Bayesian Ordinal Peer Grading

Karthik Raman
Cornell University
Ithaca, NY 14850
karthik@cs.cornell.edu

Thorsten Joachims
Cornell University
Ithaca, NY 14850
tj@cs.cornell.edu

Abstract

Massive Online Open Courses have become an accessible and affordable choice for education. This has led to new technical challenges for instructors such as student evaluation at scale. Recent work has found *ordinal peer grading*, where individual grader orderings are aggregated into an overall ordering of assignments, to be a viable alternate to traditional instructor/staff evaluation [21]. Existing techniques, which extend rank-aggregation methods, produce a single ordering as output. However, a single ordering alone may not be sufficient for instructors to confidently determine the final grades. For instance, instructors would like to have an estimate of the uncertainty of each assignment's grade as well. In this work, we tackle this problem by applying Bayesian techniques to the ordinal peer grading problem; in particular we use MCMC-based sampling techniques in conjunction with the Mallows model. Experiments are performed on real-world peer grading datasets along with an analysis of the quality of the learned posterior distributions.

1 Introduction

MOOCs (Massive Online Open Courses) offer the promise of affordable higher education, across a breadth of disciplines, for anyone with access to the Internet. The introduction of MOOCs has forced instructors to adapt conventional classroom logistics in order to scale to classrooms of 10,000+ students. One such key logistic is the *evaluation of students* in MOOCs. Given the orders of magnitude difference in scale, conventional assessment techniques such as instructor/staff-based grading are simply infeasible for MOOCs. On the other hand scalable automatic-grading schemes, such as multiple-choice questions, fall short of conventional testing standards as they are not a good measure of student learning [10, 11]. Furthermore they limit the kinds of courses offered; for instance, research-oriented classes require more open-ended testing such as essays and reports, which are very challenging to evaluate automatically.

Peer grading, where students — not instructors or staff — provide feedback on the work of other students in the class, has been proposed as a solution, since it naturally overcomes the problem of scale [9, 13]. Despite the inherent scalability of peer grading — the number of "graders" matches the number of students — a key obstacle for peer grading to work is the fact that the students are not trained graders and are just learning the material themselves. To ensure good-quality grades it is imperative that the grader feedback be simple and easy to provide. Recent work has proposed eliciting ordinal feedback from graders [21] (e.g. "project A is better than project B") rather than cardinal grades (e.g. "project A should get 87 out of 100"), since ordinal feedback has been shown to be easier to provide and more reliable than cardinal feedback [3, 22, 5] in several other settings.

This leads to the **ordinal peer grading problem**, where given the grader feedback (partial orderings over a subset of the assignments), the goal is to infer the overall ordering of all assignments. Rank-aggregation techniques have been extended to this task [21] and shown to not only be comparable to (if not better than) cardinal-grading based techniques but also traditional evaluation practices

