Bayesian Pairwise Collaboration Detection in Educational Datasets

Andrew E. Waters, Christoph Studer, and Richard G. Baraniuk
Rice University, Houston TX, USA; e-mail: {waters, studer, richb}@sparfa.com

***Abstract*—Online education affords the opportunity to revolutionize learning by providing access to high-quality educational resources at low costs. The recent popularity of so-called MOOCs (massive open online courses) further accelerates this trend. However, these exciting advancements result in several challenges for the course instructors. Among these challenges is the detection of collaboration between learners on online tests or take-home exams which, depending on the courses' rules, can be considered cheating. In this work, we propose new models and algorithms for detecting pairwise collaboration between learners. Under a fully Bayesian setting, we infer the probability of learners' succeeding on a series of test items solely based on their response data. We then use this information to estimate the likelihood that two learners were collaborating. We demonstrate the efficacy of our methods on both synthetic and real-world educational data; for the latter, we find strong evidence of collaboration for a certain pair of learners in a non-collaborative take-home exam.**

***Index Terms*—Bayesian methods, cheating, collaboration detection, hypothesis testing, online education, sparse factor analysis.**

I. Introduction

A well-known challenge for educators and course instructors is the ability to detect collaboration among learners in a course, test, or exam [1], [2]. Detecting collaboration is of particular interest in situations where it is prohibited and considered as cheating, such as for certain take-home exams or tests. In the setting of online education, e.g., MOOCs (massive open online courses), the capability of *automatically* detecting learner collaboration (or cheating-through-collaboration) becomes even more important, as potentially thousands of learners may be enrolled in a course, without ever having face-to-face interaction with an instructor [3]. In such situations, the massive number of learners simply prohibits a manual detection of collaboration.

A. Collaboration Detection via Learning Analytics

A naïve approach for computer-aided detection of collaboration in educational datasets, such as multiple-choice tests, would consist of simply comparing the answer patterns between all pairs of learners, and flagging learner pairs that exhibit a high degree of similarity. This approach, however, is prone to fail as it ignores the aptitude of the individual learners, as well as the difficulty of each test item or question [2], [4].

In [5], the authors proposed a novel framework for learning analytics (LA), referred to as SPARFA (short for SPARse Factor Analysis). This framework builds upon a principled statistical model for analyzing probability of a set of learners providing correct (or incorrect) answers to test items, solely given binary-valued (right/wrong) graded response data to multiple-choice tests. Put simply, SPARFA jointly infers the probability that each learner will answer a given question correctly or incorrectly, as well as the difficulty of each question. Armed with these capabilities, SPARFA can be used to develop powerful collaboration detection tests that take into account learner aptitude as well as question difficulties.

B. Contributions

This paper develops new collaboration detection methods that leverage SPARFA to infer information regarding learner aptitude and question difficulty. Given this information, we propose two Bayesian hypothesis tests for detecting collaboration in educational datasets obtained from multiple-choice tests. The first test examines the number of agreements between pairs of learners given the SPARFA parameters and uses this information to infer the likelihood of collaboration. The second test examines the joint answer sequence of pairs of learners using a specific collaboration model and evaluates the likelihood that such patterns would arise independently. While the first collaboration test has its main advantage in computational efficiency, the second test provides superior detection performance at higher computational complexity.

II. Learning Analytics via Sparse Factor Analysis

We start by summarizing the SPARFA framework [5] to analyze graded response data. We then outline SPARFA-B, a fully Bayesian method to extract the SPARFA parameters.

A. Sparse Factor Analysis (SPARFA)

We consider a test consisting of Q of *questions* that test the knowledge of N learners of various portions of a course's content. We assume that there are K latent factors, referred to as *concepts*, that govern the learners' responses to these questions. Let $Y_{i,j}$ denote the binary-valued (right/wrong) graded response for learner j on question i. SPARFA builds upon the following model for the graded response data [5]:

$$Y_{i,j} \sim \mathit{Ber}(\Phi(Z_{i,j})), \;\; Z_{i,j} = \mathbf{w}_i^T \mathbf{c}_j + \mu_i, \;\; \forall i,j. \tag{1}$$

Here, $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) \mathrm{d}t$ corresponds to the inverse *probit* link function, with $\Phi(Z_{i,j}) \in [0,1]$ governing

