IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 23, 2021, accepted December 10, 2021, date of publication December 16, 2021,
date of current version December 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3136352

A Graph Analysis Method to Improve Peer Grading Accuracy for Blended Teaching Courses

XING DU, XINGYA WANG, AND YAN MA
College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China

Corresponding author: Xing Du (dux@cqnu.edu.cn)

This work was supported by the Science and Technology Bureau of Chongqing under Grant cstc2019jscx-mbdxX0061.

ABSTRACT Peer grading is a tool widely used by instructors to provide fast reviews for homework that consists of open-ended questions. As the grades obtained from peer grading are not as accurate as those provided by the instructors, many methods have been proposed to improve the accuracy of peer grading. However, the current methods mainly focus on the scenario of online teaching, and they lose effectiveness in blended teaching courses because the mandatory of task and affinity among students may make the students perform irresponsibly in the grading task. This paper proposes a method based on graph analysis to improve the accuracy of peer grading. The peer grading system is modeled as a bipartite graph. In the graph, three interdependent metrics are defined to measure the dutifulness of the grader, the reliability of the rating and the true score of the submission. The stable values of the metrics are computed in an iterative way to obtain the peer grading results. Experiments demonstrate the proposed method is effective in blended teaching settings, and outperforms the current methods. Compared to the baseline of the mean value method, the proposed method decreases the root mean square error by 2.31% in the worst case and 30.72% in the best case on real-world data. It is robust to irresponsible graders, where the root mean square error keeps small even when the proportion of irresponsible graders increases to 30%.

INDEX TERMS Accuracy of grading, blended teaching, graph analysis, open-ended question review, peer grading.

I. INTRODUCTION

Blended teaching is a form of teaching that combines online teaching and traditional classroom teaching. With the popularization of online teaching, some advanced technologies employed by online platforms [1], [2] have attracted school instructors, and many traditional classroom courses have turned to blended teaching courses by moving part of the teaching procedures online. Peer grading is a tool extensively used both in online teaching and blended teaching courses to provide fast reviews for open-ended questions (e.g., essay writing, mathematical proof and short answer questions). A peer grading system requires each student to grade a small number of his peers' assignment submissions. The final grade of a submission is an aggregate of the scores rated by the peers. Peer grading benefits both instructors and students. For the instructors, it relieves their workload, and provides them an easy and fast way to evaluate the students' homework even when the number of students is huge. For the students, it enables them to learn their peers' merits, develop self-learning abilities and enhance critical thinking [1], [3].

Although peer grading has many benefits, the results of peer grading are not as accurate as those given by instructors, because the student graders vary in abilities and attitudes [4]. Numerous researches have devoted to improve the accuracy of peer grading [4], [5], [6], [7]. The current researches mainly deal with the circumstance of massive open online courses (MOOCs). In MOOCs, the students usually take a course according to their own wills, and they usually do not know each other. So almost all of the current methods are based on the assumption that the student graders are responsible, and the bias of peer grading results is mainly due to the graders' lack of knowledge or skills. There are great disparities between the MOOCs and the blended teaching courses. First, the motivations of the students are different. For a blended teaching course, although some teaching activities are conducted online, the course is essentially a school

The associate editor coordinating the review of this manuscript and approving it for publication was Chien-Ming Chen.

course. To obtain credit for the school course, the students must participate in the teaching activities even though they are not interested in it. In contrast, the students just do not take part in the teaching activities or simply quit the course if they are not interested in a MOOC, since they usually have no pressure to earn credit. We have done some preliminary experiments on blended teaching courses and MOOCs offered by our university. For blended teaching courses, we have observed that some students grade the submissions even without downloading the file of grading criteria. For MOOCs, this phenomenon is rare, but more students just do not participate in the peer grading task. So it can be concluded that there are much more randomly rated scores in blended teaching courses. Second, in blended teaching courses, students are more likely to collude due to their familiarity with each other. Though the peer grading process is double blind, we have observed some students give a full score or a near full score to all the allocated submissions in the blended teaching courses, while students in MOOCs seldom do this. Due to the two reasons, the problem of irresponsible rating is critical in blended teaching courses. The basic assumption of the current methods for MOOCs does not hold in blended teaching courses, and their performance degrades severely.

To address this problem, we propose a graph analysis method which aggregates the peer grading scores by measuring the dutifulness of the graders. The proposed method models the peer grading system as a directed bipartite graph. Three metrics are devised to measure the unobserved intrinsic properties of the graph model: the dutifulness of a grader, the reliability of a rating, and the true score of a submission. The metrics are formulated as mutually recursive equations to estimate the values of the properties. The estimated value of the true score is taken as the aggregate of peer grading scores. This method only relies on scores rated by the peer graders, and needs no extra information (e.g., text content of the submission, behavior data of the grader or calibration information of the grader), which makes it easy to implement. Experiments on both real-world and synthetic data verify the effectiveness of the proposed method in blended teaching settings, and investigate the undutiful graders' impact on the performance of several peer grading methods.

In summary, the contributions of this work are as follow.

- A graph-based method is proposed to improve the accuracy of peer grading specifically for blending teaching courses. This method addresses the problem of rating peers irresponsibly, which is critical in blended teaching courses but neglected by most of the current peer grading methods.
- Experiments on real-world and synthetic data illustrate the proposed method outperforms some popular methods when dealing with blended teaching courses, and systematically investigate the undutiful graders' impact on peer grading performance.

II. RELATED WORK

From the viewpoint of grading types, there are two types of peer grading methods, namely cardinal and ordinal. In cardinal peer grading, the evaluation for a submission is a numerical score. In ordinal peer grading, the evaluation for a submission is a rank among all the submissions. Some works [8], [9] declare that the ordinal assessment (e.g., compares which one is better between Submission A and Submission B) is easier to perform and gets more reliable result than the cardinal one (e.g., reports the score of Submission A is 80). Some other works [10], [11] argue that it is not easy for the students to rank the submissions, especially when a pair of submissions is roughly equivalent. To address the problem of the disparity between submissions is indistinct some techniques, such as fuzzy logic, are applied to ordinal peer grading [11], [12]. In general, the cardinal setting is more natural for assessing the homework and more commonly adopted in practice. This work studies the problems in cardinal peer grading. So the rest of this section gives a brief introduction of the work in cardinal peer grading.

The cardinal peer grading methods can be roughly categorized into three types: calibration-based, statistics-based, and graph-based methods.