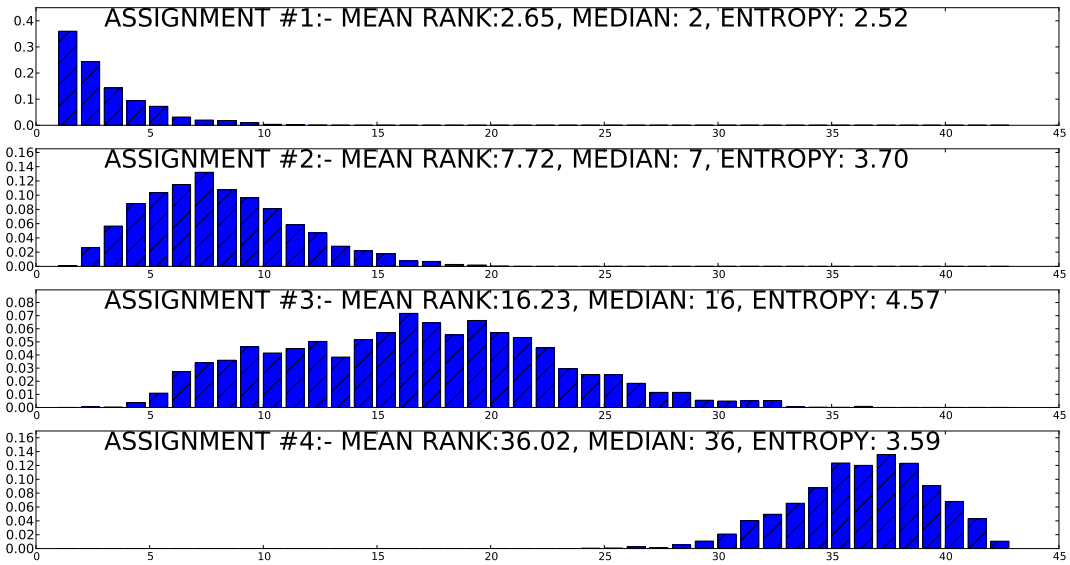

Figure 1: Having the peer grading algorithm produce more detailed information of each individual assignment's performance can be very useful for instructors when it comes to determining final grades. The above figure is one such example, where for each assignment the *posterior marginal distribution* (over position in the overall ranking) is shown (rank on x-axis, marginal probability on y-axis) along with statistics such as *posterior mean, median and entropy* of the marginal distribution.

such as course-staff (TAs) based grading. It is important to note than unlike other rank aggregation problems, peer grading requires accuracy throughout the ranking and not just at the top.

A more critical difference is the fact that unlike other rank aggregation problems, a single *aggregated* ordering alone is not always sufficient for instructors. In particular, instructors would like more information to determine the final grades for the assignments. A visual illustration of such fine-grained output produced by a peer-grading algorithm can be seen in Figure 1. Such information allows instructors to ascertain the algorithm's confidence in the grade (*i.e.,* percentile/position in ranking) of each assignment and discern the uncertainty of the underlying peer grades for each assignment. For instance, in the above example, while it is clear that assignment 1 is the best of the four assignments, it is not obvious that assignment 2 is better than assignment 3. This is because of the high uncertainty in the position of assignment 3 (as evidenced by its' high entropy of 4.57). If presented with such information, instructors could intervene and improve the final grades by having such uncertain assignments (assignments with maximum student disagreement) evaluated by staff-members. Thus while such information could be very helpful, current methods are unable to provide such information to instructors.

In this work, we look to address this problem, by employing Bayesian techniques for the ordinal peer grading problem. In particular we propose a Metropolis-Hastings [7] based Markov Chain Monte-Carlo (MCMC) method, for sampling from the posterior of a Mallows model [17]. The resulting samples allow us to empirically estimate the posterior grade distribution of each assignment, allowing us to report confidences and uncertainty information. The resulting uncertainty estimates can be used to improve the overall grading as described earlier.

We empirically study the efficacy of the proposed method on peer grading datasets, collected from a university-level class. In addition to studying the quality of the learner posterior orderings, we also analyze the resulting confidences and uncertainty information, both qualitatively and quantitatively.

2 Bayesian Methods for Ordinal Peer Grading (OPG)

In this section, we first describe the ordinal peer grading problem from a machine learning perspective. We then briefly review existing techniques for the OPG problem. Our proposed Bayesian

$G, g(\in G)$	Set of all graders, Specific grader
$D, d(\in D)$	Set of all assignments, Specific assignment
$D_g(\subset D)$	Set of items graded by grader g
$\sigma^{(g)}$	Ranking feedback (with possible ties) from g
$\eta_g(\in \Re^+)$	Predicted reliability of grader g
$r_d^{(\sigma)}$	Rank of assignment d in ordering σ (rank 1 is best)
$d_2 \succ_\sigma d_1$	d_2 is preferred/ranked higher than d_1 (in σ)
$\pi(A)$	Set of all rankings over $A \subseteq D$
$\sigma_1 \sim \sigma_2$	$\exists$ way of resolving ties in σ_2 to obtain σ_1
$\hat{\sigma}$	Estimated ordering of assignments
σ^*	(Latent) True ordering of assignments

Table 1: Notation overview and reference.

version of these techniques is then presented, followed by an empirical evaluation of these techniques in Section 3.

