Collaboration-Type Identification in Educational Datasets

ANDREW E. WATERS
Department of Electrical
and Computer Engineering
Rice University
waters@sparfa.com

CHRISTOPH STUDER
School of Electrical
and Computer Engineering
Cornell University
studer@sparfa.com

RICHARD G. BARANIUK
Department of Electrical and Computer Engineering
Rice University
richb@sparfa.com

Identifying collaboration between learners in a course is an important challenge in education for two reasons: First, depending on the courses' rules, collaboration can be considered a form of cheating. Second, it helps one to more accurately evaluate each learner's competence. While such collaboration identification is already challenging in traditional classroom settings consisting of a small number of learners, the problem is greatly exacerbated in the context of both online courses or massively open online courses (MOOCs) where potentially thousands of learners have little or no contact with the course instructor. In this work, we propose a novel methodology for *collaboration-type identification*, which both *identifies* learners who are likely collaborating and also *classifies* the type of collaboration employed. Under a fully Bayesian setting, we infer the probability of learners' succeeding on a series of test items solely based on graded response data. We then use this information to jointly compute the likelihood that two learners were collaborating and what collaboration model (or type) was used. We demonstrate the efficacy of the proposed methods on both synthetic and real-world educational data; for the latter, the proposed methods find strong evidence of collaboration among learners in two non-collaborative take-home exams.

1 INTRODUCTION

1.1 TODAY'S CHALLENGES IN IDENTIFYING COLLABORATION

A well-known challenge for educators is identifying collaboration among learners (or students) in a course, test, or exam (Frary, 1993; Wesolowsky, 2000). This task is important for a number of reasons. The first and most obvious reason is that there are many educational scenarios in which collaboration is prohibited and considered a form of cheating. Identifying collaboration, in this instance, is important for maintaining fairness and academic integrity in a course. The second reason is that collaboration among learners complicates the accurate evaluation of a

learner's true level of competence. If, for example, a group of learners work together on a set of homework problems, then it is difficult to evaluate the competence of each individual learner as opposed to the competence of the group as a whole. This aspect is especially important in scenarios where learners are simply copying the responses of a single peer. In such a scenario, a series of correct answers among collaborative group members could lead to the conclusion that all learners have mastered the material when—in reality—only one learner in the group achieved proficiency.

Manually identifying collaboration among learners is difficult enough when the class size is moderately small, say 20–30 learners, where an instructor may have a reasonable knowledge about the aptitudes and habits of each particular learner. The problem is exacerbated as the class size increases to university-level classes with hundreds of learners. In the setting of online education, such as massive open online courses (MOOCs), a manual identification of learner collaboration (or cheating-through-collaboration) becomes infeasible, as potentially thousands of learners may be enrolled in a course, without ever having face-to-face interaction with an instructor (Pappano, 2012).

1.2 Automated collaboration identification

An alternative to manually identifying learners that collaborate is to rely on statistical methods that sift through learner response data automatically. Such data-driven methods look for patterns in learner answer data in order to identify potential collaborations. A naïve approach for automated identification 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 intrinsic difficulty of each test item or question (Levine and Donald, 1979; Wesolowsky, 2000).

In order to improve on such a naïve approach, a wealth of prior work exists on developing statistically principled methods of collaboration detection. Many of these methods focus on detecting the case of simple answer copying and a variety of statistical tests have been derived for this use case (Wollack, 2003; Sotaridona and Meijer, 2002; Sotaridona and Meijer, 2003; Wesolowsky, 2000). The proposed methods typically involve two steps: First, they estimate the probability that each learner will provide the correct response to each question by fitting models to both learners and questions. Second, they examine the actual answers provided by learners and compute a statistical measure on how likely the learner response patterns are to have arisen by chance. While such methods for collaboration identification have led to promising results, they possess a number of limitations:

- The first limitation of prior work in statistical collaboration detection is the overwhelming focus on multiple-choice testing. While multiple-choice exams are a fact of life in many settings, they are very limiting. For example, creating useful multiple choice questions is non-trivial and requires careful thought and planning (Haladyna et al., 2002; Rodriguez, 1997); this is especially true when creating effective wrong answers (lures) (Butler and Roediger, 2008). Additionally, the type of knowledge that can be tested on multiple choice exams is quite limited. This is especially true in fields such as STEM (science, technology, engineering, and mathematics) as well as economics (Becker and Johnston, 1999). Hence, automated collaboration identification methods should be able to analyze more general forms of learner response

data.

- The second limitation is in the explanatory weakness of the methods used in the existing collaboration identification literature for predicting the success of each learner on each question. Learning analytics (LA) is concerned with jointly estimating learner ability and question difficulty and using these estimates to predict future observations. More advanced LA methods have the potential to further improve the performance of automated collaboration identification.

- The third limitation is the use of simplistic models for how collaborative behavior between learners manifests in learner response data. Concretely, these methods are primarily concerned with the case of one learner copying the answers of another learner. The method of (Wesolowsky, 2000), for example, proposed the combination of point-estimates of the learner's success probability (which are estimated directly from multiple-choice test results) and a basic model on the number of correspondences that should arise between learners based on these success probabilities. However, this method does not take into account the variety of complex ways that learners could collaborate, ranging from simple copying to symbiotic collaboration. By employing a variety of models for different types of collaborative scenarios, one could hope to improve overall identification performance as well as provide valuable information to educators.

1.3 CONTRIBUTIONS

This paper develops a novel methodology for *collaboration-type identification*, which jointly identifies which learners engaged in collaboration and classifies the type of collaboration employed. A block diagram of our methodology is shown in Figure 1. Our approach overcomes the limitations of existing approaches described in Section 1.2. Concretely, we make the following four contributions, each one corresponding to one of the four blocks shown in Figure 1.

- *Generic learner response data*: Our methodology relies only on simple right/wrong response data as opposed to multiple-choice responses (which usually contain multiple options per question). This response model enables our approach to be applied to a much broader range of educational datasets than existing methods.

- *Improved learning analytics*: Our methodology utilizes the recently proposed SPARFA (short for SPArse Factor Analysis) model proposed in (Lan et al., 2013), which has been shown to have state-of-the-art performance for LA. We note, however, that the algorithms employed are not tied to any particular LA method. In fact, any LA method that estimates success probabilities for each learner–question pair can be utilized. Furthermore, the LA method used in combination with our approach can either provide point estimates or full posterior distributions of the success probabilities.

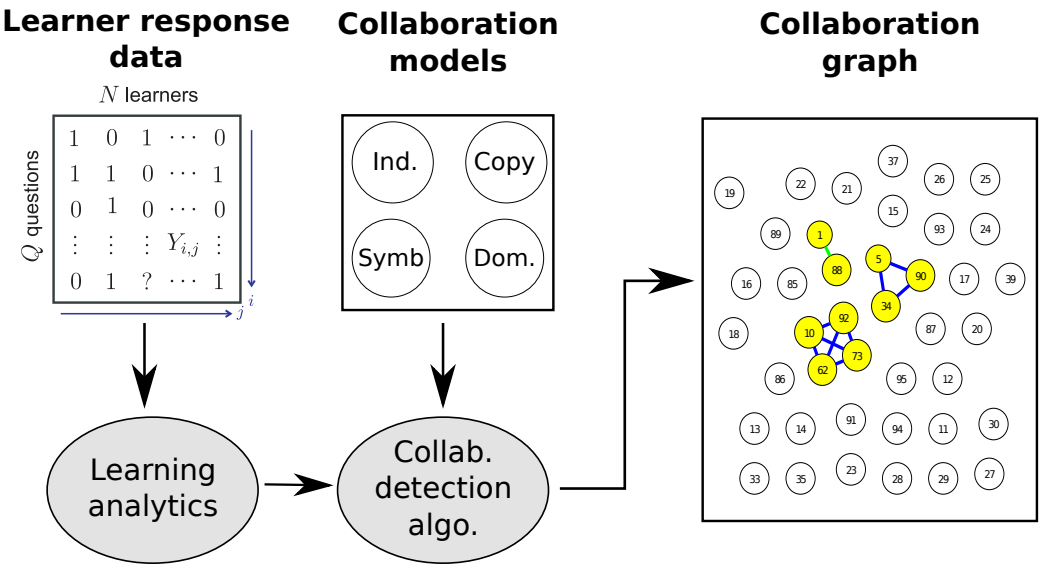

Figure 1: Block diagram for our proposed methodology for collaboration-type identification. The methodology consists of (i) *learning analytics* (Section 2) that model the success probabilities between learners and questions from *learner response data*, (ii) *collaboration models* (Section 3) for various types of real-world collaborative behavior, and (iii) *collaboration detection algorithms* (Section 4) that jointly identify collaboration and classify it according to one of the collaboration models. The *collaboration graph* summarizes the result of the collaboration detection algorithm graphically. In this example, the collaboration graph depicts collaboration on a final exam for an undergraduate electrical engineering course. Collaboration was detected among three groups of learners. In two cases, collaboration is classified as symbiotic (denoted by solid, dark blue lines). In the other case, collaboration was classified as parasitic copying (denoted by the dashed, green line). Further details of this real-world application example are given in Section 5.

- *Improved models for collaboration type*: Our methodology proposes four novel models for describing collaboration in real-world educational scenarios. By employing these models, our methodology provides superior performance and increased flexibility in representing real-world collaborative behavior.

- *Novel algorithms for collaboration-type identification*: Our methodology provides two novel algorithms for collaboration-type identification that fuse LA and collaboration models. These algorithms have superior performance compared to state-of-the-art algorithms for detecting collaboration in educational datasets.

1.4 Organization of the paper

The remainder of this paper is devoted to detailing our methodology for collaboration-type identification as depicted in Figure 1. In Section 2, we review existing algorithms for learning analytics, including a Bayesian variant of the approach of Rasch (Rasch, 1993) as well as the SPARFA framework (Lan et al., 2013). In Section 3, we develop probabilistic models for various types

of collaborative behavior between pairs of learners. In Section 4, we develop two novel algorithms for collaboration-type identification that make direct use of LA and collaboration models to search for likely pairs of learners engaged in collaboration. To demonstrate the efficacy of the proposed methodology, we validate our algorithms on both synthetic and real-world educational data in Section 5. We conclude in Section 6. The computational details of our methods are relegated to Appendices A, B, and C.

