# Leveraging Social Connections to Improve Peer Assessment in MOOCs

Hou Pong Chan[1,2] and Irwin King[1,2]
[1]Shenzhen Key Laboratory of Rich Media Big Data Analytics and Application Shenzhen Research Institute,
The Chinese University of Hong Kong, Shenzhen, China
[2]Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Hong Kong, Shatin, N.T., Hong Kong
{hpchan, king}@cse.cuhk.edu.hk

## ABSTRACT

With the advent of Massive Open Online Courses (MOOCs), students from all over the world can access to quality courses via a web browser. Due to their great convenience, a popular MOOC can easily attract tens of thousands of students to enroll. Hence, a challenging problem in MOOCs is to find an efficient way to grade a large scale of assignments. To address this problem, peer assessment was proposed to grade the assignments in a scalable way. In peer assessment, each student is asked to access a subset of his/her peers' assignments via a web interface, then all these peer grades are aggregated to predict a final grade for each submitted assignment. These peer grades are very noisy due to the fact that different students have different bias and reliability. Several probabilistic models were proposed to improve the accuracy of the predicted grades by explicitly modeling the bias and reliability of each student. However, existing methods assumed that all students are independent of each other while ignoring the social interactions among the students. In real life, students' grading bias are easily affected by their friends. For example, a student tends to have a tough grading standard if his/her friends are harsh graders. Following this intuition, we propose three probabilistic models for peer assessment by incorporating social connections to model the dependencies of bias among the students. Moreover, we evaluate our models in a new peer grading dataset, which is enhanced with the social information of users in the discussion forums of the MOOC platform. Experimental results show that our models improve the accuracy of the predicted grades by leveraging social connections of students.

**Keywords:** Massive Open Online Courses (MOOCs); Peer Assessment; Social Network

## 1. INTRODUCTION

Massive Open Online Courses (MOOCs) are courses that allow open and unlimited access, they offer a handy way for people to access university level courses via a web browser. Recently, studying in MOOCs became a popular way of learning due to its great convenience. The enrollment of a popular MOOC can reach up to tens of thousands. Hence, one of the most challenging problems in MOOCs is that it is infeasible for the teaching staffs to grade all the assignments in such a large scale. To avoid this problem, most MOOCs only offer assignments that can be graded automatically, such as multiple choice questions. However, open-ended assignments such as essays are indispensable for many courses [14], and there are no effective auto grading methods for such open-ended assignments.

Peer assessment was proposed to tackle the large-scale grading problem in MOOCs. MOOC platforms such as Coursera [1] and edX [2] allow instructors to use peer assessment to grade open-ended assignments, in which students grade a subset of their peers' assignments via a web interface. The grades given by the students (peer grades) are then aggregated by the system to compute a final score for each assignment. Since the number of graders naturally scales with the number of assignments, peer assessment provides a scalable way to grade assignments in MOOCs.

Most existing MOOC platforms simply use the median of the peer grades received by an assignment as its final score. However, each student has a different bias when grading open-ended assignments. The bias of a grader is the constant inflation or deflation of the peer grades given by that grader. For example, the true score of student $u$'s assignment is 8. Student $v_1$ thinks that student $u$'s assignment is only worth a score of 5 (a bias of -3), while student $v_2$ thinks that the same assignment is worth a score of 9 (a bias of +1). Moreover, different students have different reliability. Reliable students honestly give grades to their peers while unreliable students randomly assign grades to their peers. It will be difficult to obtain a fair or accurate score for each assignment if we do not consider the bias and the reliability of each student. Different probabilistic models were proposed to address the above issues by introducing random variables to model the bias and reliability of each student [15, 12].

Although these methods improved the accuracy of peer assessment, they assumed that students are completely independent of each other. Such assumption ignores the existence of social connections and social inferences among students. Studies showed that web users' preferences are likely to be affected by their friends' preferences [22, 27]. Fol-

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054165

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

lowing this intuition, many existing recommender system algorithms leveraged social connections to model the preferences of web users [10, 29, 26]. Similarly, we believe that the bias of graders on open-ended assignments will be influenced by the bias of their friends. For example, a student tends to have a tougher grading standard (a more negative bias) if his/her friends are harsh graders. Thus, the bias that imposed on a peer grade by a grader is a compound of the grader's own bias and the social influences received by the grader. Ignoring the social inferences among the students leads to a suboptimal model of grader bias, and in consequence, leads to a less accurate estimation of the true grades of assignments.

In order to address the above limitation of existing peer assessment methods, based on the intuition that a grader's bias will be affected by their friends, we leverage social connections to build more accurate models for peer assessment. To achieve this goal, we extend existing probabilistic models [15, 12] for peer assessment by modeling the dependencies among the students. In our solution, we estimate the true score of each assignment from peer grades by modeling the reliability as well as two types of bias for each grader: original bias and influenced bias. The original bias of a grader is the grader's own bias before affected by the grader's friends. The influenced bias of a grader is the grader's bias after affected by his/her friends, which is assumed to be a combination of a grader's own bias and the bias of his/her neighbors in the social graph. With this assumption, the knowledge gathered about the bias of a single grader can be used to infer the bias of his/her friends, i.e., learn the bias of graders in a collaborative manner. Thus, this method can better model the bias of each student imposed in the peer grades, and in consequence, leads to a more accurate estimation of the true grades of assignments from the peer grades.

To evaluate our proposed model, we conduct experiments on a new peer grading dataset from xuetangX[3], which is one of the biggest MOOC platforms in China. We enhance this peer grading dataset with social information of students in the discussion forums of the platform. Experimental results show that our models improve the accuracy of the predicted grades by leveraging social connections among students.

We summarize our contributions as follows:

1. We propose novel extensions to existing probabilistic models for peer assessment by incorporating social connections to model the dependencies of bias among the students.

2. We devise an inference algorithm based on the Gibbs sampling technique to infer the latent variables in our models, including the true score of each student, the reliability, the original bias, and the influenced bias of each grader.

3. We evaluate our proposed models in a real peer grading dataset, which is enhanced with the social information of students. Empirical results show that our models outperform existing models in terms of the accuracy of the predicted grades.

The rest of the paper is organized as follows. In Section 2, we review existing methods for peer assessments. Section 3 describes the notations and formally defines the problem that we solve. Section 4 describes our proposed probabilistic models for peer assessment. The description of the dataset we use and the experimental results are presented in Section 5, followed by the conclusions in Section 6.

