# Bayesian Ordinal Peer Grading

**Karthik Raman**
Department of Computer Science
Cornell University, Ithaca NY 14853
karthik@cs.cornell.edu

**Thorsten Joachims**
Department of Computer Science
Cornell University, Ithaca NY 14853
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 [23]. Existing techniques, which extend rank-aggregation methods, produce a single ordering as output. While these rankings have been found to be an accurate reflection of assignment quality on average, they do not communicate any of the uncertainty inherent in the assessment process. In particular, they do not to provide instructors with an estimate of the uncertainty of each assignment's position in the ranking. In this work, we tackle this problem by applying Bayesian techniques to the ordinal peer grading problem, using MCMC-based sampling techniques in conjunction with the Mallows model. Experiments are performed on real-world peer grading datasets, which demonstrate that the proposed method provides accurate uncertainty information via the estimated posterior distributions.

### Author Keywords
Peer Grading, Ordinal Feedback, Rank Aggregation

### ACM Classification Keywords
H.4 Information Systems Applications: Miscellaneous

### General Terms
Algorithms, Experimentation, Theory

## 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. While scalable automatic-grading schemes — such as multiple-choice questions — exist, they are not suitable in all settings [4, 25, 12, 13]. For instance, research-oriented classes require more open-ended testing such as essays and reports, which are very challenging to evaluate automatically. A lack of reliable assessment techniques for these types of assignments may limit the kinds of courses offered as MOOCs.

*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. Peer grading naturally overcomes the problem of scale [11, 16], since the number of "graders" matches the number of students. Despite this inherent scalability of peer grading, 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 therefore imperative that grading guidelines are easy to communicate and apply, making the feedback process a easy and unambiguous as possible. Towards this goal, recent work has proposed eliciting ordinal feedback from graders [23] (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 more reliable than cardinal feedback [15, 3, 24, 6], and avoids having to communicate absolute grading scales.

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 [23] and shown to not only be comparable to (if not better than) cardinal-grading based techniques but also traditional evaluation practices 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.

While existing ordinal peer grading techniques were shown to estimate rankings that are accurate on average, they merely output a single ranking without communicating the uncertainty inherent in the assessment process. In particular, they do not provide instructors with an estimate of the uncertainty of each assignment's position in the ranking. To overcome this limitation, this paper presents a method for inferring the posterior distribution of where each assignment falls in the overall ranking. This information can, for example, be visualized as shown in Figure 1. Most importantly, the height of the blue bars shows the probability with which each assignment falls at a specific rank. This 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 there

*L@S 2015*, March 14–18, 2015, Vancouver, BC, Canada.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3411-2/15/03 ...$15.00.
http://dx.doi.org/10.1145/2724660.2724678