2 STATISTICAL APPROACHES FOR LEARNING ANALYTICS

As discussed above, naïve methods for collaboration identification that simply compare the pattern of right/wrong learner responses are prone to fail because they do not take into account the ability of each learner and the difficulty of each question. We use the term *learning analytics* (LA) to refer to methods that estimate the probability that a given learner will be successful on a given question. LA is typically accomplished by specifying models on both the learner abilities and the question difficulties. By fusing these, one can formulate a statistical model and develop corresponding algorithms for estimating the success probability of a given learner for each question. Recent approaches to LA enable us to distinguish scenarios where two learners have highly similar response patterns due to active collaboration as opposed to simply having similar abilities or a set of very easy/hard questions, where learners jointly succeed/fail with high probability.

The collaboration identification methodology developed in this work is generic in that one may use an arbitrary LA algorithm. In this section, we briefly summarize two of our preferred approaches to LA, namely Bayesian Rasch and SPARFA. We will assume that the datasets consist of learner response data for N learners and Q questions. These questions can be administered in a variety of settings, such as during an exam or as a series of homework problems. For the sake of simplicity of exposition, we further assume that our data is fully observed; that is, each learner responds to every question. Extending our methods to the case of partially observed data is straightforward.

2.1 BAYESIAN RASCH LEARNING ANALYTICS

The Rasch model (Rasch, 1960; Rasch, 1993) is a simple, yet powerful approach to LA. This model assumes that each learner can be adequately characterized by a single latent ability parameter, $c_i \in \mathbb{R}, i = 1, \ldots, N$. Large positive values of c_i indicate strong abilities, while large negative values indicate weak ability. Questions are also modeled by a single parameter $\mu_j \in \mathbb{R}, j = 1, \ldots, Q$, with large positive values indicating easy questions and large negative values indicating difficult questions. By defining the slack variables

$$Z_{i,j} = c_i + \mu_j, \;\; \forall i, j,$$

the Rasch model expresses the probability of user i answering question j correctly (with $Y_{i,j} = 1$) or incorrectly (with $Y_{i,j} = 0$) using

$$Y_{i,j} \sim \textit{Ber}\big(\Phi(Z_{i,j})\big), \;\; \forall i, j.$$

Here, *Ber(x)* denotes a Bernoulli distribution with mean x, while the function Φ denotes a link function that maps the slack variable $Z_{i,j}$ into a probability in $[0, 1]$. The conventional Rasch

model deploys the inverse logistic link function defined as

$$\Phi_{\text{log}}(x) = \frac{\exp(x)}{1+\exp(x)}.$$

Alternatively, one can use the inverse *probit* link function defined as

$$\Phi_{\text{pro}}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)\mathrm{d}t.$$

An advantage of the inverse probit link function (over the inverse logistic link) is that, when coupled with suitable prior probability distributions (i.e., Gaussian distributions) for each parameter, it enables efficient Markov chain Monte–Carlo (MCMC) methods based on Gibbs' sampling (Gelman et al., 1995). In what follows, we exclusively make use the inverse probit link function and use the simplified notation $\Phi(x) = \Phi_{\text{pro}}(x)$.

MCMC methods enable us to sample from the posterior distribution of each Rasch parameter of interest in a computationally tractable manner. Among these parameters is the latent success probability $p_{i,j} = \Phi(Z_{i,j})$, which denotes the probability of user j correctly responding to question i. Such a Rasch MCMC sampler will produce a series of samples from the posterior distribution of $p_{i,j}$ that will be useful when developing the collaboration-type detection algorithms in Section 4. We reserve the treatment of the full sampling details of the Rasch MCMC sampler for Appendix A.

2.2 Sparse factor analysis (SPARFA) learning analytics

Like the Rasch model, SPARFA (Lan et al., 2013) characterizes the success probability of a set of learners across multiple questions. In contrast, however, the SPARFA model assumes that there are K latent factors, referred to as *concepts*, that govern the learners' responses to these questions. In particular, SPARFA deploys the following model for the graded right/wrong response data:

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

Here, 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 positive large values indicate easy questions (as in the Rasch model).

Retrieving the parameters $\mathbf{c}_j$, $\mathbf{w}_i$, and μ_i from the set of graded learner responses $Y_{i,j}$ in (1) is, in general, an ill-posed inverse problem. To enable tractable inference, SPARFA assumes that the number of concepts K is small compared to both the number of learners and questions, i.e., $K \ll N, Q$. Furthermore, to both enable interpretability and alleviate problems with model identifiability, SPARFA imposes non-negativity and sparsity on the question–concept vectors $\mathbf{w}_i$. These assumptions are imposed on the SPARFA model through the selection of the prior distributions for each parameter of interest.

The SPARFA MCMC sampler extracts samples from the posterior distribution of each parameter of interest, with the primary concern for collaboration detection being the samples of $p_{i,j} = \Phi(Z_{i,j})$. As with the Rasch approach, we reserve the treatment of the full sampling details for the SPARFA MCMC sampler for Appendix B.

3 STATISTICAL MODELS FOR REAL-WORLD COLLABORATIVE BEHAVIOR

Learners in real-world educational settings typically use a variety of strategies for providing responses to questions. In many cases, learners simply work independently (i.e., without any collaboration). In other cases, weaker learners may simply copy the responses of a stronger classmate. In yet other cases, learners may work together collaboratively such that every learner within the group both participates and benefits. Learners may also defer to one trusted learner's answer, regardless of whether or not the trusted learner is actually correct. The fact that learners may collaborate on only a subset of questions further complicates automated collaboration identification.

By explicitly modeling *collaboration type*, one could hope to both provide valuable information regarding collaboration as well as to improve detection of collaborating learners. To this end, we propose four statistical collaboration models that capture a range of different scenarios. We use the notation $\mathcal{M}_m$ for $m = 1, \ldots, 4$ to refer to each model. We express our models probabilistically for a given pair of learners, i.e., learner u and learner v, and model the joint probability distribution of observing the set of answers $(Y_{i,u}, Y_{i,v})$. This joint distribution naturally depends first on the prior success probabilities $p_{i,u}$ and $p_{i,v}$ of both learners. In practice, these probabilities can be estimated via an LA approach such as the Bayesian Rasch model (see Section 2.1) or SPARFA (see Section 2.2). All models (with the exception of the independence model) are parameterized by a scalar variable $\varepsilon_1 \in [0, 1]$, which characterizes the probability that two learners will choose to collaborate on a given question. This parametrization enables us to capture the fact that learners might only collaborate on a subset of all Q questions. Additionally, two of the collaboration-type models we propose will utilize a second parameter, $\varepsilon_2 \in [0, 1]$; the meaning of this parameter is model specific and will be explained when applicable. To simplify notation, we will use the following definitions $\bar{\varepsilon}_1 = 1 - \varepsilon_1$ and $\bar{\varepsilon}_2 = 1 - \varepsilon_2$, as well as $\bar{p}_{i,u} = 1 - p_{i,u}$ and $\bar{p}_{i,v} = 1 - p_{i,v}$.

3.1 COLLABORATIVE MODELS

INDEPENDENCE MODEL $\mathcal{M}_1$ Under the independence model, a pair of learners is not collaborating. Instead, each learner answers the assigned questions independently. Hence, there are no parameters ε for this model. The probability of observing any answer sequence for two learners working independently is simply given by the product of the individual prior probabilities. For example, the graded response pair $(1, 0)$ is achieved if learner u provides a correct response to the i^{th} question, while learner v provides an incorrect response. This case occurs with probability $p_{i,u}\bar{p}_{i,v}$. The likelihoods for each of the four possible observed set of responses under the independence model for a given question are given in Table 1.

PARASITIC MODEL $\mathcal{M}_2$ Under the parasitic model of collaboration, only one of the two learners under consideration attempts to solve the question while the other learner simply copies the solution. The parasitic model is a two-parameter model with parameters ε_1 and ε_2. The first parameter ε_1 models the rate of collaboration, with a value of $\varepsilon_1 = 1$ denoting that the learner pair collaborates on every question; $\varepsilon_1 = 0$ denotes that the learners will never collaborate (thus, collapsing to the independence model). The second parameter ε_2 denotes to the probability that each learner will be selected to answer the question. A value of $\varepsilon_2 = 0$ denotes that learner u will always be the one selected to solve the question, while $\varepsilon_2 = 1$ denotes that learner v will always be the one selected. For example, observing the graded response pair $(0, 0)$ occurs

Table 1: Independence model $\mathcal{M}_1$

$Y_{i,u}$	$Y_{i,v}$	$P(Y_{i,u}, Y_{i,v} \mid p_{i,u}, p_{i,v}, \varepsilon_1, \varepsilon_2)$
0	0	$\bar{p}_{i,u}\bar{p}_{i,v}$
0	1	$\bar{p}_{i,u}p_{i,v}$
1	0	$p_{i,u}\bar{p}_{i,v}$
1	1	$p_{i,u}p_{i,v}$

Table 2: Parasitic model $\mathcal{M}_2$

$Y_{i,u}$	$Y_{i,v}$	$P(Y_{i,u}, Y_{i,v} \mid p_{i,u}, p_{i,v}, \varepsilon_1, \varepsilon_2)$
0	0	$\bar{p}_{i,u}\bar{p}_{i,v}\bar{\varepsilon}_1 + \varepsilon_1(\bar{p}_{i,u}\bar{\varepsilon}_2 + \bar{p}_{i,u}\varepsilon_2)$
0	1	$\bar{p}_{i,u}p_{i,v}\bar{\varepsilon}_1$
1	0	$p_{i,u}\bar{p}_{i,v}\bar{\varepsilon}_1$
1	1	$p_{i,u}p_{i,v}\bar{\varepsilon}_1 + \varepsilon_1(p_{i,u}\bar{\varepsilon}_2 + p_{i,u}\varepsilon_2)$