## 2. RELATED WORK

There are several empirical studies on real peer grading datasets to discover the factors that affect peer grading performances [4, 8] and compare the performances of different peer grading algorithms [20]. Existing peer grading methods can be divided into two categories: *cardinal peer assessment* and *ordinal peer assessment*. In cardinal peer assessment, each student gives grades to their peers' assignments in absolute scale, e.g., 30. The goal is to estimate the ground truth score of each assignment from the peer grades given by the students. In ordinal peer assessment, each student is asked to rank a subset of assignments, e.g., $a_1 \succ a_3 \succ a_5$, and the goal is to aggregate all the partial rankings from the students to obtain a full ranking of all the assignments [25], e.g. $a_1 \succ ... \succ a_n$. Since the goal of our work is to improve the accuracy of the predicted grades of assignments, our work belongs to the category of cardinal peer assessment.

### 2.1 Cardinal Peer Assessment

Several iterative algorithms were proposed to learn the score of each assignment from the peer grades. De Alfaro and Shavlovsky proposed the Vancouver algorithm which iteratively updates the score of each assignment and the grading accuracy of each grader [3]. In each iteration, the algorithm weights the peer grades by the accuracy of the graders and used these weighted inputs to estimate a score for each submission; the accuracy of each grader is then updated by the estimated score of each grader. Walsh proposed another iterative algorithm called the PeerRank algorithm [23] which is inspired by the well-known PageRank algorithms [13]. On the basis of the assumption that the score of a grader reflected his/her grading ability. The PeerRank algorithm learned the scores of assignments iteratively by weighting the feedback of a grader by his/her scores. A reputation based algorithm was proposed to weight the peer grades of students by the trust they received [6].

Formulating generative models is another popular approach for peer assessment. Piech et al. proposed probabilistic models to estimate the score of each assignment by modeling the relationship between the observed peer grades of each assignment, the true score of each assignment, as well as the bias and reliability of each grader [15]. The $PG_3$ Model in their work uses a deterministic affine function to model the relationship between the reliability and the true grade of a grader, with the assumption that a grader with higher score tends to be more reliable. Mi and Yeung argued that this deterministic linear relationship might be too rigid to model the reliability and the true grade for all the graders [12]. Hence, they extended this model by relaxing this deterministic linear relationship by using a probabilistic relationship. However, all the existing cardinal peer assessment methods did not model the dependencies of bias among the students.

### 2.2 Ordinal Peer Assessment

For ordinal peer assessment models, Shah et al. generalized the Bradley-Terry model [1] to learn the latent scores of students based on the partial rankings provided by the students [21]. Raman et al. used several classical probabilis-

[3] http://www.xuetangx.com/

tic ranking models such as the Plackett-Luce model [9], the Bradley & Terry model [1], and the Mallows model [11] to learn the full ranking of all the assignments based on the partial rankings of assignments collected from students [16]. To obtain a more accurate ranking of assignments, a variability parameter was introduced into the probabilistic ranking models to estimate the grader reliability. Moreover, Bayesian techniques were employed to estimate the uncertainty of the predicted ranking [17] and allocate the ranking tasks among the students [24]. To reduce the sample complexity for the partial rankings, Chan et al. proposed a bandit style algorithm which allocates the ranking tasks among the students and learns the full ranking of all assignments in an online manner [2]. To further improve the accuracy of predicted ranking, Mi and Yeung proposed a mechanism to combine both cardinal peer assessment and ordinal peer assessment [12]. However, all the existing ordinal peer assessment methods assumed that students are independent of each other.

# 3. PROBLEM DEFINITION

In this section, we first describe the notations and concepts that will be used throughout this paper, then we formally define the problem of peer assessment that we are going to solve.

We use $u$ to denote an arbitrary student and $U$ to denote the set of all students. Then, we use $v$ to denote an arbitrary grader and $V$ to denote the set of all graders. Since the students in peer assessment also act as graders, $U$ and $V$ are actually correspond to the same set of students. We use $n$ to denote the number of graders or students involved in the peer assessment, i.e., $n = |U| = |V|$. After that, we use $v \to u$ to indicate that grader $v$ grades student $u$'s assignment. All the notations that we used are summarized in Table 1. The followings are the definitions of the concepts that will be used in this paper.

**True scores**: We assume that each assignment is associated with a true score. We use $s_u$ to denote the true score of student $u$'s submission.

**Peer grades**: Peer grades are the scores given by the students to their peers' assignments. We use $z_u^v$ to denote the score given by grader $v$ to student $u$'s submission. Then, we use $Z$ to denote the set of all peer grades that we received.

**Influence matrix**: We use $W \in \mathbb{R}^{n \times n}$ to denote the influence matrix. The $i,j$-th entry of the influence matrix, $w_{i,j}$, is defined as the influence that grader $v_j$ has on the influenced bias of grader $v_i$. $w_{i,j} = 0$ if and only if grader $v_j$ does not have any influence on grader $v_i$'s influenced bias. All entries in $W$ are nonnegative and $\sum_{j=1}^{n} w_{i,j} = 1$ for $i = 1, ..., n$, i.e., $W$ is row-wise normalized. This influence matrix can be constructed from the social graph in the social components of a MOOCs platform, such discussion forums and private messages. In addition, some MOOC platforms allow users to link their accounts with social network services such as Facebook, such social connection information can also serve as the building blocks for the construction of $W$. The details of the construction of $W$ for this paper will be discussed in Section 5.

**Influenced bias**: The influenced bias of a grader is defined as the constant inflation or deflation of peer grades given by that grader. We use $b_v$ to denote the influenced bias of grader $v$. For examples, suppose $s_u = 8$, and $b_v = 2$. Then, the mean of $z_u^v = s_u + b_v = 8 + 2 = 10$.