The authors would like to thank T. Goldstein, Mr. Lan, and D. Vats for insightful discussions on collaboration detection. This work was supported by the National Science Foundation under Cyberlearning grant IIS-1124535, the Air Force Office of Scientific Research under grant FA9550-09-1-0432, and the Google Faculty Research Award program.

the probability of learner j answering question i correctly. $Ber(z)$ designates a Bernoulli distribution with mean z. The vector $\mathbf{c}_j \in \mathbb{R}^K$, $j = 1, \ldots, N$, represents the concept mastery of the j^{th} learner, with its k^{th} entry representing the learner's mastery of concept k. The vector $\mathbf{w}_i \in \mathbb{R}^K$ models the *concept associations*, i.e., encodes how question i is related to each concept. The scalar μ_i models the *intrinsic difficulty* of question i, where large values indicate easy questions.

Retrieving the parameters $\mathbf{c}_j$, $\mathbf{w}_i$, and μ_i from the set of observations $Y_{i,j}$ in (1) is in an ill-posed inverse problem. To make the model both tractable as well as interpretable, SPARFA assumes that the number of concepts K is small relative to both the number of learners and questions. Furthermore, SPARFA imposes sparsity and non-negativity on the question–concept vectors $\mathbf{w}_i$; we refer to [5] for the details.

B. A Bayesian Method to Extract the SPARFA Parameters

Inference of the model parameters in (1) can be carried in a variety of ways [5]. One attractive solution is to use a fully Bayesian approach, which provides distributions for all model parameters of interest, rather than simple point estimates. Such distributions are capable of incorporating the estimation uncertainty of each parameter, which can then be used to improve corresponding inference tasks. In the case of collaboration detection, such Bayesian methods enable tempering both the false alarm and miss rate.

In [5], the authors develop SPARFA-B, a fully Bayesian algorithm for extracting the SPARFA parameters. By imposing specific prior distributions on the parameters of interest, such as exponentials on the non-negative, active entries in $\mathbf{w}_i$, SPARFA-B estimates posterior distributions on the model parameters via Markov Chain Monte-Carlo (MCMC) sampling; given space constraints, we omit the sampling details.

III. Collaboration Test in Educational Data

We now summarize two Bayesian likelihood tests for detecting pairwise collaboration via SPARFA. We then detail how to incorporate these tests into the SPARFA-B algorithm.

A. Collaboration Test 1: Agreement Matching

The first test to detect collaboration among pairs of learners focuses on the *number of agreements* in the graded learner response data between a pair of learners (k, ℓ). The first hypothesis $\mathcal{H}_1^1[k,\ell]$ of Test 1 corresponds to the case where learner k and ℓ decide to agree on a specific graded response with some (unknown) probability δ. The second hypothesis $\mathcal{H}_2^1[k,\ell]$ assumes that the number of agreements between the graded responses of learner k and ℓ were generated independently, given the SPARFA parameters. For the sake of simplicity of notation, we omit the learner labels k and ℓ.

1) Collaboration Hypothesis: We start by defining the probability of the first hypothesis $\mathcal{H}_1^1$, which corresponds to the probability $P(\mathcal{H}_1^1) = P(A)$ indicating that learner k and ℓ agree on A graded responses. The number of such agreements A between learner k and ℓ is given by

$$A = \sum_{i=1}^{Q} a_i \quad \text{with} \quad a_i = \begin{cases} 1, & \text{if } Y_{i,k} = Y_{i,\ell} \\ 0, & \text{if } Y_{i,k} \neq Y_{i,\ell}. \end{cases}$$

We now assume that there is an unknown probability δ indicating that learner k and ℓ agreed to provide the same graded response for question i. Under this model, the probability of having A agreements out of Q questions is given by the binomial distribution:

$$P(A \mid Q, \delta) = \binom{Q}{A} \delta^A (1-\delta)^{(Q-A)}.$$

We now assume a uniform prior on δ over the range $[0, 1]$, indicating that we have no informative belief on the value of δ. With this Bayesian approach, we obtain

$$P(\mathcal{H}_1^1) = \int_0^1 \binom{Q}{A} \delta^A (1-\delta)^{(Q-A)} \mathrm{d}\delta. \tag{2}$$

Note that this probability corresponds to the Beta function $\mathrm{B}(A+1, Q-A+1)$ as defined in [6].

