Improving Peer Assessment Accuracy by Incorporating Student Grading Behaviors

Jia Xu[1,2,3], Jing Liu[1], Pin Lv[1,2,3], Panyuan Yang[1]
[1] College of Computer, Electronics and Information, Guangxi University, Nanning, Guangxi, China
[2] Guangxi Key Laboratory of Multimedia Communications Network Technology, Nanning, Guangxi, China
[3] Guangxi Colleges and Universities Key Laboratory of Parallel and Distributed Computing, Nanning, Guangxi, China
{xujia, jingliu, lvpin, panyuan}@gxu.edu.cn

***Abstract*—Peer assessment, arranging students/peers to evaluate the other submissions via a web interface, has become a vital solution to the open-ended assignment assessment in massive open online courses (MOOCs). However, since peer scores may be biased and unreliable, the effect of simple methods such as median and mean is limited. To address this issue, some probabilistic graph models considering the graders' bias and reliability have been proposed. However, these models only focus on the grading information and ignore behavior information collected by peer assessment systems, such as consuming time, number of attempts, and comment length. A student's grading behaviors reflects the grading seriousness, which can be used to measure the reliability of the student. Therefore, we propose two new probabilistic graph models by incorporating students' grading behaviors to improve the accuracy of estimating true scores in peer assessment. Firstly, a GBDT-based regressor is built to obtain graders' grading seriousness values according to their behaviors. Secondly, utilizing the acquired grading seriousness information, each grader's reliability in the proposed models is optimized. Finally, a technique based on Gibbs sampling is used to infer the final true scores of assignments. Experimental results on a real dataset of three computer-related courses show that the proposed models improve the accuracy of estimation performance by leveraging student grading behaviors.**

***Index Terms*—peer assessment, probabilistic graph models, GBDT, online education**

I. Introduction

Massive open online courses (MOOCs) have been drawing increasingly more attention because they enable students from all over the world to freely access high-quality educational resources anytime anywhere. The cardinality of students enrolled in a popular MOOC may reach up to tens of thousands, which makes the assessment of students' assignments a great challenge. While assignments on objective questions (e.g., multiple-choice, blank-filling, and judgment) can be automatically graded by computers, assignments on open-ended questions (e.g., essay writing, programming, and design) which have no standard answers can hardly be scored in an automatic manner [1]. Considering open-ended questions have special functionalities in examining students' language expression, critical thinking, and innovation ability [2], how to assess large-scale open-ended questions in an efficient and effective way has become an important problem needed to be solved by modern education.

The mainstream approach to tackle the problem is peer assessment (or peer grading), in which students are asked to evaluate the assignments of their peers under criteria rules formulated by teachers [3], [4]. Peer assessment not only reduces the burden of assignment assessment for teachers, but also brings many benefits for students. First it helps students consolidate knowledge points with respect to the evaluated assignments, improve their learning motivation, and promote their participation in the course through learning their peers' solutions [3] [5] [6]. Second, it helps students to evaluate and reflect, develop their cognitive ability, and cultivate their sense of responsibility [7]. Moreover, in most cases getting feedback from peers is faster than getting feedback from teachers.

Therefore, representative MOOC platforms, such as Coursera[1] and edX[2], have already provided the functionality of peer assessment for teachers to evaluate massive assignments on open-ended questions submitted by students. These MOOC platforms, however, simply take the average or median of peer grades as an estimate to the true grade of each assignment, which may be inaccurate since they fail to consider each grader's grading bias and reliability when aggregating peer grades [8]. A grader's bias measures the constant inflation and deflation of peer grades given by the grader, and a grader's reliability is the variance in the difference of the peer grades given by the grader towards a group of assignments and the true grades of these assignments. Currently, a group of probabilistic graph models [9]–[13] which consider both of a grader's bias and reliability during the process of aggregating peer grades in peer assessment activities are proposed.

Although these models successfully improve accuracy of estimates to true grades of open-ended assignments, they evaluate graders' grading reliability only based on their knowledge and ability levels which are inferred by graders' historical testing results [13] or by the estimated scores they achieved in the assignment to be graded by them [9]–[12]. In this paper, we argue that a grader's grading reliability is affected not

This work is supported by the National Natural Science Foundation of China (No. 62067001), the Projects of Higher Education Undergraduate Teaching Reform Project in Guangxi (Nos. 2017JGZ103 and 2020JGA116), Innovation Project of Guangxi Graduate Education (No. JGY2021003) and the Special funds for Guangxi BaGui Scholars. This work is partially supported by the Guangxi Natural Science Foundation (No. 2019JJA170045). Pin Lv is the corresponding author (e-mail: lvpin@gxu.edu.cn).

[1]https://www.coursera.org/
[2]https://www.edx.org/

only by the grader's knowledge and ability level but also by the grader's seriousness in the grading activity. For example, a grader with high knowledge and ability level may give a low-quality score for a peer's submission without carefully evaluating the submission. On the contrary, a grader with low knowledge and ability level may give a high-quality score for a peer's submission if the grader seriously reviews the submission based on criteria rules given by teachers. Hence, without taking a grader's grading seriousness into account is inappropriate when we model the grader's reliability.

