BayesRank: A Bayesian Approach to Ranked Peer Grading

Andrew E. Waters[1], **David Tinapple**[2], **and Richard G. Baraniuk**[1]
[1] Rice University, Houston, TX
[2] Arizona State University, Tempe, AZ
[1] Houston, TX 77005, [2] Tempe, AZ 85281
{andrew.e.waters@rice.edu, dtinapple@asu.edu, richb@rice.edu}

ABSTRACT

Advances in online and computer supported education afford exciting opportunities to revolutionize the classroom, while also presenting a number of new challenges not faced in traditional educational settings. Foremost among these challenges is the problem of accurately and efficiently evaluating learner work as the class size grows, which is directly related to the larger goal of providing quality, timely, and actionable formative feedback. Recently there has been a surge in interest in using peer grading methods coupled with machine learning to accurately and fairly evaluate learner work while alleviating the instructor bottleneck and grading overload. Prior work in peer grading almost exclusively focuses on numerically scored grades – either real-valued or ordinal. In this work, we consider the implications of *peer ranking* in which learners rank a small subset of peer work from strongest to weakest, and propose new types of computational analyses that can be applied to this ranking data. We adopt a Bayesian approach to the ranked peer grading problem and develop a novel model and method for utilizing ranked peer-grading data. We additionally develop a novel procedure for adaptively identifying which work should be ranked by particular peers in order to dynamically resolve ambiguity in the data and rapidly resolve a clearer picture of learner performance. We showcase our results on both synthetic and several real-world educational datasets.

Author Keywords

Automatic grading; peer grading; Bayesian methods; rank aggregation; adaptive recommender systems

INTRODUCTION

In the shifting landscape of higher education we are seeing advances in online and hybrid forms of teaching and learning. From the flipped classroom to the MOOC, these new formats present opportunities to provide not only high quality online courses to larger and larger numbers of students, but also more engaging classroom experiences for 21st century learners.

A common challenge in these experimental formats is how to meaningfully evaluate student work. The traditional method of having the instructor evaluate and grade or score each work is increasingly untenable. In the case of the large classroom or the online MOOC, having an instructor or team grade by hand creates an enormous workload bottleneck, leading to delayed and/or inadequate feedback. Even when the grading workload can be distributed among teams of assistants, the central experience for the student is the same – a single centralized expert makes an often rapid and subjective judgement on the work, returning a grade and, at best, some qualitative feedback.

Two common approaches to overcoming the traditional grading bottleneck are machine automation and peer grading. In machine automation, software is used, most typically to determine right and wrong answers on a test. Often these methods focus on easy-to-grade question formats such as multiple choice or true/false. Increasingly, these techniques rely on natural language processing [20] to assess or estimate the quality of essays or answers. In peer grading, by contrast, the assessment is ”crowdsourced” to the group itself, usually by providing a rubric or scoring guide to students and asking them to evaluate one or more peer works, giving one or more ratings to the work. In addition to being useful in efficiently evaluating student work, it has been in several educational contexts the act of *giving* peer feedback is at least as beneficial to the giver to the receiver of the feedback [12].

While many peer review systems allow for direct written feedback from peers, they often utilize a rating system as the primary mode of evaluation. Even with a clearly communicated rubric or guide provided to raters, the ratings can vary greatly from reviewer to reviewer creating significant noise in the overall rating data. One way to combat the problem of rater reliability is to calibrate the raters themselves using a carefully created demonstration assignment. Raters responses to this example content are then used to evaluate the quality of the raters themselves. While this form of calibrated peer review (CPR) has been shown to be effective [10], it does require the creation of sample assignments and detailed multi-question rubrics for each new type of assignment. In a larger sense, CPR assumes that a students conformity to the right answers provided by the instructors determines their value as reviewers. Many kinds of creative work, however, can be viewed and reviewed from multiple valid viewpoints, which may serve as the primary value of using peer review in the classroom.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

L@S'15, March 14–15, 2015, Vancouver, BC, Canada.
Copyright © 2015 ACM 978-1-4503-3411-2/15/03...$15.00.
http://dx.doi.org/10.1145/2724660.2724672

An alternative to traditional peer grading in which graders rate items on a numerical [14] or ordinal scales [8] is to use *ranked* peer grading [18]. In this approach, graders rank the quality of a small number of items in order from best to worst. While reviewers are often less comfortable providing a ranked assessments of items [9, 5], ranked approaches have been shown to often yield significantly more reliable data with much higher discriminative power than rating data [2]. Given a set of potentially partially observed ranked data from a set of graders on a set of items, the goal of the ranked peer grading problem is to assess the quality of each item relative to all others.