A. CALIBRATION-BASED METHODS

The calibration-based methods introduce a calibration stage. Usually, calibration is performed before the peer grading task, namely a small number of submissions, which have been rated by the instructor, must have been graded before a student assesses his peers' submissions [4], [13]. The disparity between the scores rated by the student and those rated by the instructor is calculated to evaluate the reliability of the student. The more reliable a grader is, the more his rating weights in the aggregated score. The calibration technique is easy to implement and has shown effectiveness in some experiments. However, it neglects that the behavior of a student is changing over time, and the reliability measured during the calibration stage may not represent the reliability during the grading stage. Therefore, more sophisticated methods which are based on analyzing the students' behaviors during the grading process are developed.

To reduce the students' burden of grading extra submissions required by calibration, recent researches [14], [15] have proposed methods to perform calibration after the grading task. These methods pick out a small set of submissions that have been assessed by the students to send to the instructor for grading. Then the scores given by the instructor are compared with those given by the students to evaluate the graders' reliability. Since only a few of the submissions are graded by the instructor, only a small portion of the graders can be calibrated. Additional techniques are used to make the calibration more effective. In [14] the calibrated submissions are randomly selected, but the graders are allocated to the submissions in a special way so that the calibration information can propagate to all the graders. Specifically, it builds an assessment grid, which is a matrix allocating the graders for each submission. Each row of the matrix represents a submission, and each column represents a grader. It requires adjacent graders (e.g., adjacent columns in the matrix) have some common submissions allocated, which ensures the calibration information can propagate via adjacent graders. In [15] a spot-checking method is proposed to select the submissions need to be assessed by the instructor under the criterion of maximizing the grading accuracy. However, this method requires the ratings to be binary (''correct'' or ''incorrect''), which limits its application to a real-world course.

B. STATISTICS-BASED METHODS

The mainstream statistics-based peer grading methods [5], [6], [16] are based on the Bayesian Theory. The basic philosophy of the Bayesian Theory is the observed variable follows a probability distribution containing parameters, and the parameters represent the hidden variables need to be inferred. The parameters are also random variables each of which has a prior distribution obtained empirically. The posterior distribution of each parameter is inferred according to the values of the observed variable and the prior distributions of the parameters. And finally, the estimated value of a hidden variable is computed from its posterior distribution. The statistics-based peer grading methods use a Bayesian network to model the variables in the grading system. The true score of each submission, the bias and reliability of each grader are treated as hidden variables in the Bayesian network, and the peer grading scores are treated as observed variables. The hidden variables are inferred by fitting the Bayesian network on the observed variables using a Markov Chain Monte Carlo algorithm [27]. The key issue is how to model the distributions of the variables. The works in [5] and [6] have shown modelling the variables using different distributions affect the performance obviously.

The work in [17] proposes a novel method for peer grading. Different from the other methods, it treats the task of peer grading as answering multiple choice questions, and each option of a multiple choice question is a student's submission. The task of peer grading (answering multiple choice questions) and the task of finishing homework (answering open-ended questions) are combined to a uniform task. The probability of correctly finishing the uniform task is modelled using the 1-PL IRT Rasch model [18]. The aggregated result of peer grading is obtained by maximizing the probability of correctly finishing the uniform task using an expectation maximization algorithm. However, this method requires the options of the multiple choice question to be independent, which means identical or similar answers of the homework (the open-ended question) should be removed when preparing the options of the multiple choice question. Removing the identical or similar answers needs to analyse the content of the answers, which is difficult to conduct automatically.

All the above statistical models assume that students who receive higher scores from peers are more reliable, and they tend to rate their peers more accurately. This assumption does not hold in blended teaching courses. We have observed that some students are not willing to participate in the peer grading task, and thus they just rate the submissions randomly, and some students collude to give full scores to all the submissions allocated to them. The research in [19] shows similar phenomenon in the small private online courses, which have similar settings to the blended teaching courses, and demonstrates the statistical models fail to deal with this situation. To solve this problem, [19] proposes a hybrid framework that uses auto-grader to filter out the unreliable ratings. However, the auto-grader can only grade text-based questions (e.g., essay or question-and-answer questions), which limits its application.

C. GRAPH-BASED METHOD

The graph-based methods model the peer grading system as a graph. The proposed method in this paper falls into this type. The main idea of this kind of methods is to represent the graders and submissions as nodes of a graph and the rating relationship as edges, and the scores are aggregated by iteratively updating the values of the nodes and edges. Inspired by the PageRank algorithm [20] which has been used by the Google search engine, the PeerRank algorithm [7] is proposed for peer grading. It measures the reliability of a grader using the score of his submission. The score of a submission is a weighted sum of the scores given by the peer graders, where the weights are the corresponding graders' reliabilities. The score is updated iteratively until convergence. In [21] and [22], two modified versions of PeerRank are proposed, both of which apply a non-linear function to the value of the reliability to increase the difference among the graders. The same as the statistical methods, the algorithms in [7], [21], and [22] are based on the assumption that the graders are dutiful, and their reliabilities are in accordance with their mastery of the knowledges. The work in [23] proposes a trust graph algorithm to compute the weight of each grader. It needs some ground truth data, a set of submissions graded by the instructor, to determine each grader's trustiness. The method in [10] measures each student's rating accuracy by computing the difference between the scores provided by the student and the scores of the corresponding submissions given by other students. Different from the other methods, a grader's grading accuracy is used to modify his own score rather than the gradee's score. This mechanism introduces evaluation of the grading performance, which incents the students to grade the submissions more accurately. Nevertheless, as the final score contains an incentive component, the difference between the peer grading results and scores given by the instructors becomes even larger.

Due to the fact that undutiful graders impact peer grading results severely in blended teaching courses, it is necessary to consider the dutifulness of the graders when aggregating the scores. In the field of e-commerce, online marketplaces (e.g., Amazon, eBay, and Taobao) usually utilize user-based review systems to rate the quality of products. If we treat the student graders as the users, and the homework submissions as the products, it can be seen that the peer grading system for homework is similar to the review system for products. For the product review system, a variety of methods [24], [25], [26] have been proposed to improve the quality of the reviews based on evaluating the trustiness of the users. The work in [24] has devised a graph model to detect untrustworthy users utilizing the ratings given by all the users. The graph consists of two types of nodes: user nodes and product nodes. The two types of nodes are connected by the ratings that are given to the products by the users. In the graph model, three properties, namely fairness, reliability and goodness, are respectively defined for the users, ratings and products. The values of the three properties are iteratively updated until converge. Then the users with small values of fairness are removed to improve the quality of the reviews.