2.1 Ordinal Peer Grading Problem

In the *ordinal peer grading* problem, we are given a set of $|D|$ *assignments* $D = \{d_1, ..., d_{|D|}\}$ (*e.g.,* project reports, essays) which we need to grade. The grading is performed by a set of $|G|$ graders $G = \{g_1, ..., g_{|G|}\}$ (*e.g.,* student peer grader, reviewers). Each grader receives a subset of assignments $D_g \subset D$ to assess. The subsets D_g can be determined randomly, by a sequential mechanism or a deterministic policy. As feedback, each grader provides an ordering $\sigma^{(g)}$ (possibly with ties) of their assignments D_g.

The *primary goal* of OPG is *ordinal grade estimation* [21], *i.e.,* to produce an overall ordering [1] of the assignments $\hat{\sigma}$ using the individual grader orderings $\sigma^{(g)}$. While we would like this inferred ordering $\hat{\sigma}$ to accurately match some (latent) true ordering σ^*, we are faced with a couple of challenges. First, the individual grader orderings are only partial orderings, *i.e.,* the orderings only cover a small subset of the assignments ($|D_g| \ll |D|$). The second challenge is the fact that not all graders do an equally good job of grading, be it due to effort, skill or understanding of the material.

This leads to the secondary goal of *grader reliability estimation*, where we would like to estimate the accuracy/quality $\eta_g \in \Re^+$ of the feedback of each grader g. This allows us to improve the ordinal grade estimation quality by identifying unreliable graders and thus reduce the impact of their feedback on the estimated ordering $\hat{\sigma}$. Furthermore, it helps incentivize good and thorough grading by making peer grading itself part of the overall grade.

2.2 Relation to existing rank aggregation literature

The ordinal grade estimation problem in OPG can be viewed as a specific kind of rank aggregation problem. Rank aggregation [14] covers a class of problems where the goal is the combination of ordinal (ranking) information from multiple different sources. **Voting Systems** (or **Social Choice** [1]) are one of the most common applications of rank aggregation techniques. The goal of these systems is to merge the preferences of a set of individuals. Condorcet voting methods such as *Borda count* amongst others [8, 16] are commonly used to tackle these problems. **Search Result Aggregation** (also known as **Rank Fusion** or **Metasearch** [2]) is perhaps the most well-known rank-aggregation problem. Given rankings from different sources (typically different algorithms), the goal is to merge them and produce a single output ranking. Extensions of classical techniques such as the Mallows model [17] and Bradley-Terry model [4] have become popular for these problems [15, 6] and have been used to improve ranking performance in different settings [20, 23, 18].

While our work also extends the classical Mallows model, there are some fundamental differences to the these other rank aggregation problems, which make existing methods ill-suited for the OPG problem. First and foremost is the fact that while the success of search result aggregation and voting systems depend on correctly identifying the top item(s), in ordinal grade estimation it is imperative

[1]Producing an overall ordering of the assignments can be used to infer, for each assignment, a percentile rank as the grade (a common performance metric reported by standardized tests).

to accurately estimate the **full ranking**. In other words, we cannot afford to do any worse of a job identifying the 50^{th} percentile assignments than we do identifying the top assignments.

A second key difference (and the main focus of this work) is the fact that unlike other rank aggregation problems, **a single ordering** of assignments **may not suffice** for the purpose of determining grades. Before determining the final grades of assignments, instructors would like to have access to other information such as the uncertainty in the rank of an assignment. In other words, they would like to know more about the distribution of $r_d^{(\hat{\sigma})}$ (for instance a visualization such as Figure 1).

2.3 Existing Approaches to OPG

Different approaches [21] to the OPG problem include extensions of classical models such as the Mallows and Bradley-Terry model. We focus on the Mallows-based methods, as they form the basis for the techniques proposed in this work. In particular, the proposed Mallows-based peer grading model defines a distribution over rankings in terms of the **Kendall-Tau** distance [12] from the true ranking σ^* of assignments.

