Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Probabilistic Graphical Models for Boosting Cardinal and Ordinal Peer Grading in MOOCs

Fei Mi, Dit-Yan Yeung
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Hong Kong
fmi@cse.ust.hk, dyyeung@cse.ust.hk

Abstract

With the enormous scale of massive open online courses (MOOCs), peer grading is vital for addressing the assessment challenge for open-ended assignments or exams while at the same time providing students with an effective learning experience through involvement in the grading process. Most existing MOOC platforms use simple schemes for aggregating peer grades, e.g., taking the median or mean. To enhance these schemes, some recent research attempts have developed machine learning methods under either the cardinal setting (for absolute judgment) or the ordinal setting (for relative judgment). In this paper, we seek to study both cardinal and ordinal aspects of peer grading within a common framework. First, we propose novel extensions to some existing probabilistic graphical models for cardinal peer grading. Not only do these extensions give superior performance in cardinal evaluation, but they also outperform conventional ordinal models in ordinal evaluation. Next, we combine cardinal and ordinal models by augmenting ordinal models with cardinal predictions as prior. Such combination can achieve further performance boosts in both cardinal and ordinal evaluations, suggesting a new research direction to pursue for peer grading on MOOCs. Extensive experiments have been conducted using real peer grading data from a course called "Science, Technology, and Society in China I" offered by HKUST on the Coursera platform.

Introduction

Massive open online courses (MOOCs) are drawing increased attention lately because they can go well beyond the boundaries of traditional classrooms and audiences. While each offering of a traditional course delivered by a famous professor from a top university can typically benefit at most hundreds of students, and only students, an online version can easily scale it up by several orders of magnitude and make it available to people from all walks of life around the globe. However, the massive scale of MOOCs poses great challenges to student assessment. While most existing MOOCs simply give multiple-choice assignment or exam questions which can be graded automatically, open-ended and free-response exercises or essays are arguably vital for assessing learning outcomes of many courses (Paré and Joordens 2008). Unfortunately, satisfactory automatic grading of such assessment forms is beyond the current state of the art. A practical approach to tackle this problem is *peer grading* or *peer assessment* (Godlee et al. 2003; Sadler and Good 2006), in which students also play the role of graders in grading a small number of assignments submitted by other students according to the rubrics or benchmarks provided by the course instructor. The final score assigned to a submission is usually some aggregate, such as the median or mean, of the peer grades given by the graders. Due to the great diversity in student background, purpose and engagement, applying peer grading to give students a fair judgment of their learning efforts and achievements is hardly a trivial task. For example, by monitoring the registration information and accessing IP addresses of the students in one of the MOOCs offered by our university on the Coursera platform, we found the students were from around 160 different countries. Also, a recent study (Anderson et al. 2014) showed that students in a MOOC often exhibit different engagement styles probably corresponding to different purposes.

There are generally two basic rationales behind peer grading research, namely, cardinal and ordinal. In cardinal peer grading or absolute judgment (Barnett 2003; Paré and Joordens 2008), the peer evaluations for assignments are in the form of explicit numerical scores. This cardinal setting is currently the most common choice for MOOCs. Its goal is usually to minimize the absolute prediction deviation from some ground truth. However, (Shah et al. 2013; 2014) argued that it is sometimes easier for non-expert peer graders to make comparison than to give absolute scores and hence ordinal evaluation can sometimes be more effective than cardinal evaluation. As demonstrated by (Stewart, Brown, and Chater 2005; Carterette et al. 2008; Shah et al. 2014), on specific tasks, relative judgment or ordinal evaluation is more accurate than absolute judgment or cardinal evaluation. In ordinal peer grading, peer graders perform ordinal comparisons by ranking different assignments in terms of quality. A special case is pairwise comparison in which two assignments are compared at a time. The goal of ordinal peer grading is to predict the relative ranking correctly. We note that curved grading is used by many courses. As such, even absolute scores have to be converted into percentiles before the final grades are assigned. Thus we believe that

both the cardinal and ordinal aspects have important roles to play in peer grading.

Before we proceed to focus on peer grading for MOOCs, we note that reviewing or ranking items is a broad topic. Some examples include ranking search results (Aslam and Montague 2001; Joachims 2002), reviewing conference or journal papers (Peters and Ceci 1982; Harnad 2000; Rowland 2002), admitting college students, as well as voting for candidates as a social choice problem (Coughlin 1992; Arrow 2012). Without going into a detailed comparison due to page constraints, we just emphasize that peer grading for MOOCs is somewhat unique in that the peer grades are often very sparse and a total order of all the items is needed to evaluate all students. Peer grading is also related to a rapidly growing area called crowdscourcing (Surowiecki 2005; Howe 2006) which seeks to take advantage of the wisdom of the crowd. A crucial research issue in crowdsourcing is how to aggregate the results provided by individual workers to give better-quality results.