Table 1: Notations

| Notation | Description |
|---|---|
| $V$ | Set of all students |
| $U$ | Set of all graders |
| $s_u$ | The true score for the submission of student $u$ |
| $z_u^v$ | The score given by grader $v$ to student $u$'s submission |
| $\theta_v$ | The original bias of grader $v$ before influenced by her friends |
| $b_v$ | The influenced bias of grader $v$ after influenced by his/her friends |
| $\tau_v$ | The reliability of grader $v$ |
| $W$ | The influence matrix |
| $A$ | The adjacency matrix of the social graph of the graders |
| $N_G(v)$ | The set of graders that is connected to grader $v$ in the social graph of the graders (students) |

**Original bias**: We use $\theta_v$ to denote the original bias of grader $v$. We assume that the mean of $b_v$ is a linear combination of grader $v$'s original bias and the original bias of his/her friends, weighted by the influence matrix $W$. More formally, the mean of $b_v = \sum_{k:k \in N_G(v)} w_{v,k}\theta_k$, where $N_G(v)$ is the set of neighbors of grader $v$ in the social graph. We assume that only the neighbors of grader $v$ in the social graph have influence on the bias of grader $v$. If a grader $v$ did not receive any social influence, i.e., $w_{v,k} = 0$ for all $k \in V$, then, $b_v = \theta_v$.

**Reliability**: The reliability of a grader is defined as the precision (the reciprocal of the variance) of the peer grades given by that grader. We use $\tau_v$ to denote the reliability of grader $v$, i.e., $\tau_v$ is the precision of $z_u^v$. Reliability measured how close is on average is the peer grades given by that grader from the underlying true score of the assignment after correcting the bias.

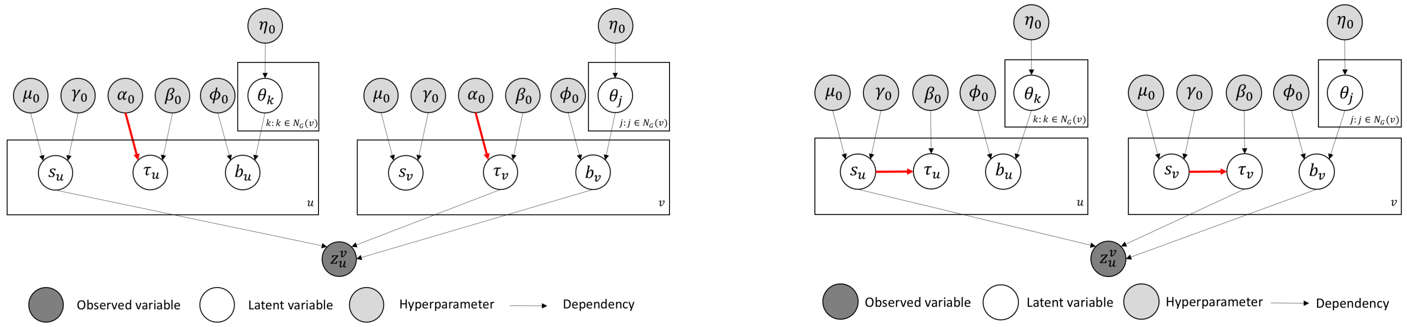
Unlike existing works that assume all graders are independent of each other, we introduce new probabilistic models by exploiting graders' social information. Our goal is to estimate the true score of each assignment by modeling the relationship between the observed peer grades, the reliability, the original bias, and the influenced bias of graders, and the true score of assignments. More formally, we define the problem as follows: Given the peer grades provided by the graders, $Z$, and the influence matrix, $W$. Our goal is to learn $\tau_v$, $b_v$, $\theta_v$, for all $v \in V$, and $s_u$ for all $u \in U$. Table 1 shows all the notations used in our proposed probabilistic models.

# 4. PROBABILISTIC MODELS

In this work, we propose three probabilistic models, $PG_6$, $PG_7$, and $PG_8$ for peer assessment. We describe the details of these models in this section.

## 4.1 The $PG_6$ Model

Our $PG_6$ Model is an extension to the $PG_1$ Model in [15] by modeling the dependencies of bias among the graders. The conditional dependence structure between the random variables in $PG_6$ are expressed by the graphical model [7]



(a) Graphical Model for $PG_6$

(b) Graphical Model for $PG_7$ and $PG_8$

Figure 1: Graphical Models

in Figure 1(a). As shown in the figure, the peer grade $z_u^v$ is the only observed random variable in the model. $s_u$, $\tau_v$, $\theta_v$, and $b_v$ are the hidden variables to be estimated. The prior distribution of these hidden variables are specified by the hyperparamters, $\alpha_0$, $\beta_0$, $\eta_0$, $\phi_0$, $\mu_0$, and $\gamma_0$. The distributions of all the random variables for the $PG_6$ Model are shown as follows.

$$\begin{aligned}
&\tau_v \sim \Gamma(\alpha_0, \beta_0),\\
&\theta_v \sim \mathcal{N}(0, \tfrac{1}{\eta_0}),\\
&b_v \sim \mathcal{N}(\textstyle\sum_{k:k \in N_G(v)} w_{v,k}\theta_k, \tfrac{1}{\phi_0}),\\
&s_u \sim \mathcal{N}(\mu_0, \tfrac{1}{\gamma_0}),\\
&z_u^v \sim \mathcal{N}(s_u + b_v, \tfrac{1}{\tau_v}).
\end{aligned}$$

We assume that the true score, $s_u$, follows a Gaussian distribution with the mean equals to $\mu_0$ and the variance equals to $1/\gamma_0$. Although different graders may have different original bias, we believe that the average original bias of all graders is 0. Hence, we assume that the original bias $\theta_v$ follows a zero-mean Gaussian distribution with a variance $1/\eta_0$. To model the dependency of the bias among the graders, we assume that the mean of $b_v$ follows a Guassian distribution, with the mean as a linear combination of grader $v$'s original bias and the original bias of his/her friends, weighted by the influence matrix $W$. The peer grades follow a Gaussian distribution with the mean equals to the true score of the assignment plus the influenced bias of the grader, and the precision equals to the reliability of the grader. The reliability follows a Gamma distribution with the shape parameter equals to $\alpha_0$ and the rate parameter equals to $\beta_0$. Thus, the mean of the reliability of every grader is $\alpha_0/\beta_0$. Since $\tau_v$ will be plugged into the variance of $s_u$, the scale parameter of the Gamma distribution also responsible for scaling the variances of the peer grades.