Definition 1 *The Kendall-τ Distance δ_K between rankings σ_1 and σ_2 is the number of incorrectly ordered pairs between the two rankings and is given by:*

$$\delta_K(\sigma_1, \sigma_2) = \sum_{d_1 \succ_{\sigma_1} d_2} \mathbb{I}[[d_2 \succ_{\sigma_2} d_1]]. \tag{1}$$

Given the grader orderings $\sigma^{(g)}$, we can define the data likelihood (if the overall ranking was σ) as:

$$P(\{\sigma^{(g)}; \forall g\}|\sigma) = \left\{ \prod_{g \in G} \frac{\sum_{\sigma' \sim \sigma^{(g)}} e^{-\delta_K(\sigma,\sigma')}}{Z_M(|D_g|)} \right\}. \tag{2}$$

where the normalization constant Z_M is easy to compute as it only depends on the ranking length.

$$Z_M(k) = \prod_{i=1}^{k} \left(1 + e^{-1} + \cdots + e^{-(i-1)}\right) = \prod_{i=1}^{k} \frac{1 - e^{-i}}{1 - e^{-1}} \tag{3}$$

Note that in Equation 2, **ties in the grader rankings** are modeled as *indifference* (*i.e.,* agnostic to either ranking), which leads to the summation in the numerator is over all total orderings σ' consistent with the weak ordering $\sigma^{(g)}$. While computing the MLE estimator of Equation 2 is NP-hard [8], a couple of simple and tractable approximations are presented in [21] that are shown to work well in practice.

While this model does not produce grader reliability estimates, an extension to the model is proposed in [21] and computed using a MAP estimator (rather than MLE estimator):

$$P(\{\sigma^{(g)}; \forall g\}|\sigma, \{\eta_g\}) = \left\{ \prod_{g \in G} \frac{\sum_{\sigma' \sim \sigma^{(g)}} e^{-\eta_g \delta_K(\sigma,\sigma')}}{Z_M(\eta_g, |D_g|)} \right\}.$$

However, both models (with and w/o reliability estimates) suffer from the same issue, in that they both produce **point estimates**, *i.e.,* a single ranking as output. In the next section, we will propose and study a Bayesian version of these models that estimates the posterior distribution of the true ranking and reliabilities.

2.4 Mallows MCMC using Metropolis-Hastings

To help provide more detailed information to instructors, we would like to have access to the posterior distribution of the orderings. In other words, instead of the data likelihood probability we have in Equation 2 (ignoring the grader reliabilities for now), we would like to know the posterior distribution of an estimated ranking σ *i.e.,* $P(\sigma|\{\sigma^{(g)}; \forall g\})$. We can safely assume a uniform prior on all orderings (for academic fairness), which gives us:

$$P(\sigma|\{\sigma^{(g)}; \forall g\}) = \frac{P(\{\sigma^{(g)}; \forall g\}|\sigma)P(\sigma)}{\sum_{\sigma' \in \pi(D)} P(\{\sigma^{(g)}; \forall g\}|\sigma')P(\sigma')} = \frac{P(\{\sigma^{(g)}; \forall g\}|\sigma)}{\sum_{\sigma' \in \pi(D)} P(\{\sigma^{(g)}; \forall g\}|\sigma')} \tag{4}$$

Algorithm 1 Sampling from Mallows Posterior using Metropolis-Hastings

1: **Input:** Grader orderings $\sigma^{(g)}$, Grader reliabilities η_g and MLE ordering $\hat{\sigma}$.
2: Pre-compute $x_{ij} \leftarrow \sum_{g \in G} \eta_g \mathbb{I}[d_i \succ_{\sigma^{(g)}} d_j] - \sum_{g \in G} \eta_g \mathbb{I}[d_j \succ_{\sigma^{(g)}} d_i[$
3: $\sigma_0 \leftarrow \hat{\sigma}$ ▷ Initialize Markov Chain using MLE estimate
4: **for** $t = 1 \ldots T$ **do**
5: Sample σ' from (**MALLOWS**) jumping distribution: $J_{MAL}(\sigma'|\sigma_{t-1})$
6: Compute ratio $r_t = \frac{P(\sigma'|\{\sigma^{(g)};\forall g\})}{P(\sigma_{t-1}|\{\sigma^{(g)};\forall g\})}$ using Equation 5
7: With probability $\min(r_t, 1)$, $\sigma_t \leftarrow \sigma'$ else $\sigma_t \leftarrow \sigma_{t-1}$
8: Add σ_t to samples (if burn-in and thinning conditions met)

However computing this posterior exactly is infeasible given the combinatorial number of possible orderings of all assignments. To help us ascertain information from the posterior, we will employ MCMC based sampling. Markov Chain Monte Carlo (or MCMC in short) are a set of techniques for sampling from a distribution by constructing a Markov Chain which converges to the desired distribution asymptotically. **Metropolis-Hastings** is a specific MCMC algorithm which is particularly common when the underlying distribution is difficult to sample from (as is the case here) especially for multi-variate distributions.