Prior Work

There has been considerable prior work in the field of rank aggregation, an excellent summary of which is available in [16]. These methods include the Mallows method (MAL [13]) which places a probabilistic distribution over a set of observed rankings given a ground truth ranking. It can be shown that using the Kendall's-τ distance metric that computation can be carried out quickly for the MAL method, allowing one to find a maximum likelihood estimate of the global rankings from a set of partially observed rankings. An augmentation of the MAL method is the score-based Mallows method (MALS, [16]) which estimates pairwise distances between item to improve the overall robustness of the inference. Another important method is the Plackett-Luce (PL) model introduced in [15], which is in turn a generalization of the Bradley-Terry method [3]. This model has the advantage of being convex and can be solved quickly using optimization techniques.

There are three important limitations in the prior rank aggregation and ranked peer grading literature. First, these methods rely on maximum likelihood estimation and output a singular point estimate of the true ranking of all items. While this is useful, it is generally difficult to assess the *reliability* of the estimate provided. This reliability information can be extremely useful to course instructors when assigning grades or to learners who wish to understand the quality of their work relative the rest of the class. Second, these methods do not explicitly model the reliability of the graders themselves, but rather focus exclusively on finding a ranking of the items that makes sense given the ranked data. Not all graders in a course are equally reliable, and by modeling grader reliability explicitly one can improve the overall inference quality while additionally relaying this reliability information to course instructors. Third, these methods typically assume a random assignment of items to graders which, especially in the case of ranked peer grading, fails to take capitalize on the ability of graders to distinguish between closely related items. This, in turn, leads to suboptimal inference of item quality.

Contributions

In this paper we develop the *BayesRank* model and method for overcoming the limitations of prior work in rank aggregation and ranked peer grading. Concretely, we make the following three contributions:

- We develop a novel Bayesian model for partially observed ranked data. This Bayesian model has the advantage of explicitly modeling reliability of each grader in addition to the intrinsic quality of each item.

- We develop a novel Markov-Chain Monte Carlo (MCMC) method for fitting the BayesRank model to partially observed ranking data. This MCMC method has the advantage of providing not only inference for all model parameters of interest but also reliability estimates for those parameters.

- We develop a novel procedure for adaptively selecting a set of items to present to a grader in order to reduce our uncertainty about item quality. We will demonstrate that this method can provide great improvements over the standard practice of assigning items to graders randomly.

THE BAYESRANK MODEL FOR RANKED PEER GRADING

We now detail the BayesRank generative model for partially observed ranked data. We assume throughout this work that we are given ranked peer grading data consisting of I items being ranked by G graders. We further assume that each grader only ranks a small subset $K \ll I$ items. Let $\Omega^g \in \{1, \ldots, I\}^K$ denote the indices of the items ranked by grader g. We are ultimately interested in determining the *quality* of each item as well as the *reliability* of each grader in a given dataset.

We model each item $i \in \{1, \ldots, I\}$ item as possessing an underlying latent quality score $s_i \in \mathbb{R}$, where items that have higher scores are said to have higher quality. We model grader g as observing a noisy version of the true score $s_i, \forall i \in \Omega^g$. Let $\mathbf{z}^g$ denote a random vector of observed item qualities observed by grader g, with z_i^g denoting the observation for item i. We model $z_i^g \sim \mathcal{N}(s_i, \sigma_g^2)$, where $\mathcal{N}(s_i, \sigma_g^2)$ denotes the standard normal distribution with mean s_i and grader-specific noise variance σ_g^2 that determines how accurately grader g observes the true score s_i. Graders with lower σ_g^2 are said to be more reliable than graders with higher σ_g^2. Grader g then returns a ranked set of indices according to the observations z_i^g. Let $\mathcal{T}_{(j)}[\mathbf{z}]$ denote the index of the j^{th} largest index of $\mathbf{z}$. Then, grader g returns the rankings $\mathbf{r}^g = [r_1^g, r_2^g, \ldots, r_K^g] = \left[\mathcal{T}_{(1)}[\mathbf{z}^g], \mathcal{T}_{(2)}[\mathbf{z}^g], \ldots, \mathcal{T}_{(K)}[\mathbf{z}^g]\right]$. As we take a Bayesian approach to our model, all that remains is to specify appropriate prior distributions for all parameters of interest. For this we choose standard prior distributions typically used in Bayesian literature. We formally write our model as:

$$\begin{aligned} s_i &\sim \mathcal{N}(0, \sigma_I^2), \\ \Omega^g &\propto 1, \\ \sigma_g^2 &\sim \mathcal{IG}(\alpha, \beta), \\ z_i^g &\sim \mathcal{N}(s_i, \sigma_g^2), \\ \mathbf{r}^g &= \left[\mathcal{T}_{(1)}[\mathbf{z}^g], \mathcal{T}_{(2)}[\mathbf{z}^g], \ldots, \mathcal{T}_{(K)}[\mathbf{z}^g]\right] \end{aligned} \tag{1}$$

where $\mathcal{IG}(\alpha, \beta)$ denotes the inverse-gamma distribution with shape parameter α and scale parameter β, and where σ_I^2, α, and β are tunable hyperparameters. We dub our generative

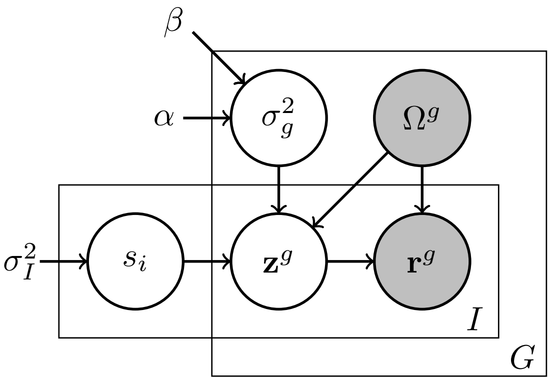


Figure 1: Graphical model for the generative BayesRank model

model BayesRank and display this model graphically in Figure 1.

We note that the BayesRank model is not strictly identifiable. As a concrete example, a set of scores $\{s_i\}$ and $\{s_i + \delta\}$ will achieve the same data likelihood. Similarly we can inversely scale the grader variance and latent score variance by any constant $a > 0$ while maintaining the same data likelihood. Consistent with prior art in Bayesian literature, we rely on our choice of prior and hyperparameters to circumvent these issues [8].

INFERENCE METHOD

Given a set of observations $\{r_k^g\}$ for $k = 1, \ldots, K$ and $g = 1, \ldots, G$, we wish to infer the latent parameters of the BayesRank model (1). There are many methods available for doing this, including expectation-maximization and variational Bayesian approaches. In this work we will use a MCMC technique [6] which is simple to implement, efficient, and provides rich posterior information for all of the model parameters of interest. This posterior information provides a wealth of statistical information unavailable under many other methods and enable us to assert not only an estimate for each model parameter but also the degree of confidence that we can assert for these parameter estimates. This posterior information can be used by course instructors to evaluate the strength of a given item or the reliability of a particular grader relative to other items and graders in the course.

Our MCMC approach is based on the Gibbs sampler [7] and estimates the posterior distributions of s_i and σ_g^2 for $i = 1 \ldots I$ and $g = 1 \ldots G$. We can considerably simplify our inference by using data augmentation [1] and additionally sampling the noisy latent observations z_i^g. Our method proceeds by sequentially sampling each random variable of interest conditioned on all other latent variables in the model. It can be shown that for our method these distributions are given by:

$$z_{r_k}^g|\cdot \sim \mathcal{N}^+(s_{r_k}, \sigma_u^2; z_{r_k-1}^g, z_{r_k+1}^g), (k,g) \in \Omega^g,$$
$$s_i|\cdot \sim \mathcal{N}(\widehat{\mu}, \widehat{\sigma}^2), i \in \{1, \ldots, I\}$$
$$\sigma_g^2|\cdot \sim \mathcal{IG}(\widehat{\alpha}, \widehat{\beta}), g \in \{1, \ldots, G\}$$