## 4.2 The $PG_7$ Model

The proposed $PG_7$ Model is an extension of the $PG_4$ Model in [12] by modeling the dependencies of bias among the graders. The conditional dependence structure between the random variables is expressed by the graphical model in Figure 1b. As shown in this figure, $PG_7$ assume that there is a dependency between the reliability of a grader and the true score of the assignment submitted by that grader, while $PG_6$ does not make such assumption. The following equations show the distributions of all the random variables for the $PG_7$ model.

$$\begin{aligned}
&\tau_v \sim \Gamma(s_v, \beta_0),\\
&\theta_v \sim \mathcal{N}(0, \tfrac{1}{\eta_0}),\\
&b_v \sim \mathcal{N}(\textstyle\sum_{k:k \in N_G(v)} w_{v,k}\theta_k, \tfrac{1}{\phi_0}),\\
&s_u \sim \mathcal{N}(\mu_0, \tfrac{1}{\gamma_0}),\\
&z_u^v \sim \mathcal{N}(s_u + b_v, \tfrac{1}{\tau_v}).
\end{aligned}$$

Unlike $PG_6$ that assumes the reliability of all graders follows the same gamma distribution, in $PG_7$, the reliability of a grader follows a gamma distribution with the true score of his/her submission, $s_v$, as the shape parameter. Thus, the mean of the reliability of grader $v$ is $s_v/\beta_0$. Since the mean of the reliability of a grader increases with her true scores, it captures the intuition that a student with a higher score tends to be a more reliable grader. The distributions of other variables are the same as that in the $PG_6$ Model.

## 4.3 The $PG_8$ Model

The proposed $PG_8$ Model is an extension of the $PG_5$ Model in [12] by modeling the dependencies of bias among the graders. $PG_8$ also assumes that there is a dependency between the reliability of a grader and the true score of the assignment submitted by that grader, thus, the conditional dependence structure of $PG_8$ is the same as that in $PG_7$. The following equations show the distributions of all the random variables for the $PG_8$ model.

$$\begin{aligned}
&\tau_v \sim \mathcal{N}(s_v, \tfrac{1}{\beta_0}),\\
&\theta_v \sim \mathcal{N}(0, \tfrac{1}{\eta_0}),\\
&b_v \sim \mathcal{N}(\textstyle\sum_{k:k \in N_G(v)} w_{v,k}\theta_k, \tfrac{1}{\phi_0}),\\
&s_u \sim \mathcal{N}(\mu_0, \tfrac{1}{\gamma_0}),\\
&z_u^v \sim \mathcal{N}(s_u + b_v, \tfrac{\lambda}{\tau_v}).
\end{aligned}$$

In $PG_8$, the probability distributions of $\tau_v$ and $z_u^v$ are different from that in $PG_7$. We assume that the reliability of grader $v$ follows a Gaussian distribution with $s_v$ as the mean and $1/\beta_0$ as the variance. Thus, the scale of $\tau_v$ will be determined by $s_v$, which is a random variable that we cannot tune. Since $\tau_v$ will be plugged into the variance of $z_u^v$, in order to scale the variance of $z_u^v$, a hyperparameter $\lambda$ is introduced. Hence, we assume $z_u^v$ follows a Gaussian distribution with the variance $\lambda/\tau_v$.

## 4.4 Model Inference

After we formulate the above probabilistic models for peer assessment, the next step is to infer the values of the latent variables including the true score of each student, the reliability, the original bias, and the influenced bias of each grader. These latent variables can be inferred by computing their posterior distribution given the peer grades, $P(\{s_u\}_{u \in U}, \{\theta_v\}_{v \in V}, \{b_v\}_{v \in V}, \{\tau_v\}_{v \in V} | Z)$. However, the latent variables in our proposed models are correlated with each other. For example, the mean of $z_u^v$ in our models is $s_u + b_v$. To estimate $s_u$ given $Z$, we need to have a good estimation of the bias of all the graders who graded submission $u$. To estimate $b_v$ given $Z$, we need to have a good estimation of the true scores of assignments graded by grader $v$. Thus, it is a chicken-and-egg problem. To tackle this problem, there are different approximate inference techniques to infer these latent variables. In this work, we use the Gibbs sampling technique [5] to generate samples of a latent variable from an approximated posterior distribution. After we generated a set of samples of a latent variables, e.g., $s_u^1, s_u^2, ..., s_u^T$, we estimate this latent variable by the empirical mean, e.g., $\hat{s}_u = \frac{1}{T}\sum_{t=1}^{T} s_u^t$. For each latent variable, we run Gibbs sampling for 300 iterations and discard the first 60 burn-in samples. For the latent variables $s_u$ in $PG_7$ and $\tau_v$ in $PG_8$, there is no closed-form distribution for the Gibbs sampling, hence, we perform a discrete approximation to these two latent variables.

Algorithm 1 shows the inference algorithm for our $PG_6$ Model, where $T$ denotes the number of iterations and $B$ denotes the number of burn-in samples. The procedure of the inference algorithm for $PG_7$ and $PG_8$ are the same as the one for $PG_6$, but with different approximated posterior distributions for the latent variables.

**Algorithm 1** $PG_6$ Inference($Z$,$W$,$T$,$B$,$\mu_0$,$\gamma_0$,$\alpha_0$,$\beta_0$,$\eta_0$,$\phi_0$)

```
$s_u \sim \mathcal{N}(\mu_0, \frac{1}{\gamma_0})$
$\tau_v \sim \Gamma(\alpha_0, \beta_0)$
$\theta_v \sim \mathcal{N}(0, \frac{1}{\eta_0})$
$b_v \sim \mathcal{N}(\sum_{k:k \in N_G(v)} w_{v,k}\theta_k, \frac{1}{\phi_0})$
for $t = 1 \to T$ do
    for each true score $s_{u_i}$ do
        Sample $s$ according to Eq. (1)
        $s_{u_i} \leftarrow s$
    for each grader reliability $\tau_{v_i}$ do
        Sample $\tau$ according to Eq. (2)
        $\tau_{v_i} \leftarrow \tau$
    for each grader original bias $\theta_{v_i}$ do
        Sample $\theta$ according to Eq. (3)
        $\theta_{v_i} \leftarrow \theta$
    for each grader influenced bias $b_{v_i}$ do
        Sample $b$ according to Eq. (4)
        $b_{v_i} \leftarrow b$
    $\xi^{(t)} \leftarrow (\{s_u\}_{u \in U}, \{\tau_v\}_{v \in V}, \{\theta_v\}_{v \in V}, \{b_v\}_{v \in V})$
$(\{\hat{s}_u\}_{u \in U}, \{\hat{\tau}_v\}_{v \in V}, \{\hat{\theta}_v\}_{v \in V}, \{\hat{b}_v\}_{v \in V}) \leftarrow$
$\frac{1}{T-B}\sum_{t=B+1}^{T} \xi^t$
return $(\{\hat{s}_u\}_{u \in U}, \{\hat{\tau}_v\}_{v \in V}, \{\hat{\theta}_v\}_{v \in V}, \{\hat{b}_v\}_{v \in V})$
```

