams during COVID-19 pandemic using data mining techniques," *Webology*, vol. 19, pp. 1–26, 2021.

[2] R. Bawarith, D. Abdullah, D. Anas, and S. Gamalel-Din, "E-exam cheating detection system," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 4, pp. 176–181, 2017.

[3] B. A. Barr and S. F. Miller, "Higher education: The online teaching and learning experience," Fac. School Adv. Stud., ERIC, Univ. Phoenix, Phoenix, AZ, USA, Tech. Rep. ED543912, 2013.

[4] R. Matos and J. Barber, "MoodleGate: Securing computer driven exam environments," in *Proc. INTED*, 2013, pp. 1–7.

[5] V. Susithra, A. Reshma, B. Gope, and S. Sankar, "Detection of anomalous behaviour in online exam towards automated proctoring," in *Proc. Int. Conf. Syst., Comput., Autom. Netw. (ICSCAN)*, Jul. 2021, pp. 1–5.

[6] G. R. Watson and J. Sottile, "Cheating in the digital age: Do students cheat more in online courses?" *Online J. Distance Learn. Admin.*, vol. 13, no. 1, pp. 1–14, 2010.

[7] C. M. Toquero, "Challenges and opportunities for higher education amid the COVID-19 pandemic: The philippine context," *Pedagogical Res.*, vol. 5, no. 4, Apr. 2020, Art. no. em0063.

[8] F. Kamalov, H. Sulieman, and D. S. Calonge, "Machine learning based approach to exam cheating detection," *PLoS ONE*, vol. 16, no. 8, Aug. 2021, Art. no. e0254340.

[9] P. R. Morales and M. L. Verde, "Cómo asegurar evaluaciones válidas y detectar falseamiento en pruebas a distancia síncronas," *Revista Digit. de Investigación en Docencia Universitaria*, vol. 14, no. 2, p. e1240, Nov. 2020.

[10] R. K. Ladyshewsky, "Post-graduate student performance in 'supervised in-class' vs. 'unsupervised online' multiple choice tests: Implications for cheating and test security," *Assessment Eval. Higher Educ.*, vol. 40, no. 7, pp. 883–897, Oct. 2015.

[11] Z. Zhang, S. Zhu, J. Mink, A. Xiong, L. Song, and G. Wang, "Beyond bot detection: Combating fraudulent online survey takers," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 699–709.

[12] H. Hu, Z. Li, and Z. Wang, "Test cheating detection method based on random forest," in *Proc. 3rd Int. Conf. Comput. Sci. Technol. Educ. (CSTE)*, May 2021, pp. 47–52.

[13] J. Ranger, N. Schmidt, and A. Wolgast, "The detection of cheating on E-exams in higher education—The performance of several old and some new indicators," *Frontiers Psychol.*, vol. 11, Oct. 2020, Art. no. 568825.

[14] M. Li et al., "Optimized collusion prevention for online exams during social distancing," *NPJ Sci. Learn.*, vol. 6, no. 1, pp. 1–9, Mar. 2021.

[15] G. J. Cizek and J. A. Wollack, *Handbook of Quantitative Methods for Detecting Cheating on Tests*. Evanston, IL, USA: Routledge, 2017.

[16] P. W. Holland, "Two measures of change in the gaps between the CDFs of test-score distributions," *J. Educ. Behav. Statist.*, vol. 27, no. 1, pp. 3–17, Mar. 2002.

[17] K. Sijtsma and R. R. Meijer, "A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model," *Appl. Psychol. Meas.*, vol. 16, no. 2, pp. 149–157, Jun. 1992.

[18] S. Sinharay, "Detection of item preknowledge using likelihood ratio test and score test," *J. Educ. Behav. Statist.*, vol. 42, no. 1, pp. 46–68, Feb. 2017.

[19] J. A. Wollack, A. S. Cohen, and C. A. Eckerly, "Detecting test tampering using item response theory," *Educ. Psychol. Meas.*, vol. 75, no. 6, pp. 931–953, Dec. 2015.

[20] D. I. Belov, "Robust detection of examinees with aberrant answer changes," *J. Educ. Meas.*, vol. 52, no. 4, pp. 437–456, Nov. 2015.

[21] J. F. George and J. R. Carlson, "Group support systems and deceptive communication," in *Proc. 32nd Annu. Hawaii Int. Conf. Syst. Sci. HICSS Abstr. CD-ROM Full Papers*, 1999, p. 10.