2) Independence Hypothesis: The probability of the second hypothesis $\mathcal{H}_2^1$ is defined as

$$P(\mathcal{H}_2^1) = P(A \mid (\mathbf{w}_i, \mu_i)\, \forall i, \mathbf{c}_k, \mathbf{c}_\ell) \tag{3}$$

and assumes that the number of agreements A are generated independently, given the SPARFA parameters. In order to expand the right-hand side of (3), we first need the probability of having an agreement of the graded responses for the i^{th} question. From (1) follows that the probability of learner j answering question i correctly or incorrectly is simply given by $\Phi(Z_{i,j})$ or $1 - \Phi(Z_{i,j})$, respectively. Hence, the probability that both learners k and ℓ *independently* achieve the same graded learner response at question i is given by

$$\begin{aligned} p_i^a = p(a_i = 1 \mid (\mathbf{w}_i, \mu_i)\, \forall i, \mathbf{c}_k, \mathbf{c}_\ell) = \\ \Phi(Z_{i,k})\Phi(Z_{i,\ell}) + \big(1 - \Phi(Z_{i,k})\big)\big(1 - \Phi(Z_{i,\ell})\big) \end{aligned} \tag{4}$$

with $Z_{i,k}$ and $Z_{i,\ell}$ as defined in (1).With (4), we can rewrite the probability $p(\mathcal{H}_2^1)$ in (3) as

$$p(\mathcal{H}_2^1) = \mathit{PoiBin}\big(A, Q, \mathbf{p}^a\big), \tag{5}$$

where $\mathit{PoiBin}(A, Q, \mathbf{p}^a)$ denotes the Poisson-Binomial probability mass function with A successful trials out of a total number of Q [7], and $\mathbf{p}^a = \{p_1^a, \ldots, p_Q^a\}$.

We note that a related model was proposed in [2]. This model relies on simple statistics for learners and questions, which are used to estimate the expected number of agreements between two learners via a normal approximation to the Poisson-Binomial model. From these parameters, the authors then calculate the probability of collaboration.

3) Log Bayes Factor: Given the probabilities (2) and (5) for the hypotheses $\mathcal{H}_1^1$ and $\mathcal{H}_2^1$, respectively, we can finally compute the *log Bayes factor* for Test 1 as follows:

$$\mathit{LBF}_1 = \log\!\left(\frac{P(\mathcal{H}_1^1)}{P(\mathcal{H}_2^1)}\right). \tag{6}$$

Note that a log Bayes factor of $\mathit{LBF}_1 > 0$ indicates that a collaboration between learner k and ℓ is more likely than independent work for the considered collaboration model.[1]

[1] Note that the log Bayes factor coincides with the log likelihood ratio (LLR) as typically used in the statistical signal processing community.

TABLE I

PROBABILITY OF OBSERVING THE GRADED RESPONSE PAIR $(Y_{i,k}, Y_{i,\ell})$

$Y_{i,k}$	$Y_{i,\ell}$	$P(Y_{i,k}, Y_{i,\ell} \,\|\, p_{i,k}, p_{i,\ell}, \varepsilon_k, \varepsilon_\ell)$
0	0	$\bar{p}_{i,k}\bar{p}_{i,\ell} + \bar{p}_{i,k}p_{i,\ell}\bar{\varepsilon}_k\varepsilon_\ell + p_{i,k}\bar{p}_{i,\ell}\varepsilon_k\bar{\varepsilon}_\ell$
0	1	$\bar{p}_{i,k}p_{i,\ell}\bar{\varepsilon}_k\bar{\varepsilon}_\ell + p_{i,k}\bar{p}_{i,\ell}\varepsilon_k\varepsilon_\ell$
1	0	$p_{i,k}\bar{p}_{i,\ell}\bar{\varepsilon}_k\bar{\varepsilon}_\ell + \bar{p}_{i,k}p_{i,\ell}\varepsilon_k\varepsilon_\ell$
1	1	$p_{i,k}p_{i,\ell} + \bar{p}_{i,k}p_{i,\ell}\varepsilon_k\bar{\varepsilon}_\ell + p_{i,k}\bar{p}_{i,\ell}\bar{\varepsilon}_k\varepsilon_\ell$