where $\widehat{\sigma}^2 = 1/\sigma_I^2 + \sum_{g:i\in\Omega^g} \in (1/\sigma_g^2)$, $\widehat{\mu} = \sum_{g:i\in\Omega^g} \frac{z_i^g}{\sigma_g^2}$, $\widehat{\alpha} = \alpha + K$, and $\widehat{\beta} = \beta + \sum_{i:i\in\Omega^g} \frac{1}{z_i^g}$. The notation $\mathcal{N}^+(\mu, \sigma^2; a, b)$ denotes a normal distribution with mean μ and variance σ^2 truncated below at a and truncated above at b. This distribution results from the ordering constraint imposed by the observations $\mathbf{r}^g$. We adopt the convention that $z_0 = -\infty$ and $z_{K+1} = \infty$, consistent with our initial modeling assumption $s_i \in \mathbb{R}$.

POST PROCESSING

After sufficient burn-in, the MCMC produces samples drawn approximately from the true posterior distribution of each parameter of interest conditioned on the data in $\{r_k^g\}$. We can obtain simple estimates of the latent score quality $\widehat{s_i}$ and grader reliability $\widehat{\sigma}_g$ via simple posterior averaging. Additionally, we can produce a global ranking of the items at each iteration of the MCMC which can be used to evaluate the rank stability of each item relative to all other items in the dataset.

ADAPTIVELY SELECTING ITEMS TO RANK

In formulating the BayesRank model in (1) we assumed explicitly that graders were assigned items uniformly and independently, consistent with the implementation of most peer grading systems. However, we can hope to do better than a purely random assignment strategy in practice. First, some learners in a class may choose not complete their ranked grading assignment. In this scenario, there may be a variable number of rankings for each item, which can lead to the scoring of certain items being more reliable than others. Second, a ranking of a randomly selected set of items may provide very little information to a rank aggregation algorithm. As a concrete example, the quality difference between the best and worst item in a dataset may be quite stark. Repeatedly asking graders to compare these two items yields little useful information since there will be little to no diversity in the rankings provided by multiple graders.

A more powerful approach than random assignment is to assign the items to graders adaptively. Such an adaptive assignment has the potential to capitalize on the ability of a grader to resolve ambiguous items. Furthermore, this approach is entirely feasible in many a practical grading scenarios. A specific example of such a scenario is when graders have some time frame (e.g., one week) to log into a computer-based system to receive and rank a set of items. In this scenario, we can use knowledge obtained from the last $g - 1$ graders to adaptively select items for grader g to rank.

Our selection strategy is based on selecting the items for grader g that maximize the entropy [4] of the ranking that grader g will provide. Given a set of K items we can have $K!$ potential rankings. Let $\mathcal{I}$ denote the current set of items and let $\mathbf{s}_{\mathcal{I}}$ denote the random variable of the latent quality scores over the items contained in $\mathcal{I}$. Now let $\mathbf{r}_{\mathcal{I}}^j$, $j \in \{1, \ldots, K!\}$ denote one of the possible ranking permutations over the set

$\mathcal{I}$. The entropy of the ranking is then given by

$$H(\mathbf{r}_{\mathcal{I}}|\mathbf{s}_{\mathcal{I}}) = \sum_{j=1}^{K!} p(\mathbf{r}_{\mathcal{I}}^{j}) \log \left(p(\mathbf{r}_{\mathcal{I}}^{j}) \right). \quad (2)$$

We then find the entropy-maximizing set $\widetilde{\mathcal{I}}$ by solving the following optimization problem

$$\widetilde{\mathcal{I}} = \arg \max_{\mathcal{I}, |\mathcal{I}|=K} \mathcal{H}(\mathbf{r}_{\mathcal{I}}|\mathbf{s}_{\mathcal{I}}). \quad (3)$$

This optimization problem, however, is intractable in most practical scenarios for two reasons. First, in the case of I items, we would be required to search over all $\binom{I}{K}$ sets. Second, evaluating the rank-entopy in (2) itself requires computation on a $K+1$-dimensional normal cumulative distribution function for which there is no closed form solution. Using full posterior information on $\mathbf{s}_{\mathcal{I}}$ renders this computation even more complicated.