The main contribution of this paper is twofold:

- We propose novel extensions to some state-of-the-art probabilistic graphical models for cardinal peer grading (Piech et al. 2013). Based on both average-case and worst-case performance criteria, our models show improvement in cardinal evaluation. Moreover, our extensions also outperform conventional ordinal models significantly in ordinal evaluation.
- We combine cardinal and ordinal models by augmenting ordinal models with the cardinal predictions as prior. Such combination can achieve further performance boosts in both cardinal and ordinal evaluations.

To our knowledge, this is the first research work which studies both the cardinal and ordinal aspects of peer grading within a common framework.

Related Work

We first review some existing methods for peer grading under the cardinal setting for predicting scores and the ordinal setting for predicting ranking or pairwise preferences.

The Vancouver algorithm (de Alfaro and Shavlovsky 2014) measures each student's grading accuracy by comparing the grades assigned by the student with the grades given to the same submissions by other students. Specifically, the grading accuracy is determined by the grading variance and higher weights are assigned to graders with higher grading accuracy. It iteratively updates the grading variance of each user from the consensus grades and computes more precise consensus grades from the updated grader variance. Inspired by the PageRank algorithm (Page et al. 1999) which was first used by the Google search engine, the PeerRank algorithm (Walsh 2014) was proposed as another iterative consensus algorithm. It relates a grader's grading ability to her performance in the course because the peer graders are also students. In (Patricia, Nardine, and Carles 2014), a trust graph is built over the referees and is used to compute the weights for assessment aggregation. Incorporated in the trust network are staff grades that are used for performance evaluation. Most related to our research is the work reported in (Piech et al. 2013) which proposed probabilistic models that model the relationships between the grader's bias and reliability, the true score of each submission, and the observed peer grades. Their PG_3 model assumes that high-scoring students tend to be more reliable as graders. The realization of this idea is extended in our proposed models to be described in the next section.

For ordinal methods, the Bradley-Terry model is used in (Shah et al. 2013) to learn the latent ability of the students from ordinal peer comparisons and perform cross-validation experiments to predict peer preferences. In (Raman and Joachims 2014), several statistical models for ordinal comparison, including the Bradley-Terry Model, Mallows Model, Thurstone Model, and Plackett-Luce Model, are applied to peer grading tasks. Performance evaluation is based on data collected from the real peer grading process of a (traditional) class consisting of 170 students.

An observation from these previous studies is worth mentioning. Not surprisingly, having more peer evaluations per submission improves the estimation accuracy for both the cardinal and ordinal settings. However, if a student is given too many assignments to grade, she may just complete the task randomly, defeating the purposes of improving the estimation accuracy and benefiting the students through the peer review process as a learning experience. Consequently, data sparsity is more of the norm than the exception and hence is a key issue to address.

Cardinal Peer Grading

We first study cardinal peer grading in this section. Our models are extensions of the probabilistic models proposed in (Piech et al. 2013). Like the models (named PG_1 to PG_3) in (Piech et al. 2013), we explicitly model the bias and reliability of each grader. While the bias corresponds to a constant grade inflation or deflation, the reliability measures how close on average is the assessment by a grader from the latent true score of the submission after eliminating the bias. Concretely, the reliability is defined as the inverse variance (or precision) of a Gaussian distribution. The experiments in (Piech et al. 2013) show that PG_3, which uses a linear relationship to relate the reliability and true score of a grader in a deterministic fashion, outperforms PG_1 which is a simpler model. We note that the idea of assigning grades by also capturing the grader's effort and grading accuracy has been exploited by other researchers, e.g., (Shah et al. 2013; Walsh 2014).

Although dependency exists between a grader's reliability and her true score, we believe that a linear deterministic relationship is overly rigid and may not provide a good enough model for many graders. Here we propose to relax it by using a probabilistic relationship instead, resulting in two models referred to as PG_4 and PG_5 in the sequel. Besides modeling the effect of the grader's true score on her reliability, we have also considered some other features such as the course forum reputation score, performance in in-video quizzes, and number of course forum views. However, none of them turns out to be more effective than the grader's true score. This is at

least true for the course considered in this study. As such, we will only focus on the influence of a grader's true score on her grading reliability in this paper.