B. Collaboration Test 2: Sequence Matching

The collaboration test detailed above solely considers the *number* of agreements in the graded responses between pairs of learners. In addition, Test 1 does not rely on the SPARFA parameters for the correlation hypothesis $\mathcal{H}_1^1$. Furthermore, the results in [8], [9] demonstrate that the use of particular cheating models is capable of improving the detection performance. One would surmise that considering the pair of response *sequences* in combination with a collaboration model that uses the SPARFA parameters for both hypotheses is likely to be more accurate in detecting collaboration. We therefore introduce a novel collaboration test that (i) uses a specific Bayesian collaboration model, (ii) relies directly on the graded responses sequences of learner k and ℓ, (iii) makes use of the SPARFA parameters for both hypotheses.

1) Collaboration Hypothesis: We start by defining the first hypothesis $\mathcal{H}_1^2$, which models the situation of observing the given pair of graded responses sequences for learner k and ℓ under a specific collaboration model. The model proposed here relies on the individual probabilities of learner k and ℓ succeeding in question i given the SPARFA parameters

$$p_{i,k} = \Phi(Z_{i,k}) \quad \text{and} \quad p_{i,\ell} = \Phi(Z_{i,\ell}). \tag{7}$$

In addition, the proposed collaboration model assumes that there are two *unknown* probabilities ε_k and ε_ℓ. The probability ε_k represents a "copy probability," which indicates the likelihood of learner k copying a graded response from learner ℓ; the probability ε_ℓ is defined analogously for learner ℓ.

With (7) and the copy probabilities ε_k and ε_ℓ, we can write the probability of observing the graded response pair $(Y_{i,k}, Y_{i,\ell})$ for question i. Table I summarizes all four cases; to simplify notation, we define $\bar{p}_{i,k} = 1 - p_{i,k}$, $\bar{p}_{i,\ell} = 1 - p_{i,\ell}$, $\bar{\varepsilon}_k = 1 - \varepsilon_k$, and $\bar{\varepsilon}_\ell = 1 - \varepsilon_\ell$. For instance, the graded response pair $(1, 1)$ can be either achieved if both learners get the i^{th} question right, or if learner k gets it wrong but copies the correct response from learner ℓ, or vice versa. The remaining cases in Tbl. I are obtained analogously.

Similar to Sec. III-A1, we follow a Bayesian approach and assume uniform priors on ε_k and ε_ℓ over the range $[0, 1]$ as

$$P(\mathcal{H}_1^2) = \int_0^1 \int_0^1 \prod_{i=1}^{Q} P(Y_{i,k}, Y_{i,\ell} \,|\, p_{i,k}, p_{i,\ell}, \varepsilon_k, \varepsilon_\ell) \mathrm{d}\varepsilon_k \mathrm{d}\varepsilon_\ell, \tag{8}$$

which corresponds to the probability of observing the pair of sequences of graded responses for all Q questions under the collaboration model specified in Tbl. I. Note that the double integral in (8) can be evaluated analytically. Numerical computation for a large number of questions Q, however, is non-trivial due to finite precision artifacts.

2) Independence Hypothesis: The probability of the second hypothesis $\mathcal{H}_2^2$ for Test 2 corresponds to the probability of the observed pair of graded response sequences, given the success probabilities (7) obtained from SPARFA, are assumed to be *independent*. This probability corresponds to

$$P(\mathcal{H}_2^2) = \prod_{i=1}^{Q} p_{i,k}^{Y_{i,k}} \bar{p}_{i,k}^{(1-Y_{i,k})} p_{i,\ell}^{Y_{i,\ell}} \bar{p}_{i,\ell}^{(1-Y_{i,\ell})}. \tag{9}$$

3) Log Bayes Factor: Given the probabilities (8) and (9) for the hypotheses $\mathcal{H}_1^2$ and $\mathcal{H}_1^2$, respectively, the log Bayes factor for Test 2 is given by:

$$\mathit{LBF}_2 = \log\left(\frac{P(\mathcal{H}_1^2)}{P(\mathcal{H}_2^2)}\right). \tag{10}$$

C. Bayesian Collaboration Detection

Both collaboration tests discussed above can be implemented directly into SPARFA-B as additional sampling steps. Concretely, we compute (6) and (10) at each iteration of the MCMC given the current estimates of $\mathbf{c}_j$, $\mathbf{w}_i$, and $\mu_i, \forall i, j$. The log Bayes factor can be equivalently converted to a posterior probability for each hypothesis, from which we can sample the hypotheses directly as part of the MCMC.