We propose two simple approximations that enable us to find a tractable approximation to the optimization problem (3) given our MCMC inference method. First, we will use the posterior mean $\widehat{\mathbf{s}}$ rather than full posterior information when calculating the entropy. Thus, $H(\mathbf{r}_{\mathcal{I}}|\mathbf{s}_{\mathcal{I}}) \approx H(\mathbf{r}_{\mathcal{I}}|\widehat{\mathbf{s}}_{\mathcal{I}})$. Second, we note that it is simple to construct the entropy maximizing set $\mathcal{I}$ for each element item i by searching for the $K-1$ nearest neighbors of item i whose quality scores s_j are closest to s_i. Once this is computed for each item, we choose the item (and neighbors) the minimize the overall square distance. We display the steps of this procedure in Algorithm 1.

Algorithm 1: Finding the approximate rank-entropy maximizing set

```
Data: Item quality means $\widehat{s}_i, i = \{1, \ldots, I\}$, Set cardinality
      $K$
Result: Approximate rank-entropy maximizing set $\mathcal{I}$
Compute pairwise distance matrix $\mathbf{D}$, s.t. $D_{ij} = (\widehat{s}_i - \widehat{s}_j)^2$
for $i \leftarrow 1$ to $I$ do
    $\mathcal{I}'_i = [i, \mathcal{T}_{(2)}[D_i], \ldots, \mathcal{T}_{(K)}[D_i]]$;
    $d_i = \sum_{k=1}^{K} D_{i,\mathcal{T}_{(k)}[D_i]}$
end
$\widehat{i} = \arg\min d_i$;
$\mathcal{I} = \mathcal{I}_{\widehat{i}}$
```

For practical purposes, we would initialize our beliefs about the item scores to 0 for all items, which would initially cause our adaptively selection strategy to behave similarly to a randomized selection strategy. As ranked comparisons begin to be made, however, and the MCMC inference method begins to infer the quality scores for the individual items, the adaptive procedure of Algorithm 1 will begin to start assigning items to graders that it believes are close together and use the grader rankings to better resolve those items.

EXPERIMENTS

Here we characterize the BayesRank model and method using both synthetic and real-world data experiments.

Synthetic Data

We first examine the performance of BayesRank on synthetic data generated according to the model described in (1) under a variety of problem configurations. We will compare our results to the known ground truth item ordering using Kendall's-τ metric, which examines the general agreement between two ordered sets of items of size I. For any two orderings σ and $\widetilde{\sigma}$, Kendall's-τ is defined by:

$$D_\tau(\sigma, \widetilde{\sigma}) = \frac{1}{\frac{1}{2}I(I-1)} \sum_{i,j} \operatorname{sgn}(\sigma_i - \sigma_j) * \operatorname{sgn}(\widetilde{\sigma}_i - \widetilde{\sigma}_j),$$

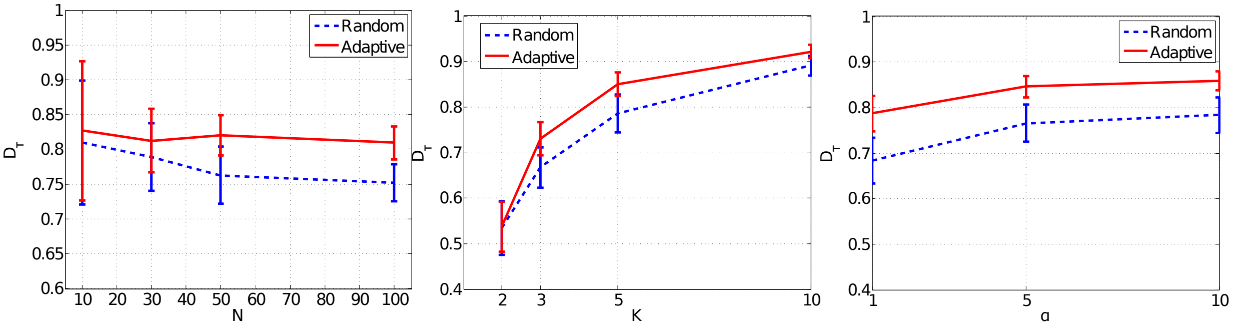
where $\operatorname{sgn}(\cdot)$ denotes the signum function. In words, Kendall's-τ looks at each pair in one ranking and compares the same items in the second ranking to check for consistency. We note that $-1 \leq D_\tau \leq 1$, with the case $D_\tau = 1$ corresponding to perfect agreement between σ and $\widetilde{\sigma}$, the case $D_\tau = -1$ corresponding to perfect disagreement between σ and $\widetilde{\sigma}$, and the case $D_\tau = 0$ corresponding to no correlation between σ and $\widetilde{\sigma}$.

