Fair Grading Algorithms for Randomized Exams

Jiale Chen* Jason Hartline† Onno Zoeter‡

Abstract

This paper studies grading algorithms for randomized exams. In a randomized exam, each student is asked a small number of random questions from a large question bank. The fair grading problem is to estimate the average grade of each student on the full question bank. The maximum-likelihood estimator for the Bradley-Terry-Luce model on the bipartite student-question graph is shown to be consistent when the number of questions asked to each student is at least the cubed-logarithm of the number of students. In an empirical study on exam data and in simulations, our algorithm based on the maximum-likelihood estimator significantly outperforms simple averaging, i.e., calculating grades by averaging scores on the questions each student is asked, even with a small class and exam size.

1 Introduction

A common approach for deterring cheating in online examinations is to assign students random questions from a large question bank. This random assignment of questions with heterogeneous difficulties leads to different overall difficulties of the exam that each student faces. Unfortunately, the predominant grading rule – simple averaging – averages all question scores equally and results in an unfair grading of the students. This paper develops a grading algorithm that utilizes structural information of the exam results to infer student abilities and question difficulties. From these abilities and difficulties fairer and more accurate grades can be estimated. This grading algorithm can also be used in the design of short exams that maintaining a desired level of accuracy.

During the COVID-19 pandemic, learning management systems (LMS) like Blackboard, Moodle, Canvas by Instructure, and D2L have benefited worldwide students and teachers, especially in remote learning [Raza et al., 2021]. The current exam module in these systems includes four steps. In the first step, the instructor provides a large question bank. In the second step, the system assigns each student a distinct random subset of the questions. (Assigning each student a distinct random subset of the questions helps mitigate cheating.) In the third step, students answer the questions. In the last step, the system grades each student proportionally to her accuracy on assigned questions, i.e., by *simple averaging.*

While randomizing questions and grading with simple averaging is ex ante fair, it is not generally ex post fair. When questions in the question bank have varying difficulty then by random chance a student could be assigned more easy questions than average or more hard questions than average. Ex post in the random assignment of questions to students, the simple averaging of scores on each question allows variation in question difficulties to manifests as ex post unfairness in the final grades.

*Department of Management Science and Engineering, Stanford University. Email: `jialec@stanford.edu`.

†Department of Computer Science, Northwestern University. Email: `hartline@northwestern.edu`.

‡Booking.com. Email: `Onno.zoeter@booking.com`.

The aim of this paper is to understand grading algorithms that are fair and accurate. Given a bank of possible questions, a benchmark for both fairness and accuracy is the counterfactual grade that a student would get if the student was asked all of the questions in the question bank. Exams that ask fewer questions to the students may be inaccurate with respect to this benchmark and the inaccuracy may vary across students and this variation is unfair. This benchmark allows for both the comparison of grading algorithms and the design of randomized exams, i.e., the method for deciding which questions are asked to which students.

The grading algorithms developed in this paper are based on the Bradley-Terry-Luce model [Bradley and Terry, 1952] on bipartite student-question graphs. This model is also studied in the psychology literature where it is known as the Rasch model [Rasch, 1993]. This model views the student answering process as a noisy comparison between a parameter of the student and a parameter of the question. Specifically, there is a merit value vector u which describes the student abilities and question difficulties and is unknown to the instructor. The probability that student i answers question j correctly is defined to be

$$f(u_i - u_j) = \frac{\exp(u_i)}{\exp(u_i) + \exp(u_j)},$$

where $f(x) = \frac{1}{1+\exp(-x)}$, and u_i, u_j represents the merit value of student i and question j respectively.

The paper develops a grading algorithm that is based on the maximum likelihood estimator $\boldsymbol{u}^*$ of the merit vector. The exam result can be represented by a directed graph, where an edge from a student to a question represents a correct answer and the opposite direction represents an incorrect answer. We prove that the maximum likelihood estimator exists and is unique within a strongly connected component (Theorem 8). Thus we use $f(u_i^* - u_j^*)$ to predict the probability that the missing edge goes from student i to question j within the same component. For missing edges across comparable components, since all directed path between the vertices have same directions, we predict a determined correct or incorrect answer consistently with the path direction. For remaining edges, we extend the heuristic of simple averaging. Compared to simple averaging which only focuses on student in-degrees and out-degrees, our grading algorithm incorporates more structural information of the exam result and, as we show, reduces ex post bias.

Results. Our theoretical analysis considers a sequence of distributions over random question assignment graphs indexed by n by setting the number of students and number of questions in the question bank to n and assigning d_n random questions uniformly and independently to each student. Let $\alpha_n = \max_{1 \le i,j \le 2n} u_i - u_j$ be the largest difference between any pair of merits. We prove that if

$$\frac{\exp(\alpha_n)\log n}{d_n} \to 0 \quad (n \to \infty),$$

then the probability that the exam result graph is strongly connected goes to 1 (Theorem 9). Thus, the existence and the uniqueness of the MLE is guaranteed under the model. We also prove that if

$$\exp\left(2(\alpha_n + 1)\right)\sqrt{\frac{\log^3 n}{d_n \log^2 d_n}} \to 0 \quad (n \to \infty), \tag{1}$$

then the MLEs are uniformly consistent, i.e., $\|\boldsymbol{u}^* - \boldsymbol{u}\|_\infty \xrightarrow{\mathbb{P}} 0$. These theoretical results complement the empirical and simulation results in the literature on the Rasch model with random missing data. Our analysis is similar to that of Han et al. [2020] which studies Erdös-Rényi random graphs.

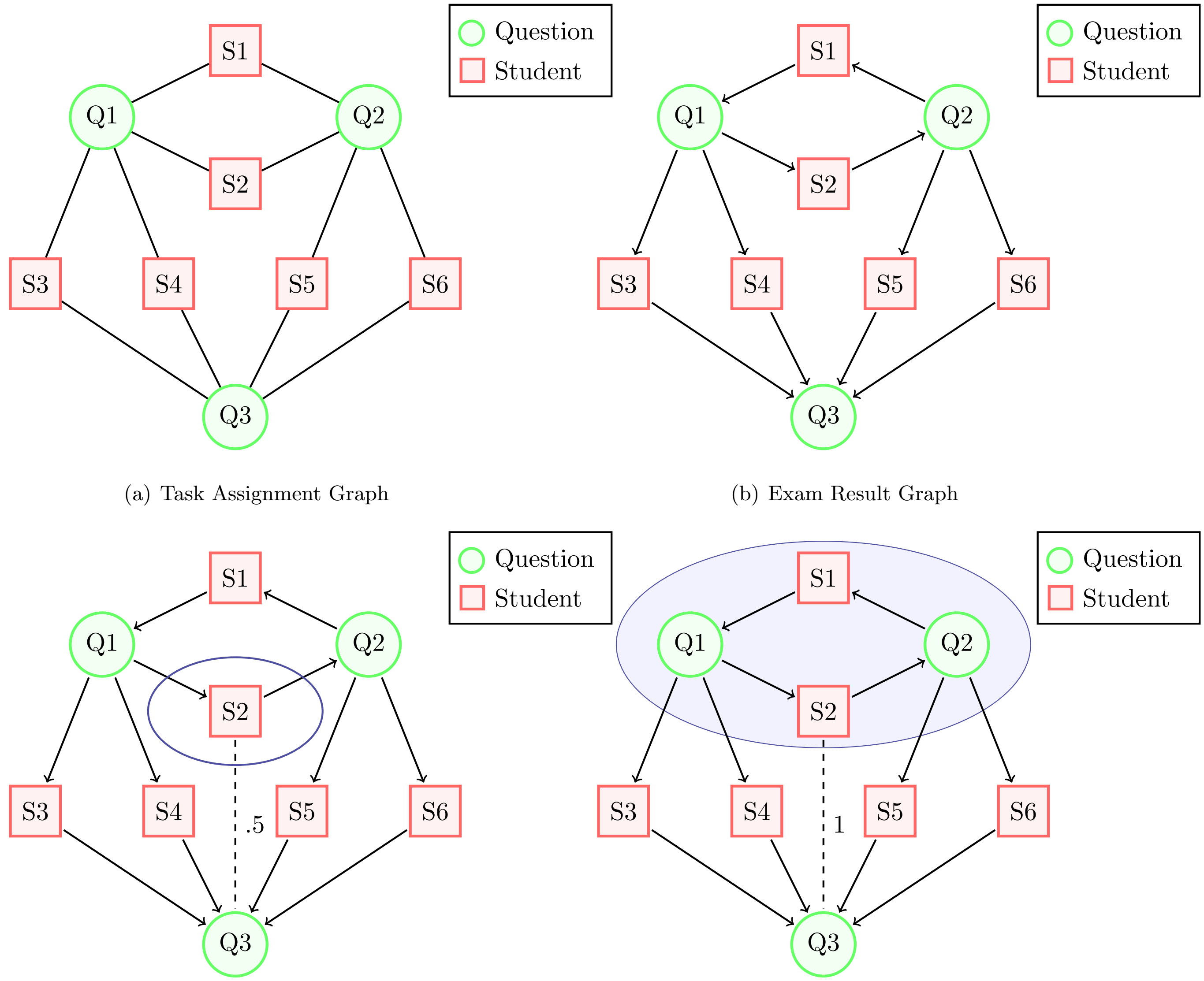
Our empirical analysis considers a study of grading algorithms on both anonymous exam data and numerical simulations. The exam data set consists of 22 questions and 35 students with all students answering all questions. From this data set, randomized exams with fewer than 22 questions can be empirically studied and grading algorithms can be compared. Our algorithm outperforms simple averaging when students are asked at least seven questions. We fit the model parameters to this real-world dataset and run numerical simulation with the resulting generative model. These simulations allow empirical estimation of the ex post bias of our algorithm in comparison to simple averaging. For example, when each of 35 students answers a random 10 of the 22 questions, we find that the maximum bias of simple averaging is more than ten times higher than the maximum bias of our algorithm. Specifically, the maximum bias of any student for simple averaging is 13% while the bias of our algorithm is 1%.

We evaluate a simple question of exam design via simulation. Given an infinite question bank, we sample a fixed number of active questions, then we create a randomized exam with five questions for each student drawn from this fixed number of active questions. When there are only five active questions our grading algorithm and simple averaging coincide as all students are asked all active questions. When there are an infinite number of active questions the algorithms again coincide as no two students are asked the same question and our algorithm and simple averaging are the same. By simulation we consider the bias as a function of the number of active questions and find that the optimal number of questions is about six to nine for maximum bias and about 10 to 15 for average bias.

Illustrative Example. Figure 1 illustrates the unfairness that can arise in simple averaging and the intuition for how our algorithm improves fairness. In the fair exam grading problem, the instructor first assigns questions to students according to an undirected student-question bipartite graph, a.k.a., the task assignment graph (Figure 1(a)). The exam result can be represented by a directed student-question bipartite graph, a.k.a., the exam result graph (Figure 1(b)), where a directed edge from a student to a question represents a correct answer and an opposite direction represents an incorrect answer. Given the exam result graph, the instructor uses a grading algorithm to estimate the average accuracy of students on the whole question bank.