To solve the limitation of existing peer assessment models, in this paper, we developed two novel probabilistic models of peer assessment (named BPG_6 and BPG_7) by making use of various behaviors of students in grading activities which help to quantify their grading seriousness. First, our self-developed teaching service system[3] which has been serving thousands of students in Guangxi University is revised to capture various behaviors of students within the process evaluating their peers' submissions. Then, a regression model based on GBDT (Gradient Boosting Decision Tree) [13] is constructed to quantify each grader's grading seriousness by exploring the collected behaviors of the grader. After that, the computed value of a grader's grading seriousness together with the estimated score of grader in the assignment to be graded by him/her are employed to optimize the modeling of the grader's reliability, based on which BPG_6 and BPG_7 are designed based on the state-of-the-art probabilistic model framework of peer assessment presented in [12]. The proposed BPG_6 and BPG_7 models are carefully evaluated with a group of peer assessment models on real-world datasets gathered from three courses which are built in our teaching service system. Experimental results show the superiority of our proposals in improving the accuracy of estimating true grades of open-ended assignments in peer assessment by leveraging the information of students grading behaviors. Contributions of this paper include:

1) We proposed to utilize various grading behaviors of students to predict their grading seriousness, which is then employed to optimize the modeling of students' grading reliability and further used to design two novel probabilistic models for peer assessment.
2) We designed inference algorithms based on Gibbs sampling technology to infer the potential variables in the models, including each student's true score, bias and reliability.
3) Our proposals are validated using a real peer assessment datasets captured in teaching practices and the experimental results show our models are superior to the state-of-the-art probabilistic models in terms of the estimation accuracy to true scores of submissions.

The rest of this paper is organized as follows. In Section II, we introduce the related methods for peer assessment. Section III describes the notations and formally defines the problem that we solve. Section IV illustrates our proposed methods for peer assessment. Then, the description of the real dataset we use and the experimental results are presented in Section V, followed by the conclusion and future work in Section VI.

[3]https://hlm.dawnlab.top/

II. Related Work

The core problem of peer assessment is to estimate the true score of open-ended assignments based on graders' feedback, including the peer grades and comments. There are lots of researches on estimating the true score of open-ended assignments in peer assessment. According to the different grading content of the grader' feedback, they can be divided into two categories: ordinal peer assessment and cardinal peer assessment.

A. Ordinal Peer Assessment

The ordinal peer assessment requires every grader to rank the quality of the assignments, and then infers the final ranking of all assignments based on the partial ranking information among assignments given by all graders.

Shah et al. [15]generalized the RBTL (referred BTL) to extend the Bradley-Terry model [16] [17] to estimate the grader's potential evaluation ability and the quality of each assignment from the ordered paired comparison collected from students. Raman et al. [18]also extended several classical ranking aggregation models with different probability distributions, such as MAL (mallows) [19], BT (Bradley Terry) [16], THUR (Thurstone) [20] and PL (Plackett Luce) [21], to learn the full ranking of all the assignments. Moreover, Bayesian techniques were employed to infer the quality of each assignment and the reliability of each grade [22]. To further improve the accuracy of predicted ranking, Mi and Yeung, considering the characteristics of homework and learners, proposed a mechanism to combine both cardinal peer assessment and ordinal peer assessment [23]. To reduce the impact of unreliable graders, Capuano et al. [24] proposed a fuzzy ordinal peer assessment model, named FOPA, utilizing the principle of fuzzy group decision-making. In their follow-up study, multiple evaluation criteria were introduced to expand FOPA to improve the effectiveness of the model [25].

B. Cardinal Peer Assessment

Different from ordinal peer assessment, cardinal peer assessment requires every grader to give a numerical grade for each assignment, and then use the numerical grades given by different graders to estimate the true score of the assignment.

Several iterative algorithms were proposed to estimate the true score of each assignment from the numerical peer grades. Alfaro et al. proposed the Vancouver algorithm [26], which iteratively weights peer grades by the accuracy of the graders to estimate a true score for each assignment. The algorithm measures the accuracy of each grader by comparing the grades of different graders on the same assignment, and gives a higher weight to the grader with a higher accuracy. Walsh proposed another iterative algorithm for peer assessment called PeerRank [27], which is inspired by Google's PageRank algorithm [28]. Assuming that a grader's score reflects his evaluation ability, the PageRank algorithm weighted the peer

grades by graders' scores to learn assignments' true scores iteratively.

Another popular approach for cardinal peer assessment is generative probabilistic graph models. Piech et al. first proposed this approach to estimate the true scores, where the true score, peer grade, graders' reliability and bias are modeled as random variables following a certain probability distribution, and then the values of the above implicit random variables are inferred by observed peer grade of each assignment [9]. Specially, the PG_3 Model in their work sets the current reliability of a grader as a random variable which depends on the linear function of the true score of the grader's assignment. Considering the deterministic linear relationship might be too strict, Mi et al. [10] proposed two extensions of PG_3, called PG_4 and PG_5, by using a probabilistic relationship to lighten up this linear relationship. Moreover, considering that a peer grader's bias is influenced by his friends' bias [29], Chan et al. [11] applied the social interactions among the students collected on a MOOC platform to optimize the modeling a grader's bias and extended the models of PG_1, PG_4 and PG_5. However, the above probabilistic models only consider absolute grades. Therefore, Wang et al. [12] introduced the relative grades, which are the difference of the absolute peer grades between the grades of different submissions given by the same grader, and constructed PG_6 and PG_7 models respectively based on PG_4 and PG_5 models. These two probability models effectively solve the problem of parameter estimation caused by data sparsity, and improve the cardinal peer assessment. Furthermore, Xu et al. [13] based on models PG_6 and PG_7, developed $CD\text{-}PG_1$ and $CD\text{-}PG_2$ respectively, which optimize reliability modeling by obtaining grader's competency information from historical tests by cognitive diagnosis. They have achieved certain improvement, but it is difficult to have enough time and prepared questions to diagnose students' competency in practice.