In each synthetic experiment we will compare BayesRank using both adaptive and non-adaptive item assignment. Random assignment is done to ensure that all items receive the same number of rankings. For the adaptive scenario we employ an additional constraint that only the items with the fewest number of rankings be considered for assignment in order to keep the total number of rankings per item equal, consistent with both the randomized approach and what would be done in a real classroom scenario.

We first look at performance as the problem size increases. We assume here that $G = I = N$ and sweep $N \in \{10, 30, 50, 100\}$. We set $K = 5$, $\sigma_I = 10$, and $\alpha = \beta = 5$ in all experiments. We repeat each configuration over 50 randomized trials and plot the average performance with error bars in Figure 2(a). We see that both methods degrade gracefully as the class size grows, but note that the adaptive item assessment considerably mitigates this effect, achieving superior performance over all class sizes.

We next examine the performance of BayesRank as the number of peer rankings, K, varies. We assume a single problem size of $G = I = 50$ with $\sigma_I = 10$ and $\alpha = \beta = 5$ and sweep $K \in \{2, 3, 5, 10\}$. We again repeat each configuration over 50 randomized trials and display our results in 2(b). As expected, performance improves as the number of peer rankings grows. Importantly, we note that performance begins to level off at around $K = 5$, meaning that we can achieve good performance with a very reasonable number of rankings that is not burdensome to learners. As before, the adaptive item selection strategy outperforms the random selection strategy.

We further examine performance as we vary the grader reliability. We again assume a problem size of $G = I = 50$ with $K = 5$ ratings per grader and with $\sigma_I = 10$. We set $\alpha = \beta$ and sweep $\alpha \in \{1, 5, 10\}$. Note that all cases correspond to



(a) D_τ vs. N (b) D_τ vs. K (c) D_τ vs. α

Figure 2: BayesRank synthetic experiments using both random and adaptive item assignment. a) D_τ as a function of the class size N, b) D_τ as a function of the number of items K ranked by each grader, c) D_τ as a function of the grader reliability parameter α.

Table 1: D_τ for various rank aggregation techniques for all four class projects. Size is measured as $G \times I$. BayesRank wins across all datasets.

Dataset	Size	BayesRank	Rank Avg	MAL	MALS	PL
Proj. 1	47 × 56	**0.2971**	0.2468	0.2390	0.2948	0.2779
Proj. 2	57 × 60	**0.4463**	0.4328	0.4034	0.4362	0.4418
Proj. 3	48 × 60	**0.4373**	0.3243	0.3096	0.4011	0.4158
Proj. 4	46 × 53	**0.4441**	0.4122	0.3570	0.4412	0.4354

an expected grader variance of 1 but that we have less prior variance as α increases. Higher values of α correspond to the case of less variability in the graders. We repeat each configuration over 100 trials and display our results in 2(c). As expected, performance improves as grader reliability improves. The adaptive selection strategy again outperforms the random selection strategy.

Real-World Educational Data

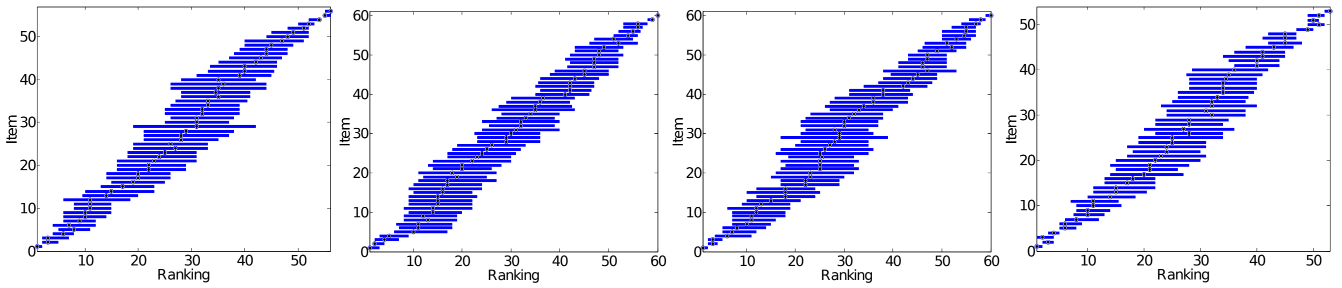
We now present the results of the analysis of a classroom of 60 learners in a computer programming class taught at Arizona State University. A part of this class consisted in learners completing four projects that were subsequently evaluated via ranked peer grading by other learners in the course. During the peer grading phase, each learner who completed their project was then asked to rank 5 other projects. Not all learners in the course completed every project, nor did all learners who completed their project complete their assigned peer grading task. Thus, for these datasets $G \neq I$ and not every item received the same number of rankings. The ranked peer grader data was collected using the CritViz peer review system [19].