Inspired by the idea of [24], we propose a method to evaluate the dutifulness of the graders, and aggregate the peer grading scores based on the dutifulness of the graders. The structure of the proposed method is similar to the graph model in [24]. It consists of grader nodes and submission nodes, and the nodes are connected by the ratings. The property of dutifulness is defined for the grader nodes, the property of true score is defined for the submission nodes, and the property of reliability is defined for the ratings. However, the proposed method is different from the model in [24]. The purpose of [24] is to find out the untrustworthy reviewers in the review system. It aims to obtain an accurate estimation of the property of the reviewers (e.g., the fairness of the user). Our purpose is to get an accurate assessment of the quality of the reviewed object. So the proposed method tries to get an estimation of the property of the reviewed object (e.g., the true score of the submission) as close as the ground truth. It can be seen in the next section, the definition of the quality of the reviewed object in our method (e.g., the true score of the submission) is completely different from that in [24] (e.g., the goodness of the product), because the definition in [24] is too rough for estimating the quality of the reviewed object.

The proposed method, the PeerRank [7] and the modifications of PeerRank [21], [22] all follow the idea of getting the aggregated results by weighting the peer grading scores. But the proposed method is different from the PeerRank and its modifications in essence, as the weights obtained by the proposed method measure a property of the graders that is totally different from those obtained by the PeerRank and its modifications. The core idea of PeerRank and its modifications is to weight the scores according to the graders' performance on finishing the homework. Due to the students' undutiful behaviors in the peer grading process of a blended teaching course, the performance on finishing the homework is not equivalent to the performance on assessing the submissions. The proposed method weights the scores according to the dutifulness of the graders which is a new metric defined by us. From the definition, we will see the dutifulness can be regarded as a measurement of the discrepancy between the scores given by the grader and the true scores of the submissions, and thus the dutifulness represents the grader's performance on assessing the submissions.

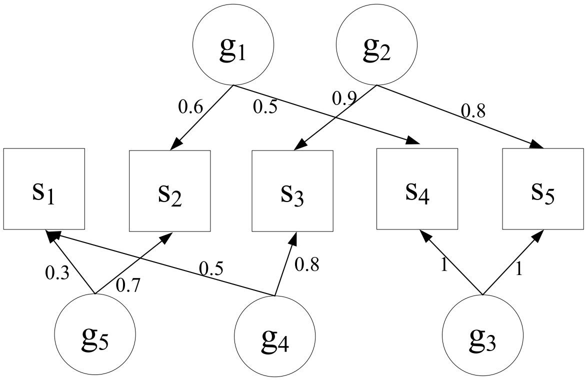

FIGURE 1. A sample model of peer grading. The nodes $s_1, s_2, \cdots, s_5$ denote submissions committed by students $g_1, g_2, \cdots, g_5$ respectively. The students $g_1, g_2, \cdots, g_5$ also act as graders, each of whom rates 2 submissions. The weights of the edges denote the rated scores.

III. THE PROPOSED METHOD FOR PEER GRADING

A graph model is proposed to improve the accuracy of peer grading by addressing the students' irresponsible behaviors in the grading process.

A. MODEL DEFINITION

Peer grading works in a way that each student generates a submission for a homework assignment, and then each submission is randomly allocated to a small number of other students to be graded (each student is also limited to grade a small number of submissions), and finally an aggregate of the scores given by the students is computed for each submission. This scenario can be modeled as a directed bipartite graph $M(G, S, E)$. As shown in Figure 1, each node $g \in G$ represents a grader (student), each node $s \in S$ represents a submission, and each directed edge $(g, s) \in E$ represents a given submission s is rated by a specific grader g. The weight of the edge represents the score of a submission rated by a grader.

Through the paper, we will use the following notation. The collection of all graders (students) is denoted by G, and a specific grader by g. The collection of all submissions is denoted by S, and a specific submission by s. The collection of all ratings is denoted by E, and a specific rating by (g, s) which represents grader g grades submission s. The score of submission s given by grader g is denoted by $C(g, s)$. The collection of graders who rate a specific submission s is denoted by $In(s)$. The collection of submissions which are rated by a specific grader g is denoted by $Out(g)$, and the number of submissions in $Out(g)$ is denoted by $|Out(g)|$.

We assume that each grader has an intrinsic value of dutifulness which determines to what extent the scores provided by him can be trusted when aggregating the scores. Based on this assumption, a model called DRT (Dutifulness, Reliability, and True score) is proposed. The DRT model assumes a peer grading system contains 3 types of latent variables:

dutifulness of a grader, reliability of a peer grading score, and true score of a submission. By inferring these variables, we can get an estimation of the ground truth score for each submission.

- **Dutifulness of a grader**: This quantity measures on average the trustiness of a grader. A grader with high dutifulness tends to grade a submission with small difference from the ground truth, while a grader with low dutifulness tends to grade a submission with large deviation from the ground truth. The dutifulness of a grader g is denoted by $D(g)$, and its value lies in the interval [0, 1]. 0 represents the grader cannot be trusted at all, while 1 represents the grader can be totally trusted.
- **Reliability of a peer grading score**: Every peer grading score $C(g, s)$ is associated with a value of reliability $R(g, s)$. This value describes to what extent the score can be trusted. As the attitude of a grader changes with time, the reliability of his rating is dynamic. The more serious the attitude is, the smaller the difference between the ground truth and rated score is. For example, a student might get bored after rating some submissions and the deviation of his rating from the ground truth will go up. $R(g, s)$ takes a value in the interval [0, 1]. 0 means a score is 100% untrustworthy, while 1 means a score is 100% trustworthy.
- **True score of a submission**: Each submission s is associated with an underlying true score $T(s)$, which indicates how well the homework is finished by a student. For simplicity of calculation, value of $T(s)$ is normalized to [0, 1], where 0 means no questions are correctly answered, while 1 means all questions are correctly answered. The peer grading score $C(g, s)$ is also normalized to [0, 1] to have the same range as $T(s)$.