### 4.4.1 Approximated Posterior Distributions

The approximated posterior distributions for the latent variables in $PG_6$ are shown as follows.

$$s \sim \mathcal{N}(\frac{\gamma_0\mu_0 + \sum_{v:v \to u_i} \tau_v(z_{u_i}^v - b_v)}{\gamma_0 + \sum_{v:v \to u_i} \tau_v}, \frac{1}{\gamma_0 + \sum_{v:v \to u_i} \tau_v}) \tag{1}$$

$$\tau \sim \Gamma(\alpha_0 + \frac{n_{v_i}}{2}, \beta_0 + \frac{1}{2}\sum_{u:v_i \to u}(z_u^{v_i} - (s_u + b_{v_i}))^2) \tag{2}$$

$$\theta \sim \mathcal{N}(\frac{\sum_{k:k \in N_G(v_i)} \phi_0(b_k w_{k,v_i} - w_{k,v_i}\sum_{j:j \in N_G(k), j \neq v_i} w_{k,j}\theta_j)}{\eta_0 + \sum_{k:k \in N_G(v_i)}}, \frac{1}{\eta_0 + \sum_{k:k \in N_G(v_i)}}) \tag{3}$$

$$b \sim \mathcal{N}(\frac{\phi_0 \sum_{k:k \in N_G(v_i)} w_{v_i,k}\theta_k + \sum_{u:v_i \to u} \tau_{v_i}(z_u^{vi} - s_u)}{\phi_0 + \sum_{u:v_i \to u} \tau_{v_i}}, \frac{1}{\phi_0 + \sum_{u:v_i \to u} \tau_{v_i}}) \tag{4}$$

The followings are the approximated posterior distributions for the latent variables in $PG_7$. However, sampling distribution for $s$ has no closed form. Hence, we use a discrete approximation to approximate this posterior distribution with intervals of width 0.1.

$$s \propto G_{u_i} \exp(\frac{-1}{2}R(s_{u_i}^2 - \frac{\gamma_0\mu_0 + \sum_{v:v \to u_i} \tau_v(z_{u_i}^v - b_v)}{\gamma_0 + \sum_{v:v \to u_i} \tau_v})^2)$$

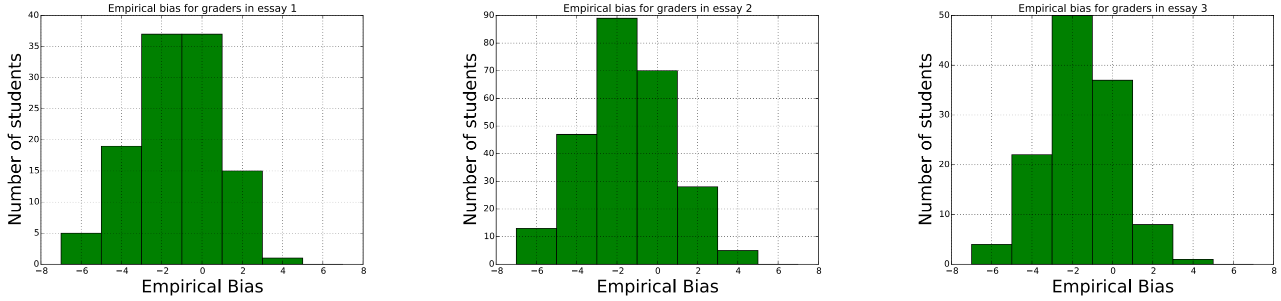
$$\tau \sim \Gamma(s_v + \frac{n_{v_i}}{2}, \beta_0 + \frac{1}{2}\sum_{u:v_i \to u}(z_u^{v_i} - (s_u + b_{v_i}))^2)$$

$$\theta \sim \mathcal{N}(\frac{\sum_{k:k \in N_G(v_i)} \phi_0(b_k w_{k,v_i} - w_{k,v_i}\sum_{j:j \in N_G(k), j \neq v_i} w_{k,j}\theta_j)}{\eta_0 + \sum_{k:k \in N_G(v_i)}}, \frac{1}{\eta_0 + \sum_{k:k \in N_G(v_i)}})$$

$$b \sim \mathcal{N}(\frac{\phi_0 \sum_{k:k \in N_G(v_i)} w_{v_i,k}\theta_k + \sum_{u:v_i \to u} \tau_{v_i}(z_u^{vi} - s_u)}{\phi_0 + \sum_{u:v_i \to u} \tau_{v_i}}, \frac{1}{\phi_0 + \sum_{u:v_i \to u} \tau_{v_i}})$$

The followings are the approximated posterior distributions for the latent variables in $PG_8$. However, sampling distribution for $\tau$ has no closed form. Hence, we use a discrete approximation to approximate this posterior distribution with intervals of width 0.1.

$$s \sim \mathcal{N}(\frac{\gamma_0\mu_0 + \sum_{v:v \to u_i} \frac{\tau_v}{\lambda}(z_u^v - b_v)}{\gamma_0 + \beta_0 + \sum_{v:v \to u_i} \tau_v}, \frac{1}{\gamma_0 + \beta_0 + \sum_{v:v \to u_i} \tau_v})$$



(a) Empirical bias of graders in essay 1

(b) Empirical bias of graders in essay 2

(c) Empirical bias of graders in essay 3

Figure 2: The empirical bias of graders in each essay question

$$\tau \propto \frac{\tau_{v_i}}{\lambda}^{\frac{n_{v_i}}{2}} \exp(\frac{-1}{2}[\beta_0 \tau_{v_i}^2 - 2(\beta_0 \tau_{v_i}^2 - \frac{\sum_{u:v_i \to u}(z_u^{v_i} - (s_u + b_{v_i}))^2}{2\lambda})\tau_{v_i}])$$

$$\theta \sim \mathcal{N}(\frac{\sum_{k:k\in N_G(v_i)} \phi_0(b_k w_{k,v_i} - w_{k,v_i} \sum_{j:j\in N_G(k), j\neq v_i} w_{k,j}\theta_j)}{\eta_0 + \sum_{k:k\in N_G(v_i)}}, \frac{1}{\eta_0 + \sum_{k:k\in N_G(v_i)}})$$

$$b \sim \mathcal{N}(\frac{\phi_0 \sum_{k:k\in N_G(v_i)} w_{v_i,k}\theta_k + \sum_{u:v_i\to u} \frac{\tau_{v_i}}{\lambda}(z_u^{vi} - s_u)}{\phi_0 + \sum_{u:v_i\to u} \frac{\tau_{v_i}}{\lambda}}, \frac{1}{\phi_0 + \sum_{u:v_i\to u} \frac{\tau_{v_i}}{\lambda}})$$

# 5. EXPERIMENTS

We conduct experiments to compare the performances of our proposed probabilistic models with other state-of-the-art probabilistic models for peer assessment. The empirical studies are intended to address the following questions:

1. Can we improve the accuracy of existing probabilistic models by incorporating the social connections among the graders?
2. How does the number of social connections among the students affect the performances of our models?