By using this fully Bayesian approach, our method leverages the full posterior distribution of the SPARFA parameters when computing the probability of each hypothesis. This method improves the robustness of our inference over classical approaches. We emphasize, however, that the proposed collaboration tests do not rely on Bayesian sampling and can easily be incorporated into classical models for educational data, such as the Rasch model [10] or item-response theory (IRT) [11].

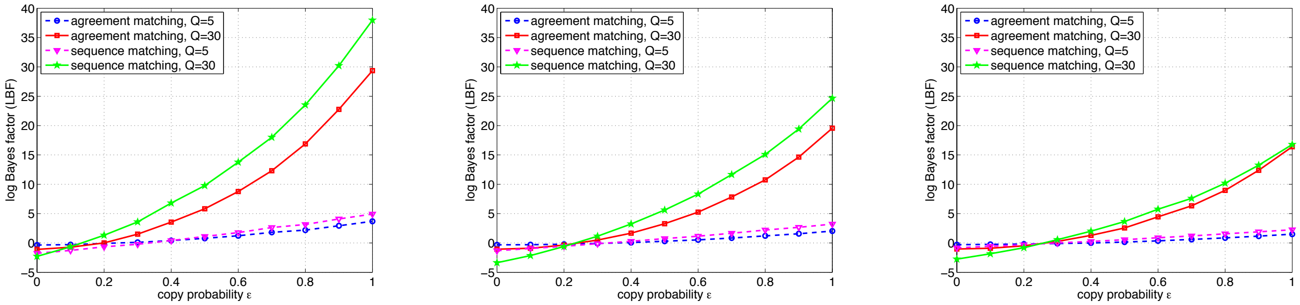
IV. EXPERIMENTS

A. Synthetic Experiments

As a first experiment, we examine the performance of Tests 1 and 2 for a simple synthetic scenario involving two learners and Q questions. Here, we avoid the use of SPARFA and characterize each learner by their probability of answering question i correctly. Let $p_{i,1}$ denote the probability that Learner 1 will answer question i correctly. We draw $p_{i,1} \sim \mathit{Beta}(\alpha, \beta)$, where α, β are tunable parameters. Now, with some copy probability ε, Learner 2 copies the answer provided by Learner 1; with probability $1 - \varepsilon$, Learner 2 correctly answers the question according to his own prior $p_{i,2}$.

We examine the ability of our tests to correctly detect the scenarios where Learner 2 is collaborating under a variety of priors. In Fig. 1(a) we consider copy probabilities ε swept in the range $[0, 1]$ in increments of 0.1 for the case where $p_{i,1} \sim \mathit{Beta}(3, 1)$ and $p_{i,2} \sim \mathit{Beta}(1, 3)$, corresponding to the case where Learner 1 has a relatively high and Learner 2 has a relatively low probability of success. In this scenario, it is relatively easy to detect collaboration. We further note that Test 2 outperforms Test 1 for all copy probabilities ε. Thus, Test 2 is preferred over Test 1 and related variants, such as [2].

We repeat the experiment for other distributions on $p_{i,1}$ and $p_{i,2}$. Figure 1(b) shows the case where $p_{i,1}, p_{i,2} \sim$



(a) $p_{i,1} \sim Beta(3,1)$ and $p_{i,2} \sim Beta(1,3)$

(b) $p_{i,1}, p_{i,2} \sim Beta(1,1)$

(c) $p_{i,1}, p_{i,2} \sim Beta(3,1)$

Fig. 1. Impact of collaboration test, prior distribution, and number of questions Q on the log Bayes factor for two synthetic learners with priors $p_{i,1}$, $p_{i,2}$. Sequence matching (Test 2) consistently outperforms agreement matching (Test 1) and larger question sets enable more accurate collaboration detection.