We take student S2 as an example to see how simple averaging, as one specific grading rule, grade students (Figure 1(c)). Simple averaging observes that student S2 has one correct answer and one incorrect answer, in other words, the corresponding vertex has in-degree one and out-degree one. Thus simple averaging predicts that the probability of student S2 answering the remaining question Q3 correctly is 0.5. The reason we believe 0.5 is not a good prediction is, in this exam, every student assigned question Q3 answers it correctly, including student S5 and S6, who give incorrect answers to question Q2. We can infer that Q3 is a relatively easy question. On the other hand, student S2 who can answer question Q2 correctly has a relatively higher ability among the class. Therefore, it is reasonable to believe that student S2 can give a correct answer to question Q3 and, thus, simple averaging underestimates her grade.

We show how our algorithm analyze the missing edge between student S2 and question Q3 in this example (Figure 1(d)). Our algorithm finds that student S2 belongs to the strongly connected component {S1,S2,Q1,Q2}, while question Q3 belongs to {Q3} and the missing edge goes across two comparable components. As a property of the graph, any directed path between student S2 and question Q3 goes from the student to the question. Our grading rule takes it as a strong evidence of the S2's ability higher than Q3's difficulty and predicts that student S2 can answer question Q3 correctly with probability one.



(a) Task Assignment Graph

(b) Exam Result Graph

(c) simple averaging

(d) Our method

Figure 1: A running example of the exam grading problem

Related Work. The literature on peer grading also compares estimation from structural models and simple averaging. When peers are assigned to grade submissions, the quality of peer reviews can vary. Structural models can be used to estimate peer quality and calculate grades on the submissions that put higher weight on peers who give higher quality reviews. Alternatively, submission grades can be calculated by simply averaging the reviews of each peer. The literature has mixed results. De Alfaro and Shavlovsky [2014] propose an algorithm based on reputation that largely outperforms simple averaging on synthetic data, and is better on real-world data when student grading error is not random. Reily et al. [2009] and Hamer et al. [2005] also point out that sophisticated aggregation improves the accuracy compared to simple averaging and also helps to avoid rogue strategies including laziness and aggressive grading. On the other hand, Sajjadi et al. [2016] show that statistical and machine learning methods do not perform better than simple averaging on their dataset. In contrast our result that structural models outperform simple averaging is replicated on several data sets. We believe this difference with the peer grading literature is due to differences in the degrees of the bipartite graphs considered. The exam grading graphs are higher degree than the peer-grading graphs.

In psychometrics, item response theory (IRT) contains mathematical models that build relationship between unobserved characteristics of respondents and items and observed outcomes of the responses. The Rasch model is a commonly used model of IRT that can be applied to psychometrics, educational research [Rasch, 1993], health sciences [Bezruczko, 2005], agriculture [Moral and Rebollo, 2017], and market research [Bechtel, 1985]. Previous simulation studies showed that among different item parameter estimation methods for the Rasch model, the joint maximum likelihood (JML) method and its variants provides one of the most efficient estimates [Robitzsch, 2021], especially with missing data [Waterbury, 2019, Enders, 2010]. In our setting, randomly assignment of questions to students can be seen as a special case of missing data. While with complete data, the condition for the consistency of the maximum likelihood estimators is analyzed [Haberman, 1977, 2004], and with missing data, plenty of works on simulation exists, there is a lack of theoretical works to prove mathematically the consistency of the maximum likelihood estimators when missing data exists.

The Rasch model can be regarded as a special case of the Bradley-Terry-Luce (BTL) model [Bradley and Terry, 1952] for the pairwise comparison of respondents with items by restricting the comparison graph to a bipartite graph. For the BTL model with Erdös-Rényi graph $G(n, p_n)$, the maximum likelihood estimator (MLE) can be solved by an efficient algorithm [Zermelo, 1929, Ford, 1957, Hunter, 2004], and is proved to be a consistent method in l_∞ norm when $\liminf_{n\to\infty} p_n > 0$ [Simons and Yao, 1999, Yan et al., 2012], and recently when $p_n \geq \frac{\log n^3}{n}$ [Han et al., 2020] which is close to the theoretical lower bound of $\frac{\log n}{n}$, below which the comparison graph would be disconnected with positive probability and there is no unique MLE.

In this paper, we follow the method of Han et al. [2020] to prove the consistency of the Rasch model with missing data, or BTL model with a sparse bipartite graph, when each vertex in the left part is assigned small number of random edges to the vertices in the right part. We also propose an extension of the algorithm that reasonably deals with the cases where the MLE does not exists.

2 Model

Consider a set of students S and a bank of questions Q. A merit vector $\boldsymbol{u}$ is used to describe the key property of the students and questions. Specifically, for any student $i \in S$, u_i represents the ability of the student; for any question $j \in Q$, u_j represents the difficulty of the question. The merit vector is unknown when the exam is designed. Denote w_{ij} as the outcome of the answering process. Then w_{ij}s are are independent Bernoulli random variables, where $w_{ij} = 1$ represents a correct answer, $w_{ij} = 0$ represents an incorrect answer, and

$$\Pr[w_{ij} = 1] = 1 - \Pr[w_{ij} = 0] = \frac{\exp(u_i)}{\exp(u_i) + \exp(u_j)} = f(u_i - u_j),$$

where $f(x) = \frac{1}{1+\exp(-x)}$. The goal of the exam design is to assign small number of questions to each student (task assignment graph), and based on the exam result (exam result graph), give each student a grade (grading rule) that accurately estimates her performance over the whole question bank (benchmark). We give a formal description of task assignment graph, exam result graph, benchmark and grading rule below.

Definition 1 (Task Assignment Graph). *The task assignment graph $G = (S \cup Q, E)$ is an undirected bipartite graph, where the left part of the vertices represents the students and the right part represents the questions, and an edge between $i \in S$ and $j \in Q$ exists if and only if the instructor decides to assign question j to student i.*

Definition 2 (Exam Result Graph). *The exam result graph $G' = (S \cup Q, E')$ is a directed bipartite graph constructed from the task assignment graph G. All directed edges are between students and questions. For any edge $(i, j) \in G$ in the task assignment graph, where $i \in S$ and $j \in Q$, if student i answers question j correctly in the exam, i.e., we observe that $w_{ij} = 1$, there is an edge $i \to j$ in G'; if the answer is incorrect, i.e., we observe that $w_{ij} = 0$, there is an edge $j \to i$ in G'. For other student-question pairs that does not occur in the task assignment graph G, there is also no edge between them in the exam result graph G'.*

To evaluate different exam designs and grading rules, we propose the following benchmark.

Definition 3 (Benchmark). *In an ideal case where we know the distribution over the outcome of the answering processes w_{ij}s, the instructor would measure the students' performance by their expected accuracy on a random question in the bank. Formally, the benchmark for any student i's grade is*

$$\mathrm{opt}_i = \mathbb{E}_{j \sim \mathcal{U}(Q)}[w_{ij}] = \frac{1}{|Q|} \sum_{j \in Q} f(u_i - u_j). \tag{2}$$

The benchmark is an ideal way to grade the student, if the instructor has complete information of all answering process. On the other hand, when the instructor only observes one sample of each w_{ij} involved in the exam, we will use a grading rule to grade the students.

Definition 4 (Grading Rule). *In an exam, the instructor gives a grade for each student based on the exam result graph. A grading rule is a mapping $g \colon G' \to \mathbb{R}^{|S|}$ from the exam result graph to the grades for each student.*

One interpretation of the grade is as an estimation of the students' expected accuracy on a random question in the bank. This benchmark combines the two important criteria of fairness and accuracy. To evaluate the exam design, we compare the performance of grading rule to the benchmark, and aggregate the error among all students. Specifically, there are two stages of the exam design (before and after the randomization of the task assignment graph), and in each stage, we care about two metrics, the maximum bias among students and the average bias among students.

Definition 5 (Ex-ante Bias). *For a given algorithm* alg*, the ex-ante bias for student i is defined as the absolute error of the algorithm's expected performance compared to the benchmark, over a random family $\mathcal{G}$ of task assignment graphs, i.e., $|\mathbb{E}_{G \sim \mathcal{G}} \mathbb{E}_w[\mathrm{alg}_i] - \mathrm{opt}_i|$.*

Definition 6 (Ex-post Bias). *For a given algorithm* alg *and a fixed task assignment graph G, the ex-post bias for student i is defined as the absolute error of the algorithm's expected performance compared to the benchmark on G, i.e., $|\mathbb{E}_w[\mathrm{alg}_i] - \mathrm{opt}_i|$.*

Example 2.1 (Simple Averaging). *Simple averaging is a commonly used grading rule in exams. It calculates the average accuracy on the questions the student receives. Formally, given a exam result graph G', the simple averaging grades student i by*

$$\mathrm{avg}_i = \frac{\deg_i^+}{\deg_i^- + \deg_i^+} = \frac{\sum_j 1_{(i,j) \in E'}}{\sum_j 1_{(i,j) \in E}}, \tag{3}$$

where $\deg^+$ *and* $\deg^-$ *represents the outdegree and indegree of the vertex in* G'*, respectively.*

Theorem 7. *The simple averaging is ex-ante fair over any family of bipartite graphs* $\mathcal{G}$ *that is symmetric with respect to the questions, i.e., its ex-ante bias is 0.*

Proof.

$$\begin{aligned}\forall i,\ \mathbb{E}_{G\sim\mathcal{G}}\mathbb{E}_w\left[\mathrm{avg_i}\right] &= \mathbb{E}_{G\sim\mathcal{G}}\mathbb{E}_w\left[\frac{\sum_j 1_{(i,j)\in E'}}{\sum_j 1_{(i,j)\in E}}\right] = \mathbb{E}_{G\sim\mathcal{G}}\mathbb{E}_w\left[\frac{\sum_j w_{ij}1_{(i,j)\in E}}{\sum_j 1_{(i,j)\in E}}\right] \\ &= \mathbb{E}_{G\sim\mathcal{G}}\left[\frac{\sum_j \mathbb{E}[w_{ij}]1_{(i,j)\in E}}{\sum_j 1_{(i,j)\in E}}\right] = \sum_j \mathbb{E}[w_{ij}]\mathbb{E}_{G\sim\mathcal{G}}\left[\frac{1_{(i,j)\in E}}{\sum_j 1_{(i,j)\in E}}\right] = \mathrm{opt_i}\end{aligned} \tag{4}$$

□

In other words, the averaging grading rule can be seen as an ex-ante unbiased estimator of the benchmark. However, ex-post, i.e., on one specific task assignment graph, the averaging grading rule is unfair. Intuitively, some unlucky students might be assigned harder questions and receive a significant lower averaging grade than the benchmark, and the opposite happens to some lucky students. We will visualize this phenomena in Section 5.3.