To sum up, none of the existing cardinal peer assessment methods takes into account the current grading behaviors of the graders. In fact, leveraging the grading behaviors of a grader can measure whether the scorer is serious in scoring, which helps to model the reliability of the scorer more comprehensively, thereby improving the accuracy of peer assessment estimation. To the best of our knowledge, this is the first work to introduce students' grading behaviors into cardinal peer assessment to achieve improved estimation.

III. Problem Definition

Through this paper, we will use the following notations. We use U to denote the collection of all students who have submitted their open-ended assignments on an online platform and u_i represents an arbitrary student in U. Then, we let G represents the collection of graders who grade for those open-ended assignments and g represents an arbitrary grader in G. Note that the students with submissions in peer assessment are also required to grade their peers' submissions, U and G are actually correspond to the same set of students, i.e., $|U| = |G|$. The following are definitions of important concepts which are either observed or unobserved(latent) variables to be estimated. Table 1 summarizes all the notations of variables we will be used in this paper.

- **True scores**: In the proposed models, every assignment submitted by student u is associated with a true score, denoted s_u, which is unobserved and modeled as a random variable following a Gaussian distribution.
- **Peer grades**: the notation z_u^g is used to denote the peer grade given by grader v to student u's submission. Peer grades are observed scores, and the collection of peer grades is denoted as $Z = \{z_i^g | u_i \in U, g \in G\}$.
- **Relative peer grades**: the relative peer grade, denoted by d_{ij}^g, is defined the difference between two observed peer grades given by grader g for the submissions of students u_i and u_j (i.e., z_i^g and z_j^g, respectively).
- **Grader bias**: The bias of grader g, denoted by b_g, is defined as the constant inflation and deflation of peer grades given by the grader. For example, suppose $s_i = 7$, and $b_g = 2$. Then, the mean of the peer grades given by g is $z_i^g = s_i + b_g = 7 + 2 = 9$.
- **Grader reliability**: the reliability of grader g, denoted by τ_g, is defined as the precision of the peer grades given by the grader, reflecting how close on average a grader's peer assessment tend to land near the corresponding submission's true score after having corrected bias. In this work, the reliability of grader g will be determined by his knowledge level and the seriousness value of his behaviors. We assume that the grader with higher ability in the open-ended assignment is the more reliable rater in the assignment.

Different from the existing models assuming the reliability of a grader is only determined by his knowledge level and ability, we propose two new probability graph models by introducing the grader's grading seriousness quantified from student grading behaviors. Our goal is to effectively estimate the true score of each open-ended assignment by modeling the relationship among peer grades, relative grades, the reliability, the bias, and the seriousness of graders, and the true score of assignments. More formally, we define the problem as follows: Given the peer grades, Z, relative peer grades, D, and grading seriousness vector, P. Our goal is to learn τ_g, b_g, for all $g \in G$, and s_i for all $u_i \in U$.

IV. Method

In this section, we describe the details of our solution for peer assessment. Firstly, a GBDT-based regressor learning from student grading behaviors is built to acquire the current grading seriousness of the grader. Then, we propose two probabilistic graph models, namely BPG_6 and BPG_7, which use grading seriousness to extend the PG_6 and PG_7 models in [12] respectively.

A. GBDT-based Regressor to model student behaviors

To collect student grading behaviors, we optimized our university's MOOC platform to support recording student behaviors. When a grader evaluates a submission, he or she needs

TABLE I
NOTATIONS

Notation	Description
U	The set of students
G	The set of graders
s_i	The true score for the submission of student u_i
τ_g	The grading reliability of grader g
b_g	The grading bias of grader g
p_g	The quantified grading seriousness value of grader g
P	The grading seriousness vector composed by grading seriousness values of all graders in G
z_i^g	The grade given by grader g to student u_i's submission
Z	The set of all observed peer grades
d_{ij}^g	Observed relative peer grade between z_i^g and z_j^g
D	The set of all relative peer grades

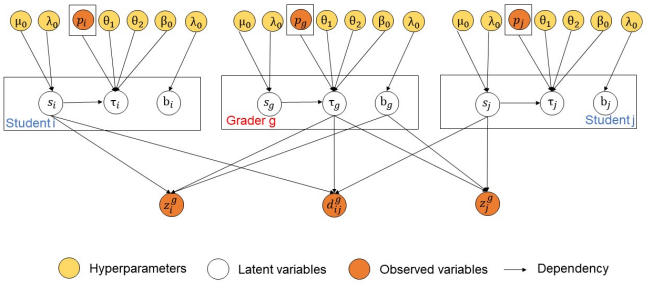

Fig. 1. Graphical Model for BPG_6 and BPG_7

to grade and write comments according to each rule specified by the teacher, and finally choose an overall impression level. As shown in the following table II, the recorded information of each student in an evaluation process includes time consumed, number of reviewing times, grading click track for each rule, word count track for each rule, overall level of the submission, etc. We extract features from the above behavior information and these features are learned by a GBDT regression model to predict a grader's grading seriousness value, which enable us to capture students' reliability.

TABLE II
NOTATIONS

Student's behaviors	Description
Time Consumed	The time of consuming to complete a submission evaluation.
Submit Number	The number of grader g attempts on evaluating a submission.
Grade track	The sequence of each rule recorded by clicking on the optional score. If the grader changes the selected score frequently, it means that the grader may be hesitant.
Comment track	The sequence of the number of words in the comment box per 10 seconds. It will mainly explore the characteristics of input frequency changes and the final number of comments.
The overall level track	the overall level of the submission is different from the grading track. It includes five levels of excellent, good, fair, passing, and failing. If the overall level differs too much from the total grade of all the rules, the grader is most likely to be considered not serious.