## 5.1 Real Dataset

### 5.1.1 *Peer Grading Dataset*

The real peer grading dataset is collected from a massive open online course called “Art History: A look into masters and classics” offered by the Tsinghua University on the xuetangX Platform. This dataset contains both the peer grades given by the students as well as the grades given by the teaching staff.

The dataset consisted of the peer grading results of three essay questions. The summary statistics of these peer grading results are presented in Table 2. For each essay question, each student was asked to evaluate three submissions of his/her peers according to the rubrics specified by the instructor. Some students eventually evaluated more than three submissions while some students evaluated less than three submissions. The system automatically assigned peer grading tasks to the students such that the each submission was graded by a similar number of graders. The gradee identities were concealed from the graders, and vice versa, throughout the whole peer grading process. At the end of the peer assessment process, the system used the median of the scores given by the graders as the predicted score of a submission.

Besides the grades given by the students, this dataset also contains grades given by the teaching staffs. A very little proportion of the assignments did not receive any peer grades, we eliminate such staff grades in our work. As shown in Table 2, 99% of this dataset are staff-graded submissions. These staff grades are considered as the ground truth grades for the assignments. One of the advantages of this peer grading dataset is that it contains a much higher percentage of staff-graded submissions than other related peer grading dataset in MOOCs, e.g., 2.6% in [12], and 0.21% to 0.35% in [15]. Thus, this dataset allows us to evaluate the quality of most of the predicted grades of assignments.

Next, we give an overview of the influenced bias of the students in this peer grading dataset. We estimate the influenced bias of students by their empirical bias. The empirical bias of a student is the average of inflation of grades given by that grader. For example, the peer grades given by grader $v_i$ are $z^{v_i}_{u_1}$, $z^{v_i}_{u_2}$, and $z^{v_i}_{u_2}$. Then the empirical bias of grader $v_i$ is $[(z^{v_i}_{u_1} - s_{u_1}) + (z^{v_i}_{u_2} - s_{u_2}) + (z^{v_i}_{u_3} - s_{u_3})]/3$.

The histograms of the empirical bias of students are shown in Figure 2. From these histograms, we can see the bias of most of the students is more negative than -1. Moreover, a significant amount of their bias is more negative than -3. Hence, the students in this dataset tend to have tough grading standards and we cannot ignore the bias of students when predicting the true score of each assignment from peer grades.

Table 2: Summary statistics of the essay questions for peer grading

| | Essay 1 | Essay 2 | Essay 3 |
|---|---|---|---|
| Submissions | 126 | 288 | 141 |
| Staff grades | 126 | 286 | 516 |
| Peer grades | 493 | 1121 | 516 |
| Full scores | 15 | 15 | 15 |

Table 3: Summary statistics of students' forum activities

| | Students in essay 1 | Students in essay 2 | Students in essay 3 |
|---|---|---|---|
| # Threads | 11 | 80 | 6 |
| # Comments | 1704 | 3233 | 2078 |
| # Upvotes | 75 | 219 | 74 |
| # Implicit social links | 12 | 36 | 8 |

### 5.1.2 Forum Activities Data

To extract the social connections among the students, we investigate students' social activities in the discussion forums of the xuetangX platform. For each essay question, we show the summary statistics of the forum activities of the involved students in Table 3. The discussion forums in xuetangX allow students to give an upvote to a particular comment or thread. These upvote activities actually express the preferences of the students. Studies showed that there is a strong correlation between preference similarities and trust relationships in a social network [30]. Hence, we assume that users with similar preferences tend to form a community. Based on this assumption, we use the upvote similarity between the students to infer their social connections, i.e., we build the homophily-based implicit social graph [28] for the students by using their upvote similarity. More specifically, we first represent the upvote activities of each student. Secondly, we measure the upvote similarity between the students. Thirdly, we infer the weights for the social links between the students.

To represent the upvote activities of each student, we use a vector, $l_i \in \mathbb{R}^n$, to be the upvote vector of student $i$. The j-th entry of $l_i$ will be 1 if student $i$ gives at least one upvote to student $j$'s threads or comments. Otherwise, the j-th entry of $l_i$ will be 0. For example, there are three students, $v_1$, $v_2$, and $v_3$. Student 1 only gives an upvote to student 3's thread, so $l_1 = [0, 0, 1]$.

Next, we measure the upvote similarity between two students, $v_i$ and $v_j$, by the Jaccard similarity coefficient of their upvote vectors, $J(l_i, l_j)$. Jaccard similarity coefficient is a well-known metric to measure the similarity between two sample sets, with a range from 0 to 1. The higher the Jaccard similarity coefficient, the more similar the two sample sets. Then, we assign the value of $J(l_i, l_j)$ as the weight for the social link between $v_i$ and $v_j$ if $J(l_i, l_j) > threshold$. In this work, we set $threshold = 0.4$. More formally, let $A \in \mathbb{R}^{n\times n}$ be the adjacency matrix for the social graph. For $i, j = 1, ..., n$,

$$A_{i,j} = \begin{cases} J(l_i, l_j), & \text{if } J(l_i, l_j) > 0.4 \\ 0, & \text{otherwise.} \end{cases}$$

The fourth row of table 3 shows the number of links in the homophily-based implicit social graph for the students in each essay question after excluding all the loops (a link that connect a student to itself).

After we have extracted the social graph for the students, we construct the social influence matrix, $W$, by normalizing the rows of matrix $A$, such that $\sum_{j=1}^{n} W_{i,j} = 1$ for $i = 1, ..., n$.