Based on above definitions, we now formalize the procedure and goal of the exam grading problem.

i. The instructor chooses a task assignment graph G.

ii. The students receive questions according to G and give their answer sheet back, thus the instructor receives the exam result graph G'.

iii. The instructor uses a grading rule g to grade the students based on G'.

iv. We want the grade $g(G')$ to have small maximum (average) ex-post bias.

3 Method

In this section, we propose our method for the exam grading problem. According to our formalization of the problem, any method contains two parts: generating the task assignment graph G, and choosing the grading rule g. We describe each of them respectively.

3.1 Task Assignment Graph

For simplicity, we assume both the student set S and the question set Q is finite. To generate the task assignment graph, we sample m different questions u.a.r. from the question bank, and independently assign each student d different questions u.a.r. from those m questions.

Algorithm 1 Task assignment graph generation

```
Require: finite sets $S$ and $Q$, question sample size $1 \le m \le |Q|$, degree constraint $1 \le d \le m$
Ensure: a task assignment graph $G = (S \cup Q, E)$
  $\tilde{Q} \leftarrow$ a set of $m$ questions sampled u.a.r. without replacement from $Q$
  for all $i \in S$ do
      $J \leftarrow$ a set of $d$ questions sampled u.a.r. without replacement from $\tilde{Q}$
      $E \leftarrow E \cup \{(i,j) | j \in J\}$
```

3.2 Grading Rule

Recall that a grading rule maps from a exam result graph G' to a vector of probabilities. In contrast with simple averaging which only considers the local information (the in-degrees and out-degrees of the students), we use structural information of the exam result graph for analysis. Our grading rule is an aggregation of a prediction matrix $h \in [0,1]^{S \times Q}$, where h_{ij} represents the algorithm's prediction on the probability that student i answers correctly question j. The grade for student i will be the average of $h_{ij}s$ over all $j \in Q$, i.e. $\text{alg}_i = \frac{1}{|Q|}\sum_{j \in Q} h_{ij}$. We use $u \rightsquigarrow v$ to represent the existence of a directed path in G' that starts with u and ends with v, and $u \not\rightsquigarrow v$ for nonexistence. The algorithm classifies the elements h_{ij}s into four cases: existing edge $(i,j) \in E$, same component $i \rightsquigarrow j \wedge j \rightsquigarrow i$, comparable components $i \rightsquigarrow j \oplus j \rightsquigarrow i$, and incomparable components $i \not\rightsquigarrow \wedge i \not\rightsquigarrow j$.

Existing Edge For $(i,j) \in E$, we observe w_{ij} from the exam result graph G', hence $h_{ij} = w_{ij}$.

Same Component For student $i \in S$ and question $j \in Q$ satisfy $i \rightsquigarrow j \wedge j \rightsquigarrow i$, they are in the same strongly connected component in G'. We make all predictions in the component simultaneously, by inferring the student abilities and question difficulties from the structure of the component. Formally, denote V' as the vertex set of the component. From Theorem 8, the strong connectivity guarantees the existence of the maximum likelihood estimators (MLEs) $\boldsymbol{u^*} \in \mathbb{R}^{V'}$. We can use an minorization–maximization algorithm from Hunter [2004] to calculate the MLEs and set $h_{ij} = f(u_i^* - u_j^*)$ for any missing edge (i,j) between students and questions inside this component.

Comparable Components W.l.o.g., we assume $i \rightsquigarrow j$ and $j \not\rightsquigarrow i$, thus every directed path between those two vertices starts with the student and ends with the question, showing a strong evidence of a determinate correct answer or a determinate incorrect answer. In other words, consider the strongly connected components they belong to, the component that contains the student has a "higher level" in the condensation graph of G' and can reach the component that contains the question, i.e., they belong to comparable components in the condensation graph. In this case, we set $h_{ij} = 1$. Similarly, if $j \rightsquigarrow i$ and $i \not\rightsquigarrow j$, we set $h_{ij} = 0$

Incomparable Components For a student i and question j that satisfy $i \not\rightsquigarrow \wedge i \not\rightsquigarrow j$, i.e., they lie in incomparable components, we use the average of the predictions in the above three cases as the prediction for h_{ij}.

4 Theory

In this section, we show several properties of our algorithm. Recall that the Bradley-Terry-Luce model describes the outcome of pairwise comparisons as follows. In a comparison between subject i and subject j, subject i beats subject j with probability

$$p_{ij} = \frac{\exp(u_i)}{\exp(u_i) + \exp(u_j)} = f(u_i - u_j),$$

where $\boldsymbol{u} = (u_1 \ldots, u_{2n})$ represents the merit parameters of $2n$ subjects and $f(x) = \frac{1}{1+\exp(-x)}$. We consider the Bradley-Terry-Luce model under a family of random bipartite task assignment graphs $\mathcal{B}(n, d_n)$. Specifically, a task assignment graph $G(L \cup R, E)$ with n vertices in each part is

Algorithm 2 Grade Generation

```
Require: an exam result graph $G'(S \cup Q, E')$
Ensure: a grade vector $g \in [0,1]^S$ for students
  From the exam result graph $G'$, we can get the task assignment graph $G(S \cap Q, E)$.
  for all $(i,j) \in E$ do                                                    ▷ Case 1: Existing Edge
      $h_{ij} \leftarrow w_{ij}$.
  for all Strongly Connected Component $\tilde{G}(\tilde{S} \cup \tilde{Q}, \tilde{E})$ do    ▷ Case 2: Same Component
      $\boldsymbol{u^*} \leftarrow$ the MLEs of the merit parameters of $\tilde{S} \cup \tilde{Q}$.
      for all $(i,j) \in \tilde{(}S \times \tilde{Q}) \setminus E$ do
          $h_{ij} \leftarrow f(u_i^* - u_j^*)$.
  for all $(i,j) \in (S \times Q) \setminus E \wedge i \leadsto j \wedge j \not\leadsto i$ do      ▷ Case 3: Comparable Component
      $h_{ij} \leftarrow 1$.
  for all $(i,j) \in (S \times Q) \setminus E \wedge j \leadsto i \wedge i \not\leadsto j$ do
      $h_{ij} \leftarrow 0$.
  for all $(i,j) \in (S \times Q) \setminus E \wedge i \not\leadsto j \wedge j \not\leadsto i$ do      ▷ Case 4: Incomparable Component
      $h_{ij} \leftarrow$ the average of the existing $h_{i\cdot}$ in previous steps.
  for all $i \in S$ do                                                        ▷ Grade Aggregation
      $g_i \leftarrow$ the average of $h_{ij}$ for all $j$s.
```

constructed by linking d_n different random vertices in R to each left vertex in L, which means L is regular but R is not.

Given a task assignment graph G, denote A as its adjacency matrix. For any two subjects i and j, the number of comparisons between them follows $A_{ij} \in \{0,1\}$. We define A'_{ij} as the number of times that subject i beats subject j, thus $A'_{ij} + A'_{ji} = A_{ij} = A_{ji}$. In other words, A' is the adjacency matrix of the exam result graph G'. Based on the observation of G', the log-likelihood function is

$$\mathcal{L}(\boldsymbol{u}) = \sum_{1 \le i \ne j \le 2n} A'_{ij} \log p_{ij} = \sum_{1 \le i \ne j \le 2n} A'_{ij} \log f(u_i - u_j). \tag{5}$$

Denote $\boldsymbol{u}^* = (u_1^*, u_1^*, \ldots, u_{2n}^*)$ as the maximum likelihood estimators (MLEs) of $\boldsymbol{u}$. Since $\mathcal{L}$ is additive invariant, w.l.o.g. we assume $u_1 = 0$ and set $u_1^* = 0$. Since $(\log f(x))' = 1 - f(x)$ the likelihood equation can be simplified to

$$\sum_{j=1}^{2n} A'_{ij} = \sum_{j=1}^{2n} A_{ij} f(u_i^* - u_j^*), \forall\ i. \tag{6}$$

4.1 Existence and Uniqueness of the MLEs

Zermelo [1929] and Ford [1957] gave a necessary and sufficient condition for the existence and uniqueness of the MLEs in (6).

Condition A. For every two nonempty sets that form a partition of the subjects, a subject in one set has beaten a subject in the other set at least once.

To provide an intuitive understanding of Condition A, we show its equivalence to the strong connectivity of the exam result graph G'.

Theorem 8. *Condition A holds if and only if the exam result graph G' is strongly connected.*

Proof. Condition A says that for any partition (V_1, V_2) of the vertices $L \cup R$, there exists an edge from V_1 to V_2 and also an edge from V_2 to V_1. If G' is strongly connected, Condition A directly holds by definition of strong connectivity. Otherwise, if G' is not strongly connected, the condensation of G' contains at least two SCCs. We pick one strongly connected component with no indegree as V_1 and remaining vertices as V_2, then there is no edge from V_2 to V_1, i.e., Condition A fails. □

Theorem 9 (Existence and Uniqueness of MLEs)**.** *If*

$$\frac{\exp(\alpha_n)\log n}{d_n} \to 0 \quad (n \to \infty), \tag{7}$$

where $\alpha_n = \max_{1 \le i,j \le 2n} u_i - u_j$ is the largest difference between all possible pair of merits, then $\Pr[\textit{Condition A is satisfied}] \to 1 \quad (n \to \infty)$.

To prove Theorem 9, we first show the edge expansion property of the task assignment graph G then bound the probability that G' fails Condition A given the edge expansion property.

Lemma 10 (Edge Expansion)**.** *Under condition* (7)*,*

$$\Pr\left[\forall S \subset V,\ s.t.\ |S| \le n, \quad \frac{|\partial S|}{|S|} > \frac{d_n}{4}\right] \to 1 \quad (n \to \infty),$$

where $\partial S = \{(u,v) \in E : u \in S, v \in V \setminus S\}$ for the task assignment graph $G(V,E)$.