The dutifulness of a grader is an average measurement of the trustiness of all the ratings given by him. We define the dutifulness of a grader as Eq. (1).

$$D(g) = \frac{\sum\limits_{s \in Out(g)} R(g, s)}{|Out(g)|}, \quad \text{for } g \in G \tag{1}$$

We assume that the reliability of a specific rating is determined by two factors, the dutifulness of the grader and his attitude when rating the submission. The dutifulness reflects the overall bias of the grader, while the attitude reflects the fluctuation when grading a submission. We use the absolute distance between the true score $T(s)$ and the rated score $C(g, s)$ to measure the grader's attitude. The reliability is formulated as Eq. (2), where α is a parameter in (0, 1), which tunes the weights of the dutifulness and the attitude. The larger α is, the more the reliability of a specific rating is determined by the grader's overall bias on all the submissions rated by him. The smaller α is, the more the reliability of a specific rating is determined by the grader's attitude to this specific submission. In practice, α should be set to a suitable value to strike a balance between the overall bias and specific attitude.

$$R(g, s) = \alpha D(g) + (1 - \alpha)(1 - |C(g, s) - T(s)|), \quad \text{for } (g, s) \in E \tag{2}$$

The true score $T(s)$ of a specific submission s is computed by Eq. (3), which is a weighted sum of the peer grading scores. The more dutiful a grader is, the more the score from him can be trusted, and thus the score from a grader with lager $D(g)$ weights more in the aggregated result. When calculating the weight, we feed $D(g)$ to a power function $f(x) = x^{\beta}$, where $\beta > 1$. The power function makes each grader's impact on the aggregated result more different. A larger value of β can more remarkably maximize the contributions of dutiful users while minimizing the contributions of undutiful users. It also should be noted that β cannot be too large, otherwise all the graders except the one with the highest value of dutifulness will contribute little to the aggregated score.

$$T(s) = \frac{\sum\limits_{g \in In(s)} D(g)^{\beta} C(g, s)}{\sum\limits_{g \in In(s)} D(g)^{\beta}}, \quad \text{for } s \in S \tag{3}$$

B. MODEL SOLVING

Eq. (1)~(3) are interdependent, so they cannot be solved directly. We use an iterative algorithm similar to the method in [24] to solve the DRT model. Let $D^t(g)$, $R^t(g, s)$, and $T^t(s)$ denote the dutifulness, reliability and true score at the t-th iteration respectively. For all the graders, peer grading scores, and submissions, initialize $D^0(g)$, $R^0(g, s)$, and $T^0(s)$ to 1 which is the highest value they can achieve. Update the three equations iteratively by Algorithm 1. When the algorithm stops, $T^t(s)$ is taken as the aggregated result, which is an estimation of the ground truth score of submission s. Although convergence is not theoretically proven, experiments show the algorithm converges to a stable value within a small error bound fairly quickly. We have conducted a large number of experiments to test the convergence, and find the number of iterations to attain convergence is affected by the total number of students, where more iterations are needed when the number of students is increasing, and when the number of students is 500, only about 40 iterations are required to achieve convergence.

IV. EXPERIMENTAL EVALUATION

We conduct experiments both on real-world and synthetic data to evaluate the performance of the DRT model.

A. EXPERIMENTS ON REAL-WORLD DATA

In this subsection, we use a real-world dataset to test whether the DRT model can improve peer grading performance in blended teaching courses.

The dataset is collected from a course named "Introduction to Computer Science" held in Fall 2020 at our university. It is a blended teaching course. Some teaching activities, including homework, quiz, discussion and

Algorithm 1 The Algorithm for Solving the DRT Model

```
Initialize $D^0(g) = 1, R^0(g, s) = 1, T^0(s) = 1$ for all graders, all peer grading scores, and all submissions
Set error bound $\varepsilon = 10^{-6}$, maximum number of iterations $K = 500$, iteration counter $t = 1$
do
    Update true score using Eq. (3):
    $T^t(s) = \frac{\sum_{g\in In(s)} D^{t-1}(g)^{\beta} C(g,s)}{\sum_{g\in In(s)} D^{t-1}(g)^{\beta}}$, for all submissions $s \in S$
    Update reliability using Eq. (2):
    $R^t(g, s) = \alpha D^{t-1}(g) + (1-\alpha)(1 - |C(g, s) - T^t(s)|)$, for all ratings $(g, s) \in E$
    Update dutifulness using Eq. (1):
    $D^t(g) = \frac{\sum_{s\in Out(g)} R^t(g,s)}{|Out(g)|}$, for all graders $g \in G$
    $t = t + 1$
    Calculate *error*:
        $error = Max(\sum_{s\in S} |T^t(s) - T^{t-1}(s)|, \sum_{(g,s)\in E} |R^t(g, s) - R^{t-1}(g, s)|, \sum_{g\in G} |D^t(g) - D^{t-1}(g)|)$
while ($error > \varepsilon$ and $t < K$)
return $T^t(s)$ for all $s \in S$
```

pre-class reading, are conducted on an online platform (https://www.teachermate.com.cn/), and the rest are conducted in a form of traditional classroom teaching. To finish the homework, the students should submit their answers to the platform. After the deadline of submission, the students are given 5 days to finish peer grading on the platform. After the deadline of peer grading, the system computes the average score for each submission, which is the final assessment of the submission. The grading task is mandatory. If a student does not finish the grading task, the system will subtract 10 points (10% of the full score) from his homework. To make the students perform more fairly, the grading process is set to be double blind. Only students who have committed the submission can participate in the peer grading task. Namely, for a specific homework assignment, if a student does not commit his submission before the deadline, he is not allowed to rate his peers.

The course is divided to 4 sections, and each section has a homework assignment. Each assignment consists of question-and-answer questions and computational questions. 55 students are enrolled in the course. For each assignment, each submission is randomly allocated to 5 other students, and each student grades 5 submissions. Each submission is also graded by the instructor. The full score of each assignment is normalized to 100. It should be noted that for the students who have committed the submissions but not finished the grading task, the system automatically subtracts 10 points from their final scores. This subtraction does not affect the following experiments. Because only the original scores rated by the students are collected for the experiments, while the final scores computed by the platform are discarded.

The statistics of the data set is shown in Table 1. To get a ground truth score for each submission, the instructor also grades each submission, so the number of submissions is equal to the number of instructor grading scores in Table 1. Data missing is a common situation in a real-world dataset. Some students may have not finished the homework, so the number of submissions may be less than the number of students (e.g., the number of submissions is less than 55 in Assignment 1 and 3). Some students may have finished the homework but not participate in the grading task, so the number of graders may be less than the number of students who have successfully submitted the homework (e.g., the number of graders is less than the number of submissions in Assignment 1, 2 and 3). Some graders may only rate a part of the assigned submissions, so the number of peer grading scores may be less than the number of graders multiplies the number of submissions assigned to each grader (e.g., the number of peer grading scores is only 188, less than 47 × 5, in Assignment 3). For Assignment 1, it was the first time to use the grading system and the students were not familiar with it, so only 40 students successfully submitted their homework and only 37 students finished peer grading. For Assignment 3, the deadline of the peer grading task happened to be in the middle of a 3-day holiday, so some students forgot to finish the task. Among the 52 students who had successfully submitted the homework, 5 students did not participate in peer grading, and 16 students did not rate all the assigned submissions (e.g., they rated less than 5 submissions). So we only got 188 peer grading scores for Assignment 3. Due to some students did not finish homework and/or did not finish peer grading, the problem of data missing is relatively significant for Assignment 1 and 3. But when observing from the view of the submissions, most of them still have received at least 4 ratings, so the experiment results will not be affected a lot.