B. the BPG_6 model

Our BPG_6 model, an extension to the PG_6 model in [12], introduces grading seriousness acquired based on student grading behaviors. The conditional dependence structure among the random variables in BPG_6 are expressed by graphical model shown in Fig. 1. As shown in the figure, peer grade z_i^g, relative peer grade d_{ij}^g and the evaluation ability p_g for grader g are the observed random variables in the model. The true score for student u_i's submission s_i, grader g's reliability τ_g, and bias b_g are the latent variables in the model to be estimated. The prior distribution of these latent variables is specified by the hyper-parameters μ_0, γ_0, η_0, and β_0. The distributions of all the random variables for BPG_6 model are shown in following (1).

$$\begin{aligned} \tau_g &\sim \Gamma(\theta_1 s_g + \theta_2 p_g, \beta_0) \\ b_g &\sim N(0, \frac{1}{\eta_0}) \\ s_i &\sim N(\mu_0, \frac{1}{\gamma_0}) \\ z_i^g &\sim N(s_i + b_g, \frac{1}{\tau_g}) \\ d_{ij}^g &\sim N(s_i - s_j, \frac{2}{\tau_g}) \end{aligned} \tag{1}$$

In the BPG_6 model, we assume that true score, s_i, follows a Gaussian distribution with the mean equals to μ_0 and the variance equals to $1/\gamma_0$. Though graders may have different biases in peer assessment, we believe that the average bias of all graders is 0. Hence, the grader bias b_g is assumed to follow a zero-mean Gaussian distribution with the variance equals to $1/\eta_0$. Because a grader's reliability in an open-ended assignment is influenced by the knowledge ability and grading seriousness, the reliability of grader τ_g is modeled as a random variable that follows a Gamma distribution where the shape parameter equals to g's comprehensive value $\theta_1 s_g + \theta_2 p_g$ in the assignment, and the rate parameter equals to β_0. The peer grades z_i^g, which is given by grader g to student u_i's submission, follow a Gaussian distribution with the mean equals to the true score of submission s_i plus grader g's bias b_g, and the variance is inversely proportional to grader g's reliability. The relative peer grade d_{ij}^g, corresponding to grader g of grading student u_i's submission and student u_j's submission, is assumed to follow a Gaussian distribution with the mean equals to the difference between the true score for student u_i's submission and the true score for student u_j's submission (i.e., $s_i - s_j$), and the variance equals $2/\tau_g$.

C. the BPG_7 model

The proposed BPG_7 Model is an extension to the PG_7 model in [12] by incorporating the graders' grading seriousness from their behaviors. The conditional dependence

structure of BPG_7 is the same as that in BPG_6, which was shown in Fig.1. The distributions of all the random variables for BPG_7 model is shown as (2).

$$\begin{aligned}
\tau_g &\sim N(\theta_1 s_g + \theta_2 p_g, \beta_0) \\
b_g &\sim N(0, \frac{1}{\eta_0}) \\
s_i &\sim N(\mu_0, \frac{1}{\gamma_0}) \\
z_i^g &\sim N(s_i + b_g, \frac{\lambda}{\tau_g}) \\
d_{ij}^g &\sim N(s_i - s_j, \frac{2\lambda}{\tau_g})
\end{aligned} \tag{2}$$

In BPG_7, the probability distributions of τ_g, z_i^g and d_{ij}^g are different from those in BPG_6. The reliability of grader g in BPG_7 is assumed to follow a Gaussian distribution while the reliability in PG_7 follows Gamma distribution. The scale of τ_g will be determined by $\theta_1 s_g + \theta_2 p_g$, which is a random variable that we cannot tune. Since τ_g will be plugged into the variance of z_i^g, in order to scale the variance of z_i^g, a hyper-parameter λ is introduced. Hence, we assume z_i^g follows a Gaussian distribution with the variance λ/τ_g. In the same way, relative score d_{ij}^g is assumed to follow a Gaussian distribution with $2\lambda/\tau_g$ as the variance.

D. model inference

After formulating the above probabilistic models for peer assessment, the next step is to infer the values of latent variables including the true score of each student, the reliability, the bias of each grader. Based on the observed random variables, involving peer grades, relative peer grades and graders' grading seriousness, the latent variables can be inferred by computing their posterior distribution, $P(\{s_i\}_{u_i \in U}, \{b_g\}_{g \in G}, \{\tau_g\}_{g \in G} | Z, D, P)$. Considering the correlation between latent variables, we apply Gibbs sampling technique [30] to generate samples of a latent variable from an approximated posterior distribution to estimate the latent variables. Specifically, after generating a set of samples of latent variables, we take the empirical mean as the estimated value of the latent variables. Since the burn-in samples are not accurate enough, we run Gibbs sampling for 600 iterations and discard the first 60 burn-in samples for estimating each latent variable.