Proof. Consider any subset of vertices S with size $r \le n$. Denote $X = S \cap L, Y = S \cap R, |X| = x$, thus $|Y| = r - x, |L \setminus X| = n - x, |R \setminus Y| = n + x - r$. ∂S is a random variable that can be expressed as

$$|\partial S| = \sum_{u \in X}\sum_{v \in R \setminus Y} A_{uv} + \sum_{u \in L \setminus X}\sum_{v \in Y} A_{uv},$$

where A is the adjacency matrix of the task assignment graph G. Recall that the task assignment graph G is generated by linking d_n random different vertices in R to each vertex in L. Thus for different $u_1 \ne u_2 \in L$, $A_{u_1\cdot}$ is independent with $A_{u_2\cdot}$, while for a fixed $u \in L$, $A_{u\cdot}$ is chosen randomly without replacement. Chernoff bound applies under such conditions, i.e.,

$$\Pr\left[|\partial S| \le \frac{1}{2}\mathbb{E}\left[|\partial S|\right]\right] \le \exp\left(-\frac{\mathbb{E}[|\partial S|]}{8}\right).$$

Then we lower bound $\mathbb{E}[|\partial S|]$ by

$$\begin{aligned}\mathbb{E}[|\partial S|] &= \frac{d_n}{n}(|X||R \setminus Y| + |L \setminus X||Y|) \\ &= \frac{d_n}{n}(2x^2 - 2xr + nr) \ge \frac{d_n}{n}\left(-\frac{1}{2}r^2 + nr\right) \ge \frac{d_n r}{2}.\end{aligned}$$

Thus for any fixed set S with size r,

$$\Pr\left[|\partial S| \le \frac{d_n r}{4}\right] \le \Pr\left[|\partial S| \le \frac{1}{2}\mathbb{E}\left[|\partial S|\right]\right] \le \exp\left(-\frac{\mathbb{E}[|\partial S|]}{8}\right) \le \exp\left(-\frac{d_n r}{16}\right).$$

Finally, by union bound,

$$\begin{aligned}
\Pr\left[\forall S \subset V,\ \text{s.t. } |S| \le n, \quad \frac{|\partial S|}{|S|} > \frac{d_n}{4}\right] &= 1 - \Pr\left[\exists S \subset V,\ \text{s.t. } |S| \le n, \quad \frac{|\partial S|}{|S|} \ge \frac{d_n}{4}\right] \\
&\ge 1 - \sum_{r=1}^{n} \binom{2n}{r} \exp\left(-\frac{d_n r}{16}\right) \\
&\ge 1 - \sum_{r=1}^{n} \exp\left(-\frac{d_n r}{16} + r\log(2n)\right) \\
&\ge 1 - \sum_{r=1}^{n} \exp\left(-\frac{d_n r}{48}\right) \\
&\ge 1 - \exp\left(-\frac{d_n}{48} + \log n\right) \\
&\ge 1 - \exp\left(-\frac{d_n}{192}\right)
\end{aligned}$$

The third-to-last inequality holds when $d_n > 24\log(2n)$ and the last inequality holds when $d_n > 64\log n$. Note that condition (7) implies $\log n/d_n \to 0 \quad (n \to \infty)$ since $\alpha_n \ge 0$. Thus for large enough n and d_n,

$$\Pr\left[\forall S \subset V,\ \text{s.t. } |S| \le n, \quad \frac{|\partial S|}{|S|} > \frac{d_n}{4}\right] \ge 1 - \exp\left(-\frac{d_n}{192}\right) \to 1 \quad (n \to \infty).$$

□

Proof of Theorem 9. For an edge between vertex i and j in the task assignment graph G, i.e. $A_{ij} = 1$, the corresponding directed edge in the exam result graph G' goes from i to j with probability

$$\Pr[A'_{ij} = 1] = f(u_i - u_j) \le \max_{1 \le i,j \le 2n} f(u_i - u_j) \le \frac{1}{1 + \exp(-\alpha_n)} \le 2^{-\exp(-\alpha_n)}.$$

By Lemma 10, under condition (7),

$$\Pr\left[\forall S \subset V,\ \text{s.t. } |S| \le n, \quad \frac{|\partial S|}{|S|} > \frac{d_n}{4}\right] \to 1 \quad (n \to \infty).$$

Now consider any subset of vertices $S \subset V$ s.t. $|S| = r \le n$. The probability that all edges between S and $V \setminus S$ go in the same direction in G' is no more than $2\left(2^{-\exp(-\alpha_n)}\right)^{d_n r/4}$. Thus by union bound, the probability that Condition A holds is at least

$$\begin{aligned}
&1 - 2\sum_{1 \le r \le n} \binom{2n}{r} \left(2^{-\exp(-\alpha_n) d_n r/4}\right) \\
\ge &1 - 2\left(\sum_{0 \le r \le 2n} \binom{2n}{r} \left(2^{-\exp(-\alpha_n) d_n r/4}\right) - 1\right) \\
\ge &1 - 2\left(\left(1 + \left(2^{-\exp(-\alpha_n) d_n/4}\right)\right)^{2n} - 1\right),
\end{aligned}$$

which converges to 1 when $n \to \infty$ under condition (7). □

4.2 Uniform Consistency of the MLEs

Based on condition (7), Theorem 9 shows the existence and uniqueness of the MLEs. We will only discuss the situation when condition (7) is satisfied. In this part, we will prove the uniform consistency of the MLEs.

Theorem 11 (Uniform Consistency of MLEs)**.** *If*

$$\exp\left(2(\alpha_n + 1)\right)\Delta_n \to 0 \quad (n \to \infty), \tag{8}$$

where $\Delta_n = \sqrt{\frac{\log^3 n}{d_n \log^2 d_n}}$, *then the MLEs are uniformly consistent, i.e.,* $\|\boldsymbol{u}^* - \boldsymbol{u}\|_\infty \xrightarrow{\mathbb{P}} 0$.

To bound $\|\boldsymbol{u}^* - \boldsymbol{u}\|_\infty$, we consider the two subjects

$$i_0 = \arg\min_i u_i^* - u_i \quad \text{and} \quad i_1 = \arg\max_i u_i^* - u_i.$$

Since we assume $u_1 = 0$ and set $u_1^* = 0$, we have

$$u_{i_0}^* - u_{i_0} \le u_1^* - u_1 = 0 \quad \text{and} \quad u_{i_1}^* - u_{i_1} \ge u_1^* - u_1 = 0,$$

thus

$$\|\boldsymbol{u}^* - \boldsymbol{u}\|_\infty = \max\{-(u_{i_0}^* - u_{i_0}), u_{i_1}^* - u_{i_1}\} \le (u_{i_1}^* - u_{i_1}) - (u_{i_0}^* - u_{i_0}).$$

Thus instead of directly proving $\|\boldsymbol{u}^* - \boldsymbol{u}\|_\infty$ tends to 0, we bound the difference of $u_{i_1}^* - u_{i_1}$ and $u_{i_0}^* - u_{i_0}$. We first introduce some notations,

$$K_n = \left\lfloor \frac{2\log n}{\log d_n} + 1 \right\rfloor, \quad c_n = \frac{\exp(-(\alpha_n + 1))}{4}, \quad q_n = \frac{c_n \log d_n}{4 \log n}, \quad z_n = \sqrt{\frac{32 \log n}{d_n}}.$$

We also define a sequence of increasing numbers $\{D_k\}_{k=0}^{K}$ with size K_n, representing the "distance" in the difference of the MLE and merit parameter,

$$D_k = \frac{4k}{c_n}\sqrt{\frac{2(1+z_n)\log n}{(1-z_n)d_n}} \quad \text{for } k = 0, 1, \ldots, K_n - 1,$$

$$D_{K_n} = \frac{80K_n}{c_n^2}\sqrt{\frac{2(1+z_n)\log n}{(1-z_n)d_n}},$$

and two growing sets $\{B_k\}_{k=0}^{K}$ and $\{C_k\}_{k=0}^{K}$ which contains the subjects that is D_k close to i_0 or i_1 respectively in terms of the difference,

$$B_k = \{j : (u_j^* - u_j) - (u_{i_0}^* - u_{i_0}) \le D_k\}$$
$$C_k = \{j : (u_{i_1}^* - u_{i_1}) - (u_j^* - u_j) \le D_k\}.$$

To bound the difference between $u_{i_1}^* - u_{i_1}$ and $u_{i_0}^* - u_{i_0}$, imagine in every round, we increase the "distance" from D_k to D_{k+1} to incorporate more subjects into B_{k+1} and C_{k+1} until $k = K_n$. If $|B_{K_n}| > n$ and $|C_{K_n}| > n$, then B_{K_n} and C_{K_n} share at least one common subject. This common subject links i_0 and i_0 through bounding the difference between $u_{i_1}^* - u_{i_1}$ and $u_{i_0}^* - u_{i_0}$ by $2D_{K_n}$, which tends to 0 under condition (7) and (8).

The main issue in the proof is to demonstrate the growth of B_k and C_k. By symmetry, we only need to focus on B_k. Define the neighborhood of a set of subjects S as $N(S) = \{j : \exists\ i \in$

S, s.t. $A_{ij} = 1\}$. An intuitive thought on the growth of B_k is to incorporate additional subjects in the neighborhood of B_k into B_{k+1}. Thus firstly, we check the vertex expansion of the task assignment graph $G \sim \mathcal{B}(n, d_n)$. By Chernoff bound and union bound, we can easily prove the concentration of vertex degree, i.e.,

$$\forall\ i \in V, \quad \Pr\left[(1 - z_n)d_n \leq |N(\{i\})| \leq (1 + z_n)d_n\right] \geq 1 - n^{-4}, \tag{9}$$

where z_n is defined above as $\sqrt{\frac{32\log n}{d_n}}$ that tends to 0 when $n \to \infty$ under condition (7). More generally, we have the following lemma.

Lemma 12 (Vertex Expansion)**.** *Regarding the task assignment graph $G(L \cup R, E) \sim \mathcal{B}(n, d_n)$, for a fixed subset of vertex in the left part $X \subset L$ with $|X| \leq \frac{n}{2}$, it holds with probability $1 - n^{-4|X|}$ that*

- *If $1 \leq |X| < n/d_n$,*
$$\frac{|N(X)|}{|X|} > (1 - z_n)\left(1 - \frac{d_n|X|}{n}\right)d_n;$$
- *If $|X| \geq n/d_n$,*
$$\frac{|N(X)|}{n} > 1 - z_n - e^{-1},$$

where $z_n = \sqrt{\frac{32\log n}{d_n}}$. Similar arguments exist for a fixed subset of vertex in the right part $Y \subset R$.