TABLE 1. Dataset statistics.

	Assignment 1	Assignment 2	Assignment 3	Assignment 4
Graders	37	53	47	55
Submissions	40	55	52	55
Peer grading scores	185	265	188	274
Instructor grading scores	40	55	52	55

1) ACCURACY OF AGGREGATED SCORES

To test the performance on estimating the ground truth scores, we run experiments on the collected dataset. The scores graded by the instructor are taken as the ground truth. Four exiting methods are taken for comparison. The mean and the median of peer grading scores are most often used by peer grading systems, so the two methods are taken as baselines. The PeerRank [7] is also a method based on graph analysis, so we compare our method with it. Our method is developed

TABLE 2. RMSE of the aggregated scores.

	Mean	Median	PeerRank	REV2	DRT
Assignment 1	7.80	8.49	7.86	8.23	**7.62**
Assignment 2	6.57	5.42	6.61	8.32	**5.09**
Assignment 3	7.08	7.08	7.21	6.86	**6.70**
Assignment 4	11.59	9.02	12.75	11.39	**8.03**

from REV2 [24], so we also take it for comparison. In the peer grading scenario, due to the number of graders for each submission is small, and no behavior information is recorded, we use the basic form of REV2 which drops the prior belief term and behavior feature term. The parameters of our DRT model are set to be $\alpha = 0.5$ and $\beta = 3$. As the aggregated scores directly obtained from the 5 methods are in different scales, all the results reported below are normalized to have a full score of 100.

The root mean square errors (RMSE) are listed in Table 2. It can be seen our DRT model achieves the lowest RMSE on all the 4 assignments. Compared to the baseline of the mean value method, our method decreases the RMSE by 2.31% in the worst case (e.g., Assignment 1) and by 30.72% in the best case (e.g., Assignment 4). The results verify the effectiveness of our method. It is surprising that the two elaborately designed methods, PeerRank and REV2, perform even worse than the baselines on some assignments. PeerRank is based on the assumption that the reliability of a grader is determined by the performance on his own homework. We will show in the next experiment that this assumption does not hold in this blended teaching course. So PeerRank fails. REV2 was originally designed for online marketplaces. Actually, a product in an online marketplace does not have a definite score of quality, and the score is just the consensus of a group of people. On the other hand, REV2 mainly focuses on evaluating the trustiness of users, but not the quality of products.

To further compare the performance of the methods, we investigate the distribution of the absolute error between the aggregated score and the ground truth. Table 3 and Table 4 respectively report the percentage of aggregated scores having error within 5 points (5% of the full score) and 10 points (10% of the full score). For a given method, the higher percentage means a better performance, as it means more estimated scores are located near the ground truth. It can be seen that our method achieves the highest percentage among the 5 methods most of the time. This result demonstrates the aggregated scores obtained by our method are more concentrated in the vicinity of the ground truth. This experiment again verifies the effectiveness of our method.

2) FACTORS AFFECTING PEER GRADING ERROR

Most of the current methods are based on the assumption that a grader's grading ability is related to his performance on finishing his own homework, which means the higher a grader's submission is graded, the more accurate he grades his peers' submissions. We run experiments to investigate whether this assumption holds. We calculate the mean absolute error (MAE) by Eq. (4) to evaluate the average peer grading error of a grader.

TABLE 3. Percentage of aggregated scores having error within 5 points.

	Mean	Median	PeerRank	REV2	DRT
Assignment 1	45.00	47.50	47.50	50.00	**52.50**
Assignment 2	69.09	**80.00**	65.45	61.82	**80.00**
Assignment 3	**55.77**	50.00	53.85	51.92	53.85
Assignment 4	32.73	45.45	32.73	29.09	**47.27**

TABLE 4. Percentage of aggregated scores having error within 10 points.

	Mean	Median	PeerRank	REV2	DRT
Assignment 1	75.00	72.50	75.00	72.50	**82.50**
Assignment 2	83.64	**94.55**	83.64	76.36	92.73
Assignment 3	86.54	82.69	82.69	86.54	**88.46**
Assignment 4	61.82	74.55	58.18	61.82	**78.18**

$$MAE(g) = \frac{1}{|Out(g)|} \sum_{s \in Out(g)} |C(g,s) - GT(s)|, \quad \text{for } g \in G \tag{4}$$

where $GT(s)$ denotes the ground truth score of submission s given by the instructor. The MAE can be used to measure the grader's grading ability, as a smaller value of MAE means doing better in peer grading. The Pearson correlation coefficient between the MAE and the ground truth score is computed. Table 5 shows the correlation coefficients and the p-values. It can be seen that the MAE is irrelevant to the ground truth score, which means the grading ability of a grader is irrelevant to his performance on finishing his own homework. The assumption of the current methods does not hold. This result is in accordance with the findings in [19] that due to affinity among students, they trend to assign random scores to other submissions without seriously evaluating their peers' homework. This result explains why the existing methods are sometimes ineffective in the real world.

To verify the dutifulness of a grader devised in our DRT model (defined by Eq. (1)) is a good metric to measure a grader's grading ability, the Pearson correlation coefficient between the MAE and the value of dutifulness is computed. The results are reported in Table 5. It shows that the MAE has a significant negative correlation with the dutifulness, which supports our assumption that the more dutiful a grader is, the more accurately he rates his peers' submissions. The results also verify that weighting the observed scores by the graders' dutifulness (defined by Eq. (3)) is a rational way to estimate the ground truth score.

B. EXPERIMENTS ON SYNTHETIC DATA

We use synthetic data, which simulate a course with a large number of enrolled students, to measure the scalability of the

TABLE 5. Pearson correlation coefficients and p-values between the MAE and the affecting factor.

Affecting factor	Assignment 1	Assignment 2	Assignment 3	Assignment 4
Ground truth score	0.13 (p=0.44)	-0.04 (p=0.76)	-0.10 (p=0.52)	-0.07 (p=0.62)
Dutifulness	-0.75 (p<0.01)	-0.97 (p<0.01)	-0.79 (p<0.01)	-0.96 (p<0.01)

DRT model and investigate the undutiful graders' impact on peer grading accuracy.