In the BPG_6 model, because there is no closed-form distribution for student u_i's true score s_i, we apply a discrete approximation to approximate its posterior distribution with intervals of width 0.1. The approximated posterior distributions for the latent variables in BPG_6 are shown in the following (3).

$$\begin{aligned}
&s \propto \frac{\beta_0^{\theta_1 s_i} \tau_i^{(\theta_1 s_i - 1)}}{\Gamma(\theta_1 s_i + \theta_2 p_g)} \times exp(\frac{R}{2}(s_i - \frac{Y}{R})^2) \\
&where\ R = \gamma_0 + \sum_{g \in G_{u_i}} \tau_g + \sum_{g \in G_{u_i}} \sum_{u_j \in U_g} \frac{\tau_g}{2}, \\
&Y = \mu_0 \gamma_0 + \tau_g (\sum_{g \in G_{u_i}} (z_i^g - b_g) + \sum_{g \in G_{u_i}} \sum_{u_j \in U_g} \frac{(d_{ij}^g + s_i)}{2}) \\
&\tau \sim \Gamma(\theta_1 s_i + \theta_2 p_g + \frac{|U_g|^2}{2}, \\
&\beta_0 + \frac{\sum_{u_i \in U_g} (z_i^g - s_i - b_g)^2 + \sum_{u_i, u_j \in U_g} \frac{1}{2} (d_{ij}^g - s_i + s_j)^2}{2}) \\
&b \sim N(\frac{\sum_{u_i \in U_g} \tau_g (z_i^g - s_i)}{\eta_0 + |U_g| \tau_g}, \frac{1}{\eta_0 + |U_g| \tau_g})
\end{aligned} \tag{3}$$

In the BPG_7 model, because there is no closed-form distribution for the true score τ_g, we use a discrete approximation to approximate its posterior distribution with intervals of width 0.1. The approximated posterior distributions for the latent variables in BPG_7 are shown in the following (4).

$$\begin{aligned}
&s \propto \frac{\beta_0^{\theta_1 s_i} \tau_i^{(\theta_1 s_i - 1)}}{\Gamma(\theta_1 s_i + \theta_2 p_g)} \times exp(\frac{R}{2}(s_i - \frac{Y}{R})^2) \\
&where\ R = \gamma_0 + \sum_{g \in G_{u_i}} \frac{\tau_g}{\lambda} + \sum_{g \in G_{u_i}} \sum_{u_j \in U_g} \frac{\tau_g (|U_g| - 1)}{2\lambda}, \\
&Y = \mu_0 \gamma_0 + \frac{\tau_g}{\lambda} (\sum_{g \in G_{u_i}} (z_i^g - b_g) + \sum_{g \in G_{u_i}} \sum_{u_j \in U_g} \frac{(d_{ij}^g + s_i)}{2}) \\
&\tau \propto \tau_g^{\frac{|U_g|^2}{2}} \times exp(-\frac{\beta_0}{2}(\tau_g - Y)^2), \\
&where\ Y = \theta_1 s_i + \theta_2 p_g + \sum_{u_i \in U_g} \frac{(z_i^g - s_i - b_g)^2}{\lambda \beta_0} + \\
&\sum_{u_i, u_j \in U_g} \frac{(d_{ij}^g - s_i + s_j)^2}{2\lambda\beta_0}) \\
&b \sim N(\frac{\sum_{u_i \in U_g} \tau_g (z_i^g - s_i)}{\eta_0 + |U_g| \frac{\tau_g}{\lambda}}, \frac{1}{\eta_0 + |U_g| \frac{\tau_g}{\lambda}})
\end{aligned} \tag{4}$$

Algorithm 1 shows the inference algorithm of our BPG_6 model, where T is the number of iterations and B is the number of burn-in samples. The inference algorithms of BPG_7 is the same as that of BPG_6, but with different approximated posterior distributions for the latent variables.

V. EXPERIMENTS

A. real dataset

The dataset used in our experiments is collected from three courses, including "IT English", "Computer System Structure" and "Database Principles", provided by our university's MOOC platform. The peer assessment settings for these courses are as follows:

- There are multiple assignments for each course, and each assignment includes an open-ended question.

Algorithm 1 BPG_6 Inference

Input: $Z, P, D, U, G, T, B, \mu_0, \gamma_0, \beta_0, \eta_0, \theta_1, \theta_2$
Output: (s_i, τ_g, b_g) for all $u_i \in U$ and $g \in G$

$P =$ GDBT(graders' reviewing behaviors)
$\tau_g \sim \Gamma(\theta_1 s_g + \theta_2 p_g, \beta_0)$
$b_g \sim N(0, \frac{1}{\eta_0})$
$s_i \sim N(\mu_0, \frac{1}{\gamma_0})$
$z_u^g \sim N(s_i + b_g, \frac{1}{\tau_g})$
$d_{ij}^g \sim N(s_i - s_j, \frac{2}{\tau_g})$
for $t = 1 \to T$ **do**
  for each s with $u_i \in U$ **do**
    Sample s according to (3)
    $s_{u_i} \leftarrow s$
  end for
  for each τ with $g_i \in G$ **do**
    Sample τ according to (3)
    $\tau_{g_i} \leftarrow \tau$
  end for
  for each b with $g_i \in G$ **do**
    Sample b according to (3)
    $b_{g_i} \leftarrow b$
  end for
  $\xi^{(t)} \leftarrow (\{s_i | u_i \in U\}, \{\tau_g | g \in G\}, \{b_g | g \in G\})$
end for
$(\{\hat{s}_i | u_i \in U\}, \{\hat{\tau}_g | g \in G\}, \left\{\hat{b}_g | g \in G\right\}) \leftarrow \frac{1}{T-B} \sum_{t=B+1}^{T} \xi^{(t)}$
return $(\{\hat{s}_i | u_i \in U\}, \{\hat{\tau}_g | g \in G\}, \left\{\hat{b}_g | g \in G\right\})$

- For each assignment, every student who has submitted an open-ended assignment is asked to evaluate three other submissions, according to rules specified by the teacher. Grader assignment is done randomly by the platform to ensure that each submission is evaluated by three graders. And the whole peer assessment process is double-blind.
- After receiving the peer grades, the median of peer grades is used as the final grades for submissions.