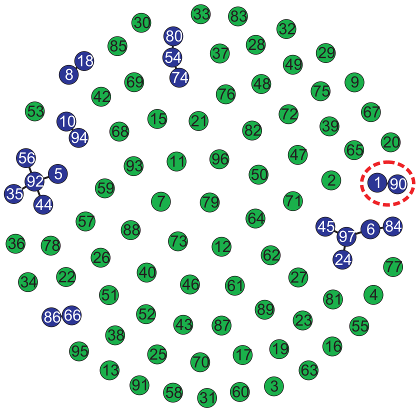


Fig. 2. Collaboration detection result for a take-home exam in an undergraduate electrical engineering course consisting of 97 learners answering 38 questions. The connected nodes correspond to learners for which the collaboration hypothesis under Test 2 was accepted with probability 0.98 or higher. Manual inspection of the open-form responses provided by Learners 1 and 90 (surrounded by the red dashed oval) reveals obvious collaboration.

$Beta(1,1)$, assuming a uniform distribution for both learners, and $p_{i,1}, p_{i,2} \sim Beta(3,1)$, assuming that both learners have a generally high probability of answering each question correctly. In all cases, the both methods accurately detect collaboration when applicable, while Test 2 outperforms Test 1.

B. Real-World Experiment

We finally analyze a real-world educational dataset taken from a low-level undergraduate course in electrical engineering administered on OpenStax Tutor[2]. This course consists of 97 learners answering a total of 203 questions, distributed over various homework assignments and exams. We examine collaboration among students in the final exam, which consisted of 38 questions. The final exam was administered as a take-home examination where learners were instructed not to collaborate with their peers.

In order to learn the SPARFA parameters, we use all questions in the entire course and then use Test 2 to extract the log Bayes factors for each pair of learners on the subset of questions corresponding to the final exam and use these to sample one of the two hypotheses at each iteration. We display the resulting collaboration graph in Fig. 2. Each learner is represented by a node; green nodes designate learners for which no significant evidence of collaboration was found; blue nodes correspond to learners for which collaboration is strongly suspected; edges between pairs of nodes indicate likely collaborations. We accept the collaboration hypothesis only if it is sampled in over 99% of the MCMC iterations.

A specific example concerns Learners 1 and 90. These two learners exhibited identical answer patterns on the exam, but have very different priors on successfully answering the questions. In addition, they both respond incorrectly to one question that is labeled by SPARFA as relatively easy. In order to prevent false accusations [12], we manually inspected the open-form responses available in OpenStax Tutor; these reveal that Learner 1 consistently provides a shortened response of all the responses provided by Learner 90. These preliminary results demonstrate that Test 2 provides accurate information regarding collaboration in real datasets.

References

[1] R. B. Frary, "Statistical detection of multiple-choice answer copying: Review and commentary," *Applied Measurement in Education*, vol. 6, no. 2, pp. 153–165, 1993.

[2] G. O. Wesolowsky, "Detection excessive similarity in answers on multiple choice exams," *Journal of Applied Statistics*, vol. 27, no. 7, pp. 909–921, 2000.

[3] L. Pappano, "The year of the MOOC," *The New York Times*, Nov. 4 2012.

[4] M. V. Levine and B. R. Donald, "Measuring the appropriatemess of multiple-choice test scores," *Journal of Educational Statistics*, vol. 4, no. 5, pp. 269–290, 1979.

[5] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk, "Sparse factor analysis for learning and content analytics," *Journal of Machine Learning Research*, 2013, submitted.

[6] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin, *Bayesian Data Analysis*, Chapman and Hall/CRC, 2003.

[7] M. Fernandez and S. Williams, "Closed-form expression for the poisson-binomial probability density function," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 46, no. 2, pp. 803–817, 2010.

[8] P. H. Kvam, "Using exam scores to estimate the prevalence of classroom cheating," *The American Statistican*, vol. 50, no. 3, pp. 238–242, 1996.

[9] S. W. Link and R. B. Day, "A theory of cheating," *Behavior Research Methods, Instruments, and Computers*, vol. 24, no. 2, pp. 311–316, 1992.

[10] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests*, MESA Press, 1993.

[11] Y. Bergner, S. Droschler, G. Kortemeyer, S. Rayyan, D. Seaton, and D. Pritchard, "Model-based collaborative filtering analysis of student response data: Machine-learning item response theory," in *Proceedings of the 5th International Conference on Educational Data Mining*, Chania, Greece, Jun. 2012, pp. 95–102.

[12] W. W. Chaffin, "Dangers in using the Z index for detection of cheating on tests," *Psychological Reports*, vol. 45, pp. 776–778, Dec. 1979.

[2] http://www.openstaxtutor.org