Table 3: Dominance model $\mathcal{M}_3$

$Y_{i,u}$	$Y_{i,v}$	$P(Y_{i,u}, Y_{i,v} \mid p_{i,u}, p_{i,v}, \varepsilon_1, \varepsilon_2)$
0	0	$\bar{p}_{i,u}\bar{p}_{i,v} + \bar{p}_{i,u}p_{i,v}\varepsilon_1\varepsilon_2 + p_{i,u}\bar{p}_{i,v}\varepsilon_1\bar{\varepsilon}_2$
0	1	$\bar{p}_{i,u}p_{i,v}\bar{\varepsilon}_1$
1	0	$p_{i,u}\bar{p}_{i,v}\bar{\varepsilon}_1$
1	1	$p_{i,u}p_{i,v} + \bar{p}_{i,u}p_{i,v}\varepsilon_1\bar{\varepsilon}_2 + p_{i,u}\bar{p}_{i,v}\varepsilon_1\varepsilon_2$

Table 4: OR model $\mathcal{M}_4$

$Y_{i,u}$	$Y_{i,v}$	$P(Y_{i,u}, Y_{i,v} \mid p_{i,u}, p_{i,v}, \varepsilon_1, \varepsilon_2)$
0	0	$\bar{p}_{i,u}\bar{p}_{i,v}$
0	1	$\bar{p}_{i,u}p_{i,v}\bar{\varepsilon}_1$
1	0	$p_{i,u}\bar{p}_{i,v}\bar{\varepsilon}_1$
1	1	$p_{i,u}p_{i,v} + \bar{p}_{i,u}p_{i,v}\varepsilon_1 + p_{i,u}\bar{p}_{i,v}\varepsilon_1$

in the event that (i) both learners do not collaborate on the question and both provide incorrect responses independently or (ii) both learners are collaborating on the question and that the learner chosen to solve the question does so incorrectly. The probability of this event is given by $\bar{p}_{i,u}\bar{p}_{i,v}\bar{\varepsilon}_1 + \varepsilon_1(\bar{p}_{i,u}\bar{\varepsilon}_2 + \bar{p}_{i,u}\varepsilon_2)$. The likelihood table for each of the four possible observed set of responses under the parasitic model is given in Table 2.

DOMINANCE MODEL $\mathcal{M}_3$ Under the dominance model, each learner works a question independently, after which each pair of learners discusses which of the two answers will be used. Under this model, each of the two learners attempts to convince the other to accept their response. The parameter ε_1 denotes the probability that the pair will collaborate on a given question (analogous to the parasitic model), while the second parameter ε_2 denotes the probability that learner u will convince learner v to adopt their response. For example, $\varepsilon_2 = 1$ implies that learner u will always convince learner v, while $\varepsilon_2 = 0$ indicates the opposite scenario. Under this model, observing the graded response pair $(0, 0)$ occurs in either the event that (i) both learners get the incorrect response (regardless of which learner dominates) or (ii) only one learner produces an incorrect responses, but convinces the other learner to accept the response. The probability of this event is given by $\bar{p}_{i,u}\bar{p}_{i,v} + \bar{p}_{i,u}p_{i,v}\varepsilon_1\varepsilon_2 + p_{i,u}\bar{p}_{i,v}\varepsilon_1\bar{\varepsilon}_2$. The likelihood table for each of the four possible observed set of responses under the dominance model is given in Table 3.

OR MODEL $\mathcal{M}_4$ Under the OR model, each learner may not be able to provide the correct response to a given question. However, they can identify the correct response if at least one of them is able to provide it. Thus, the learner pair will jointly provide the correct response if at least one of the learners succeeds. The name of this model derives from the Boolean OR function, which is 1 if either one or both inputs to the function are 1 and 0 otherwise. This model only uses a single parameter ε_1, which denotes the probability that the pair will collaborate on a given question (analogous to the Models $\mathcal{M}_2$ and $\mathcal{M}_3$). As an example, the graded response pair $(1, 1)$ occurs if (i) both learners produce the correct response (regardless of whether or not

they are collaborating on the question) or (ii) only one learner produces the correct response and the pair is actively collaborating on the given question. This probability of this scenario is given by $p_{i,u}p_{i,v} + \bar{p}_{i,u}p_{i,v}\varepsilon_1 + p_{i,u}\bar{p}_{i,v}\varepsilon_1$. The likelihood table for each of the four possible observed sets of responses under the OR model for question i is given in Table 4.

3.2 Discussion of pairwise collaboration models

Many more models can be developed to emulate various collaboration types. Such new models can be easily integrated into our methodology. We further note a number of correspondences that exist between the collaboration models detailed above. For example, the models $\mathcal{M}_2$, $\mathcal{M}_3$, and $\mathcal{M}_4$ are equivalent to $\mathcal{M}_1$ whenever $\varepsilon_1 = 0$, i.e., with the collaboration probability equal to zero. Further, models $\mathcal{M}_2$ and $\mathcal{M}_4$ are equivalent under the same value of ε_1 and whenever ε_2 is either 0 or 1.

One limitation of the collaboration models proposed above is that we have entirely decoupled the collaboration rate parameter ε_1 from the success probabilities $p_{i,u}$ and $p_{i,v}$. In real educational scenarios, learners might choose to collaborate when they perceive a large potential benefit. Learners may, for example, be less likely to collaborate on questions that they are likely to answer correctly (i.e., $p_{i,u}, p_{i,v}$ are large) and more likely to collaborate on questions that they are likely to answer incorrectly (i.e., $p_{i,u}, p_{i,v}$ are small). The development of such collaboration models is an interesting topic for future work.

4 Algorithms for Collaboration-Type Indentification

We now develop two novel algorithms for pairwise collaboration-type identification. Both algorithms jointly utilize learner–response data, an LA method, and a set of collaboration models to jointly detect and classify different types of collaboration in educational datasets (recall Figure 1).

The first algorithm, referred to as *sequential hypothesis testing* (SHT), uses a Bayesian hypothesis test first introduced in (Waters et al., 2013). This algorithm examines the joint answer sequence of a pair of learners and evaluates the likelihood that such patterns would arise independently (under model $\mathcal{M}_1$) or under one of the other collaboration model ($\mathcal{M}_2$, $\mathcal{M}_3$, or $\mathcal{M}_4$). The second algorithm, referred to as *collaborative model selection* (CMS), uses Bayesian model selection (Hoff, 2009) in order to *jointly* compute posterior distributions on the probability of learner response data arising under various collaboration models.

4.1 Sequence hypothesis testing (SHT)

SHT compares two hypotheses. The first hypothesis $\mathcal{H}_1$ corresponds to the case where learner u and v collaborate under a pre-defined collaboration-type model $\mathcal{M}_m, m \neq 1$, given the LA parameters. The second hypothesis $\mathcal{H}_2$ of SHT assumes that the number of agreements between the graded responses of learner u and v are a result of the independence model $\mathcal{M}_1$, given the LA parameters.

4.1.1 Collaboration hypothesis

We start by defining the first hypothesis $\mathcal{H}_1$, which models the situation of observing the given pair of graded responses sequences for learner u and v under the chosen collaboration model

$\mathcal{M}_m$, $m \neq 1$. Note that the SHT method can be utilized with any of the collaborative models introduced in Section 3. The model proposed here relies on the individual probabilities $p_{i,u}$ and $p_{i,v}$, which are the probabilities of learner u and v succeeding in question i given the LA parameters. For ease of exposition, we will proceed with our derivation for a two-parameter model such as $\mathcal{M}_2$ or $\mathcal{M}_3$ with parameters ε_1 and ε_2; the reduction to a single parameter model (such as the OR model) or the extension to a model with additional parameters is straightforward.

Assuming uniform priors on ε_1 and ε_2 over the range $[0, 1]$ our collaboration hypothesis for a given model $\mathcal{M}_m$ is simply given by

$$P(\mathcal{H}_1 \mid \mathcal{M}_m) = \int_0^1 \int_0^1 \prod_{i=1}^{Q} P(Y_{i,u}, Y_{i,v} \mid p_{i,u}, p_{i,v}, \varepsilon_1, \varepsilon_2, \mathcal{M}_m) \mathrm{d}\varepsilon_1 \mathrm{d}\varepsilon_2, \tag{2}$$

which corresponds to the probability of observing the pair of sequences of graded responses for all Q questions under the collaboration model $\mathcal{M}_m$. The quantity in (2) can be computed efficiently via convolution; we reserve the computation details for Appendix C.

4.1.2 Independence hypothesis

The probability of the second hypothesis $\mathcal{H}_2$ for SHT corresponds to the probability of the observed pair of graded response sequences, given the success probabilities $p_{i,u}$ and $p_{i,v}$ obtained under the independence model $\mathcal{M}_1$, i.e.,

$$P(\mathcal{H}_2) = \prod_{i=1}^{Q} p_{i,u}^{Y_{i,u}} \bar{p}_{i,u}^{(1-Y_{i,u})} p_{i,v}^{Y_{i,v}} \bar{p}_{i,v}^{(1-Y_{i,v})}. \tag{3}$$

Given the probabilities (2) and (3) for the hypotheses $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively, we can finally compute the *log Bayes factor*[1] for SHT for a given pair of learners as follows:

$$\textit{LBF} = \log\left(\frac{P(\mathcal{H}_1)}{P(\mathcal{H}_2)}\right). \tag{4}$$

A log Bayes factor greater than 0 indicates more evidence for the collaborative hypothesis (under the chosen model $\mathcal{M}_m$) than the independent hypothesis, while a log Bayes factor smaller than 0 indicates the reverse scenario. In general, however, a large value of the log Bayes factor is required when asserting that the evidence of collaboration is strong.

4.1.3 Discussion of SHT

The primary advantage of the SHT method is computational efficiency and flexibility. It can be used with simple point estimates of the learner success probabilities. Thus, it can be easily incorporated into classical approaches for LA, such as the standard (non-Bayesian) Rasch model (Rasch, 1993) or item-response theory (IRT) (Bergner et al., 2012). When utilized in this way, the log Bayes factor needs only be computed once for each pair of students, making it computationally very efficient.

SHT can also be used with a fully Bayesian LA approach (such as those detailed in Section 2 that provide full posterior distributions of the learner success probabilities). This is done by

[1] Under a uniform prior, the log Bayes factor is called the log likelihood ratio (LLR) in the statistical signal processing community.