Thus to help us estimate properties of the posterior we will design a Markov Chain whose stationary distribution is the distribution of interest: $P(\sigma|\{\sigma^{(g)};\forall g\})$. Along with the theoretical guarantees accompanying these methods, an added advantage is the fact that we can control the desired estimation accuracy (by selecting the number of samples).

The resulting algorithm is a simple and efficient algorithm shown in Algorithm 1. To begin with we pre-compute statistics of the net cumulative weighted total each assignment d_i is ranked above another assignment d_j. We then initialize the Markov Chain using the MLE estimate of the ordering: $\hat{\sigma}$. At each timestep, to propose a new sample σ' given the previous sample σ_{t-1}, we sample from a jumping distribution (Line 5). In particular, we use a **Mallows**-based jumping distribution:
$\rightarrow \quad J_{MAL}(\sigma'|\sigma) \propto e^{-\delta_K(\sigma',\sigma)}$.
This is a simple distribution to sample from and can be done efficiently in $|D|log|D|$ time. Furthermore as this is a symmetric jumping distribution (*i.e.,* $J_{MAL}(\sigma'|\sigma) = J_{MAL}(\sigma|\sigma')$), the acceptance ratio computation is simplified.

When it comes to computing the (acceptance) ratio r_t (Line 6), we can rely on the pre-computed statistics to do so efficiently. In particular, we can simplify the expression for the ratio to:

$$\frac{P(\sigma_a|\{\sigma^{(g)};\forall g\})}{P(\sigma_b|\{\sigma^{(g)};\forall g\})} = \prod_{g \in G} e^{\delta_K(\sigma^{(g)},\sigma_b) - \delta_K(\sigma^{(g)},\sigma_a)} = \prod_{i,j} e^{x_{ij}(\mathbb{I}[d_i \succ_{\sigma_a} d_j] - \mathbb{I}[d_i \succ_{\sigma_b} d_j])} \tag{5}$$

This expression is again simple to compute and can be done in time proportional to the number of flipped pairs between σ_a and σ_b, which in the worst case is $O(|D|^2)$. Overall, the algorithm has a **worst-case time complexity** of $O(T|D|^2)$. The resulting samples produced by the algorithm can be used to *estimate* the posterior distributions including the marginal posterior of the rank of each assignment *i.e.,* $P(r_d|\{\sigma^{(g)};\forall g\}$, as well as statistics such as the entropy of the marginal, the posterior mean and median etc.

In order to improve the quality of the resulting estimates, we ensure proper mixing by targeting a moderate acceptance rate and by thinning samples (in our experiments we thin every 10 iterations). Furthermore we draw samples only one the chain has started converging *i.e.,* we use a burn-in of around 10,000 iterations.

We also derive a Metropolis-Hastings based extension of the Mallows model with grader reliabilities. In addition to sampling the orderings, we also sample the reliabilities using a Gaussian jumping distribution (also symmetric). However the acceptance ratio computation is now more involved and hence less efficient than that for Algorithm 1, but nonetheless can be computed fairly efficiently. We omit the precise equation and computations for the purpose of brevity.

Data Statistic	PO	FR
Number of Assignments	42	44
Number of Peer Reviewers	148	153
Total Peer Reviews	996	586

Table 2: Statistics for the two datasets (PO=Poster, FR=Report)

.

3 Experiments

In this section, we shall empirically evaluate the performance of the Bayesian Mallows-based peer grading method. In particular, we shall study a) the quality of the learned posterior, as measured with regards to conventional instructor grades; and b) the quality of the confidence intervals and uncertainty information.