Besides the peer grades given by the students, this dataset contains grades given by two experienced teachers for each assignment. The average grades assigned by the two teachers are considered as the ground truth scores in evaluation. It is worth mentioning that the dataset also records student grading behaviors of all graders. Table 2 gives summary statistics for the three courses involved in peer assessment.

B. evaluation metrics

The RMSE (Root Mean Square Error) is used to measure the deviations of the predicted grades from the ground truth scores assigned by teachers. RMSE is a widely used metric for evaluating the performance of cardinal peer assessment methods [9]–[13]. The formal definition of RMSE is given by (5), where s_i denotes the ground truth score of student u_i's submission; $\hat{s}_i$ represents the estimated true score by a peer assessment method, and U denotes the set of graded submissions. The lower the RMSE of the estimated scores, the higher the accuracy of the estimated scores.

$$RMSE = \sqrt{\frac{1}{|U|} \sum_{u_i \in U} (s_i - \hat{s}_i)^2} \tag{5}$$

TABLE III
SUMMARY STATISTICS OF ASSIGNMENTS FOR PEER ASSESSMENT

	IT English	Comp. Sys. Structure	DB Principles
Assignments	7	9	5
Submissions	427	442	516
Teacher grades	854	884	1032
Peer grades	1279	1297	1526
Student behavior records	1279	1297	1526
Full grades	20	20	20

C. comparison methods

In order to evaluate the performance of our proposed models, BPG_6 and BPG_7, we compare them with 4 baseline methods discussed as follows.

- **Mean**: This method simply takes the mean value of peer grades as the final grade.
- **Median**: This method takes the median of peer grades as the final grade.
- PG_6: This is a probabilistic model, which is the advanced method to solve the peer assessment problem of open-ended assignments. This model introduces the relative peer grades and assumes the prior distribution of grader reliability satisfies a Gamma distribution. The BPG_6 model is an extension of this model.
- PG_7: Similar to the PG_6 model, the PG_7 also introduces the relative peer grades, but assumes that a grader's reliability follows a Gaussian distribution. The BPG_7 model is an extension of this model.

D. experimental settings

Due to the differences in the content characteristics of different courses, we train a GBDT-based regressor for each course to predict a grader's grading seriousness value. The input of the regression model is the behavior features extracted from the student grading behaviors, and the output is the students' grading seriousness for one's submission defined by the following (6). It can be seen from the formula that the value range of a grader's grading seriousness is [0,1]. The larger the value is, the closer the score assigned by the grader is to the teacher's score, and the more serious the grader's evaluation is.

$$p_{g \to s_i} = \frac{full\ score - |teacher\ score - Z_i^g|}{full\ score} \tag{6}$$

To verify the effectiveness of the proposed model, for each course, we choose to leave a certain assignment data as the test set, and other assignments as the training and verification data

to train the GBDT-based regressor. For example, the GBDT regressor for assignment 1 in IT English is built based on the data of assignment 2-5. Then the trained GBDT regressor is used to predict the grading seriousness of all graders in assignment 1. In the same way, each assignment of each course is processed to get the grading seriousness values of the corresponding graders.

After obtaining the grading seriousness by GBDT-based regressor, the model inference algorithm will be carried out. As described before, there are many hyper-parameters used in the proposed probabilistic models and baselines, and it is important to set reasonable values for these hyper-parameters to ensure the accurate model estimation. Since the proposed model is an extension of PG_6 and PG_7 models in [11], in order to verify the effect of using grading seriousness information, we set the same values for the shared hyper-parameters in the proposed models and PG_6 and PG_7 models. For the most important latent variable s_i, which denotes the true score of student u_i's submission to an assignment, we use the mean and variance of the peer grades as the mean (μ_0) and variance ($\frac{1}{\beta_0}$) of its prior distribution. As claimed in [12], the β_0 in the $PG6$ and BPG_6, which is the rate parameter in the gamma distribution for the grader's reliability, and λ in BPG_7 and PG_7, which determines the variance of the Gaussian distribution for peer grades, are the most critical hyper-parameters. From the probabilistic models, these two hyper-parameters have a significant influence on the estimation accuracy of the true scores, while other hyper-parameters influence the estimation accuracy slightly with set in a reasonable range. Therefore, we mainly tune β_0 in PG_6 and BPG_6 and λ in PG_7 and BPG_7 in our experiments by following the tuning idea proposed in [10] [12] [13]. Specifically, We search for these two hyper-parameters in the range of [100, 400] with the interval of 50 to get the best performance. We set η_0 to 0.1 in our experiment, and β_0 is set to 0.1 in the PG_7 and BPG_7 model. Meanwhile, we try multiple combinations for θ_1 and θ_2 in the range [0.5, 2] with the interval of 0.4 in BPG_6 and BPG_7. For every model, the model inference algorithm was executed 10 times to infer the values of latent variables, and the average estimated results over 10 runs were reported. During each execution, every latent variable was sampled based on the Gibbs sampling method for 600 iterations, and the first 60 iterations are burn-in iterations that will be discarded.

E. performance on real data

The experimental results are shown in Table IV. The RMSE reported in this table is the average of ten repetitions per assignment per course. STD represents the standard deviation of the RMSE. For all the above probabilistic models, we tune the hyper-parameters to the ones which achieved the lowest RMSE on the dataset. Overall, all the probabilistic graph models are more accurate than the Mean and the Median method, which fail to consider the reliability and the bias of graders. The proposed BPG_6 and BPG_7 methods are the most accurate methods compared with the other advanced solutions.