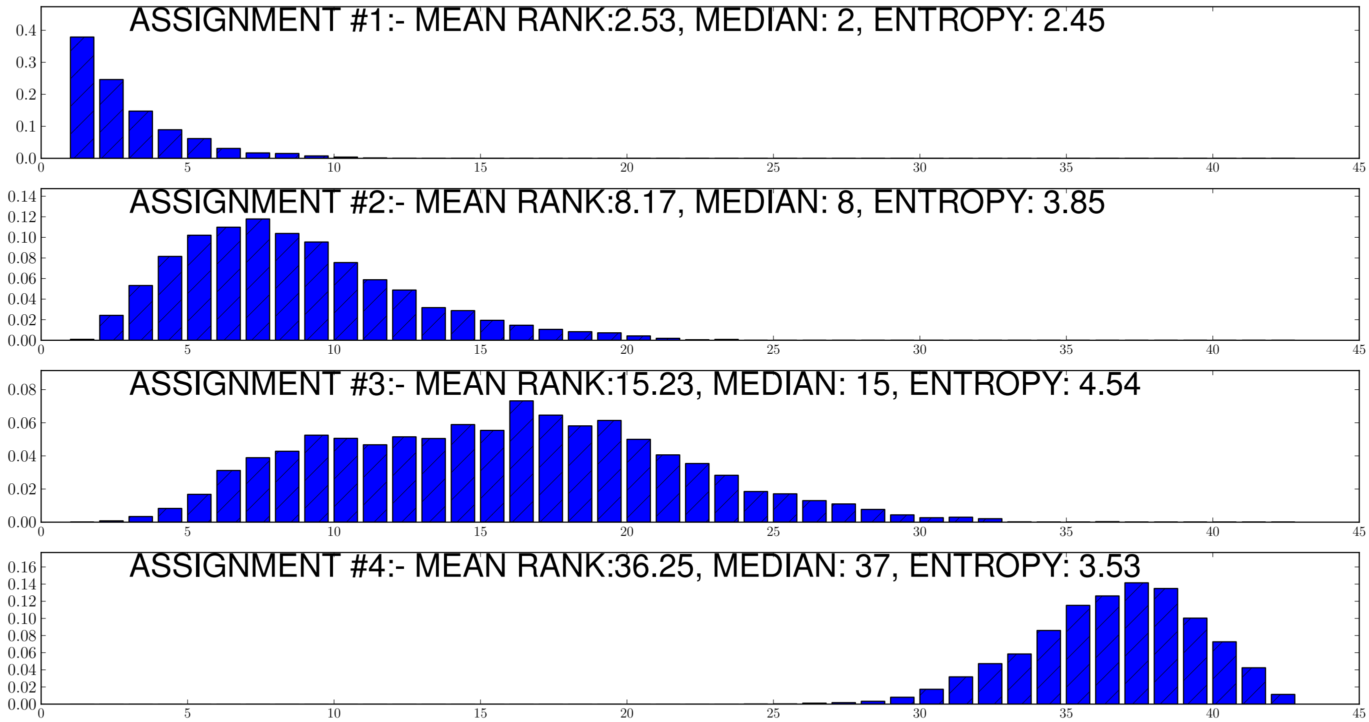

**Figure 1. Output of the Bayesian ordinal peer grading method proposed in this paper. Having the peer grading algorithm produce more detailed information of each individual assignment's performance can be very useful for instructors when it comes to determining final grades. The above figure is one 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.**

is a high probability that assignment 1 is the best of the four assignments, it is less certain 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 4.54). If presented with such information, instructors could intervene and improve certainty by soliciting additional reviews for specific assignments, or at least by accounting for the uncertainty when deriving their grades from the ranking.

In this work, we address the problem of uncertainty modeling by employing Bayesian techniques for the ordinal peer grading problem. In particular we propose a Metropolis-Hastings [8] based Markov Chain Monte-Carlo (MCMC) method, for sampling from the posterior of a Mallows model [20]. The resulting samples allow us to empirically estimate the posterior rank distribution of each assignment, allowing us to report confidences and uncertainty information.

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 learned posterior orderings, we also analyze the resulting confidences and uncertainty information, both qualitatively and quantitatively.

## BAYESIAN METHODS FOR ORDINAL PEER GRADING

In this section, we first describe the ordinal peer grading problem from a machine learning perspective. We then briefly review existing techniques for the ordinal peer grading problem. Our proposed Bayesian version of these techniques is then presented, followed by an empirical evaluation of these techniques in the Experiments section.

### Ordinal Peer Grading (OPG) 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* [23] *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 $\sigma^*$, 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.

[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).

| | |
|---|---|
| $G, g(\in G)$ | 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 $\sigma$ (rank 1 is best) |
| $d_2 \succ_\sigma d_1$ | $d_2$ is preferred/ranked higher than $d_1$ (in $\sigma$) |
| $\pi(A)$ | Set of all rankings over $A \subseteq D$ |
| $\sigma_1 \sim \sigma_2$ | $\exists$ way of resolving ties in $\sigma_2$ to obtain $\sigma_1$ |
| $\hat{\sigma}$ | Estimated ordering of assignments |
| $\sigma^*$ | (Latent) True ordering of assignments |

**Table 1. Notation overview and reference.**

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 should allow 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, the ability to identify unreliable graders enables the instructor to incentivize good and thorough grading by making peer grading itself part of the overall grade.