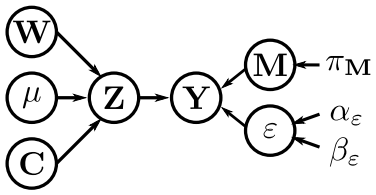

Figure 2: Graphical model for collaborative model selection (CMS).

adding the computation of (4) as an additional sampling step of the MCMC sampler. Concretely, we compute (4) at each iteration of the MCMC sampler given the current estimates of $p_{i,u}$ and $p_{i,v}$, $\forall i, u, v$. 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 sampler. This approach has the advantage of improving the robustness of our inference over classical approaches, albeit at higher computational cost.

One restriction of our method is that SHT compares the independence model $\mathcal{M}_1$ against exactly one other collaboration model $\mathcal{M}_m$. One could, however, consider testing multiple models simultaneously by using a form of Bonferroni correction to control the family-wise error rate (Westfall et al., 1997). The approach proposed in the next section avoids such necessary corrections by means of Bayesian model selection.

4.2 Fully Bayesian Collaborative Model Selection

We now turn to a collaboration-type indentification method based on Bayesian model selection (Hoff, 2009). This method allows us to *jointly* explore multiple collaboration models (types) and to extract the associated model parameters in an efficient way in order to find configurations that best explain the observed data. The result will provide estimates of the full posterior distributions for each collaboration model and each parameter of interest. We dub this method *collaborative model selection* (CMS).

4.2.1 Generative model for CMS

We first present the complete generative model for the pairwise collaborative model and state all necessary prior distributions. This will enable efficient MCMC sampling methods for estimating the relevant posterior distributions.

The full generative model is illustrated in Figure 2 for the case of the SPARFA LA model (the equivalent Rasch-based model is obtained by removing the node $\mathbf{W}$ and replacing $\mathbf{C}$ with the vector $\mathbf{c}$). By symmetry of the proposed collaboration models, collaboration between each pair of N learners can be specified with $D = (N^2 - N)/2$ total models and corresponding sets of the associated model parameters. We will use the quantity M_d to denote the random variable that indexes the collaboration model for learner pair d; the notation $\boldsymbol{\varepsilon}_d$ denotes the random vector of model parameters for learner pair d. For the collaborative model index M_d we assume a discrete prior $\pi_{m,d}$ such that $\sum_{m=1}^{4} \pi_{m,d} = 1$ for all d. For the elements of the parameter vector $\boldsymbol{\varepsilon}_d$, we assume a Beta-distributed prior $\mathit{Beta}(\alpha_\varepsilon, \beta_\varepsilon)$. Generation of the latent variables in $\mathbf{Z}$ is done for either the Rasch or SPARFA LA model as discussed in Sections 2.1 and 2.2, respectively. Finally, the observed learner–response matrix $\mathbf{Y}$ for learner u and v is generated jointly as detailed in Section 3 given the model type index M_d and the associated

model parameters $\boldsymbol{\varepsilon}_d$.

4.2.2 MCMC sampling for collaboration type detection

Given the graded response data matrix $\mathbf{Y}$ along with the prior distribution on M_d and $\boldsymbol{\varepsilon}_d$, we wish to estimate the posterior distribution of each model index along with its respective parameters for each pair of learners $d = 1, \ldots, D$. Doing this will allow us to infer (i) which pairs of learners are collaborating, (ii) what type of collaborative model are they using, and (iii) how strong the evidence is for these assertions.

We use Bayesian model selection techniques (Hoff, 2009) to efficiently search the space of possible models and model parameters for configurations that best explain the observed data $\mathbf{Y}$. Full conditional posteriors for the models and model parameters, however, are not available in closed form, rendering Gibbs' sampling infeasible. Thus, we make use of a suitable Metropolis-Hastings step (Gelman et al., 1995). Specifically, assume that at iteration t of the MCMC sampler and for a specific pair of learners d, we have a model sample M_d^t parametrized by $\boldsymbol{\varepsilon}_d^t$. The Metropolis-Hastings step proceeds by proposing a new model M_d^{t+1} with parameters $\boldsymbol{\varepsilon}_d^{t+1}$ via some proposal distribution $q(M_d^{t+1}, \boldsymbol{\varepsilon}_d^{t+1} | M_d^t, \boldsymbol{\varepsilon}_d^t)$. We will utilize a proposal distribution of the following form:

$$\begin{aligned} q(M_i^{t+1}, \boldsymbol{\varepsilon}_i^{t+1} | M_i^t, \boldsymbol{\varepsilon}_i^t) &= q_\varepsilon(\boldsymbol{\varepsilon}_i^{t+1} | M_i^{t+1}, M_i^t, \boldsymbol{\varepsilon}_i^t) q_M(M_i^{t+1} | M_i^t, \boldsymbol{\varepsilon}_i^t) \\ &= q_\varepsilon(\boldsymbol{\varepsilon}_i^{t+1} | M_i^{t+1}, M_i^t, \boldsymbol{\varepsilon}_i^t) q_M(M_i^{t+1} | M_i^t). \end{aligned}$$

In words, we (i) split the proposal into a model component and model parameters component and (ii) make use of a proposal for the model M_d that is independent of the model parameters $\boldsymbol{\varepsilon}_d$. We implement this proposal in two steps: First, we propose $M_d^{t+1} \sim q_M(M_d^{t+1} | M_d^t)$. Note that there are many choices for this proposal; we will make use of the following simple one given by

$$p(M_d^{t+1} = \mathcal{M}^{t+1} | M_d^t = \mathcal{M}^t) = \begin{cases} \gamma, & \text{if } \mathcal{M}^{t+1} = \mathcal{M}^t \\ \frac{1-\gamma}{|\mathcal{M}|-1}, & \text{if } \mathcal{M}^{t+1} \neq \mathcal{M}^t. \end{cases} \tag{5}$$

Here, $\gamma \in (0, 1)$ is a user-defined tuning parameter. In words, with probability γ the MCMC sampler will retain the previous model; otherwise, one from the remaining $|\mathcal{M}| - 1$ models is proposed uniformly. The proposal for the parameters $\boldsymbol{\varepsilon}_d^{t+1}$ takes the following form:

$$q(\boldsymbol{\varepsilon}_d^{t+1} | M_d^{t+1}, M_d^t, \boldsymbol{\varepsilon}_d^t) = \begin{cases} \delta_0, & \text{if } M_d^{t+1} = \mathcal{M}_1 \\ q_\beta(\boldsymbol{\varepsilon}_d^{t+1} | M_d^{t+1}, M_d^t, \boldsymbol{\varepsilon}_d^t), & \text{if } M_d^{t+1} = \mathcal{M}^t, \\ \pi_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon} | \alpha_\varepsilon, \beta_\varepsilon), & \text{otherwise,} \end{cases} \tag{6}$$

where δ_0 corresponds to a point-mass at 0; the distribution q_β corresponds to a random walk proposal on the interval $[0, 1]$ defined by

$$q_\beta(a|b) = \textit{Beta}(cb, c(1-b)), \tag{7}$$

where $c > 0$ is a tuning parameter. In words, the sampling of $\boldsymbol{\varepsilon}_d^{t+1}$ (i) is performed via a random walk when the model remains unchanged, (ii) is drawn directly from the prior $\pi_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon} | \alpha_\varepsilon, \beta_\varepsilon)$ when a new model non-independent model is proposed, and (iii) is set to 0 when the model changes to the independence model (since this model has no parameters $\boldsymbol{\varepsilon}$, it is simply set to 0 for

convenience). Note that it can be shown that the mean of the proposal distribution q_β is simply b (the previous value used) while the variance is $\frac{c^2(b-b^2)}{c^2+c+1}$, which tends to zero as b approaches either 0 or 1.

After proposing the new model $\{M_d^{t+1}, \boldsymbol{\varepsilon}_d^{t+1}\}$ via (5)–(7) we accept the proposal with a probability r as

$$r = \min\left\{1, \frac{p(\mathbf{Y}|M_d^{t+1}, \boldsymbol{\varepsilon}_d^{t+1})\pi(M_d^{t+1})\pi(\boldsymbol{\varepsilon}_d^{t+1}|\alpha_\varepsilon, \beta_\varepsilon)q_\varepsilon(\boldsymbol{\varepsilon}_d^{t}|M_d^{t}, M_d^{t+1}, \boldsymbol{\varepsilon}_d^{t+1})q_M(M_d^{t}|M_d^{t+1}))}{p(\mathbf{Y}|M_d^{t}, \boldsymbol{\varepsilon}_d^{t})\pi(M_d^{t})\pi(\boldsymbol{\varepsilon}_d^{t}|\alpha_\varepsilon, \beta_\varepsilon)q_\varepsilon(\boldsymbol{\varepsilon}_d^{t+1}|M_d^{t+1}, M_d^{t}, \boldsymbol{\varepsilon}_d^{t})q_M(M_d^{t+1}|M_d^{t}))}\right\}.$$

The accept/reject decision is computed individually for each learner pair $d = 1, \ldots, D$.

4.2.3 Discussion of CMS

The primary advantage of CMS is that it can jointly search across all collaborative models for each pair of learners in the dataset. This comes at the price of additional computational complexity, as a new set of models and model parameters must be proposed at each iteration of the MCMC.

We note that while our fully Bayesian method for collaboration detection uses the success probability matrix $\Phi(\mathbf{Z})$ when exploring new collaboration models, those models do *not influence* the sampling of $\mathbf{Z}$ itself. This is due to the structure of the model we have proposed, as $\mathbf{Y}$ separates the LA portion of the MCMC from the CMS portion. This assumption is similar to the work in (Wesolowsky, 2000) which computes success probabilities for each learner–question pair based only on the data $\mathbf{Y}$ regardless of the evidence of collaboration.

The model depicted in Figure 2 could be augmented in a way that enables us to propose a new posterior distribution for $\mathbf{Z}$ where the current belief about the collaborative models will influence our beliefs about learner ability. For example, such a model could be accomplished by proposing an additional latent variable for the answer that a learner would have provided had they not been in collaboration; this would enable us to automatically temper our beliefs about learner ability in the event that we believe that they are involved in collaboration. We will leave such an approach for future work.