## 5.2 Evaluation Metrics

We use the root-mean-square error (RMSE) to measure the deviations of the predicted grades from the ground truth grades. The RMSE is a widely used metric in cardinal peer assessment literatures [15, 12]. The formal definition of the RMSE is in Eq. (5). We use $s_u^*$ to denote the ground truth score of student $u$'s submission and $U^*$ to denote the set of staff-graded submissions.

$$\text{RMSE} = \sqrt{\sum_{u \in U} (s_u - s_u^*)^2}. \tag{5}$$

The lower the RMSE of the predicted grades, the higher the accuracies of the predicted grades.

## 5.3 Comparison Methods

In order to demonstrate the advantages of our models, we compare our models with some baseline methods. Since the problem we solve in this paper is to predict the true score of each assignment, we will not compare our models with ordinal peer assessment models. We consider the following peer assessment models as baseline methods.

- Median: Simply taking the median of the peer grades given to an assignment as the predicted grade.
- $PG_1$: The first probabilistic model for peer assessment [15]. Our $PG_6$ Model is an extension of this model.
- $PG_3$: A probabilistic model which assumes that the reliability of a grader is linearly correlated with the true grade of a grader [15].
- $PG_4$: A probabilistic model which assumes a probabilistic relationship between the reliability of a grader and the true grade of a grader [12]. Our $PG_7$ Model is an extension of this model.
- $PG_5$: Another probabilistic model which also assumes a probabilistic relationship between the reliability of a grader and the true grade of a grader [12]. Our $PG_8$ Model is an extension of this model.

## 5.4 Performance on Real Data

Table 4 shows the accuracies for the grades predicted by different models. The RMSE reported in this table is the average of RMSE over ten repetitions. STD stands for the standard deviation of the RMSE. For all the above probabilistic models, we tune the hyperparameters to the ones which achieved the lowest RMSE on the dataset. Overall, our $PG_7$ Model is the most accurate model among all the above methods, while the median baseline is the least accurate method. The $PG_5$ Model is the least accurate probabilistic model in this dataset although it is the more accurate than $PG_4$ and $PG_3$ in [12]. Thus, the $PG_5$ Model does not fit this data set well.

Since our $PG_6$, $PG_7$, and $PG_8$ are extensions of $PG_1$, $PG_4$, and $PG_5$ respectively by incorporating social connections of graders, we compare these three pairs of models as follows:

- **$PG_8$ vs. $PG_5$ :** From the 3-rd and 4-th rows of Table 4, we can see that the RMSE of $PG_8$ is significantly lower than that of $PG_5$ in essay 1, essay 2, and essay

3. The standard deviation of RMSE of $PG_8$ is also significantly lower than that of $PG_5$ in all the 3 essays.

- $PG_6$ **vs.** $PG_1$ **:** From the 5-th and 6-th rows of Table 4, we can see that the RMSE of $PG_6$ is slightly lower than that of $PG_1$ in essay 2 and essay 3. Both methods achieved the same RMSE of 2.30 in essay 1. The standard deviations of RMSE of $PG_6$ and that of $PG_1$ are similar in all the 3 essays.

- $PG_7$ **vs.** $PG_4$ **:** From the 7-th and 8-th rows of Table 4, we can see that the RMSE of $PG_7$ is slightly lower than that of $PG_4$ in essay 1, essay 2, and essay 3. The standard deviations of RMSE of $PG_6$ and that of $PG_1$ are similar in all the 3 essays.

Even through $PG_5$ does not fit this data set very well, by extending this model with social connections of graders, we can significantly improve its accuracy and stability. For $PG_4$ and $PG_1$, incorporating social connections of graders can slightly improve their accuracy. In sum up, by leveraging the social connections of graders, we can improve the accuracy of peer assessment.

### 5.5 Simulated Studies

We use synthetic data to analyze the performances of our models, $PG_6$, $PG_7$, and $PG_8$ under different numbers of social connections among the students. First, we simulate six adjacent matrices, $A'_1, ..., A'_6$ among 100 students, with 0, 20, 40, 60, 80, 100 social connections respectively. Similar to the real data set, we construct the social influence matrices, $W'_1, ..., W'_6$ from these adjacent matrices. Then, for each of our peer grading models, we follow its probabilistic assumptions to simulate six different sets of ground truth grades and peer grades of all students by using six different social influence matrices.

Figure 3 illustrates the performances of our models on synthetic datasets under different numbers of social connections. For $PG_7$ and $PG_8$, we can see that their errors decrease gradually as the number of social connections increases. Thus, the accuracy of our $PG_7$and $PG_8$ model increases with the number of social connections. For $PG_6$, its error has some fluctuations but does not show any trends as the number of social connections increases. Hence, the accuracy of our $PG_6$ model is less sensitive to the number of social connections compare to $PG_7$ and $PG_8$.

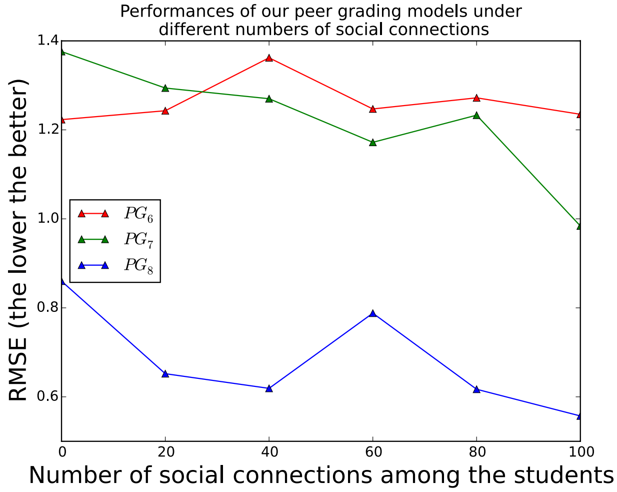


Figure 3: The performances of our peer grading models under different numbers of social connections.