Since our BPG_6 and BPG_7 are extensions of PG_6 and PG_7 respectively by incorporating student grading behaviors, we compare these two pairs of models as follows:

- From Table IV we can see that BPG_6 and PG_6 have similar STDs of RMSE for all assignments in three courses, and all their STD values are small. This indicates that both models act quite stably in estimating the true scores. It can also be observed from the table that the RMSE of BPG_6 is significantly lower than that of PG_6. Moreover, BPG_7 achieved the best performance in both IT English and Database Principles.
- As can be seen from Table IV, the RMSE of BPG_7 is also significantly lower than that of PG_7 for all assignments in three courses. The maximum STD value of the two models is 0.02, indicating that the two models are also very stable in estimating the true score. In particular, BPG_7 performed better than other models for assignments of Computer System Structure.

In sum up, by incorporating student grading behaviors in the open-ended assignments to optimize the modeling of graders' reliability, the BPG_6 and BPG_7 methods successfully improve the accuracy of peer assessment.

TABLE IV
EVALUATION OF DIFFERENT PEER ASSESSMENT MODELS (THE BEST RESULTS ARE IN BOLD AND THE NEXT BEST RESULTS ARE UNDERLINED)

	IT English		**Comp. Sys. Structure**		**DB Principles**	
	MEAN	STD	MEAN	STD	MEAN	STD
Mean	2.68	–	2.78	–	2.36	–
Median	2.86	–	3.07	–	2.43	–
PG_6	1.69	0.01	2.39	0.01	2.13	0.01
BPG_6	**1.12**	**0.01**	<u>2.12</u>	<u>0.01</u>	**1.28**	**0.01**
PG_7	1.35	0.02	2.37	0.02	2.16	0.02
BPG_7	<u>1.23</u>	<u>0.02</u>	**2.00**	**0.01**	<u>1.58</u>	<u>0.02</u>

F. sensitivity of hyper-parameters

To show how hyper-parameter β_0 in the BPG_6 model and hyper-parameter λ in the BPG_7 model will influence the performance, we conduct experiments using different values of these two hyper-parameters with all other parameters being fixed. In particular, the value of β_0 in the BPG_6 model was set in the range of [100, 400] with an interval of 50, and the value of λ in the BPG_7 model was set in the range of [50, 300] with an interval of 50. The results in Figure 4 indicate that in a reasonable range these two models are robust to the settings of the hyper-parameters and achieve acceptable performance.

VI. CONCLUSION AND FUTURE WORK

In this paper, two new probabilistic graph models are proposed for peer assessment by incorporating students' grading behaviors. We first build a GBDT-based regressor to predict the graders' grading seriousness based on student grading behaviors in every assignment, and then leverage such information to optimize the modeling of graders' reliability. Moreover, an effective inference algorithm is proposed to infer

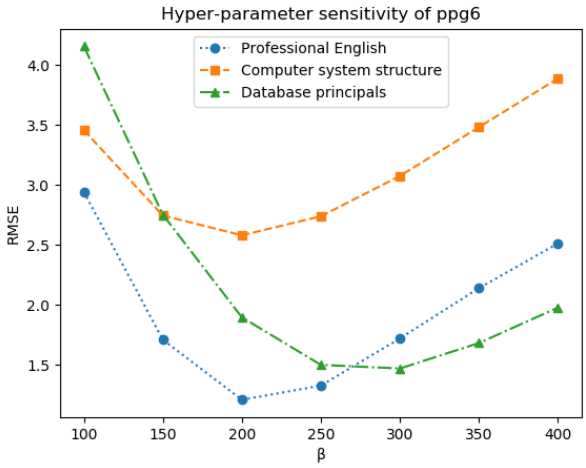

Fig. 2. Sensitivity analysis of hyper-parameter β_0 for BPG_6

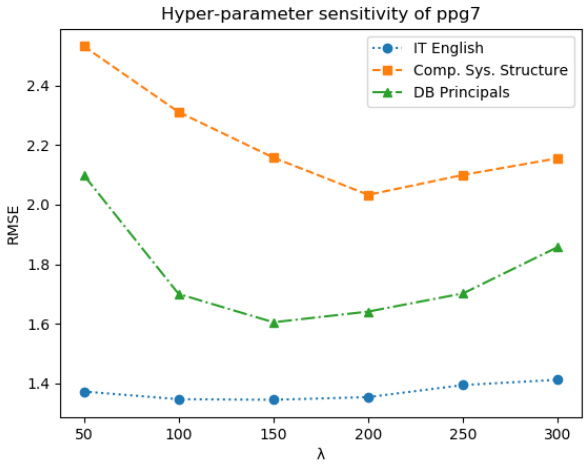

Fig. 3. Sensitivity analysis of hyper-parameter λ for BPG_7

both model parameters and the true scores of students' assignments. Experimental results based on a real peer assessment dataset demonstrate that the two proposed models improve the accuracy of estimating true score for peer assessment. It can also be observed that the evaluation ability acquired from students behaviors indeed contribute to the improvement in the accuracy of estimating true score for peer assessment.

In addition to the area of MOOCs, our proposed models can be applied to crowdsourcing, where a crowdsourcing task needs to predict a metric based on the behaviors of crowdworkers. In the future, we will attempt to introduce other factors that affect the reliability and bias of graders, such as answering behaviors and comments, to further improve the peer assessment estimation.

Acknowledgment

We would like to thank Fei Mi from the Hong Kong University of Science and Technology and Hou Pong Chan from The Chinese University of Hong Kong for providing codes of related probabilistic graph models to them.