5 Experiments

We now validate the performance our proposed methodology. We first examine the identification capabilities using synthetic data with a known ground truth. Following this, we showcase the capabilities of our methods on several real-world educational datasets.

5.1 Synthetic experiments

We first validate the performance of our methodology using synthetic test data using both the SHT and CMS algorithms. We furthermore compare against two other methods. The first is the state-of-the-art method collaboration identification method developed in (Wesolowsky, 2000), which was designed specifically for handling responses to multiple choice exams, where one is interested in the specific option chosen by the pair of learners. Since we are interested in detecting collaboration given only binary (right/wrong) graded responses, we need to first modify the method (Wesolowsky, 2000) accordingly. Concretely, we set their term v_i indicating the number of wrong options for question i to 1, meaning that all wrong answers are treated equally. The remaining aspects of this method are left unchanged. The second method that we compare against

is the *agreement hypothesis testing* method proposed in (Waters et al., 2013), which we will call AHT for short. This method utilizes a Bayesian hypothesis test similar to SHT that compares the likelihood of the independence model $\mathcal{M}_1$ relative to a simple collaboration model in which two learners choose to agree in their answer patterns with some arbitrary probability δ. We refer the interested reader to (Waters et al., 2013) for further details.

5.1.1 Performance metrics

In order to evaluate collaboration identification performance, we will examine how well the considered methods identify learners who collaborate relative to learners who work independently. Each of the four methods naturally outputs a *collaboration metric* related to the probability of collaboration between each pair of learners:

- *Bayesian Hypothesis Tests (AHT and SHT):* For each learner pair, the collaboration metric is given by the log Bayes' factor.

- *Bayesian Model Selection (CMS):* For each learner pair, we first threshold on the posterior probability that two learners worked under a collaborative model and then, we rank them according to the posterior mean of ε_1.

- *Wesolowsky's Method:* For each learner pair, the collaboration metric is given by the Z-score (see (Wesolowsky, 2000) for the details).

By sorting the output metrics in ascending order, a course instructor can easily see which pairs of learners in a class are most likely engaging in collaborative behavior. To this end, let ξ denote the output vector of pairwise metrics for each learner pair for a given algorithm. Sorting the entries of ξ from smallest to largest, we can compute a normalized percentile ranking for each pair of learners. Let $\mathcal{I}_d$ denote the index of learner pair d in this sorted vector. The normalized percentile ranking is then given simply by

$$\mathcal{P}_d = \frac{\mathcal{I}_d}{D}, \ d = 1, \ldots, D, \tag{8}$$

with larger values of $\mathcal{P}_d$ denoting higher likelihood of collaboration relative to the rest of the entire learner population.

5.1.2 Algorithm comparison on synthetic data

As a first synthetic experiment, we consider a class of $N = 30$ learners (with $D = 435$ unique learner pairs) answering $Q = 50$ questions. Learner abilities are initially generated via the SPARFA model. We select three pairs of learners who will work together collaboratively, one pair for each model $\mathcal{M}_2$, $\mathcal{M}_3$, and $\mathcal{M}_4$ as defined in Section 2. Each pair has a per-question collaboration probability $\varepsilon_1 = 0.75$, while the value for ε_2 for each of the two-parameter models is set to 0 for simplicity. The answers for the collaborating learners are generated according the appropriate collaboration model. The remainder of the learner pairs work independently, and

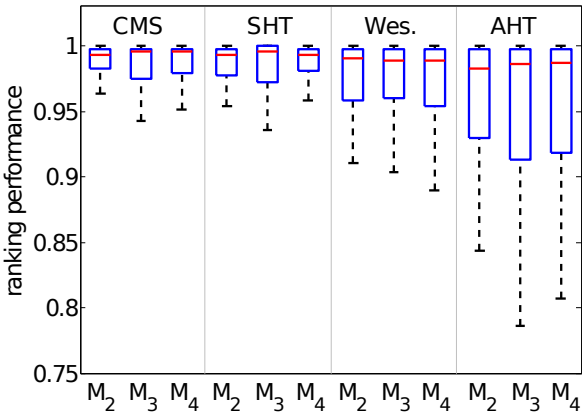

Figure 3: Normalized percentile ranking performance for all four collaboration methods with a synthetic dataset consisting of $N = 30$ learners and $Q = 50$ questions. Three learner pairs are engaged in collaboration, one for each of the collaborative models, with a per-question collaboration probability $\varepsilon_1 = 0.75$. Larger values indicate better identification performance. The CMS method achieves the best collaboration identification performance, followed by SHT, Wesolowsky's method (denoted by "Wes."), and AHT, respectively.

their answers are generated according to the SPARFA model via (1). We then deploy CMS, SHT, Wesolowsky's method (denoted by "Wes." in Figure 3), and AHT, and we compute the normalized percentile ranking for each learner pair according to (8). We repeat this experiment over 100 trials and present the normalized percentile ranking statistics for the collaborating learner pairs for each collaboration model and each algorithm as a box-whisker plot.

From Figure 3, we see that CMS outperforms all other methods, both in the average and standard deviation of the normalized percentile ranking. CMS is followed by SHT, Wesolowsky's method, and AHT. Since our proposed CMS method shows the best collaboration identification performance, we focus exclusively on this method in the remaining synthetic experiments.

5.1.3 Performance evaluation for CMS over multiple parameters

We now examine the performance trends of the CMS method for a varying number of questions as well as the collaboration probability ε_1. First, we generate data for $N = 50$ learners with $Q = 50$ questions and sweep the collaboration probability $\varepsilon_1 \in \{0.25, 0.5, 0.75, 1.0\}$ and repeat this experiment over 100 trials. We again examine performance using the normalized percentile ranking (8) and display the results in Figure 4(a). We can see that the performance is excellent for collaboration probabilities as low as $\varepsilon_1 = 0.5$, meaning that learners were expected to collaborate on only every-other question. Second, we fix $\varepsilon_1 = 0.5$ and sweep $Q \in \{25, 50, 75, 100\}$. The result is displayed in Figure 4(b). We see that the proposed CMS method achieves excellent identification performance for $Q \geq 50$ questions.

5.2 Real-world experiments

We now turn to two real-world educational datasets. Specifically, we analyze datasets taken from undergraduate courses in electrical and computer engineering administered on OpenStax

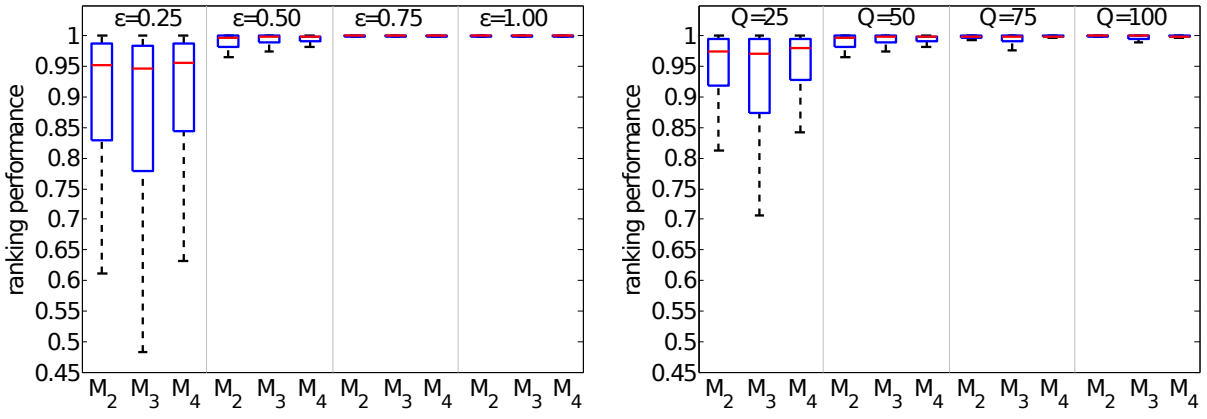

Figure 4: Collaboration-type identification performance for the collaborative model selection (CMS) approach with a synthetic $N = 50$ learner dataset. (a) impact of varying collaboration probabilities $\varepsilon_1 \in \{0.25, 0.5, 0.75, 1.0\}$ for $Q = 50$ questions; (b) impact of numbers of questions for $Q \in \{25, 50, 75, 100\}$ with collaboration probability $\varepsilon_1 = 0.5$.

Tutor.[2]

5.2.1 Undergraduate signal processing course

We first identify collaboration on homework assignments in the course up to the first midterm examination. One interesting aspect of this course is that learners were encouraged to work together on all homework assignments, albeit with some restrictions. Concretely, each learner was assigned into a group of 2-to-4 learners with whom they were allowed to actively collaborate on homework assignments. Learners within each group were free to discuss each homework problem as well as its solution with any members of their group, though each learner was required to submit their own homework solutions for final grading. Learners were, however, not required to collaborate; the only restriction was that any collaboration with other learners was to be confined to the assigned homework group. Collaborating outside of the assigned homework group was considered cheating.

This particular setting presents an interesting test case for our method since we have a rough ground truth with which to compare our results. We examine the performance of CMS and the method of Wesolowsky on all homework assignments up to the first midterm exam; a total of $Q = 50$ questions and $N = 38$ learners. We further include all question–responses to the midterm (14 additional responses) in extracting the SPARFA parameters, though these questions were excluded from the collaboration detection algorithm. The data is especially challenging since learners were given ample time to solve and discuss homework problems. Because of this, most responses given on homework problems were correct. As a consequence, an extremely high degree of similarity between answer patterns is required for collaboration to be considered probable.