## 6. CONCLUSION AND DISCUSSION

Table 4: The accuracies for the grades predicted by different models

| | Essay 1 | | Essay 2 | | Essay 3 | |
|---|---|---|---|---|---|---|
| | RMSE | STD | RMSE | STD | RMSE | STD |
| Median | 3.14 | 0.00 | 3.91 | 0.00 | 3.17 | 0.00 |
| $PG_3$ | 2.27 | 0.01 | 2.35 | 0.00 | 2.32 | 0.01 |
| $PG_5$ | 3.00 | 0.05 | 3.62 | 0.06 | 3.01 | 0.17 |
| $PG_8$ | **2.46** | **0.01** | **2.47** | **0.01** | **2.25** | **0.01** |
| $PG_1$ | 2.30 | 0.00 | 2.29 | 0.00 | 2.06 | 0.00 |
| $PG_6$ | **2.30** | **0.00** | **2.19** | **0.00** | **2.02** | **0.00** |
| $PG_4$ | 2.28 | 0.01 | 2.14 | 0.00 | 1.97 | 0.01 |
| $PG_7$ | **2.27** | **0.01** | **2.13** | **0.00** | **1.96** | **0.01** |

In this paper, we propose new probabilistic models for peer assessment by incorporating the social influences among the students, on this basis of the intuition that the bias of graders will be affected by their social connections. To the best of our knowledge, this is the first work to leverage the social information of students to improve the accuracy of peer assessment. To verify the performances of our models, we conduct experiments on a new peer grading dataset which is enhanced by the social information of the students in the discussion forum of the course. Experimental results show that our proposed models outperform previous work in terms of the accuracies of the predicted grades. The implication of this work is that by leveraging the social information of the students, we can improve the accuracy of peer assessment.

Our proposed models can be easily extended to model the dependencies of the reliability among the students. It is also possible that we can further improve the accuracy of the predicted grade by modeling the dependencies of the true scores among the students. However, such models will be unfair to the good students who often interact with students with lower grades.

Besides the area of MOOCs, our models can also be applied to the area of crowdsourcing [18, 19]. If a crowdsourced task requires subjective evaluations from the crowdworkers, e.g., evaluate the quality of an image. Then, our model can be easily extended to learn the bias, the reliability of the crowdworkers, and the true quality of the images.

## 7. ACKNOWLEDGMENTS

The work described in this paper was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK 14208815 of the General Research Fund) and Ministry of Education of China (D.01.16.00101). We would like to thank the xuetangX platform for letting us use the learning data for this research. We would also like to thank Prof. Jie Tang and Mr. Xia Jing for their help in accessing and processing the data.

## 8. REFERENCES

[1] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952.

[2] H. P. Chan, T. Zhao, and I. King. Trust-aware peer assessment using multi-armed bandit algorithms. In *Proceedings of the 25th International Conference on World Wide Web, Companion Volume*, pages 899–903, 2016.

[3] 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*, pages 415–420, 2014.

[4] L. de Alfaro and M. Shavlovsky. Dynamics of peer grading: An empirical study. In *Proceedings of the 9th International Conference on Educational Data Mining*, pages 62–69, 2016.

[5] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, 1984.

[6] P. Gutierrez, N. Osman, and C. Sierra. Collaborative assessment. In *Proceedings of the 17th International Conference of the Catalan Association for Artificial Intelligence*, pages 136–145, 2014.

[7] M. I. Jordan. Graphical models. *Statistical Science*, pages 140–155, 2004.

[8] C. E. Kulkarni, P. W. Wei, H. Le, D. J. hao Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer. Peer and self assessment in massive online classes. *ACM Trans. Comput.-Hum. Interact.*, 20(6):33, 2013.

[9] R. D. Luce. *Individual choice behavior: A theoretical analysis.* Courier Corporation, 2005.

[10] H. Ma, H. Yang, M. R. Lyu, and I. King. Sorec: social recommendation using probabilistic matrix factorization. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 931–940, 2008.

[11] C. L. Mallows. Non-null ranking models. i. *Biometrika*, pages 114–130, 1957.

[12] 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*, pages 454–460, 2015.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[14] D. E. Paré and S. Joordens. Peering into large lectures: examining peer and expert mark agreement using peerscholar, an online peer assessment tool. *J. Comp. Assisted Learning*, 24(6):526–540, 2008.

[15] 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*, pages 153–160, 2013.

[16] K. Raman and T. Joachims. Methods for ordinal peer grading. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1037–1046, 2014.

[17] K. Raman and T. Joachims. Bayesian ordinal peer grading. In *Proceedings of the Second ACM Conference on Learning @ Scale*, pages 149–156, 2015.

[18] F. P. Ribeiro, D. A. F. Florêncio, and V. H. Nascimento. Crowdsourcing subjective image quality evaluation. In *18th IEEE International Conference on Image Processing*, pages 3097–3100, 2011.

[19] F. P. Ribeiro, D. A. F. Florêncio, C. Zhang, and M. L. Seltzer. CROWDMOS: an approach for crowdsourcing mean opinion score studies. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2416–2419, 2011.

[20] M. S. M. Sajjadi, M. Alamgir, and U. von Luxburg. Peer grading in a course on algorithms and data structures: Machine learning algorithms do not improve over simple baselines. In *Proceedings of the Third ACM Conference on Learning @ Scale*, pages 369–378, 2016.

[21] N. B. Shah, J. K. Bradley, A. Parekh, M. Wainwright, and K. Ramchandran. A case for ordinal peer-evaluation in MOOCs. *In NIPS Workshop on Data Driven Education*, 2013.

[22] P. Singla and M. Richardson. Yes, there is a correlation: - from social networks to personal behavior on the web. In *Proceedings of the 17th International Conference on World Wide Web*, pages 655–664, 2008.

[23] T. Walsh. The peerrank method for peer assessment. In *European Conference on Artificial Intelligence*, pages 909–914, 2014.

[24] 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*, pages 177–183, 2015.

[25] F. L. Wauthier, M. I. Jordan, and N. Jojic. Efficient ranking from pairwise comparisons. In *Proceedings of the 30th International Conference on Machine Learning*, pages 109–117, 2013.

[26] Q. Wu, H. Wang, Q. Gu, and H. Wang. Contextual bandits in a collaborative environment. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 529–538, 2016.

[27] 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*, pages 537–546, 2011.

[28] T. Zhao, J. Hu, P. He, H. Fan, M. R. Lyu, and I. King. Exploiting homophily-based implicit social network to improve recommendation performance. In *2014 International Joint Conference on Neural Networks*, pages 2539–2547, 2014.

[29] T. Zhao, J. J. McAuley, and I. King. Leveraging social connections to improve personalized ranking for collaborative filtering. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 261–270, 2014.

[30] C. Ziegler and J. Golbeck. Investigating interactions of trust and interest similarity. *Decision Support Systems*, 43(2):460–475, 2007.