3.1 Experimental Setting

We used the peer-grading datasets introduced in [21]. These datasets were collected in a real-classroom setting from a large university class. The class which consisted of about 170 students and 9 Teaching Assistants (TAs), used peer grading to evaluate the course projects (done in groups of 3-4 students). The advantage of this class size is the availability of conventional instructor and TA based grades for assignments, in addition to the peer grades (performed individually by each student). Having these instructor grades allows us to provide a more robust evaluation of the educational impact of these techniques, beyond what previous work has done.

We used both the **Poster (PO)** and **Final Report (FR)** datasets in this work. The two datasets correspond to different parts of the course. Students were incentivized to do a good job grading by incorporating their peer grading performance into their overall grade for the course. While the peer grading was done on a 10-point (cardinal) Likert scale (so as to compare cardinal and ordinal peer grading methods) in this work we simply use the implied orderings of the assignment. Table 2 summarizes some of the key statistics of the two datasets.

The Bayesian Mallows MCMC method was run with identical (fixed) parameters for both datasets. In total, 5000 sample orderings were drawn from the Markov Chain using Algorithm 1. These samples were used to estimate the posterior distributions and for obtaining the statistics in the following subsections.

3.2 How good are the posterior orderings learned

Although we are inferring a distribution over rankings, rather than a single point estimate, the hope is that the overall quality of these rankings is good. To verify this, we computed the *expected* Kendall-Tau error of the learned posterior. We also compute the expectation of a *weighted* version of the Kendall-Tau error where misordering items with a larger (instructor) score difference leads to a worse performance measure. We compared the following techniques:

- **MLE:** MLE Estimate of the Mallows model. This is a single point estimate and is used to initialize the Markov Chain.
- **Mode:** (One of the) Modes of the posterior of the Mallows distribution (as discovered during the sampling process). This is a single point estimate.
- **Bayes-MAL**: This is an estimate of the expected value of the τ_{KT} over the posterior learned by Alg 1.
- **Bayes-MAL+G**: This is an estimate of the expected value of the τ_{KT} over the posterior learned by the Metropolis-Hastings version of the Mallows model with grader reliability estimates.

The results are shown in Figure 2. Interestingly, we observe that the expected performance of the Bayesian posteriors, on both the Kendall-Tau error and the weighted variant, is nearly as good (if not better) than the MLE and MODE estimates for the **POSTER** dataset. Even on the **REPORT**

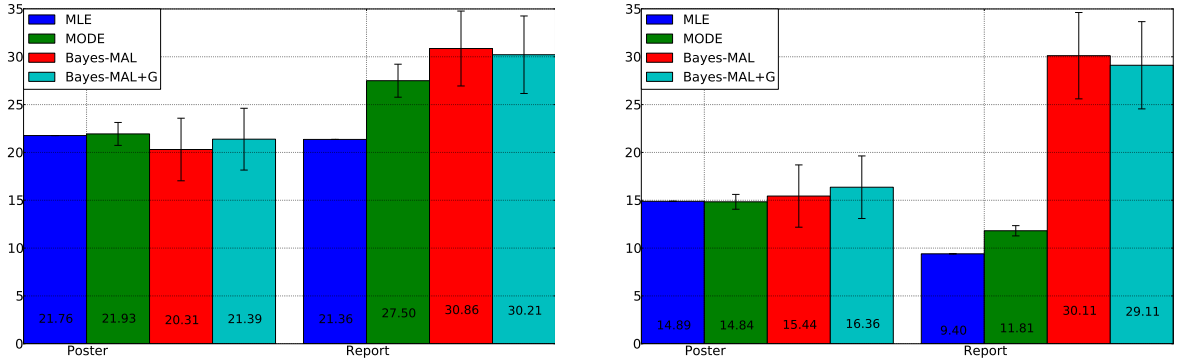

Figure 2: (*Left*) Kendall-Tau Performance of the MCMC Mallows methods (with and without grader reliability estimation) against the instructor grades. Lower the better. **Note:** Performance of a *random baseline* would be 50%. Figure on the *Right* is similar but reports a *weighted* version of the Kendall-Tau error.

dataset, though performance worsening is unsurprising, we still find that the performance is quite competitive and far better than random ordering. In both cases, the performance reported is better than the **NCS** cardinal grading method [19] as reported in [21]. We should also point out that while the diminished performance of the MODE compared to the MLE may be a bit surprising, this is strongly determined by which mode was selected (as the distribution tends to be multi-modal).