For CMS we posit collaborative connections between learner pairs for whom $M_d \neq 1$ (i.e., from which the independence model $\mathcal{M}_1$ is excluded) in more than 90% of MCMC it-

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

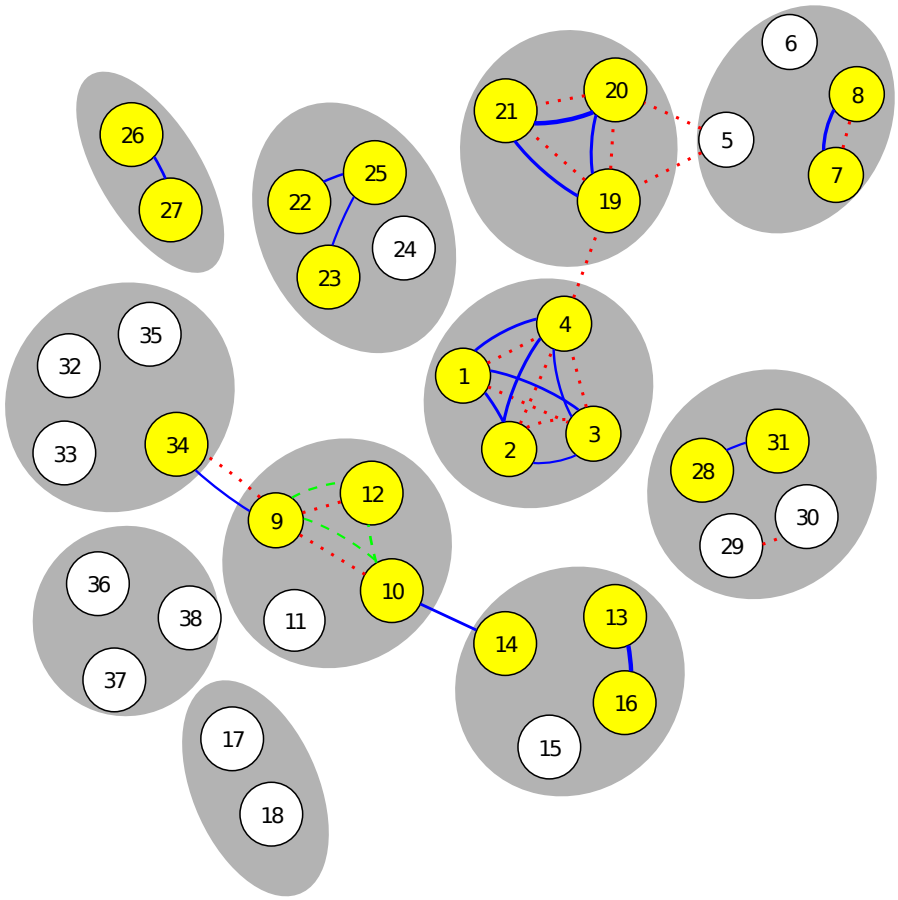

Figure 5: Collaboration-type identification result for the Bayesian model selection method for the first set of homework assignments in the undergraduate signal processing class dataset. The data consists of 38 learners answering 50 homework questions plus 14 midterm exam questions. Grey ellipses designate the assigned homework groups. Dashed green lines denote parasitic collaborations, while solid blue lines denote symbiotic collaborations detected by CMS. Dotted red lines denote the connections found using Wesolowsky's method, which, in general, finds fewer ground truth connections than the CMS method.

erations and for whom the posterior mean of ε_1 was greater than 0.4. The Z-score threshold for Wesolowsky's method was adjusted manually to provide the best match to the ground truth. We display the corresponding results in Figure 5. Dotted red lines denote connections detected under Wesolowsky's method. For the Bayesian model selection method, blue solid lines denote detections under symbiotic (OR) model, whereas dashed green lines show detections under the parasitic model.

Most collaborative types found using CMS for this dataset are of the OR type. An exception is the group $\{9, 10, 12\}$, for which the parasitic copying model was proposed most frequently. Examination of the answer pattern for these learners show that while the joint answer patterns for these learners are very similar on the homework assignments, Learners 10 and 12 perform poorly on the midterm relative to Learner 9. Thus, the algorithm assumes that their success in the

homework is more a consequence of copying the responses of Learner 9 rather than because of their mastery of the subject material. We additionally note the collaborative connection between Learners 9 and 24 as well as between Learners 10 and 14. These connections arise due to high similarity in the homework answering patterns which are also quite different from the rest of the collaborative group. Interestingly, Wesolowsky's method also found strong evidence of collaborations between Learners 9 and 24; this method, however, failed to reveal three of the intra-group collaborations found by our proposed CMS method. In the following, we omit further comparisons with Wesolowsky's method for the sake of brevity.

As a second experiment with the same undergraduate signal processing course, we consider collaboration identification on the final exam, which was administered as a take-home test. During this examination, learners were instructed *not to collaborate* or discuss their results with any other learners in the class. The final exam consisted of 24 questions. We deploy CMS using all questions in the course (a total of 147 questions) to extract the SPARFA parameters and search for collaboration only on the questions on the final exam. We jointly threshold on the posterior mean of ε_1 and the proportion of MCMC samples that indicated a non-independent collaborative model for each learner pair to arrive at the collaboration graph of Figure 6. We find strong evidence of collaboration between the learner pair $\{14, 38\}$ under the symbiotic collaboration model. Interestingly, Learner 14 was also detected in the previous experiment as a learner working outside of his collaborative group on the course homework assignments. Both learners provided correct responses to each question on the final exam, although their previous performance in the course would lead one to expect otherwise. To prevent false accusations (see, e.g., (Chaffin, 1979) for a discussion on this matter), we examined their open form responses available in OpenStax Tutor and found a remarkable similarity in the text of their answers; this observation further strengthens our belief about their collaboration.

5.2.2 Final exam of an undergraduate computer engineering course

This course consists of 97 learners who completed the course answering a total of 203 questions, distributed over various homework assignments and three exams. We examine collaboration among learners in the final exam, which consists of 38 questions. As was the case with the signal processing final exam, the computer engineering final exam was administered as a take-home examination where learners were instructed not to collaborate with their peers. In order to extract the SPARFA parameters, we use all questions administered during the entire course and then use CMS to extract the posterior distributions for each pair of learners on the subset of questions corresponding to the final exam. We jointly threshold the posterior probability of non-independent collaboration as well as the posterior mean of ε_1. We display the result in Figure 7, where dashed green lines correspond to parasitic collaboration ($\mathcal{M}_2$) and solid blue lines denoting symbiotic collaboration ($\mathcal{M}_4$). Note that no collaborations were detected under the dominance model ($\mathcal{M}_3$).

All three groups of learners in Figure 7 for whom we identify collaboration have identical answer patterns. The group $\{10, 92, 62, 73\}$ provides the correct response to every question on the exam. The group $\{5, 34, 90\}$ jointly miss only one question which is estimated by SPARFA to be in the mid to high range of difficulty. The group $\{1, 88\}$ jointly miss the same two questions, one of these being found by SPARFA to exhibit low intrinsic difficulty. We manually inspected the open-form responses available in OpenStax Tutor of all learners in the identified groups to prevent false accusations. We found that there is some diversity in the responses provided by the

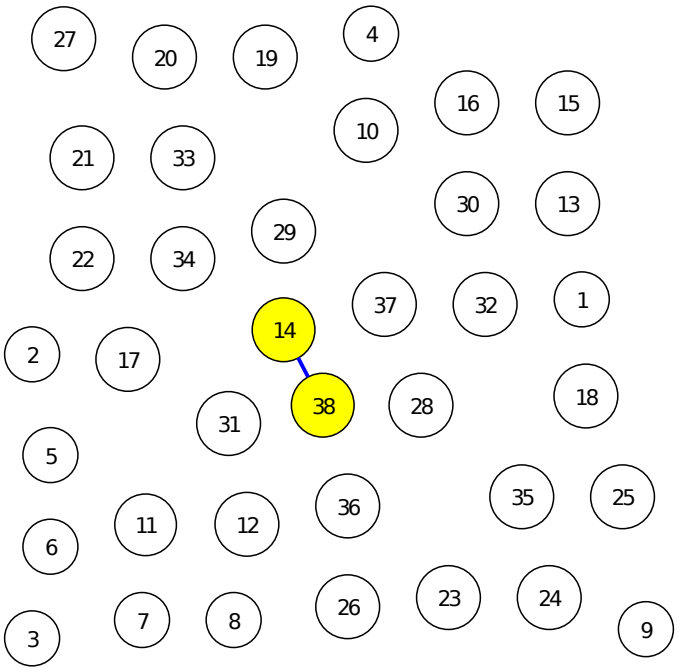

Figure 6: Collaboration-type identification result for a take-home exam in an undergraduate electrical engineering course consisting of 38 learners answering 24 questions. The connected nodes correspond to learners for which the collaboration hypothesis. Manual inspection of the open-form responses provided by Learners 14 and 38 further strengthens the collaboration hypothesis.

group $\{10, 62, 73, 92\}$. This, coupled with the fact that each of the learners have only managed to perform slightly better than what SPARFA would predict, allows us to reasonably exclude this group from further scrutiny. By contrast, the answer patterns for the other groups reveal strong evidence of collaboration due to very similar wording and grammar. This is especially true for the pair $\{1, 88\}$; as it can be seen from Table 5, Learner 1 consistently provides a shortened version of the responses provided by Learner 88, including those answered incorrectly.[3]

6 Conclusions

We have developed new methods for pairwise collaboration-type identification in large educational datasets, where the objective is to both identify which learners work together and classify the type of collaboration employed. Our framework combines sophisticated approaches to learning analytics (LA) with new models for real-world collaboration and employs powerful algorithms that fuse the two to search for active collaborations among all pairs of learners. We have validated our methodology on both synthetic and real-world educational data and have shown

[3]We analyzed the same computer engineering dataset using a different collaboration detection framework in (Waters et al., 2013). We omitted two learners in this work due to their failure to submit all homework assignments. Thus, learner indices in these two papers differ. As an example, Learners 1 and 88 in this work thus correspond to Learners 1 and 90 in (Waters et al., 2013).