Notations

Let v denote an arbitrary grader and V the collection of all graders. Similarly, let u denote an arbitrary student (a.k.a. gradee) whose submissions are graded and U the collection of all students. Table 1 shows the variables used in the graphical model. We note that the only observed variable z_u^v represents the score or peer assessment assigned to a submission of student u by grader v.

Notation	Description
τ_v	Reliability of grader v
b_v	Bias of grader v
s_u	True score for a submission of student u
z_u^v	Score assigned to a submission of u by v

Table 1: Notations used in our models

Graphical Model

As depicted in Figure 1, the plate notation is used to represent separately the gradee and grader roles for clarity. For the peer grading setting, however, the gradees and graders actually correspond to (roughly) the same group of students. The first three variables in Table 1 are latent variables in the graphical model for the reliability, bias, and true score of a student, and the last one is the only observed variable involving both a gradee and a grader. The hyperparameters $\mu_0, \gamma_0, \beta_0, \eta_0$ are used for specifying the probability distributions of the latent variables. A major difference between this model and the model in (Piech et al. 2013) is the way in which the relationship between the reliability and true score of a grader is modeled. Here we impose a probabilistic relationship between them. Moreover, unlike PG_3, we represent the reliability explicitly as a random variable in the graphical model to make it easier for future extension to incorporate other factors that can influence the reliability.

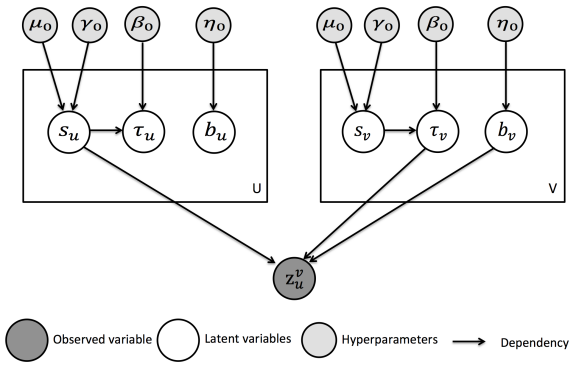

Figure 1: Graphical model

Two Model Variants: PG_4 and PG_5

$\boldsymbol{PG_4}$ **Model**	$\boldsymbol{PG_5}$ **Model**
$\tau_v \sim \mathcal{G}(s_v, \beta_0)$	$\tau_v \sim \mathcal{N}(s_v, \frac{1}{\beta_0})$
$b_v \sim \mathcal{N}(0, \frac{1}{\eta_0})$	$b_v \sim \mathcal{N}(0, \frac{1}{\eta_0})$
$s_u \sim \mathcal{N}(\mu_0, \frac{1}{\gamma_0})$	$s_u \sim \mathcal{N}(\mu_0, \frac{1}{\gamma_0})$
$z_u^v \sim \mathcal{N}(s_u + b_v, \frac{1}{\tau_v})$	$z_u^v \sim \mathcal{N}(s_u + b_v, \frac{\lambda}{\tau_v})$

To realize the graphical model formulation in Figure 1, we propose two model variants called PG_4 and PG_5. In PG_4, a grader's reliability τ_v follows the gamma distribution with rate parameter β_0 and the grader's true score s_v as the shape parameter. Consequently, as desired, the mean reliability of a grader is s_v/β_0, which increases with her latent true score.

On the other hand, PG_5 assumes that τ_v follows the Gaussian distribution with s_v as the mean and $1/\beta_0$ as the variance. We note that the β_0 parameter in PG_5 is different from that in PG_4, but we want to describe both PG_4 and PG_5 using the same graphical model in Figure 1. For both PG_4 and PG_5, the mean reliability of grader v is positively correlated with her true score s_v and both the bias and true score follow the Gaussian distributions. The mean bias across all graders is assumed to be zero and the mean true score μ_0 is set to the average score of all the submissions. The observed score z_u^v follows the Gaussian distribution with the mean equal to the true score plus the grader's bias and the variance inversely related to the grader's reliability. In this sense, the hyperparameter β_0 in PG_4 plays an important role in scaling τ_v to a proper range before plugging into the variance of the Gaussian density. In PG_5, the scale of τ_v depends on s_v or the grading scheme for a specific grading task, so we have a hyperparameter λ, similar to the role of β_0 in PG_4, to scale the variance of the Gaussian density. Due to page constraints, details of the model inference procedures for PG_4 and PG_5 are described in the appendix. They are mostly based on Gibbs sampling by running for 300 iterations with the first 60 burn-in samples eliminated. For the latent variables s_u in PG_4 and τ_v in PG_5, however, closed-form distributions are not available for performing Gibbs sampling and hence discrete approximation is applied with fine separating intervals. Moreover, for a small group of graders with no submission and hence no score for an assignment, we assume that they have the lowest scores and relatively low reliability among all graders for a particular assignment.