Proof. We need to define another family of random bipartite graph $\tilde{\mathcal{B}}$. Each graph in $\tilde{\mathcal{B}}(n, d_n)$ contains n vertices in each part and assigns d_n random neighbors to each vertex in the left part (multi-edges are allowed). For any $X \subset L$, it's easy to see that $|N(X)|$ in $G \sim \mathcal{B}(n, d_n)$ stochastically dominates $|N(X)|$ in $G \sim \tilde{\mathcal{B}}(n, d_n)$. Thus it's sufficient to prove the theorem under $\tilde{\mathcal{B}}(n, d_n)$. On the other hand, counting $|N(X)|$ under $\tilde{\mathcal{B}}(n, d_n)$ is the same random process as counting the number of non-empty bins after independently throwing $d_n|X|$ balls u.a.r. into n bins. By linearity of expectation over every bin, we know

$$\mathbb{E}[|N(X)|] = n\left(1 - \left(1 - \frac{1}{n}\right)^{d_n|X|}\right).$$

We need several lower bounds of $\mathbb{E}[|N(X)|]$ here. With the fact of

$$\frac{x}{2} \leq 1 - \exp(-x) \leq x, \quad \forall\ 0 \leq x < 1,$$

we have

$$\mathbb{E}[|N(X)|] = n\left(1 - \left(1 - \frac{1}{n}\right)^{d_n|X|}\right) \geq n\left(1 - \exp\left(\frac{d_n|X|}{n}\right)\right) \geq \frac{d_n|X|}{2}.$$

Therefore, using Azuma's inequality, we can lower bound $|N(X)|$, i.e.,

$$\Pr\left[|N(X)| \leq (1 - z_n)\mathbb{E}[|N(X)|]\right] \leq \exp\left(-\frac{z_n^2\left(\mathbb{E}[|N(X)|]\right)^2}{2d_n|X|}\right) = \exp\left(-\frac{z_n^2 d_n|X|}{8}\right) \leq n^{-4|X|}.$$

Also, when $|X| < n/d_n$, we have

$$\left(1 - \left(1 - \frac{1}{n}\right)^{d_n|X|}\right) \geq \frac{d_n|X|}{n}\left(1 - \frac{d_n|X|}{n}\right),$$

thus with probability $1 - n^{-4|X|}$,

$$|N(X)| \geq (1 - z_n)\mathbb{E}[|N(X)|] \geq (1 - z_n) d_n |X| \left(1 - \frac{d_n|X|}{n}\right);$$

Similarly when $|X| \leq n/d_n$, we have

$$\left(1 - \left(1 - \frac{1}{n}\right)^{d_n|X|}\right) \geq 1 - e^{-1},$$

and

$$|N(X)| \geq (1 - z_n)\mathbb{E}[|N(X)|] \geq (1 - z_n)\left(1 - e^{-1}\right) n \geq \left(1 - z_n - e^{-1}\right) n.$$

The proof for $Y \subset R$ is almost the same except that it's sufficient to use Chernoff bound rather than Azuma's inequality since the independence among the subjects in $N(Y)$. □

Lemma 12 describes the vertex expansion, an important structural property we need, of the task assignment graph G. In the next step, we will analyze the first order conditions (6) and show a local property of B_k. More specifically, Lemma 13 cares about for any subject in B_k, how many of its neighbors are in B_{k+1}.

Lemma 13 (Local Growth of B_k)**.** *For n large enough, round $k < K_n$ and a fixed subject $i \in B_k$, it holds with probability* $1 - 3n^{-4}$ *that*

- *If $k < K_n - 1$,*
$$|N(\{i\}) \cap B_{k+1}| \geq q_n |N(\{i\})|,$$
where $q_n = \frac{c_n \log d_n}{4 \log n}$ and $c_n = \frac{\exp(-(\alpha_n + 1))}{4}$;
- *If $k = K_n - 1$,*
$$|N(\{i\}) \cap B_{k+1}| \geq \frac{75}{81} |N(\{i\})|.$$

Proof. Pick a subject $i \in B_k$. From (9) we know that with probability $1 - n^{-4}$,

$$(1 - z_n) d_n \leq |N(\{i\})| \leq (1 + z_n) d_n,$$

where $z_n = \sqrt{\frac{32 \log n}{d_n}}$. For any task assignment graph G and its adjacency matrix A, the corresponding adjacency matrix A' of the exam result graph is a random variable of A. Specifically, for any $A_{ij} = 1$, A'_{ij}s are independent Bernoulli random variables with probability $f(u_i - u_j)$ to be 1. In other words, $\mathbb{E}[A'_{ij}] = A_{ij} f(u_i - u_j)$. By Chernoff bound,

$$\Pr\left[\left|\sum_j A'_{ij} - \sum_j A_{ij} f(u_i - u_j)\right| \geq \sqrt{2(1 + z_n) d_n \log n}\right] \leq 2 \exp\left(-\frac{4(1 + z_n) d_n \log n}{|N(i)|}\right) \leq 2n^{-4}.$$

Thus according to equation (6), we have, with probability $1 - 3n^{-4}$,

$$\left|\sum_j A_{ij} \left(f(u_i - u_j) - f(u_i^* - u_j^*)\right)\right| < \sqrt{2(1 + z) d_n \log n}.$$

Below we use the above inequality and some analysis of function f to count the number of subjects in $N(\{i\}) \cap B_{k+1}$. The fact we use about function f is

$$\begin{aligned} f'(x) &= \frac{\exp(-x)}{(1+\exp(-x))^2} \leq \frac{1}{4} \quad \text{and} \\ f'(x) &\geq \frac{\exp(-(\alpha_n+1))}{(1+\exp(-(\alpha_n+1)))^2} \geq \frac{\exp(-(\alpha_n+1))}{4} = c_n, \ \forall |x| \leq \alpha_n + 1. \end{aligned}$$

Thus for another subject j such that $u_j^* - u_j \leq u_i^* - u_i$, by mean value theorem, we have

$$f\left(u_i^* - u_j^*\right) - f\left(u_i - u_j\right) = f'(\xi_{ij})\left[\left(u_i^* - u_i\right) - \left(u_j^* - u_j\right)\right] \leq \frac{1}{4}\left[\left(u_i^* - u_i\right) - \left(u_{i_0}^* - u_{i_0}\right)\right] \leq \frac{D_k}{4},$$

where $\xi_{ij} \in \left[u_i - u_j, u_i^* - u_j^*\right]$. Similarly, for a subject j with $u_j^* - u_j > u_i^* - u_i + D_{k+1} - D_k$, we have

$$f\left(u_i - u_j\right) - f\left(u_i^* - u_j^*\right) = f'(\xi'_{ij})\left[\left(u_j^* - u_j\right) - \left(u_i^* - u_i\right)\right] \geq c_n(D_{k+1} - D_k),$$

where $\xi'_{ij} \in \left[u_i^* - u_j^*, u_i - u_j\right]$. Since

$$u_i - u_j - D_{K_n} \leq u_i - u_j - \left[\left(u_j^* - u_j\right) - \left(u_i^* - u_i\right)\right] \leq \xi'_{ij} \leq u_i - u_j,$$

and $D_{K_n} \to 0$ as $(n \to \infty)$ under condition (8), $|\xi'_{ij}|$ is bounded by $\alpha_n + 1$ when n is large enough, thus $f'(\xi'_{ij}) \geq c_n$. Therefore, on the one hand,

$$\begin{aligned} &\sum_{u_j^* - u_j > u_i^* - u_i} A_{ij}\left(f(u_i - u_j) - f(u_i^* - u_j^*)\right) \\ =&\sum_j A_{ij}\left(f(u_i - u_j) - f(u_i^* - u_j^*)\right) - \sum_{u_j^* - u_j \leq u_i^* - u_i} A_{ij}\left(f(u_i - u_j) - f(u_i^* - u_j^*)\right) \\ \leq&\sqrt{2(1+z)d_n \log n} + \frac{1}{4} D_k \sum_{u_j^* - u_j \leq u_i^* - u_i} A_{ij}. \end{aligned} \tag{10}$$

On the other hand,

$$\begin{aligned} &\sum_{u_j^* - u_j > u_i^* - u_i} A_{ij}\left(f(u_i - u_j) - f(u_i^* - u_j^*)\right) \\ \geq&\sum_{u_j^* - u_j > u_i^* - u_i + D_{k+1} - D_k} A_{ij}\left(f(u_i - u_j) - f(u_i^* - u_j^*)\right) \\ \geq& c_n(D_{k+1} - D_k) \sum_{u_j^* - u_j > u_i^* - u_i + D_{k+1} - D_k} A_{ij}. \end{aligned} \tag{11}$$

Combining (10) and (11), we have

$$|N(\{i\}) \cap B_{k+1}| \geq \sum_{u_j^* - u_j \leq u_i^* - u_i + D_{k+1} - D_k} A_{ij} \geq \frac{c_n(D_{k+1} - D_k) - \sqrt{\frac{2(1+z_n)\log n}{(1-z_n)d_n}}}{c_n(D_{k+1} - D_k) + \frac{1}{4}D_k} |N(\{i\})|.$$

For $k < K_n - 1$,

$$\frac{c_n(D_{k+1} - D_k) - \sqrt{\frac{2(1+z_n)\log n}{(1-z_n)d_n}}}{c_n(D_{k+1} - D_k) + \frac{1}{4}D_k} |N(\{i\})| \geq q_n |N(\{i\})|.$$

For $k = K_n - 1$,

$$\frac{c_n(D_{k+1} - D_k) - \sqrt{\frac{2(1+z_n)\log n}{(1-z_n)d_n}}}{c_n(D_{k+1} - D_k) + \frac{1}{4}D_k}|N(\{i\})| \geq \frac{75}{81}|N(\{i\})|.$$

□

With the help of Lemma 12 and 13, we are able to show the growth from B_k to B_{k+1}. Specifically, we consider two subsets of B_k, $X_k = B_k \cap L$ and $Y_k = B_k \cap R$, and use the following inequalities

$$\begin{aligned}
|Y_{k+1}| \geq |N(X_k) \cap B_{k+1}| &= |N(X_k)| - |N(X_k) \cap \overline{B_{k+1}}| \\
&= |N(X_k)| - \sum_{i \in X_k} |N(\{i\}) \cap \overline{B_{k+1}}| \\
&= |N(X_k)| - \sum_{i \in X_k} \left(|N(\{i\})| - |N(\{i\}) \cap B_{k+1}|\right), \\
|X_{k+1}| \geq |N(Y_k) \cap B_{k+1}| &= |N(Y_k)| - |N(Y_k) \cap \overline{B_{k+1}}| \\
&= |N(Y_k)| - \sum_{i \in Y_k} |N(\{i\}) \cap \overline{B_{k+1}}| \\
&= |N(Y_k)| - \sum_{i \in Y_k} \left(|N(\{i\})| - |N(\{i\}) \cap B_{k+1}|\right),
\end{aligned} \tag{12}$$

to show the growth of X_k and Y_k respectively.