In the following experiments, the scores (both ground truth and peer grading scores) of a submission are integers in the range of [0, 100]. The ground truth score of a submission is simulated by a normal distribution with a fixed standard deviation of 4 and a mean of μ, so the ground truth score of a submission s is drawn from:

$$GT(s) \sim N(\mu, 4^2) \tag{5}$$

The obtained score is rounded to an integer. If the score is outside [0, 100], it is clipped to the bounds of the range.

To simulate the existence of undutiful graders, we randomly divide the students to two types: the dutiful ones and undutiful ones. A dutiful grader rates a submission based on the ground truth score of his own submission, and an undutiful grader rates a submission randomly. For a dutiful grader, we take the rating model in [7] and [22] to generate the scores given by him. According this rating model, if the ground truth score of a submission is $GT(s)$, and the ground truth score of a grader's own submission is $GT(g)$, then the peer grading score $C(g, s)$ is the sum of two binomial distributions:

$$\begin{aligned} C_1(g, s) &\sim B(GT(s), GT(g)/100) \\ C_2(g, s) &\sim B(100 - GT(s), 1 - GT(g)/100) \\ C(g, s) &= C_1(g, s) + C_2(g, s) \end{aligned} \tag{6}$$

where $B(n, p)$ is a binomial distribution of n trails with probability p. The undutiful graders are further randomly divided to 2 halves. An undutiful grader in the first half simply gives a high score to a submission, so the peer grading score is drawn from a discrete uniform distribution:

$$C(g, s) \sim U(90, 100) \tag{7}$$

where $U(a, b)$ is a discrete uniform distribution with parameter a and b. An undutiful grader in the second half simply gives a low score to a submission, so we draw the peer grading score from the following discrete uniform distribution:

$$C(g, s) \sim U(10, 20) \tag{8}$$

In the following experiments, submissions are randomly assigned to graders and the scores are generated using Eq. (5)~Eq. (8). Each experiment is repeated 50 times, and the average result is reported.

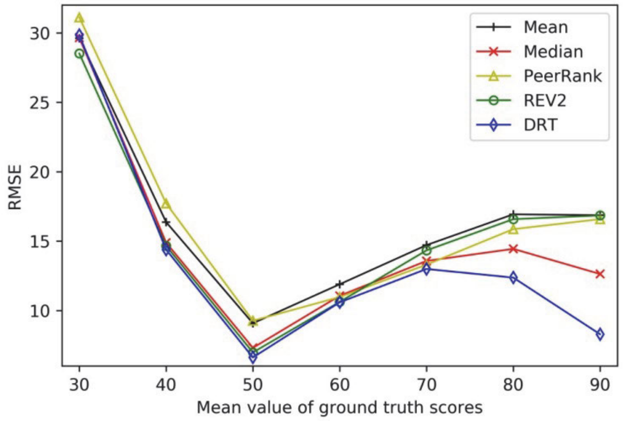

FIGURE 2. RMSE of 5 methods on a normal distribution of ground truth scores, with the mean value of the normal distribution varying from 30 to 90.

1) VARIABLE MEAN VALUE OF GROUND TRUTH SCORES

This experiment investigates the grading accuracy when the mean value of the ground truth scores (μin Eq. (5)) varies. The number of students is set to be 100, 20% of the students are undutiful graders, and each student grades 5 submissions. Figure 2 shows the RMSE when μvaries between 30 and 90. We can see that the advantage of the DRT model is significant when μ is large. Specifically, when μ is larger than 70, the DRT obtains a significantly lower RMSE than the other methods, but when μ is less than 70, the DRT, Median, and REV2 methods obtain almost the same RMSE. When μ is small, the dutiful graders have poor grading ability. Although they try to rate the submissions as possible as they can, the scores given by them are still unreliable due to the limitation of grading ability. Since a majority of the peer grading scores are unreliable, there is not enough information for the DRT to detect the undutiful graders, and the DRT shows no advantage. Indeed, when the unreliable scores are too many, no method can obtain a good result due to the lack of information, and we can see no method can perform better than the DRT.

In Figure 2, it can be seen that all the methods achieve the lowest RMSE when $\mu = 50$. This does not imply the best performance can be obtained when $\mu = 50$, it is just a coincidence caused by the data generating models. We can see that when $\mu = 50$, the mean value of the ground truth scores is 50 (from Eq. (5)), the mean value of the scores given by dutiful graders is 50 (from Eq. (6)), and the mean value of the scores given by undutiful graders is near 50 (from Eq. (7) and Eq. (8)). So the distance between the ground truth scores and the peer grading scores are small, and we observe a valley for each curve at $\mu = 50$ in Figure 2.

Since the DRT model needs the peer grading scores informative enough to distinguish dutiful graders and undutiful graders, μ is set to be 80 in the rest experiments.

2) VARIABLE NUMBER OF UNDUTIFUL GRADERS

This experiment explores the impact of undutiful graders to the accuracy of aggregated scores by changing the quantity

]

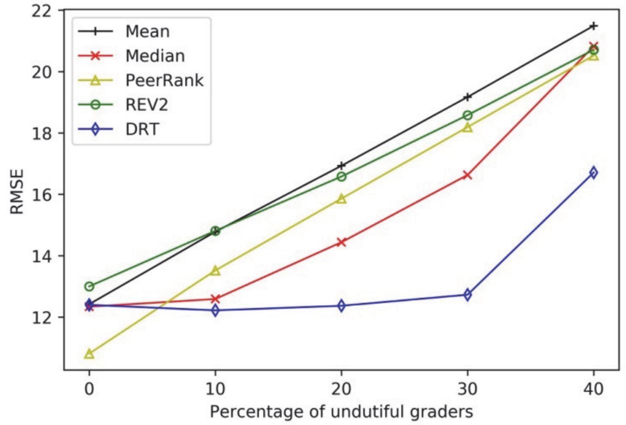

FIGURE 3. **RMSE of 5 methods with the percentage of undutiful graders varying from 0 to 40%.**

of undutiful graders. The experiment setting is as follow: the percentage of undutiful graders among the students varies from 0 to 40%, the total number of students is 100, each student grades 5 submissions, and the mean value of the ground truth scores is 80. The result is shown in Figure 3. We can see that the DRT model achieves the lowest RMSE among the 5 methods most of the time. Actually, this experiment demonstrates the superiority of the DRT model from two aspects: one is the low RMSE, which represents the high accuracy of the aggregated scores, and the other is the low growth rate of the RMSE, which represents the robustness to undutiful graders' disturbances. It can be seen that when the percentage of the undutiful graders is less than 30%, the RMSE of the DRT model increases very slow. The slow increase verifies the DRT model is effective to detect the undutiful graders and remove their disturbances from the aggregated scores. Although the RMSE of the DRT mode starts to increase fast when the percentage of the undutiful graders exceeds 30%, the DRT model still obtains the lowest RMSE due to its advantages accumulated in the early stage.