Dataset

The peer grading dataset used in our experiments is from a course called "Science, Technology, and Society in China I" offered by our university on the Coursera platform. Table 2 gives summary statistics of the three assignments involving peer grading. The peer grading method adopted by the course is detailed below:

- There are three assignments, each of which asks the students to write a short essay with suggested word limit (250 for the first and 500 each for the second and third ones).

- For each assignment, each student was asked to evaluate three other submissions although some ended up evaluating more and some less. A few even evaluated more than 10. Grader assignment was done automatically and randomly by the system with the goal of maintaining a similar number of graders for each submission.
- Grading was based on three rubrics provided: 0-7, 0-7, 0-7 for assignment 1 and 0-7, 0-7, 0-11 for assignments 2 and 3. The median of the peer grades for each rubric was used to compute the assignment score for each submission.

For each assignment, as in Table 2, around 20 submissions were graded by the course instructor, and on average, there are 4 peer graders per staff-graded assignment. We treat these staff grades as ground truth for those submissions involved. This is a versatile dataset for peer grading research. Not only is it reasonably large with 7546 peer grades, but it also contains more submissions with staff grades than some other related datasets, e.g., 3 to 5 staff-graded submissions per assignment in (Piech et al. 2013). As such, both cardinal and ordinal evaluations using ground-truth data can be conducted.

	Assignment 1	**Assignment 2**	**Assignment 3**
# finished students	1202	845	724
# peer grades	3201	2261	2084
# staff grades	23	19	23
Full score	21	25	25
Mean score	14.8 (70%)	17.2 (69%)	16.5 (58%)

Table 2: Summary statistics of assignments for peer grading

Results for Cardinal Experiments

Table 3, Table 4, and Figure 2 show experimental results for cardinal evaluation on the three assignments with the staff grades acting as ground truth for the instructor-graded submissions. From the probabilistic models, we can see that the precision parameter of the Gaussian distribution for the observed score is determined by the grader's reliability. So in PG_4, we mainly tune the rate parameter β_0 which is the rate parameter in the gamma distribution for the grader's reliability. For PG_5, we mainly tune λ in the density function for the observed score. During experiments, we found that PG_3 is more sensitive to hyperparameters due to their linearity assumption, compared with β_0 in PG_4 and λ in PG_5. Moreover, the hyperparameters η_0 and γ_0 in PG_4 and PG_5, which control the variance of the grader's bias and that of the true score respectively, need to be tuned slightly. During the parameter tuning process, we try multiple combinations of η_0 and γ_0 in the range $[0.04, 0.2]$. For each combination, we perform line search within a certain range on β_0 for PG_4 and on λ for PG_5. The search range is $[100, 600]$ with a fixed interval of 50. For completeness, sensitivity analysis of the hyperparameters is provided in the appendix.

Average-Case Performance Here we use the root-mean-square error (RMSE) as an average measure for the difference between the score predicted by a model and that assigned by the instructor. Table 3 reports the best results, averaged over 10 runs, among hyperparameter settings in the range considered. Both the mean and standard deviation over 10 runs are shown for each assignment. We can see that PG_3, PG_4, and PG_5 have lower RMSE than the Median baseline and PG_1, showing that coupling a grader's reliability and true score is effective. Moreover, compared with PG_3, PG_4 and PG_5 can achieve quite significant RMSE improvement for assignments 2 and 3 although they are slightly worse than PG_3 for assignment 1. On average, PG_5 gives the best performance with 33% improvement for assignments 1 and 3 and 15% improvement for assignment 2 over the Median baseline which is used by Coursera as the default scheme. Relatively speaking, PG_4 performs slightly worse than PG_5 but it still outperforms PG_3 on average. Although PG_3 performs the best for assignment 1, its performance for assignments 2 and 3 is much worse than PG_4 and PG_5. We believe the probabilistic dependency between a grader's reliability and true score as adopted by PG_4 and PG_5 plays a crucial role in achieving the performance boost.