Proof of Theorem 11. Denote $X_k = B_k \cap L$ and $Y_k = B_k \cap R$. We inductively prove the following fact that, for n large enough, with probability $1 - n^{-2}$,

- For $1 \leq k \leq K_n - 2$,
$$|X_k|, |Y_k| \geq d_n^{(k-1)/2};$$
- for $k = K_n - 1$,
$$|X_k|, |Y_k| \geq \frac{n}{d_n};$$
- for $k = K_n$,
$$|X_k|, |Y_k| > \frac{n}{2}.$$

From now on we only consider n large enough. Since $i_0 \in B_0$, w.l.o.g. we assume $|X_0| = 1$. if X_0 contains other subjects, we take a subset with size 1. Then by fact (12), (9) and Lemma 13, we know with probability $1 - 4n^{-4}$ that

$$|Y_1| \geq |N(X_0) \cap B_{k+1}| \geq q_n |N(X_0)| \geq q_n(1 - z_n)d_n > 0.$$

For $1 < k \leq K_n - 2$, we prove inductively. We assume $|X_k| = d_n^{(k-1)/2}$. If X_k is larger, we pick any subset with size $d_n^{(k-1)/2}$. Fact (12) show that

$$|Y_{k+1}| \geq |N(X_k)| - \sum_{i \in X_k} \left(|N(\{i\})| - |N(\{i\}) \cap B_{k+1}|\right).$$

By Lemma 12 and union bound over all subset of L with size $d_n^{(k-1)/2}$, it holds with probability $1 - n^{-3|X_k|}$ that,

$$|N(X_k)| > (1 - z_n)\left(1 - \frac{d_n|X_k|}{n}\right) d_n|X_k|.$$

By Lemma13 and union bound over all possible subject $i \in X_k$, it holds with probability $1 - 3n^{-3}$ that,

$$\forall i \in X_k,\ |N(\{i\}) \cap B_{k+1}| \geq q_n |N(\{i\})|.$$

By (9) and union bound over all possible subject $i \in X_k$, it holds with probability $1 - n^{-3}$ that,

$$\forall i \in X_k,\ (1 - z_n)d_n \leq |N(\{i\})| \leq (1 + z_n)d_n.$$

Therefore, with probability $1 - 4n^{-3}$ we have

$$\begin{aligned}
|Y_{k+1}| &\geq |N(X_k)| - \sum_{i \in X_k} \left(|N(\{i\})| - |N(\{i\}) \cap B_{k+1}|\right) \\
&\geq |N(X_k)| - (1 - q_n) \sum_{i \in X_k} |N(\{i\})| \\
&\geq (1 - z_n)\left(1 - \frac{d_n|X_k|}{n}\right) d_n|X_k| - (1 - q_n)(1 + z_n)d_n|X_k| \\
&\geq |X_k| d_n^{1/2} \left((1 + z_n)q_n d_n^{1/2} - 2z_n d_n^{1/2} - (1 - z_n)\frac{d_n^{3/2}|X_k|}{n}\right) \\
&\geq |X_k| d_n^{1/2} \left(q_n d_n^{1/2} - 2z_n d_n^{1/2} - \frac{d_n^{3/2}|X_k|}{n}\right) \\
&\geq |X_k| d_n^{1/2} \left(q_n d_n^{1/2} - 2z_n d_n^{1/2} - 1\right),
\end{aligned}$$

where the last inequality holds because we assume $|X_k| = d_n^{(k-1)/2}$, thus the last term satisfies

$$d_n^{3/2}|X_k|/n = d_n^{(k+1)/2}/n \leq d_n^{(K_n-1)/2}/n \leq 1.$$

Finally, under condition (8), we have for large enough n, $q_n d_n^{1/2} - 2z_n d^{1/2} - 1 > 1$, which finishes the proof of the fact for $k \leq K_n - 1$. The proof for $k = K_n$ is similar with the help of the same lemmas and inequalities. Thus with probability $1 - 4n^{-3}$, for large enough n,

$$\begin{aligned}
|Y_{K_n}| &\geq |N(X_{K_n-1})| - \sum_{i \in X_{K_n-1}} \left(|N(\{i\})| - |N(\{i\}) \cap B_{K_n}|\right) \\
&\geq |N(X_{K_n-1})| - \left(1 - \frac{75}{81}\right) \sum_{i \in X_{K_n-1}} |N(\{i\})| \\
&\geq (1 - z_n - e^{-1})n - \frac{6}{81}(1 + z_n)n \\
&> \frac{n}{2}.
\end{aligned}$$

The same proof applies for $|Y_k|$ and $|X_{k+1}|$. To summarize, with probability $1 - n^{-2}$, $|X_{K_n}|, |Y_{k_n}| > n/2$, thus $|B_{K_n}| > n$. By symmetry, $|C_{K_n}| > n$ with probability $1 - n^{-2}$. Then with probability $1 - 2n^{-2}$, at least one subject $i \in B_{K_n} \cap C_{K_n}$ lies in both B_{K_n} and C_{K_n}. By definition, subject i satisfies

$$(u_i^* - u_i) - (u_{i_0}^* - u_{i_0}) \leq D_{K_n} \quad \text{and} \quad (u_{i_1}^* - u_{i_1}) - (u_i^* - u_i) \leq D_{K_n},$$

thus

$$\|\boldsymbol{u}^* - \boldsymbol{u}\|_\infty \leq (u_{i_1}^* - u_{i_1}) - (u_{i_0}^* - u_{i_0}) \leq 2D_{K_n},$$

which tends to 0 under condition (8). □

4.3 Analysis of Our Algorithm

Our algorithm uses the MLEs to predict the student's performance within the component. Based on the consistency of the MLEs, we show the bias of our algorithm when Condition A is satisfied.

Theorem 14. *When Condition A is satisfied, the exam result graph is strongly connected. In this case, the MLE is unique and we have*

$$|\mathrm{alg}_i - \mathrm{opt}_i| \leq \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{u}^*\|_\infty.$$

Proof. When the exam result graph is strongly connected, the algorithm calculates the MLEs $\boldsymbol{u}^*$ and gives student i a grade of $\mathrm{alg}_i = \frac{1}{|Q|}\sum_{j\in Q} f(u_i^* - u_j^*)$, while the ground truth probability of answering a random question correctly is $\mathrm{opt}_i = \frac{1}{|Q|}\sum_{j\in Q} f(u_i - u_j)$. Thus we have

$$\begin{aligned}
|\mathrm{alg}_i - \mathrm{opt}_i| =& \left|\frac{1}{|Q|}\sum_j f(u_i^* - u_j^*) - \frac{1}{|Q|}\sum_j f(u_i - u_j)\right| \\
\leq& \frac{1}{|Q|}\sum_j \left|f(u_i^* - u_j^*) - f(u_i - u_j)\right| \\
=& \frac{1}{|Q|}\sum_j \left|f'(\xi_{ij})\right| \left|(u_i^* - u_i) - (u_j^* - u_j)\right| \\
\leq& \frac{2}{n}\|\boldsymbol{u} - \boldsymbol{u}^*\|_\infty \sum_j \left|f'(\xi_{ij})\right| \\
\leq& \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{u}^*\|_\infty,
\end{aligned} \tag{13}$$

where the third-to-last equality is because of the mean value theorem, the next-to-last inequality is because $\left|(u_i^* - u_i) - (u_j^* - u_j)\right| \leq 2\|\boldsymbol{u} - \boldsymbol{u}^*\|_\infty$, and the last inequality is because $|f'(x)| \leq \frac{1}{4}$. □

Next we discuss the performance of our algorithm on several extreme cases of the task assignment graph. For example, the extremely sparse cases when $N(i)$ is mutually disjoint for each student i or each student receives only $d = 1$ question. Another example is that the task assignment graph is a complete bipartite graph. In all of the above cases, our algorithm gives the same grade as simple averaging.

Theorem 15. *When the task assignment graph satisfies that $N(i)$ is mutually disjoint for each student i or each student receives only $d = 1$ question, our algorithm gives the same grade as simple averaging.*

Proof. In both cases, the exam result graph satisfies that every SCC is a single point, thus the algorithm's output totally relies on cross-component predictions. For each student, the comparable components for each student are exactly the questions that student receives. Thus the algorithm gives the same prediction as the student's correctness on those questions. The prediction for remaining questions is the average accuracy on the assigned questions by the algorithm's rule for incomparable components. Therefore, the algorithm's grade for the student is exactly the same as simple averaging. □

Theorem 16. *When the task assignment graph is a complete bipartite graph, our algorithm gives the same grade as simple averaging.*

Proof. In this case, the output of the algorithm only relies on existing edges. It directly follows that the algorithm gives the same grade as simple averaging. □

5 Experiments

5.1 Real-World Data

We collected the anonymous answer sheets from one exam with $|S| = 35$ students and $|Q| = 22$ questions. The task assignment graph of the exam is a complete bipartite graph, i.e., each student is assigned with all questions. The corresponding exam result graph is strongly connected, which means we can infer about student abilities and question difficulties under the Bradley-Terry-Luce model (Figure 2).

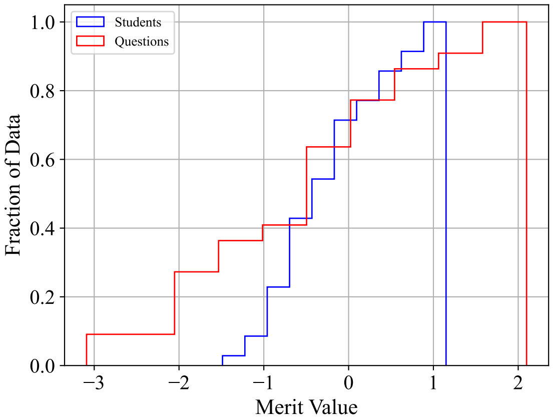


Figure 2: Empirical Cumulative Distribution of Merit Value. We analyze all students and questions under the Bradley-Terry-Luce model and show the empirical cumulative density function of inferred student abilities and question difficulties. The abilities ranges from -1.486 to 1.149 while the difficulties ranges from -3.090 to 2.099.

5.2 Algorithms

Simple Averaging The grade for student i is its average correctness on assigned questions. See a formal definition in Example 2.1.

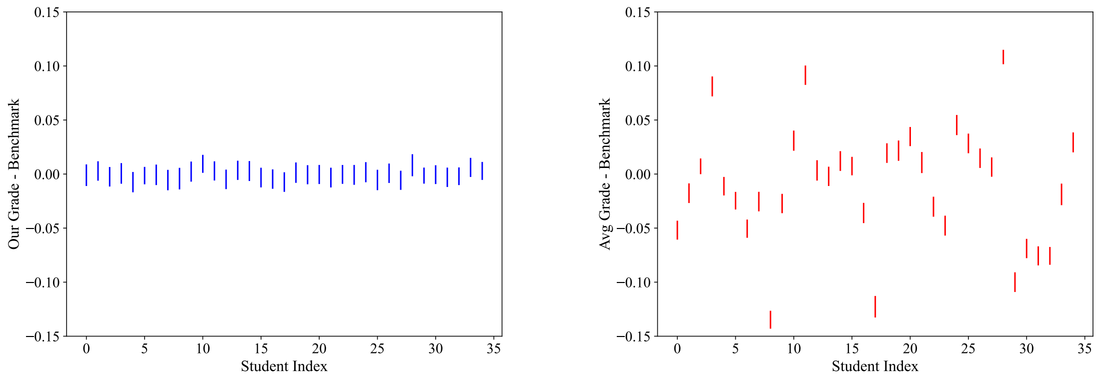
Our Algorithm The grade for student i is an aggregation of the algorithm's prediction on her performance on each question. All predictions can be classified into four cases, including existing edges (keep the fact as prediction), same component (maximum likelihood estimators), comparable components (determined answer in line with the path direction) and incomparable components (heuristic as simple averaging). See its formal definition in Section 3.2.