[22] F. Noorbehbahani, A. Mohammadi, and M. Aminazadeh, "A systematic review of research on cheating in online exams from 2010 to 2021," *EduC. Inf. Technol.*, vol. 27, pp. 8413–8460, Mar. 2022.

[23] S. Janke, S. C. Rudert, Ä. Petersen, T. M. Fritz, and M. Daumiller, "Cheating in the wake of COVID-19: How dangerous is ad-hoc online testing for academic integrity?" *Comput. Educ. Open*, vol. 2, Dec. 2021, Art. no. 100055.

[24] A. Fask, F. Englander, and Z. Wang, "Do online exams facilitate cheating? An experiment designed to separate possible cheating from the effect of the online test taking environment," *J. Academic Ethics*, vol. 12, no. 2, pp. 101–112, Jun. 2014.

[25] E. T. Dille, *A Multi-Institutional Investigation Into Cheating on Tests in College Online Courses*. Columbia, SC, USA: Univ. South Carolina, 2011.

[26] K. Man, J. R. Harring, and S. Sinharay, "Use of data mining methods to detect test fraud," *J. Educ. Meas.*, vol. 56, no. 2, pp. 251–279, Jun. 2019.

[27] F. Drasgow, M. V. Levine, and M. E. McLaughlin, "Detecting inappropriate test scores with optimal and practical appropriateness indices," *Appl. Psychol. Meas.*, vol. 11, no. 1, pp. 59–79, Mar. 1987.

[28] C. Zopluoglu, "Similarity, answer copying, and aberrance: Understanding the status quo," in *Handbook of Quantitative Methods for Detecting Cheating on Tests*. Evanston, IL, USA: Routledge, 2016, pp. 25–46.

[29] J. A. Ruipérez-Valiente, S. Joksimović, V. Kovanović, D. Gašević, P. J. Muñoz-Merino, and C. D. Kloos, "A data-driven method for the detection of close submitters in online learning environments," in *Proc. 26th Int. Conf. World Wide Web Companion WWW Companion*, 2017, pp. 361–368.

[30] A. Balderas, M. Palomo-Duarte, J. A. Caballero-Hernández, M. Rodriguez-Garcia, and J. M. Dodero, "Learning analytics to detect evidence of fraudulent behaviour in online examinations," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 7, no. 2, pp. 241–249, 2021.

[31] Y. Cai, G. Fragkos, E. E. Tsiropoulou, and A. Veneris, "A truth-inducing sybil resistant decentralized blockchain Oracle," in *Proc. 2nd Conf. Blockchain Res. Appl. Innov. Netw. Services (BRAINS)*, Sep. 2020, pp. 128–135.

[32] E. Bilen and A. Matros, "Online cheating amid COVID-19," *J. Econ. Behav. Org.*, vol. 182, pp. 196–211, Feb. 2021.

[33] L. C. O. Tiong and H. J. Lee, "E-cheating prevention measures: Detection of cheating at online examinations using deep learning approach—A case study," 2021, *arXiv:2101.09841*.

[34] G. Alexandron, J. A. R. Valiente, and D. E. Pritchard, "Towards a general purpose anomaly detection method to identify cheaters in massive open online courses," in *Proc. EDM*, 2019, pp. 1–4.

[35] K. Krueger. (Aug. 2015). *How to Catch Students Cheating on Online Tests*. [Online]. Available: http://mediashift.org/2015/08/how-to-catch-students-cheating-on-online-tests/

[36] R. F. Parks, P. B. Lowry, R. T. Wigand, N. Agarwal, and T. L. Williams, "Why students engage in cyber-cheating through a collective movement: A case of deviance and collusion," *Comput. Educ.*, vol. 125, pp. 308–326, Oct. 2018.

[37] S. S. Chua, J. B. Bondad, Z. R. Lumapas, and J. D. L. Garcia, "Online examination system with cheating prevention using question bank randomization and tab locking," in *Proc. 4th Int. Conf. Inf. Technol. (InCIT)*, Oct. 2019, pp. 126–131.

[38] A. Vegendla and G. Sindre, "Mitigation of cheating in online exams: Strengths and limitations of biometric authentication," in *Biometric Authentication in Online Learning Environments*. Hershey, PA, USA: IGI Global, 2019, pp. 47–68.