While we typically observe better performance on the POSTER dataset than the REPORT dataset, this can largely be attributed to the larger number of peer grades per assignment for the posters. Lastly we also note that the performance does not vary much on adding grader reliability estimation. This observation agrees with that made in [21] where a similar finding was made. The most likely reason for observing this behavior is the explicit incentive the students were given for doing a good job grading.

3.3 How good are the confidence intervals learned

While the previous experiment indicated the overall quality of the orderings tends to be quite good (with regards to instructor grades), it does not tell us how good the confidence intervals estimates are. In particular, we would like to estimate how good are the Bayesian confidence intervals (*i.e.,* credible intervals) of the inferred posterior marginal distributions (over position in the overall ranking) for individual assignments. To evaluate these uncertainty estimates, we again utilize the instructor grades [2]. In particular we evaluate the quality of the 50% and 80% credible intervals.

For each assignment, we first compute the (posterior) marginal distribution over the ranking positions as shown in Figure 1 from the introduction. We then compute the overlap of the credible intervals of these marginals with the instructor ranking distribution *i.e.,* an assignment whose credible interval contains (all) the instructor-provided ranks has a 100% overlap, whereas an interval with no overlap scores a 0%. We report this overlap averaged over all assignments. In addition to this, we also report the size of these intervals (as a percentage of the overall ranking length).

The results are shown in Figure 3. Encouragingly we find that the 50% and 80% interval cover roughly that percentage of the instructor grades, which indicates that the intervals are meaningful and of good quality. Furthermore we find that the overlap is far more than the size of the interval (indicating that the performance is far better than random). As in the previous experiment, we do not find a significant difference in performance when using grader reliability estimation.

Lastly we find that the interval quality tends to be slightly better on the reports than the posters, but that is largely due to the intervals being significantly larger than for the posters, *i.e.,* there is far more

[2] Since these also have ties, we treat ties as indifference and hence have a uniform probability distribution over all possible *valid* rank positions.

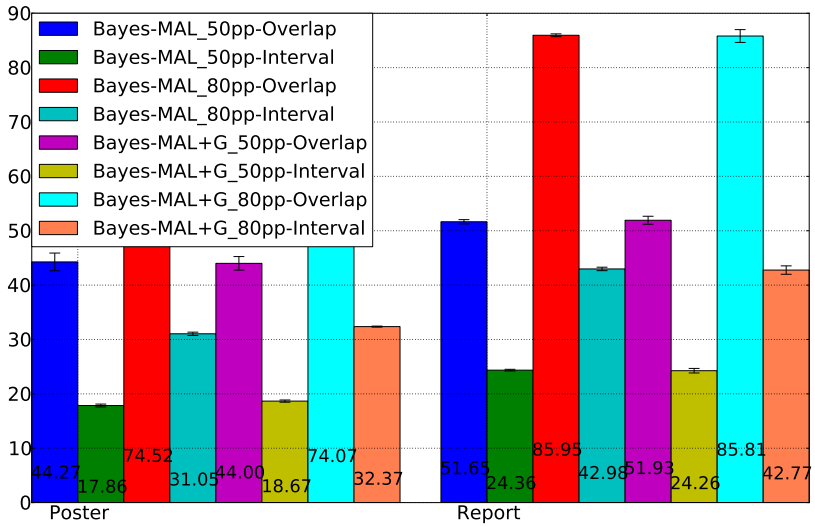

Figure 3: Overlap of the 50% and 80% Bayesian credible intervals with the instructor rank distribution, for both the Bayes-MAL and Bayes-MAL+G methods. For each interval, we report the average overlap followed by the average size of the interval (as a percentage) of overall ranking length.