References

[1] I. Caragiannis, G. A. Krimpas, and A. A. Voudouris, "Aggregating partial rankings with applications to peer grading in massive online open courses," *computer science*, pp. 675–683, 2014.

[2] D. Paré and S. Joordens, "Peering into large lectures: examining peer and expert mark agreement using peerscholar, an online peer assessment tool," *Journal of Computer Assisted Learning*, vol. 24, no. 6, pp. 526–540, 2008.

[3] K. Topping, "Peer assessment between students in colleges and universities," *Review of Educational Research*, vol. 68, no. 3, pp. 249–276, 1998.

[4] P. M. Sadler, "The impact of self- and peer-grading on student learning: Educational assessment: Vol 11, no 1," *Educational Assessment*, 2006.

[5] N. Falchikov, *Learning together: Peer tutoring in higher education*. Learning Together: Peer Tutoring in Higher Education, 2001.

[6] E. F. Gehringer, "A survey of methods for improving review quality," in *New Horizons in Web Based Learning - ICWL 2014 International Workshops, SPeL, PRASAE, IWMPL, OBIE, and KMEL, FET, Tallinn, Estonia, August 14-17, 2014, Revised Selected Papers*, 2014, pp. 92–97.

[7] J. W. Strijbos and D. Sluijsmans, "Unravelling peer assessment: Methodological, functional, and conceptual developments," *Learning and Instruction*, vol. 20, no. 4, pp. 265–269, 2010.

[8] F. G. Loro, S. Martín, J. A. R. Valiente, E. S. Ruiz, and M. Castro, "Reviewing and analyzing peer review inter-rater reliability in a MOOC platform," *Comput. Educ.*, vol. 154, p. 103894, 2020.

[9] C. Piech, J. Huang, Z. Chen, C. B. Do, A. Y. Ng, and D. Koller, "Tuned models of peer assessment in moocs," in *Proceedings of the 6th International Conference on Educational Data Mining, Memphis, Tennessee, USA, July 6-9, 2013*, 2013, pp. 153–160.

[10] F. Mi and D. Yeung, "Probabilistic graphical models for boosting cardinal and ordinal peer grading in moocs," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, 2015, pp. 454–460.

[11] H. P. Chan and I. King, "Leveraging social connections to improve peer assessment in moocs," in *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, 2017, pp. 341–349.

[12] T. Wang, Q. Li, J. Gao, X. Jing, and J. Tang, "Improving peer assessment accuracy by incorporating relative peer grades," in *Proceedings of the 12th International Conference on Educational Data Mining, EDM 2019, Montréal, Canada, July 2-5, 2019*, 2019.

[13] J. Xu, Q. Li, J. Liu, P. Lv, and G. Yu, "Leveraging cognitive diagnosis to improve peer assessment in moocs," *IEEE Access*, vol. 9, pp. 50 466–50 484, 2021.

[14] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[15] N. B. Shah, J. K. Bradley, A. Parekh, and K. Ramchandran, "A case for ordinal peer-evaluation in moocs," 2013.

[16] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, 1952.

[17] D. R. Luce, "Individual choice behavior: A theoretical analysis," *Journal of the American Statistical Association*, vol. 67, no. 293, pp. 1–15, 2005.

[18] K. Raman and T. Joachims, "Methods for ordinal peer grading," in *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, 2014, pp. 1037–1046.

[19] C. L. Mallows, "Non-null ranking models. i," *Biometrika*, vol. 44, no. 1/2, pp. 114–130, 1957.

[20] Thurstone and L. L., "The method of paired comparisons for social values." *Journal of Abnormal & Social Psychology*, vol. 21, no. 4, pp. 384–400, 1927.

[21] B. R. Dansie, "The analysis of permutations," 1988.

[22] A. E. Waters, D. Tinapple, and R. G. Baraniuk, "Bayesrank: A bayesian approach to ranked peer grading," in *Proceedings of the Second ACM Conference on Learning @ Scale, L@S 2015, Vancouver, BC, Canada, March 14 -18, 2015*, 2015, pp. 177–183.

[23] O. Luaces, J. Díez, A. Alonso-Betanzos, A. T. Lora, and A. Bahamonde, "A factorization approach to evaluate open-response assignments in moocs using preference learning on peer assessments," *Knowl. Based Syst.*, vol. 85, pp. 322–328, 2015.
[24] N. Capuano, V. Loia, and F. Orciuoli, "A fuzzy group decision making model for ordinal peer assessment," *IEEE Trans. Learn. Technol.*, vol. 10, no. 2, pp. 247–259, 2017.
[25] N. Capuano, S. Caballé, G. Percannella, and P. Ritrovato, "FOPA-MC: fuzzy multi-criteria group decision making for peer assessment," *Soft Comput.*, vol. 24, no. 23, pp. 17 679–17 692, 2020.
[26] L. de Alfaro and M. Shavlovsky, "Crowdgrader: a tool for crowdsourcing the evaluation of homework assignments," in *The 45th ACM Technical Symposium on Computer Science Education, SIGCSE 2014, Atlanta, GA, USA, March 5-8, 2014*, 2014, pp. 415–420.
[27] T. Walsh, "The peerrank method for peer assessment," in *ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic - Including Prestigious Applications of Intelligent Systems (PAIS 2014)*, 2014, pp. 909–914.
[28] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," *Stanford Digital Libraries Working Paper*, 1998.
[29] S. Yang, B. Long, A. J. Smola, N. Sadagopan, Z. Zheng, and H. Zha, "Like like alike: joint friendship and interest propagation in social networks," in *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, 2011, pp. 537–546.
[30] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, 1984.