The PeerRank performs best among all the methods when there is no undutiful grader, but its performance decreases immediately when a small number of undutiful ones are added to the graders. This result is in line with the expectations. When the dataset contains no undutiful graders, the assumption of the PeerRank is well satisfied, and it obtains a good performance. But when undutiful graders are added to the dataset, the assumption is broken, and the performance decreases immediately. The performances of all the methods except the DRT deteriorate rapidly with the percentage of undutiful graders increasing. This reminds us that the undutiful graders should be treated seriously when designing a peer grading method, otherwise the method may lose effect even when a small number of undutiful ones are contained in the graders.

3) VARIABLE NUMBER OF GRADERS PER SUBMISSION

This experiment inquires the impact of the number of graders assigned to each submission by varying the number from 3 to 10. In the experiment, the total number of students is 100, 20% of students are undutiful graders, and the mean value of the ground truth scores is 80. As is shown in Figure 4, the DRT always gets the lowest RMSE among the methods. In addition, we can see that the DRT reaches a stable RMSE at 6 graders per submission, while the RMSE of the other methods still decreases after 6 graders per submission. This result implies that our DRT model can get a good result with a small number of graders per submission. This is a good property in a real-world peer grading system. Because only grading a few submissions will not significantly increase the graders' burden, and when the burden of the task is light, the graders tend to review the submissions more carefully and seriously.

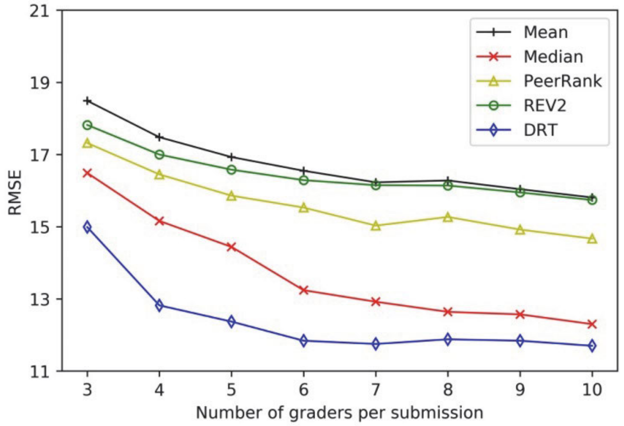

FIGURE 4. **RMSE of 5 methods with the number of graders for each submission varying from 3 to 10.**

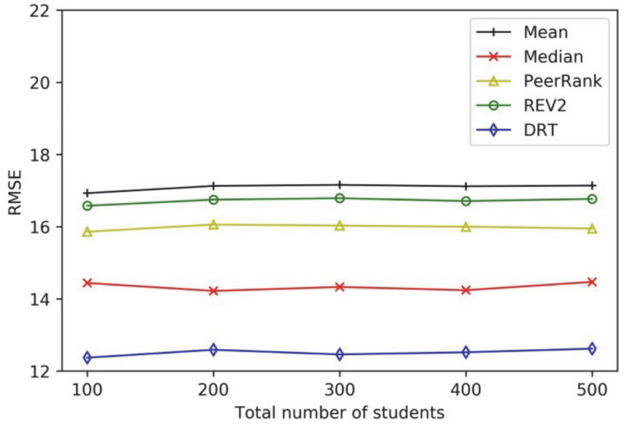

FIGURE 5. **RMSE of 5 methods with the total number of students increasing from 100 to 500.**

4) VARIABLE TOTAL NUMBER OF STUDENTS

This experiment tests whether the DRT model can be applied to a course with even larger scale. In the experiment, the total number of students varies from 100 to 500, 20% of the students are undutiful graders, each student grades 5 submissions, and the mean value of the ground truth scores is 80. The result is shown in Figure 5. The RMSEs of all the

5 methods show no obvious change when the number of students increases. The DRT model performs best among all the methods. This implies that our method is applicable to a large-scale course.

V. CONCLUSION AND FUTURE WORK

This paper proposes the DRT model to improve the accuracy of peer grading for blended teaching courses. Starting from the analysis that the rating errors are mainly due to the graders' undutiful behaviors when grading the submissions, we define three types of variables, which are the dutifulness of a grader, the reliability of a peer grading score, and true score of a submission, to quantitatively describe the intrinsic relationship among the grader, submission and rated score. By solving these variables in an iterative manner, an aggregated score for each submission is obtained. The effectiveness of the DRT model is verified by extensive experiments.

Experiments are conducted on both real-world and synthetic datasets. From the experiments we can get the following conclusions. (1) In blended teaching courses, there are indeed a significant number of irresponsible students whose grading performance is irrelevant to the performance on finishing the homework. With the proportion of irresponsible students increasing, the accuracy of many current peer grading methods decreases severely, since their assumption that the students are responsible is broken. (2) The DRT model can achieve high peer grading accuracy in blended teaching courses. This model is based on the assumption that there are many irresponsible ones mixed in the students, and tries to alleviate the adverse effect of the irresponsible students. Since its assumption well fits the reality and the dutifulness designed in the model is a good metric to evaluate the trustiness of the students, the model obtains good performance. (3) The DRT model is robust to the irresponsible students. Experiments show that even when the proportion of irresponsible students reaches 30%, the accuracy keeps almost unchanged. (4) The DRT model shows some other good properties. A high accuracy can be obtained with each student grading only a small number of submissions. This can lighten the students' work burden while keeping high grading performance. The DRT model is not limited by the scale the class. It can be applied to both small classes with dozens of students and large classes with hundreds of students.

Though the DRT model has achieved good performance, future work can be done from many aspects to further improve it. The current work has illustrated the dutifulness in the DRT model is a good evaluation of the students' grading performance. So the value of dutifulness can be used to incent the students to perform more carefully and reliably in the grading task. How to incorporate an incentive mechanism into the current model can be studied in the future. The DRT model only considers the students' irresponsible behavior, while the methods for MOOCs only consider the students' mastery of knowledge. Theoretically, both factors impact the performance of peer grading. Future work can study how to model the two factors simultaneously. A majority of peer grading methods, including the DRT model, assign the submissions to the graders randomly. To improve the performance of peer grading by allocating the submissions according to more rational rules is another direction of the future work. An important reason of the students' irresponsible behaviors in peer grading is that the grading task is tedious. Many researches have devoted to improve learning by gamification. The learning process becomes interesting via gamification. The future work can study reshaping the peer grading task in a game-based form to make it more attractive.