	Assignment 1		**Assignment 2**		**Assignment 3**	
	Mean	*Std*	*Mean*	*Std*	*Mean*	*Std*
Median	4.94		5.54		4.12	
PG_1	3.77 (23%)	0.02	4.93 (11%)	0.03	3.66 (11%)	0.01
PG_3	**3.22 (35%)**	0.02	5.24 (5%)	0.04	3.15 (23%)	0.02
PG_4	3.35 (32%)	0.05	4.75 (14%)	0.06	2.83 (31%)	0.09
PG_5	3.31 (33%)	0.05	**4.69** (15%)	0.05	**2.76 (33%)**	0.09

Table 3: Experimental results for cardinal models. Median represents taking the medium of the peer grades. PG_1 and PG_3 are models proposed in (Piech et al. 2013) and PG_4 and PG_5 are our models described above.

Worst-Case Performance and Model Sensitivity We also assess the worst-case performance of each method by measuring the maximum prediction deviation from the instructor's score. Table 4 shows that PG_4 and PG_5 have lower worst-case prediction errors than PG_3 for all three assignments, with PG_5 being the better. Moreover, from Figure 2 in which the submissions are sorted according to the instructor's grades, we can see that the predicted scores of PG_4 and PG_5 fluctuate less from the instructor's grades compared with PG_3. For some submissions with similar instructor's grades, the predicted scores by PG_3 differ quite significantly. For example, the 7th and 8th submissions of assignment 1 have the same score from the instructor but the predicted scores by PG_3 differ by around 8 points. Too much fluctuation is undesirable and can flag concerns about grading fairness as the ranking of the submissions according to the scores will be affected more seriously.

	Assignment 1	**Assignment 2**	**Assignment 3**
PG_3	6.52	11.10	6.77
PG_4	5.84	9.86	6.70
PG_5	**5.81**	**9.85**	**5.79**

Table 4: Maximum prediction deviation from the ground truth for the optimal settings in Table 3.

Although Table 3 shows that PG_3 performs the best for assignment 1 under an average-case performance measure,

considering also its worst-case performance and the fairness concern does not favor PG_3 as the preferred choice. In PG_3, the latent true score of a student is inferred directly from both the observed score when she plays the role of a gradee and the observed scores when she plays the role of a grader. However, in PG_4 and PG_5, the observed scores by a grader only infer her reliability, which then indirectly infers her true score. Consequently, the student's score in PG_4 and PG_5 is less sensitive to her own performance as a grader than PG_3. Considering various factors related to both average-case and worst-case performance assessment, it is fair to conclude that PG_4 and PG_5 outperform PG_3 under cardinal evaluation.

Ordinal Peer Grading

We now study the other aspect of peer grading, ordinal peer grading, and propose a novel approach to combining cardinal and ordinal models for enhanced performance. An ordinal peer grading task could be formulated as a rank aggregation problem (Dwork et al. 2001) or a preference learning problem (Chu and Ghahramani 2005; Fürnkranz and Hüllermeier 2010). In this paper, we consider a specific form of ordinal peer grading in terms of pairwise preferences (Fürnkranz and Hüllermeier 2003). In pairwise preference learning, a classical model is the Bradley-Terry model (Bradley and Terry 1952) which was recently applied to peer grading (Shah et al. 2013; Raman and Joachims 2014). Specifically, pairwise ordinal peer grading takes as training data examples with partial and possibly inconsistent pairwise preferences. The learning task is to predict from the data a total order or ranking of all the submissions.

Our proposition is to augment ordinal models, e.g., the Bradley-Terry model, with prior information from predictions of cardinal models studied in the previous section. For the sake of referencing, we refer to this extension as "Cardinal + Ordinal" models in the sequel. To gain a deeper understanding of this approach, we consider several combinations of cardinal and ordinal models in our experiments.