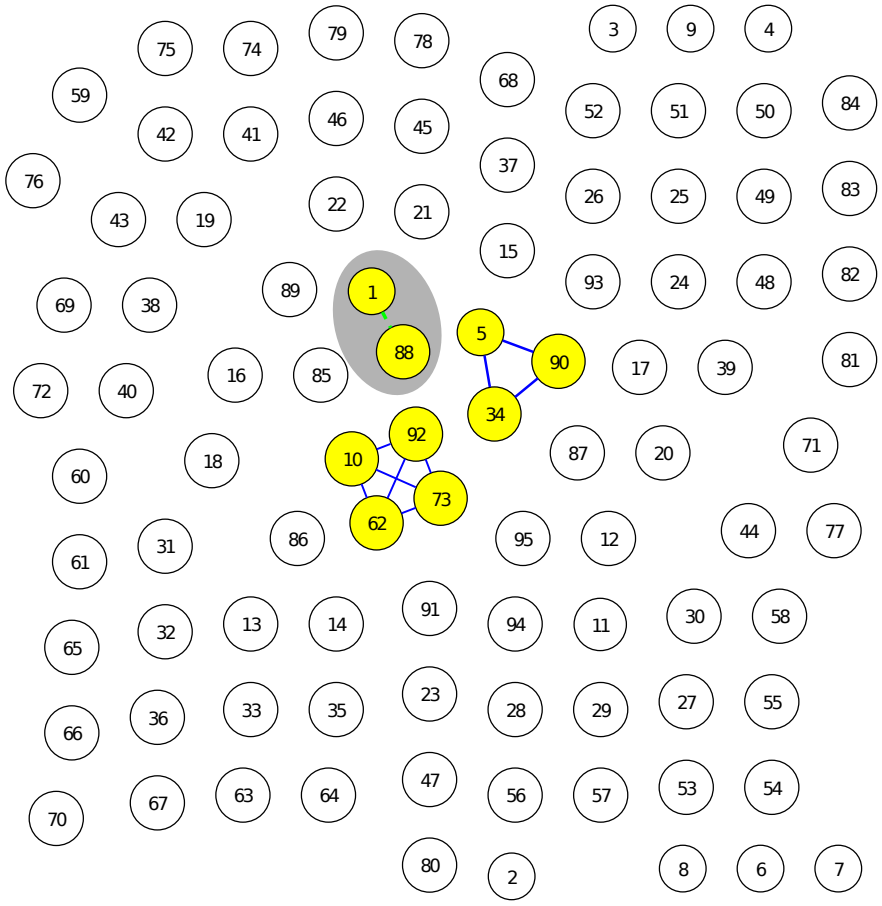

Figure 7: Collaboration identification result for a take-home exam in an undergraduate electrical engineering course consisting of 97 learners answering 38 questions. The connected nodes correspond to learners identified by CMS to be collaborating, with dashed green lines denoting one-side copying and solid blue lines denoting symbiotic collaboration. Manual inspection of the open-form responses provided by Learners 1 and 88 (highlighted by a gray oval) reveals obvious collaboration.

Table 5: Selected responses of Learners 1 and 88 in the non-collaborative take-home exam. Answer similarities are highlighted in black. Both learners responded to the last item incorrectly.

Learner 1	Learner 88
double char integral	**double** cannot be used to base a switch decision int and **char** can be used as they are of **integral** type
When the **name field is defined**.	student.name would correctly access the **name field** if the name of the student struct object **declared** is student.
A	This prints the ASCII character associated with the decimal value 65, which is **A**
5	The value of x would be **5** ; it would truncate the digits right of the decimal point.
−2147483648−+2147483647	Its potential range is **−2,147,483,648− +2,147,483,647** (using the typical 32 bit representation).

that they significantly outperform the state-of-the-art methods available in the open literature. Additionally, we detected several cases of non-permissible collaboration (which is considered cheating) on both homework assignments and examinations in two undergraduate-level courses.

The collaboration rankings that our method provides can greatly facilitate collaboration identification as it provides a good (and small) set of candidates that need to be evaluated in greater depth (with respect to collaborative behavior) by an instructor. This advantage reduces the instructor's workload and promotes fairness in educational settings.

One interesting avenue for future work involves modeling more complicated social groups among learners. In particular, extending the capability of collaboration detection methods beyond pairwise collaboration is useful in real-world educational scenarios in which learners often work in larger groups with complicated social dynamics. Another avenue for future work consists of using collaboration detection methods to "denoise" or, more colloquially, "decollaborate" LA. Such an application is crucial in the deployment of intelligent tutoring systems (Nwana, 1990), as it could use its beliefs about collaboration to estimate the true learner ability (i.e., without collaboration).

Additionally, fusing data across different modalities is an open problem that could deliver impressive results. Concretely, in addition to the learner's answer patterns, we often possess a wealth of side information, such as their open-form responses. Incorporating this side information directly into our collaboration detection approach has the potential to dramatically improve the identification of collaborative behavior, as well as reduce the rate of false detections that can

arise when only the answer patterns themselves are considered.

A Derivation of the MCMC sampler for Bayesian Rasch approach to Learning Analytics

Here, we derive the sampling steps for the Rasch MCMC sampler. Recall that the generative model for the data $\mathbf{Y}$ under the Rasch approach is given by

$$Y_{i,j} \sim \textit{Ber}(\Phi(Z_{i,j})) \quad \text{with} \quad Z_{i,j} = \mathbf{c}_i + \mu_i, \, \forall i, j.$$

It can be shown (see, e.g., (Chib and Greenberg, 1998)) that this model is equivalent to

$$Y_{i,j} \sim \textit{sign}(\Phi(Z'_{i,j})) \quad \text{with} \quad Z'_{i,j} = \mathbf{c}_i + \mu_i + e_{i,j}, \, \forall i, j.$$

where $\textit{sign}(\cdot)$ is the signum function and $e_{i,j} \sim \mathcal{N}(0, 1)$. This latter representation is more convenient for the purposes of MCMC.

By specifying the following prior distributions

$$\pi(c_j) \sim \mathcal{N}(0, \sigma_c^2) \quad \text{and} \quad \pi(\mu_i) \sim \mathcal{N}(0, \sigma_\mu^2),$$

we can perform Gibbs' sampling on each of the variables c_j, μ_j by augmenting with the latent variable $Z'_{i,j}$. The sampling steps at each MCMC iteration are given by

(1) For all $i = 1, \ldots, Q$ and $j = 1, \ldots, N$ sample $Z'_{i,j} \sim \mathcal{N}(c_i + \mu_j, 1)$, truncating above 0 if $Y_{i,j} = 1$, and truncating below 0 if $Y_{i,j} = 0$.

(2) For all $i = 1, \ldots, Q$ sample $\mu_i \sim \mathcal{N}(\widetilde{\sigma}_\mu^2 \sum_{j=1}^{N} (Z'_{i,j} - c_j), \widetilde{\sigma}_\mu^2)$, where $\widetilde{\sigma}_\mu^2 = \left(\frac{1}{\sigma_\mu^2} + N\right)^{-1}$.

(3) For all $j = 1, \ldots, N$ sample $c_j \sim \mathcal{N}(\widetilde{\sigma}_c^2 \sum_{i=1}^{N} (Z'_{i,j} - \mu_i), \widetilde{\sigma}_c^2)$, where $\widetilde{\sigma}_c^2 = \left(\frac{1}{\sigma_c^2} + Q\right)^{-1}$.

By repeating this sampling scheme over several iterations, we assemble a set of samples from the posterior distribution of the Rasch parameters c_j, $\forall j$ and μ_i, $\forall i$. In addition, the values of $p_{i,j} = \Phi(c_j + \mu_i)$ are samples of the probability of learner j answering item i correctly, which are then used by the collaboration-type identification algorithms of Section 4.

B Derivation of the MCMC Sampler for the SPARFA Approach to Learning Analytics

Here, we discuss the sampling scheme for the SPARFA-based MCMC sampler. Like the Rasch MCMC sampler of Appendix A, we can introduce the latent variable Z' and use the equivalent generative model

$$Y_{i,j} \sim \textit{sign}(\Phi(Z'_{i,j})) \quad \text{with} \quad Z'_{i,j} = \mathbf{w}_i^T \mathbf{c}_j + \mu_i + e_{i,j}, \, \forall i, j.$$

where $e_{i,j} \sim N(0, 1)$ as with the Rasch MCMC.

In order to comply with the constraints discussed in Section 2.2, Bayesian SPARFA imposes the following prior distributions

$$W_{i,k} \sim r_k \, \textit{Exp}(\lambda_k) + (1 - r_k) \, \delta_0, \;\; \lambda_k \sim \textit{Ga}(\alpha, \beta), \;\; \text{and} \;\; r_k \sim \textit{Beta}(e, f)$$
$$\mathbf{c}_j \sim \mathcal{N}(0, \mathbf{V}), \;\; \mathbf{V} \sim \textit{IW}(\mathbf{V}_0, h), \;\; \text{and} \;\; \mu_i \sim \mathcal{N}(\mu_0, v_{\boldsymbol{\mu}}),$$

it can be shown that the posterior samples can be computed via a Gibbs' sampler with the following updates

(1) For all $i = 1, \ldots, Q$ and $j = 1, \ldots, N$, sample $Z'_{i,j} \sim \mathcal{N}\big((\mathbf{WC})_{i,j} + \mu_i, 1\big)$, truncating above 0 if $Y_{i,j} = 1$, and truncating below 0 if $Y_{i,j} = 0$.

(2) For all $i = 1, \ldots, Q$, draw $\mu_i \sim \mathcal{N}(m_i, v)$ with $v = (v_{\boldsymbol{\mu}}^{-1} + N)^{-1}$, $m_i = \mu_0 + v \sum_{(} i, j) \left(Z'_{i,j} - \mathbf{w}_i^T \mathbf{c}_j\right)$.

(3) For all $j = 1, \ldots, N$, draw $\mathbf{c}_j \sim \mathcal{N}(\mathbf{m}_j, \mathbf{M}_j)$ with $\mathbf{M}_j = (\mathbf{V}^{-1} + \mathbf{W}^T\mathbf{W})^{-1}$, and $\mathbf{m}_j = \mathbf{M}_j \mathbf{W}^T(\mathbf{z}_j - \boldsymbol{\mu})$, where $\mathbf{z}_j$ denotes the j^{th} column of $\mathbf{Z}$.

(4) Draw $\mathbf{V} \sim \mathit{IW}(\mathbf{V}_0 + \mathbf{CC}^T, N + h)$.