REFERENCES

[1] C. Kulkarni, K. P. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer, "Peer and self assessment in massive online classes," *ACM Trans. Comput.-Hum. Interact.*, vol. 20, no. 6, pp. 1–31, Dec. 2013.

[2] L. Zhang and K. VanLehn, "Evaluation of auto-generated distractors in multiple choice questions from a semantic network," *Interact. Learn. Environ.*, vol. 29, no. 6, pp. 1019–1036, Aug. 2021.

[3] P. Sadler and E. Good, "The impact of self- and peer-grading on Student learning," *Educ. Assessment*, vol. 11, no. 1, pp. 1–31, Feb. 2006.

[4] P. A. Carlson and F. C. Berry, "Calibrated peer review TM and assessing learning outcomes," in *Proc.33rd Annu. Frontiers Educ. Conf.*, 2003, pp. F3E1–F3E6.

[5] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, "Tuned models of peer assessment in MOOCs," in *Proc.6th Int. Conf. Educ. Data Mining*, 2013, pp. 153–160.

[6] F. Mi and D. Y. Yeung, "Probabilistic graphical models for boosting cardinal and ordinal peer grading in MOOCs," in *Proc. 19th AAAI Conf. Artif. Intell.*, 2015, pp. 454–460.

[7] T. Walsh, "The PeerRank method for peer assessment," in *Frontiers in Artificial Intelligence*, vol. 263. Amsterdam, The Netherlands: IOS Press, 2014, pp. 909–914.

[8] N. B. Shah, J. K. Bradley, A. Parekh, M. Wainwright, and K. Ramchandran, "A case for ordinal peer-evaluation in MOOCs," in *Proc. NIPS, Work. Data Driven Educ.*, 2013, pp. 1–8.

[9] K. Raman and T. Joachims, "Methods for ordinal peer grading," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 1037–1046.

[10] L. de Alfaro and M. Shavlovsky, "CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments," in *Proc. 45th ACM Tech. Symp. Comput. Sci. Educ.*, Mar. 2014, pp. 415–420.

[11] 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, Apr. 2017.

[12] 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. 17679–17692, Dec. 2020.

[13] R. Alcarria, B. Bordel, D. M. de Andrés, and T. Robles, "Enhanced peer assessment in MOOC evaluation through assignment and review analysis," *Int. J. Emerg. Technol. Learn.*, vol. 13, no. 1, pp. 206–219, 2018.

[14] Y. Wang, H. Fang, Q. Jin, and J. Ma, "SSPA: An effective semi-supervised peer assessment method for large scale MOOCs," *Interact. Learn. Environ.*, early access, pp. 1–19, Jul. 2019.

[15] W. Wang, B. An, and Y. Jiang, "Optimal spot-checking for improving the evaluation quality of crowdsourcing: Application to peer grading systems," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 4, pp. 940–955, Aug. 2020.

[16] T. Sunahase, Y. Baba, and H. Kashima, "Probabilistic modeling of peer correction and peer assessment," in *Proc. 12th Int. Conf. Educ. Data Mining*, 2019, pp. 426–431.

[17] I. Labutov and C. Studer, "JAG: A crowdsourcing framework for joint assessment and peer grading," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1010–1016.

[18] L. Wolins, B. D. Wright, and G. Rasch, "Probabilistic models for some intelligence and attainment tests," *J. Amer. Stat. Assoc.*, vol. 77, no. 377, p. 220, Mar. 1982.

[19] Y. Han, W. Wu, S. Ji, L. Zhang, and H. Zhang, "A human-machine hybrid peer grading framework for SPOCs," in *Proc. 12th Int. Conf. Educ. Data Mining*, 2019, pp. 300–305.

[20] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," *World Wide Web Internet Web Inf. Syst.*, vol. 54, nos. 66–1999, pp. 1–17, 1998.

[21] N. Capuano and S. Caballe, "Towards adaptive peer assessment for MOOCs," in *Proc. 10th Int. Conf. P2P, Parallel, Grid, Cloud Internet Comput. (3PGCIC)*, Nov. 2015, pp. 64–69.

[22] N. Capuano, S. Caballé, and J. Miguel, "Improving peer grading reliability with graph mining techniques," *Int. J. Emerg. Technol. Learn.*, vol. 11, no. 7, pp. 24–33, 2016.

[23] P. Gutierrez, N. Osman, and C. Sierra, "Collaborative assessment," in *Frontiers in Artificial Intelligence*, vol. 269. Amsterdam, The Netherlands: IOS Press, 2014, pp. 136–145.

[24] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. S. Subrahmanian, "REV2: Fraudulent user prediction in rating platforms," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, Feb. 2018, pp. 333–341.

[25] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Identify online store review spammers via social review graph," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, pp. 1–21, Sep. 2012.

[26] B. Hooi, N. Shah, A. Beutel, S. Günnemann, L. Akoglu, M. Kumar, D. Makhija, and C. Faloutsos, "BIRDNEST: Bayesian inference for ratings-fraud detection," in *Proc. SIAM Int. Conf. Data Mining*, Jun. 2016, pp. 495–503.

[27] S. Brooks, "Markov chain Monte Carlo method and its application," *J. Royal Stat. Soc., D Statistician*, vol. 47, no. 1, pp. 69–100, Apr. 1998.

XING DU received the B.S. degree in information engineering, in 2006, and the Ph.D. degree in instrument science and technology from Chongqing University, China, in 2012. He is currently an Associate Professor with the College of Computer and Information Science, Chongqing Normal University. His research interests include data mining, computer vision, and technology-enhanced learning.

XINGYA WANG is currently pursuing the master's degree with Chongqing Normal University, major in educational technology. Her research interest includes educational data mining.

YAN MA received the B.S. degree in physics and electronics from Yunnan University, China, in 1982, and the M.S. and Ph.D. degrees in information technology from Huazhong Normal University, China, in 1993 and 2008, respectively. He is currently a Professor with the College of Computer and Information Science, Chongqing Normal University, China. He is an Advanced Member of the China Computer Society. His research interests include artificial intelligence and wisdom education.