Combining Cardinal and Ordinal Models

We first briefly review the Bradley-Terry model which is referred to as BTL in (Shah et al. 2013) and BT in (Raman and Joachims 2014). In the ordinal setting, $u_i \succ_{\rho^{(v)}} u_j$ indicates that grader v prefers u_i over u_j. The probability of observing $u_i \succ_{\rho^{(v)}} u_j$ is defined using a logistic function with parameter $s_{u_i} - s_{u_j}$, where s_{u_i} and s_{u_j} are the latent true scores of submissions by u_i and u_j respectively:

$$\text{hypothesis} = P(u_i \succ_{\rho^{(v)}} u_j) = \frac{1}{1+\exp(-(u_i - u_j))}$$

Besides using the Bradley-Terry model, some recent attempts such as RBTL (Shah et al. 2013) and BT+G (Raman and Joachims 2014) incorporate the grader's reliability to boost the performance of ordinal peer grading. The rationale for incorporating the grader's reliability in RBTL is similar to PG_3 while that for BT+G is similar to PG_1. If we assume a Gaussian prior on the latent true scores of all students with mean μ and variance σ^2, the cost function of the Bradley-Terry model is defined as follows:

$$\mathcal{L} = \frac{\lambda}{2\sigma^2}\sum_{u\in U}(s_u - \mu)^2 - \sum_{v\in V}\sum_{u_i \succ_{\rho^{(v)}} u_j} \log(\text{hypothesis})$$

where the first term is a regularization term to avoid overfitting and the second term is the data likelihood term. Since the cost function above is jointly convex with respect to all the latent variables, stochastic gradient descent (SGD) can be used for solving the optimization problem.

Due to the data sparsity issue as discussed before, maximizing the data likelihood alone may lead to inconsistent and bad estimations. The regularization term defined based on the prior latent score distribution thus plays an important role. The original ordinal peer grading models (Shah et al. 2013; Raman and Joachims 2014) assume that all students have the same prior score distribution. Here, we propose to augment the ordinal models with the prior given by the predictions of the cardinal models. Alternatively, we may also interpret it as tuning the predictions of the cardinal models with the ordinal peer preferences. So the parameter μ in previous cost function is changed to the predicted score μ_u for submission u obtained from a cardinal model, with the modified cost function given by:

$$\mathcal{L} = \frac{\lambda}{2\sigma^2}\sum_{u\in U}(s_u - \mu_u)^2 - \sum_{v\in V}\sum_{s_{u_i} \succ_{\rho^{(v)}} s_{u_j}} \log(\text{hypothesis})$$

This formulation offers a principled approach to combining both cardinal and ordinal evaluations.

Results for Ordinal Experiments

We first generate pairwise preferences from the cardinal evaluations of each peer grader to form the training data. Ties are excluded as they indicate lack of preference. Our underlying assumption is that the absolute evaluations of a grader on a set of assignments also reflect her ordinal preferences. Table 5 shows the results based on an accuracy measure that reflects the percentage of correctly evaluated pairs, which is similar to Kendall's tau rank correlation coefficient (Litchfield Jr and Wilcoxon 1955), except that we exclude ties and discordant pairs from consideration. For example, if there are 100 pairwise preferences based on the instructor's grading (with ties ignored) and a model correctly predicts 75 of them, then its accuracy is 0.75. The results in Table 5 are for the best hyperparameter settings averaged over 10 runs. For the "Cardinal + Ordinal" models, the best results also consider different cardinal priors corresponding to different hyperparameter settings explored by the cardinal models. Since preliminary investigation showed that the accuracy during testing converged fast, so we only report the average accuracy at iteration 30 for simplicity. We also report the results in Table 6 using the evaluation metric in Table 3 for cardinal evaluation.

We first note that two minor results about the pure ordinal models are consistent with those reported previously (Shah et al. 2013; Raman and Joachims 2014). First, the pure ordinal models have accuracy comparable with the Median baseline even though converting numerical scores to pairwise preferences incurs information loss. Second, modeling grader reliability in pure ordinal models (RBTL, BT+G) can slightly improve the performance of BT. Besides, we have found that simply maximizing the data likelihood as in BT Same Initial or BT Random Initial does not perform well,

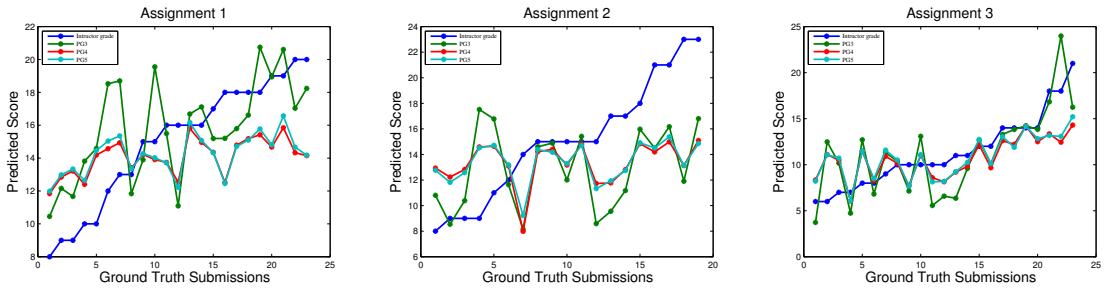

Figure 2: For the results in Table 3, the figures above show the exact predicted scores of PG_3, PG_4 and PG_5 compared with the instructor grades for the optimal settings reported in Table 3, with instructor grades sorted in ascending order.

showing that peer preferences alone have drawbacks in preference prediction.