In addition to the ranking scores provided by peer graders, the course instructor and assistants also assigned numerical quality scores to each project on a (non-integer) scale of 0—3. This additional numerical scores provide us a ground truth that can be used for comparison — by sorting the numerical grades in order we can compare the ranking estimated by BayesRank to the ground truth under the Kendall's-τ metric.

We compare BayesRank against four other rank aggregation models: Rank Averaging (in which the final score is simply the average of the ranks given to each item), MAL, MALS, and PL. Our results are displayed in Table 1, where we see that the BayesRank method achieves superior performance for all projects under consideration. To put these results in perspective, recall that the Kendall's-τ metric counts the number of ranking agreements between two sets of I items (a total of $I(I-1)$ comparisons). For many of our datasets, the global improvement is relatively small. As an example, the improvement in BayesRank over MALS for Project 1 is only 7 rankings. For Project 3, however, the improvement of BayesRank over MALS is much larger with 76 additional rankings matching the ground truth.

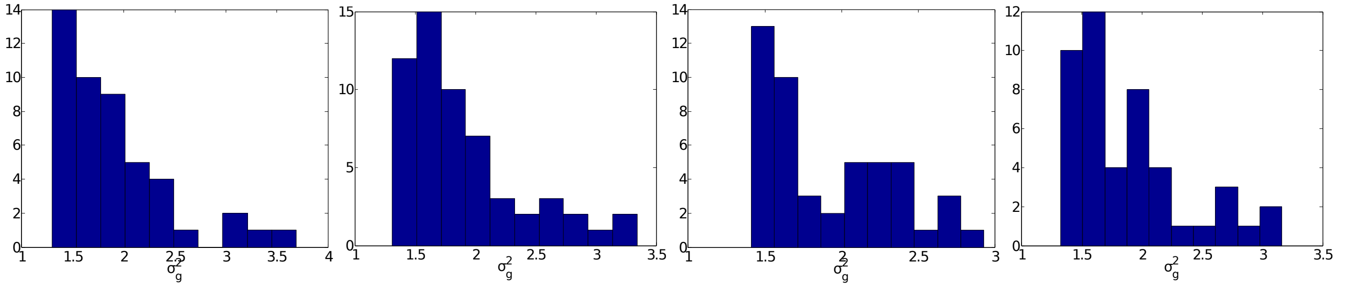
As discussed previously, our method enables us to make useful inference not only about the final quality of each project but also how reliably we can make such assertions. We compute this reliability information for each project by computing the ranking of each item at each iteration of the MCMC. We examine the statistics of these rankings in Figure 3. Here we have sorted the items in order by their average ranking (denoted by the black dot). We additionally show a 50% Bayesian credible interval as a blue bar showing which rankings are most likely for each item. We note that there is quite a bit of certainty for objects with very high and very low ranking, while there is considerable variation for item in the middle rankings.

Finally we show histograms for posterior mean grader reliability parameters σ_g^2 for all graders and each project in Figure 4. Here we find that most graders are highly reliable, with a small number being less reliable. As discussed previously, an advantage of the BayesRank approach is that the rankings



(a) Project 1 (b) Project 2 (c) Project 3 (d) Project 4

Figure 3: Item ranking credible intervals all four projects. Items are sorted in order of decreasing rank, with rank 1 corresponding to the highest quality project. The credible is small for high-tier and bottom-tier items, with larger variance for mid-tier items.



(a) Project 1 (b) Project 2 (c) Project 3 (d) Project 4

Figure 4: Histogram of average grader reliability σ_g^2 for each grader across all projects. Most graders are highly reliable, but a small number of graders are less reliable. The BayesRank model naturally uses this information to better assess the quality of all items.

from the less-reliable graders are naturally de-weighted when determining the latent item quality parameters.