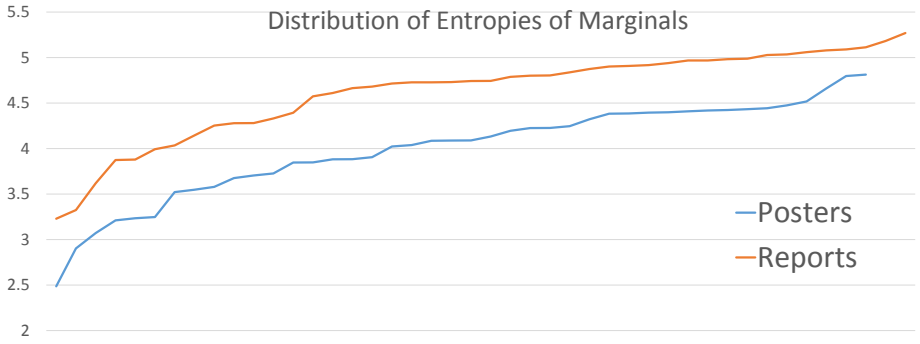

Figure 4: Distribution of the entropies of the marginals for the Bayes-MAL method.

uncertainty in the marginals of the reports than the posters. This is confirmed by Figure 4 which shows the distribution of the marginal entropies for the two datasets.

Together these two experiments indicate that the posterior marginals learned are meaningful and convey the uncertainty information fairly accurately and can hence be used to report such information to instructors. In fact, the examples from Fig 1 are the actual inferred distributions of assignments from the poster dataset using the Bayes MCMC Mallows model.

4 Future Work

In addition to further empirical studies into the quality of the learned posteriors, we are also exploring other Bayesian techniques for the OPG problem. Furthermore we are also exploring studying the quality of the credible intervals of the estimated grader reliabilities.

Acknowledgments

This research was funded in part by NSF Awards IIS-1217686 and IIS-1247696, the JTCII Cornell-Technion Research Fund and a Google PhD Fellowship.

References

[1] K. J. Arrow. *Social Choice and Individual Values*. Yale University Press, 2nd edition, Sept. 1970.

[2] J. A. Aslam and M. Montague. Models for metasearch. In *SIGIR*, pages 276–284, 2001.

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

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

[5] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there: Preference judgments for relevance. In *ECIR*, pages 16–27, 2008.

[6] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz. Pairwise ranking aggregation in a crowd-sourced setting. In *WSDM*, pages 193–202, 2013.

[7] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335, 1995.

[8] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW*, pages 613–622, 2001.

[9] S. Freeman and J. W. Parks. How accurate is peer grading? *CBE-Life Sciences Education*, 9(4):482–488, 2010.

[10] J. Haber. `http://degreeoffreedom.org/between-two-worlds-moocs-and-assessment`.

[11] J. Haber. `http://degreeoffreedom.org/mooc-assignments-screwing/`, Oct. 2013.

[12] M. Kendall. *Rank correlation methods*. Griffin, London, 1948.

[13] C. Kulkarni, K. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. Klemmer. Peer and self assessment in massive online classes. *ACM Trans. CHI*, 20(6):33:1–33:31, Dec. 2013.

[14] T.-Y. Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, Mar. 2009.

[15] T. Lu and C. Boutilier. Learning mallows models with pairwise preferences. In *ICML*, pages 145–152, June 2011.

[16] T. Lu and C. E. Boutilier. The unavailable candidate model: A decision-theoretic view of social choice. In *EC*, pages 263–274, 2010.

[17] C. L. Mallows. Non-null ranking models. *Biometrika*, 44(1/2):pp. 114–130, 1957.

[18] S. Niu, Y. Lan, J. Guo, and X. Cheng. Stochastic rank aggregation. *CoRR*, abs/1309.6852, 2013.

[19] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in MOOCs. In *EDM*, 2013.

[20] T. Qin, X. Geng, and T.-Y. Liu. A new probabilistic model for rank aggregation. In *NIPS*, pages 1948–1956, 2010.

[21] K. Raman and T. Joachims. Methods for ordinal peer grading. In *KDD*, KDD '14, pages 1037–1046, New York, NY, USA, 2014. ACM.

[22] N. Stewart, G. D. A. Brown, and N. Chater. Absolute identification by relative judgment. *Psychological Review*, 112:881–911, 2005.

[23] M. N. Volkovs and R. S. Zemel. A flexible generative model for preference aggregation. In *WWW*, pages 479–488, 2012.