On the other hand, our extensive experiments show two major results:

- Cardinal models can perform much better than pure ordinal models even for ordinal evaluation;
- Combining cardinal and ordinal models can further boost the accuracy in both cardinal and ordinal evaluations.

From results in Table 5, cardinal models outperform pure ordinal ones because they use fine-grained numerical information rather than binary comparisons (better or worse) in the ordinal setting. For example, with absolute evaluations, we can interpret the numerical scores as degree of preferences, and we could also model grader's bias from numerical evaluations. With the help of prior from cardinal models, combining cardinal and ordinal models ("Cardinal + Ordinal") gives the best ordinal accuracy for all three assignments among all the models. As it may be interpreted as tuning the cardinal predictions as well, we also evaluate the combined models in a cardinal way in Table 6, which shows consistent cardinal performance boost against all corresponding pure cardinal models for all three assignments in Table 3. For now, the ordinal comparisons are generated from cardinal evaluations. If ordinal data are generated directly, combining cardinal and ordinal models may be used to further tune and reinforce each other so that further improvement may be possible.

From another perspective, the technique to combine cardinal and ordinal models may be used to alleviate the data sparsity problem in peer grading as well. In other words, with cardinal evaluations, we can generate ordinal comparisons and combine the models using two types of input to boost accuracy. If we are forced to make only one choice, our recommendation would be a cardinal peer grading method which can obtain more accurate grade prediction as demonstrated above. In addition, from (Shah et al. 2013), ordinal models need more comparisons per grader to get a reasonable accuracy. For typical peer grading task with open-ended exercises of several hundred words, it is hard for peer graders to compare submissions in detail. At last, we maintain that it is easier for students to perform cardinal grading for typical peer grading tasks because the instructor's detailed cardinal rubrics can clarify students' thoughts, while quantizing the subjective and vague evaluation standard when making ordinal comparisons.

	Assignment 1	**Assignment 2**	**Assignment 3**
Cardinal Models			
PG_3	0.7526	0.6155	0.7775
PG_4	0.6928	0.6552	0.7854
PG_5	0.6979	0.6616	0.7889
"Cardinal + Ordinal" Models			
PG_3+BT	0.7577	0.6110	0.7892
PG_4+BT	0.7221	0.6484	0.7931
PG_5+BT	0.7191	0.6646	0.8000
PG_3+BT+G	0.7645	0.6587	0.7879
PG_4+BT+G	0.7145	0.7032	0.7896
PG_5+BT+G	0.7170	**0.7065**	**0.8013**
PG_3+RBTL	**0.7660**	0.6494	0.7979
PG_4+RBTL	0.7064	0.6745	0.7835
PG_5+RBTL	0.7201	0.6845	0.8009
Pure Ordinal Models			
BT (or BTL)	0.6536	0.6329	0.6896
RBTL	0.6583	0.6432	0.6996
BT+G	0.6547	0.6535	0.7009
BT Same Initial	0.6387	0.6194	0.6407
BT Random Initial	0.6381	0.6416	0.6667
Baseline Method			
Median	0.6043	0.6610	0.6753

Table 5: Ordinal evaluation results for different models. Cardinal models: described above but evaluated differently here. "Cardinal + Ordinal" models: ordinal models with predictions from different cardinal models as prior. Pure ordinal models: BTL and RBTL from (Shah et al. 2013), BT and BT+G from (Raman and Joachims 2014); BT Same Initial or BT Random Initial, similar to BT (or BTL), has no prior on the student scores for regularization with the same or random scores initially for all students. Median: baseline that takes the median of the peer grades.

Conclusion

In this paper, we have proposed a new probabilistic model for peer grading and a novel mechanism for combining cardinal and ordinal models. Extensive experiments on both cardinal and ordinal evaluations show promising results.