CONCLUSIONS

Peer grading is a valuable tool for alleviating instructor workload in large courses while maintaining a fair standard of grading. We have developed the BayesRank model and method for ranked peer grading data which jointly infers the quality of student as well as the reliability of each peer grader. Additionally, we have developed a simple method for adaptively recommending items to graders that improves the overall reliability of the ranked peer grading method.

Our findings point toward several exciting directions for future work. First is determining the optimal number of items to be graded in a single ranking session. The optimal number is a practical question and must balance not only the needs of the model but also the cognitive limits of human graders. Ranking too many items might easily lead to sloppy ranking and, thus, bad data.

Next, we will pilot a study of an adaptive peer review assignment strategy in which during the review period, reviewers are assigned to look at specific clusters of work rather than randomized selections. Our work here points towards the possibility of greatly increasing the effectiveness of reviews by using reviewers to tease out the subtle differences between closely ranked works. In addition, we seek to understand how students would respond to such a system.

Similarly, the instructors can be asked to perform a small number of timely and carefully selected reviews that might serve as a real-time calibration, improving overall estimation not only by direct exposure to instructor review but by anchoring the entire network of review activity. Additionally, the model for item quality considered in this work is single dimensional with every item is modeled by a single quality parameter. Extensions to the multi-dimensional case [11, 17], where graders are asked to evaluate over multiple criteria, is a yet unexplored area. Finally, there may be useful ways to blend ranking and rating, fusing both ranked (comparing works) and numerically rated (isolated works) grading data, allowing for a balance between these two distinct cognitive activities.

Acknowledgments

The authors would like to thank Pavan Turaga and Mr. Lan for many helpful discussions regarding this work. This research was supported in part by NSF grant # 1431856.

REFERENCES

1. Albert, J. H., and Chib, S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*, 422 (1993), 669–679.
2. Bass, B. M., and Avolio, B. J. Potential biases in leadership measures: How prototypes, leniency, and general satisfaction relate to ratings and rankings of transformational and transactional leadership constructs. *Educational and Psychological Measurement 49*, 3 (1989), 509–527.
3. Bradley, R. A., and Terry, M. E. Rank analysis of incomplete block designs: I: The method of paired comparisons. *Biometrika* (1952), 324–345.
4. Cover, T. M., and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.
5. Elig, T. W., and Frieze, I. H. Measuring causal attibutions for success and failure. *Journal of Personality and Social Psychology 37*, 4 (1979), 621.
6. Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. *Bayesian Data Analysis*. CRC press, 2013.
7. Hoff, P. D. *A First Course in Bayesian Statistical Methods*. Springer, 2009.
8. Johnson, V. E., and Albert, J. H. *Ordinal Data Modeling*. Springer, 1999.
9. Krosnick, J. A. Survey research. *Annual review of psychology 50*, 1 (1999), 537–567.
10. Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., and Klemmer, S. R. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction (TOCHI) 20*, 6 (2013), 33.
11. Lan, A. S., Waters, A. E., Studer, C., and Baraniuk, R. G. Sparse factor analysis for learning and content analytics. *J. Machine Learning Research 15* (June 2014), 1959–2008.
12. Lundstrom, K., and Baker, W. To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing 18*, 1 (2009), 30–43.
13. Mallows, C. L. Non-null ranking models. i. *Biometrika* (1957), 114–130.
14. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. Tuned models of peer assessment in MOOCs. *arXiv preprint arXiv:1307.2579* (2013).
15. Plackett, R. L. The analysis of permutations. *Applied Statistics* (1975), 193–202.
16. Raman, K., and Joachims, T. Methods for ordinal peer grading. *arXiv preprint arXiv:1404.3656* (2014).
17. Reckase, M. D. *Multidimensional Item Response Theory*. Springer, 2009.
18. Shah, N. B., Bradley, J. K., Parekh, A., Wainwright, M., and Ramchandran, K. A case for ordinal peer-evaluation in MOOCs. In *NIPS Workshop on Data Driven Education* (2013).
19. Tinapple, D. *The CritViz Peer Review System*, 2014 (accessed October 21, 2014). `http://critviz.com/`.
20. Xiong, W., Litman, D., and Schunn, C. Improving research on and instructional quality of peer feedback through natural language processing. *Journal of Writing Research 4*, 2 (2012), 155–176.