Maximum a Posteriori Estimation This algorithm maximizes the posterior probability or equivalently $\Pr[G'|G, \boldsymbol{u}]\Pr[\boldsymbol{u}]$, where G' is the exam result graph, G is the task assignment graph, $\boldsymbol{u}$ is the merit value vector and $\Pr[\boldsymbol{u}]$ is the prior distribution of $\boldsymbol{u}$. The prior distribution is defined to be a i.i.d. normal distribution on each merit value, where its mean and variance is fitted by the real-world data shown by Figure 2. It then use the maximizer $\boldsymbol{u}^*$ to predict the missing edge between student i and question j by $f(u_i^* - u_j^*)$, where $f(x) = \frac{1}{1+\exp(-x)}$.

5.3 Simulation 1: A Visualization of Simple Averaging's Ex-post Unfairness

We compare the ex-post bias (Definition 6) between our algorithm and simple averaging given a fixed random task assignment graph. The simulation setting is the following. There are $|S| = 35$ students and $|Q| = 22$ questions, whose abilities and difficulties respectively are shown in Figure 2. The task assignment graph is generated by Algorithm 1 with the parameter $m = |Q|$ and $d = 10$, i.e. each student is assigned 10 random questions from the whole question bank. The exam result graph is repeatedly generated according to the model.

Figure 3 shows the performance of two algorithms. The left plot corresponds to our algorithm and the right plot corresponds to simple averaging. In each plot, there are 35 confidence intervals, each corresponds to the difference between the student's expected grade and her benchmark, i.e. $\text{alg}_i - \text{opt}_i$. The confidence intervals in the left plot are significantly closer to 0, compared to the right plot, which visualizes the intuition that students are facing different overall question difficulties under the random assignments and simple averaging fails to adjust their grades. Instead, our algorithm infers the question difficulties and the student abilities and adjust their grades accordingly, thus could largely reduce the ex-post unfairness.



(a) Ex-post Grade Error of Our Algorithm

(b) Ex-post Grade Error of Simple Averaging

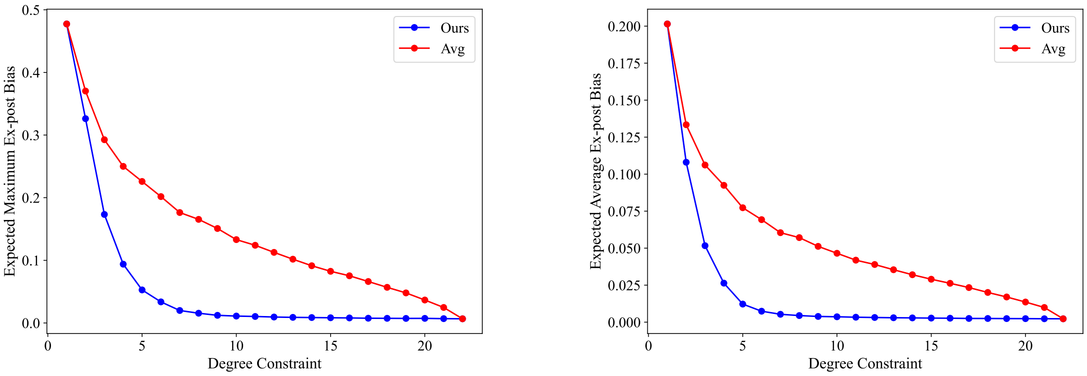
Figure 3: A Visualization of the Ex-post Grade Error with Degree Constraint $d = 10$

5.4 Simulation 2: The Effect of the Degree Constraint

We compare the expected maximum ex-post bias among students and the expected average ex-post bias among students between our algorithm (Ours) and simple averaging (Avg). The simulation setting is the following. There are $|S| = 35$ students and $|Q| = 22$ questions, whose abilities and difficulties respectively are showed in Figure 2. For each degree constraint d, we repeatedly generate task assignment graphs using algorithm Algorithm 1 with the other parameter

$m = |Q|$. For each task assignment graph, the exam result graph is repeatedly generated according to the model.

Figure 4 shows two algorithms' expected maximum ex-post bias (Figure 4(a)) and expected average ex-post bias (Figure 4(b)) under different degree constraints. It can be seen that our algorithm (blue curve) outperforms simple averaging (red curve) on every degree constraint d. Our algorithm's expected ex-post bias with the degree constraint $d = 6$ is close to simple averaging's with the degree constraint $d = 20$, which means our algorithm can ask 14 less questions to each student to achieve a same grading accuracy as simple averaging.



(a) Expected Maximum Ex-post Bias

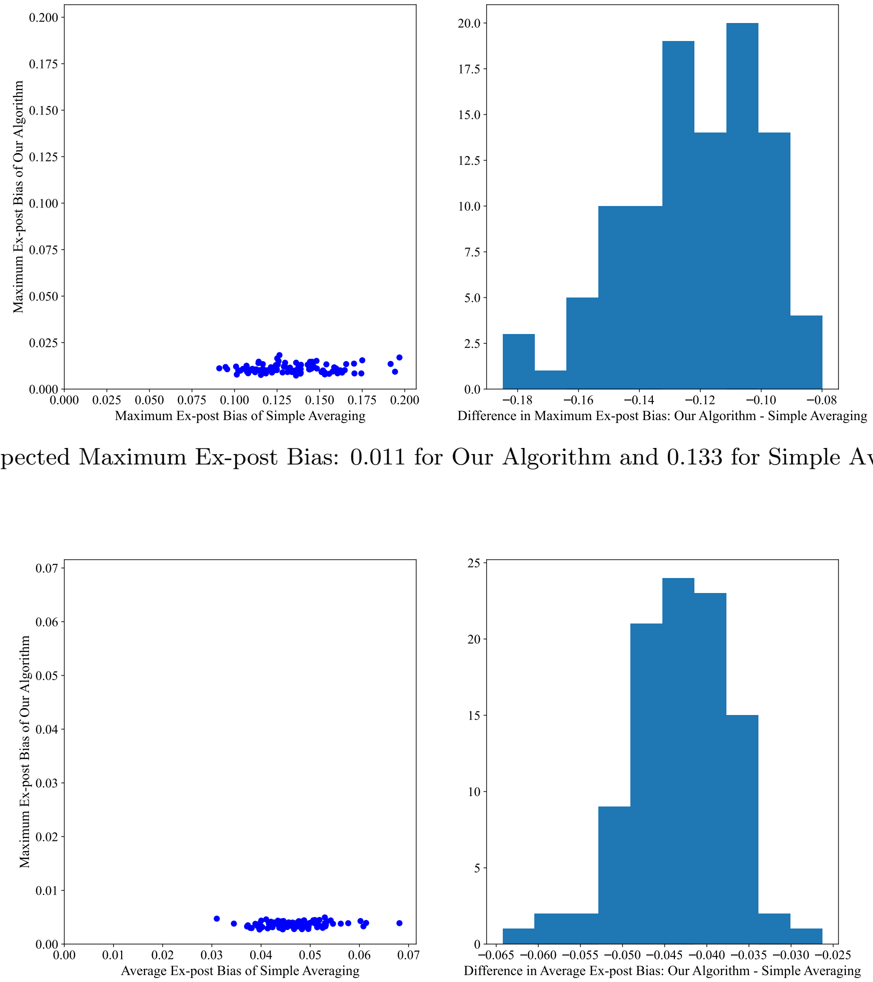
(b) Expected Average Ex-post Bias

Figure 4: Expected Aggregated Ex-post Bias v.s. Degree Constraint

If we zoom in on a specific degree constraint, we can see two algorithms' maximum ex-post bias and average ex-post bias on every different task assignment graph. Figure 5 shows the case with the degree constraint $d = 10$. Figure 5(a) corresponds to the maximum ex-post bias and Figure 5(b) corresponds to the average ex-post bias. In each case, the left plot contains 100 points, corresponding to a different task assignment graph, whose x-axis is the aggregated ex-post bias of simple averaging and whose y-axis is that of our algorithm; the right plot is the histogram of the difference of the aggregated ex-post bias between our algorithm and simple averaging. The simulation results show that our algorithm has negligible aggregated ex-post bias compared to simple averaging on every task assignment graph.

5.5 Simulation 3: The Effect of the Question Sample Size

We now consider the case where question bank is infinite and compare the expected maximum ex-post bias and the expected average ex-post bias between our algorithm and simple averaging. We sampled and fixed $|S| = 5$ random students among the previous 35 students. The question difficulties are i.i.d. distributed according to the linear interpolation of the difficulties shown in Figure 2. For each question sample size m, we repeatedly generate task assignment graphs using Algorithm 1 with the other parameter $d = 5$. It can be expected that when $m = d$ and $m \to \infty$, two algorithms should have the same performance. Figure 6 shows a consistantly smaller expected maximum ex-post bias (Figure 6(a)) and expected average ex-post bias (Figure 6(b)) of our algorithm (blue curve) than simple averaging (red curve). As the question sample size grows, the expected aggregated ex-post bias of our algorithm first decreases and then increases. The turning point is about 6-9 for expected maximum ex-post bias and 10-15 for expected average ex-post bias.



(a) Expected Maximum Ex-post Bias: 0.011 for Our Algorithm and 0.133 for Simple Averaging

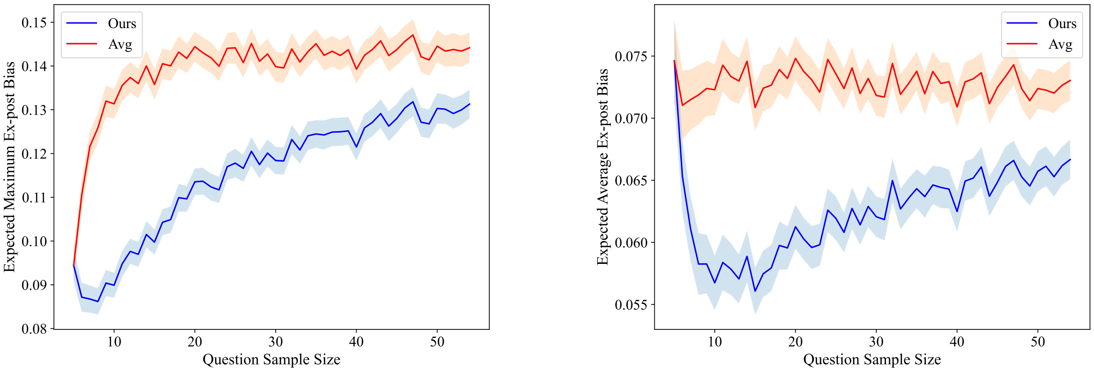
(b) Expected Average Ex-post Bias: 0.004 for Our Algorithm and 0.047 for Simple Averaging

Figure 5: Aggregated Ex-post Bias on Task Assignment Graphs with Degree Constraint $d = 10$