Acknowledgments

This research has been supported by *General Research Fund 621310* from the Research Grants Council of Hong Kong. We would also like to thank the instructor of the Coursera course, Naubahar Sharif, for letting us use the course data for our research and the Coursera support team in our uni-

	Assignment 1	Assignment 2	Assignment 3
PG_3+BT	3.04	5.30	3.18
PG_3+BT+G	3.01	**4.95**	**3.10**
PG_3+RBTL	**3.00**	5.04	3.15
PG_4+BT	3.47	4.87	3.03
PG_4+BT+G	**3.31**	**4.52**	2.91
PG_4+RBTL	3.44	4.70	**2.77**
PG_5+BT	3.30	4.77	2.93
PG_5+BT+G	3.35	**4.50**	2.74
PG_5+RBTL	**3.24**	4.62	**2.70**

Table 6: Cardinal evaluation (RMSE) results for "Cardinal+Ordinal" models

versity, particularly Tony Fung, for their help in accessing and processing the data.

References

Anderson, A.; Huttenlocher, D.; Kleinberg, J.; and Leskovec, J. 2014. Engaging with massive online courses. In *Proceedings of the 23rd International Conference on World Wide Web*, 687–698.

Arrow, K. J. 2012. *Social Choice and Individual Values*, volume 12. Yale University Press.

Aslam, J. A., and Montague, M. 2001. Models for metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 276–284.

Barnett, W. 2003. The modern theory of consumer behavior: Ordinal or cardinal? *The Quarterly Journal of Austrian Economics* 6(1):41–65.

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

Carterette, B.; Bennett, P. N.; Chickering, D. M.; and Dumais, S. T. 2008. Here or there. In *Advances in Information Retrieval*, 16–27.

Chu, W., and Ghahramani, Z. 2005. Preference learning with Gaussian processes. In *Proceedings of the 22nd International Conference on Machine learning*, 137–144.

Coughlin, P. J. 1992. *Probabilistic Voting Theory*. Cambridge University Press.

de Alfaro, L., and Shavlovsky, M. 2014. Crowdgrader: Crowdsourcing the evaluation of homework assignments. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, 415–420.

Dwork, C.; Kumar, R.; Naor, M.; and Sivakumar, D. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web*, 613–622.

Fürnkranz, J., and Hüllermeier, E. 2003. Pairwise preference learning and ranking. In *Proceedings of the 14th European Conference on Machine Learning*, 145–156.

Fürnkranz, J., and Hüllermeier, E. 2010. *Preference Learning*. Springer.

Godlee, F.; Jefferson, T.; Callaham, M.; Clarke, J.; Altman, D.; Bastian, H.; Bingham, C.; and Deeks, J. 2003. *Peer Review In Health Sciences*. BMJ Books London.

Harnad, S. 2000. The invisible hand of peer review. *Exploit Interactive* 5(April).

Howe, J. 2006. The rise of crowdsourcing. *Wired Magazine* 14(6):1–4.

Joachims, T. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 133–142.

Litchfield Jr, J. T., and Wilcoxon, F. 1955. Rank correlation method. *Analytical Chemistry* 27(2):299–300.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab.

Paré, D. E., and Joordens, S. 2008. Peering into large lectures: examining peer and expert mark agreement using peerScholar, an online peer assessment tool. *Journal of Computer Assisted Learning* 24(6):526–540.

Patricia, G.; Nardine, O.; and Carles, S. 2014. Collaborative assessment. In *Extended Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014)*.

Peters, D. P., and Ceci, S. J. 1982. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences* 5(02):187–195.

Piech, C.; Huang, J.; Chen, Z.; Do, C.; Ng, A.; and Koller, D. 2013. Tuned models of peer assessment in MOOCs. In *Proceedings for the 6th International Conference on Educational Data Mining.*, 153–160.

Raman, K., and Joachims, T. 2014. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*.

Rowland, F. 2002. The peer-review process. *Learned Publishing* 15(4):247–258.

Sadler, P. M., and Good, E. 2006. The impact of self- and peer-grading on student learning. *Educational Assessment* 11(1):1–31.

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

Shah, N. B.; Balakrishnan, S.; Bradley, J.; Parekh, A.; Ramchandran, K.; and Wainwright, M. 2014. When is it better to compare than to score? *arXiv preprint arXiv:1406.6618*.

Stewart, N.; Brown, G. D.; and Chater, N. 2005. Absolute identification by relative judgment. *Psychological Review* 112(4):881–911.

Surowiecki, J. 2005. *The Wisdom of Crowds*. Random House LLC.

Walsh, T. 2014. The peerrank method for peer assessment. In *Proceedings of the 21th European Conference on Artificial Intelligence*.