(5) For all $i = 1, \ldots, Q$ and $k = 1, \ldots, K$, draw $W_{i,k} \sim \widehat{R}_{i,k}\, \mathcal{N}^r(\widehat{M}_{i,k}, \widehat{S}_{i,k}) + (1 - \widehat{R}_{i,k})\delta_0$, where $\mathcal{N}^r(a, b)$ is a rectified Normal distribution (Schmidt et al., 2009) and:

(a) $\widehat{R}_{i,k} = p(W_{i,k} = 0 | \mathbf{Z}', \mathbf{C}, \boldsymbol{\mu}) = \frac{\frac{\mathcal{N}^r(0|\widehat{M}_{i,k}, \widehat{S}_{i,k}, \lambda_k)}{Exp(0|\lambda_k)}(1 - r_k)}{\frac{\mathcal{N}^r(0|\widehat{M}_{i,k}, \widehat{S}_{i,k}, \lambda_k)}{Exp(0|\lambda_k)}(1 - r_k) + r_k}$,

(b) $\widehat{M}_{i,k} = \frac{\sum_{(i,j)} \left((Z'_{i,j} - \mu_i) - \sum_{k' \neq k} W_{i,k'} C_{k',j}\right) C_{k,j}}{\sum_{(i,j)} C_{k,j}^2}$, and

(c) $\widehat{S}_{i,k} = \left(\sum_{(i,j)} C_{k,j}^2\right)^{-1}$.

(6) For all $k = 1, \ldots, K$, let b_k define the number of active (i.e., non-zero) entries of $\mathbf{w}_k$. Draw $\lambda_k \sim \mathit{Ga}(\alpha + b_k, \beta + \sum_{i=1}^{Q} W_{i,k})$.

(7) For all $k = 1, \ldots, K$, draw $r_k \sim \mathit{Beta}(e + b_k, f + Q - b_k)$, with b_k defined as in Step 6.

We refer the interested reader to the work in (Lan et al., 2013) for further details regarding the derivation of these sampling steps.

C NUMERICAL EVALUATION OF (2)

Here, we detail the efficient numerical evaluation of the SHT collaboration hypothesis (2). We do this specifically for the case of a two-parameter collaboration model such as $\mathcal{M}_2$ or $\mathcal{M}_3$. Reduction to a single parameter model such as $\mathcal{M}_4$ or to the extension to a model with additional parameters is straightforward.

First, it is important to notice that the product term in (2) is a polynomial in the variables ε_1 and ε_2 of the form

$$\prod_{i=1}^{Q} P(Y_{i,k}, Y_{i,\ell} \,|\, p_{i,k}, p_{i,\ell}, \varepsilon_1, \varepsilon_2, \mathcal{M}_j) = \\ g_{0,0}\varepsilon_1^0\varepsilon_2^0 + g_{0,1}\varepsilon_1^0\varepsilon_2^1 + \ldots + g_{0,Q}\varepsilon_1^0\varepsilon_2^Q + g_{1,0}\varepsilon_1^1\varepsilon_2^0 + \ldots + g_{Q,Q}\varepsilon_1^Q\varepsilon_2^Q. \tag{9}$$

The coefficients $g_{a,b}$ of the polynomial expansion in (9) can be evaluated efficiently using a 2-dimensional convolution. In particular, consider the matrix expansion

$$\mathbf{G} = \bigotimes_{i=1}^{Q} \mathbf{G}_i(Y_{i,k}, Y_{i,\ell} \,|\, p_{i,k}, p_{i,\ell}, \varepsilon_1, \varepsilon_2, \mathcal{M}_j), \tag{10}$$

where $\circledast$ is the (2-dimensional) convolution operator. The term $\mathbf{G}_i(\cdot) \in \mathbb{R}^{2\times 2}$ is a matrix polynomial in the variables ε_1 and ε_2 of the form

$$\mathbf{G}_i(\cdot) = \begin{bmatrix} \widetilde{G}^i_{0,0} & \widetilde{G}^i_{0,1} \\ \widetilde{G}^i_{1,0} & \widetilde{G}^i_{1,1} \end{bmatrix},$$

where $\widetilde{G}^i_{a,b}$ is the coefficient associated with $\varepsilon_1^a \varepsilon_2^b$ corresponding to the i^{th} question: $P(Y_{i,k}, Y_{i,\ell} \,|\, p_{i,k}, p_{i,\ell}, \varepsilon_1, \varepsilon_2, \mathcal{M}_j)$. For example, $\mathbf{G}_i(0,0 \,|\, p_{i,k}, p_{i,\ell}, \mathcal{M}_2)$ is given by

$$\mathbf{G}_i(0,0 \,|\, p_{i,k}, p_{i,\ell}, \mathcal{M}_2) = \begin{bmatrix} \bar{p}_{i,k}\bar{p}_{i,\ell} & -\bar{p}_{i,k}\bar{p}_{i,\ell} + \bar{p}_{i,k} \\ 0 & -\bar{p}_{i,k} + \bar{p}_{i,\ell} \end{bmatrix}.$$

The result of (10) is a matrix $\mathbf{G} \in \mathbb{R}^{(Q+1)\times(Q+1)}$ where $G_{a,b} = g_{a,b}$. Since

$$\int_0^1 \int_0^1 g_{a,b} \varepsilon_1^a \varepsilon_2^b \mathrm{d}\varepsilon_1 \mathrm{d}\varepsilon_2 = \frac{g_{a,b}}{(a+1)(b+1)},$$

we can evaluate (2) by computing

$$P(\mathcal{H}_1^2) = \textstyle\sum_{ij} H_{ij} \quad \text{with} \quad \mathbf{H} = \mathbf{G} \circ \mathbf{F}, \tag{11}$$

where the entries of the matrix $\mathbf{F}$ correspond to $F_{a,b} = \frac{1}{(a+1)(b+1)}$, and $\circ$ denotes the Hadamard (element-wise) matrix product. Simply put, $P(\mathcal{H}_1^2)$ is given by the sum of all elements in the matrix $\mathbf{H} = \mathbf{G} \circ \mathbf{F}$.

It is important to note that finite precision artifacts in the computation of (10) and (11) become non-negligible as Q becomes large, i.e., if Q exceeds around 35 items with double-precision floating-point arithmetic. In order to ensure numerical stability while evaluating of (10) and (11) for large Q, we deploy specialized high-precision computation software packages. Specifically, for our experiments, we used Advanpix's Multiprecision Computing Toolbox for MATLAB.[4]

D Acknowledgments

The authors would like to thank J. Cavallaro, T. "Smash" Goldstein, Mr. Lan, and D. Vats for insightful discussions. 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.

References

Becker, W. E. and Johnston, C. 1999. The relationship between multiple choice and essay response questions in assessing economics understanding. *Economic Record 75,* 4 (Dec.), 348–357.

Bergner, Y., Droschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., and Pritchard, D. 2012. Model-based collaborative filtering analysis of student response data: Machine-learning item response theory. In *Proc. 5th Intl. Conf. Educational Data Mining*. Chania, Greece, 95–102.

[4] `http://www.advanpix.com`

BUTLER, A. C. AND ROEDIGER, H. L. 2008. Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition 36,* 3 (Apr.), 604–616.

CHAFFIN, W. W. 1979. Dangers in using the Z index for detection of cheating on tests. *Psychological Reports 45*, 776–778.

CHIB, S. AND GREENBERG, E. 1998. Analysis of multivariate probit models. *Biometrika 85,* 2 (June), 347–361.

FRARY, R. B. 1993. Statistical detection of multiple-choice answer copying: Review and commentary. *Applied Measurement in Education 6,* 2, 153–165.

GELMAN, A., ROBERT, C., CHOPIN, N., AND ROUSSEAU, J. 1995. *Bayesian Data Analysis*. CRCPress.

HALADYNA, T. M., DOWNING, S. M., AND RODRIGUEZ, M. C. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education 15,* 3, 309–333.

HOFF, P. D. 2009. *A First Course in Bayesian Statistical Methods*. Springer Verlag.

LAN, A. S., WATERS, A. E., STUDER, C., AND BARANIUK, R. G. 2013. Sparse factor analysis for learning and content analytics. *submitted to Journal of Machine Learning Research,*.

LEVINE, M. V. AND DONALD, B. R. 1979. Measuring the appropriatemess of multiple-choice test scores. *Journal of Educational Statistics 4,* 5 (Winter), 269–290.

NWANA, H. S. 1990. Intelligent tutoring systems: an overview. *Artificial Intelligence Review 4,* 4, 251–277.

PAPPANO, L. 2012. The year of the MOOC. *The New York Times*.

RASCH, G. 1960. *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests*. Nielsen & Lydiche.

RASCH, G. 1993. *Probabilistic Models for Some Intelligence and Attainment Tests*. MESA Press.

RODRIGUEZ, M. C. 1997. The art & science of item writing: A meta-analysis of multiple-choice item format effects. In *annual meeting of the American Education Research Association, Chicago, IL*.

SCHMIDT, M. N., WINTHER, O., AND HANSEN, L. K. 2009. Bayesian non-negative matrix factorization. In *Independent Component Analysis and Signal Separation*. Vol. 5441. 540–547.

SOTARIDONA, L. AND MEIJER, R. 2002. Statistical properties of the k-index for detecting answer copying. *Journal of Educational Measurement 39,* 2, 115–132.

SOTARIDONA, L. S. AND MEIJER, R. R. 2003. Two new statistics to detect answer copying. *Journal of Educational Measurement 40,* 1, 53–69.

WATERS, A. E., STUDER, C., AND BARANIUK, R. G. 2013. Bayesian pairwise collaboration detection in educational datasets. In *Proc. IEEE Global Conf. on Sig. and Info. Proc. (GlobalSIP)*. Austin, TX.

WESOLOWSKY, G. O. 2000. Detection excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics 27,* 7 (Aug.), 909–921.

WESTFALL, P. H., JOHNSON, W. O., AND UTTS, J. M. 1997. A bayesian perspective on the bonferroni adjustment. *Biometrika 84,* 2, 419–427.

WOLLACK, J. A. 2003. Comparison of answer copying indices with real data. *Journal of Educational Measurement 40,* 3, 189–205.