5.6 Real-World Data Experiment: Cross Validation

We cannot repeat an exam in real world and check the ex-post bias of the algorithms. Thus, we sample part of the data we have as a new exam result, and use them to predict the students' actual average on the data. We randomly split the real-world data into training data and test data. Specifically, for a fixed student sample size d_1 and a degree constraint d_2, in each repetition, we randomly sample d_1 students and randomly choose d_2 questions and corresponding answers for each student as the training data, use our algorithm (Ours), simple averaging (Avg) and maximum a posteriori estimation (MAP) to predict every student's average accuracy on the whole question bank, and calculate the mean squared error. Formally, the mean squared error MSE is defined as

$$\mathrm{MSE} = \mathbb{E}_{X,\tilde{S}}\left[\frac{1}{|\tilde{S}|}\sum_{i\in\tilde{S}}\left(\mathrm{alg}_i - \frac{1}{|Q|}\sum_{j\in Q} w_{ij}\right)^2\right],$$



(a) Expected Maximum Ex-post Bias

(b) Expected Average Ex-post Bias

Figure 6: Expected Aggregated Ex-post Bias v.s. Question Sample Size

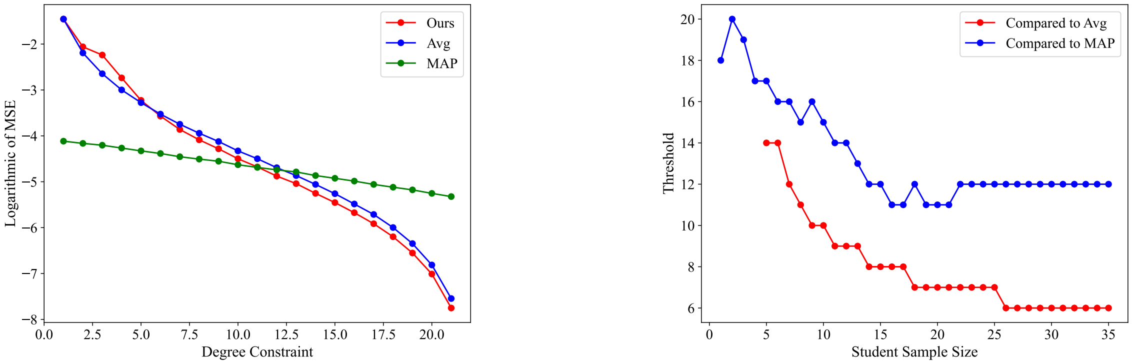
where X is the training set given a fixed degree d, $\tilde{S}$ is the sampled student set, alg_i is student i's grade given by the algorithm and w_{ij} is the correctness of student i's answer to question j.

In Figure 7(a), we fix the student sample size $d_1 = |S|$, i.e., $\tilde{S} = S$ and change the degree constraint d_2 from 1 to $|Q|$ and show the curve of the logarithmic of MSE. Maximum a posteriori estimation as a regularized algorithm performs the best at a low data environment when the degree constraint d_2 is less than 12. Our algorithm performs better than simple averaging when the degree constraint d_2 is larger than 5 and has 16% to 20% smaller MSE compared to simple averaging when the degree constraint d_2 is larger than 10. In Figure 7(b), we consider for every possible student sample size d_1, what the smallest degree constraints d_2 is for our algorithm to perform better than simple averaging or maximum a posteriori estimation. It provides a reference for choosing the grading rule in different situations.

We additionally run a same cross validation in numerical simulation. We assume two different Normal prior distributions for student abilities and question difficulties, fitted by the data in Figure 2. For a fixed student sample size d_1 and degree constraint d_2, in every repetition, we first draw d_1 i.i.d. student abilities and $|Q| = 22$ i.i.d. question difficulties from those two prior distributions. We use a complete task assignment graph to generate the exam result graph according to the model. Then we randomly choose d_2 questions and the corresponding answers of each student as the training set, and use our algorithm, simple averaging and maximum a posteriori estimation to predict each student's average accuracy over the whole question bank. Note that we do exact same things in the cross validation, so when calculating the MSE, we compare the algorithms' grades to the students' average accuracy given by the exam result graph instead of the benchmark. From Figure 8 we observe that the simulation result is quite similar to that of the real-world cross validation, which suggests that the numerical simulation result could be a good reference for the practical use of our algorithm.

6 Conclusions

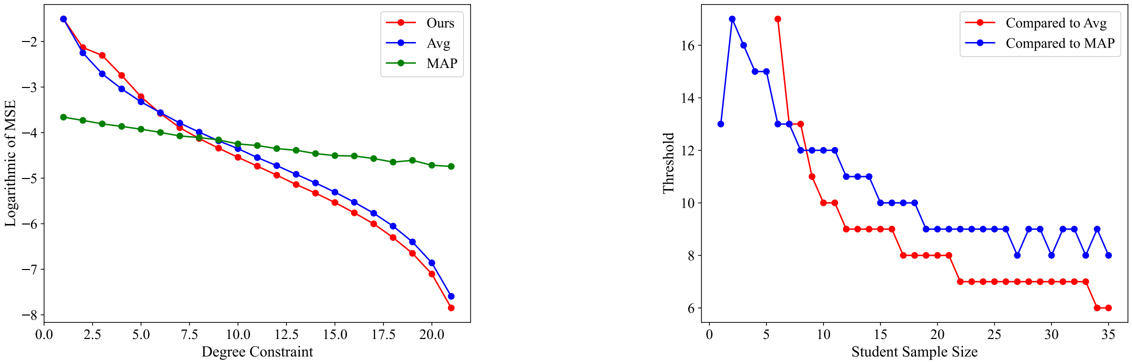
We formulate and study the fair exam grading problem under the Bradley-Terry-Luce model. We propose an algorithm that is a generalization of the maximum likelihood estimation method. To theoretically validate our algorithm, we prove the existence, uniqueness, and the uniform consistency of the maximum likelihood estimators under the Bradley-Terry-Luce model on sparse



(a) Logarithmic of MSE v.s. Degree Constraint

(b) Threshold v.s. Student Sample Size

Figure 7: Cross Validation



(a) Logarithmic of MSE v.s. Degree Constraint

(b) Threshold v.s. Student Sample Size

Figure 8: Simulated Cross Validation

bipartite graphs. Our algorithm significantly outperforms simple averaging in numerical simulation. On real-world data, our algorithm is better when the students are assigned a sufficient number of questions (i.e., on sufficiently long exams). In the low-degree environment where the students are assigned a small number of questions, we observe that the maximum a posteriori estimation (with a fitted prior distribution), which can be seen as a regularization of the maximum likelihood estimation, performs the best. We provide guidelines for how to choose the grading rule given certain number of students and a fixed exam length.

Our model in this paper mainly considers true-or-false questions, which can be extended to multiple choice questions and to the case where it can be assumed that students would guess if they cannot solve the question. Our method to treat missing edges across comparable components needs to be improved especially in the low-degree environment (i.e., short exam lengths), since little information across the components should not result in a determined prediction. It is likely that our theoretical analysis can be extended to scenarios where the number of students and the size of the question bank are different.

References

Gordon G. Bechtel. Generalizing the Rasch Model for Consumer Rating Scales. *Marketing Science*, 4(1):62–73, February 1985. ISSN 0732-2399. doi: 10.1287/mksc.4.1.62.

Nikolaus Bezruczko. *Rasch Measurement in Health Sciences.* JAM Press, Maple Grove, Minn, 2005. ISBN 978-0-9755351-2-7 978-0-9755351-3-4.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons, 1952.

Luca De Alfaro and Michael Shavlovsky. Crowdgrader: A tool for crowdsourcing the evaluation of homework assignments. In *Proceedings of the 45th ACM technical symposium on Computer science education*, pages 415–420, 2014.

Craig K. Enders. *Applied Missing Data Analysis.* Methodology in the Social Sciences. Guilford Press, New York, 2010. ISBN 978-1-60623-639-0.

L. R. Ford. Solution of a Ranking Problem from Binary Comparisons. *The American Mathematical Monthly*, 64(8):28–33, 1957. ISSN 0002-9890. doi: 10.2307/2308513.

Shelby J. Haberman. Maximum Likelihood Estimates in Exponential Response Models. *The Annals of Statistics*, 5(5):815–841, September 1977. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176343941.

Shelby J. Haberman. Joint and Conditional Maximum Likelihood Estimation for the Rasch Model for Binary Responses. *ETS Research Report Series*, 2004(1):i–63, June 2004. ISSN 23308516. doi: 10.1002/j.2333-8504.2004.tb01947.x.

John Hamer, Kenneth T. K. Ma, Hugh H. F. Kwong, Kenneth T. K, Ma Hugh, and H. F. Kwong. A Method of Automatic Grade Calibration in Peer Assessment. In *Of Conferences in Research and Practice in Information Technology, Australian Computer Society*, pages 67–72, 2005.

Ruijian Han, Rougang Ye, Chunxi Tan, and Kani Chen. Asymptotic theory of sparse Bradley–Terry model. *The Annals of Applied Probability*, 30(5):2491–2515, October 2020. ISSN 1050-5164, 2168-8737. doi: 10.1214/20-AAP1564.

David R. Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1), February 2004. ISSN 0090-5364. doi: 10.1214/aos/1079120141.

Francisco J. Moral and Francisco J. Rebollo. Characterization of soil fertility using the Rasch model. *Journal of soil science and plant nutrition*, 17(2):486–498, June 2017. ISSN 0718-9516. doi: 10.4067/S0718-95162017005000035.

Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests.* MESA Press, 5835 S, 1993. ISBN 978-0-941938-05-1.

Syed A. Raza, Wasim Qazi, Komal Akram Khan, and Javeria Salam. Social Isolation and Acceptance of the Learning Management System (LMS) in the time of COVID-19 Pandemic: An Expansion of the UTAUT Model. *Journal of Educational Computing Research*, 59(2):183–208, April 2021. ISSN 0735-6331. doi: 10.1177/0735633120960421.

Ken Reily, Pam Finnerty, and Loren Terveen. Two peers are better than one: Aggregating peer reviews for computing assignments is surprisingly accurate. In *GROUP'09 - Proceedings of the 2009 ACM SIGCHI International Conference on Supporting Group Work*, pages 115–124, January 2009. doi: 10.1145/1531674.1531692.

Alexander Robitzsch. A Comprehensive Simulation Study of Estimation Methods for the Rasch Model. *Stats*, 4(4):814–836, December 2021. ISSN 2571-905X. doi: 10.3390/stats4040048.

Mehdi S. M. Sajjadi, Morteza Alamgir, and Ulrike von Luxburg. Peer Grading in a Course on Algorithms and Data Structures: Machine Learning Algorithms do not Improve over Simple Baselines, February 2016.

Gordon Simons and Yi-Ching Yao. Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *The Annals of Statistics*, 27(3):1041–1060, June 1999. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1018031267.

Glenn Waterbury. Missing Data and the Rasch Model: The Effects of Missing Data Mechanisms on Item Parameter Estimation. *Journal of applied measurement*, 20:1–12, May 2019.

Ting Yan, Yaning Yang, and Jinfeng Xu. Sparse Paired Comparisons in the Bradley-Terry Model. *Statistica Sinica*, 22(3):1305–1318, 2012. ISSN 1017-0405.

E. Zermelo. Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460, December 1929. ISSN 0025-5874, 1432-1823. doi: 10.1007/BF01180541.