### 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 [17] 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 [10, 19] 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 [20] and Bradley-Terry model [5] have become popular for these problems [18, 7] and have been used to improve ranking performance in different settings [22, 26, 21].

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 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).

### Existing Approaches to OPG

Different approaches [23] 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 [14] from the true ranking $\sigma^*$ of assignments.

DEFINITION 1. *The Kendall-$\tau$ Distance $\delta_K$ between rankings $\sigma_1$ and $\sigma_2$ is the number of incorrectly ordered pairs between the two rankings and is given by*

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

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

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

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

$$Z_M(k) = \prod_{i=1}^{k} \Big(1 + e^{-1} + \cdots + e^{-(i-1)}\Big) = \prod_{i=1}^{k} \frac{1 - e^{-i}}{1 - e^{-1}} \qquad (3)$$

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 $\sigma'$ consistent with the weak ordering $\sigma^{(g)}$. While computing the Maximum-Likelihood Estimator (MLE) of Equation 2 is NP-hard [10], several simple and tractable approximations that are shown to work well in practice are presented in [23].

While this model does not produce grader reliability estimates, an extension to the model is proposed in [23] 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 predicted ranking and reliabilities.

### 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

**Algorithm 1** Sampling from Mallows Posterior using Metropolis-Hastings

```
1: Input: Grader orderings $\sigma^{(g)}$, Grader reliabilities $\eta_g$ 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 $\sigma'$ 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 $\sigma_t$ to samples (if burn-in and thinning conditions met)
```

for now), we would like to know the posterior distribution of the inferred rankings $\sigma$ *i.e.,* $P(\sigma|\{\sigma^{(g)}; \forall g\})$. We can safely assume a uniform prior on all orderings (for academic fairness), which gives us

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

With the posterior distribution in hand, we can derive the desired marginal rank distributions of each assignment, or we can predict a single ranking that minimizes posterior expected loss.

However, exact computations with this posterior are 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).

This results in 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 $\sigma'$ given the previous sample $\sigma_{t-1}$, we sample from a jumping distribution (Line 5). In particular, we use a **Mallows**-based jumping distribution:

$$\rightarrow \qquad J_{MAL}(\sigma'|\sigma) \propto e^{-\delta_K(\sigma',\sigma)}.$$

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.

| Data Statistic | Poster | Report |
|---|---|---|
| Number of Assignments | 42 | 44 |
| Number of Peer Reviewers | 148 | 153 |
| Total Peer Reviews | 996 | 586 |

**Table 2. Statistics for the two datasets "Poster" and "Report"**

.

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:

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

This expression is again simple to compute and can be done in time proportional to the number of flipped pairs between $\sigma_a$ and $\sigma_b$, 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 once 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.

Software and an online service that implements these methods is available at `http://www.peergrading.org/`.

## EXPERIMENTS

In this section, we empirically evaluate the performance of the Bayesian Mallows-based peer grading method. In particular, we study a) the quality of its predicted rankings in comparison with existing peer-grading methods as measured with regards to conventional instructor grades; and b) the accuracy of the confidence intervals and uncertainty information.

### Experimental Setting

We used the peer-grading datasets introduced in [23]. 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 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** and **Final Report** 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. The peer grading was done on a 10-point (cardinal) Likert scale so as to compare cardinal and ordinal peer grading methods. The ordinal peer grading methods merely used the ordering implies by the cardinal grades.

Table 2 provides some of the key statistics of the two datasets. On average each poster and final report received roughly 24 and 13 peer reviews respectively. For both datasets there was a single instructor grade for each assignment. As described in [23], the instructor grades for the reports were determined completely independent of the peer grades. For the posters the instructor grades utilized the TA grades, which were partly influenced by student grades.

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.

### Are the inferred orderings accurate?

A key benefit of the Bayesian approach is that the posterior distribution of the orderings provides uncertainty information. But we can also use the posterior distribution to predict a single ordering of the assignments. How does the accuracy of the orderings predicted by the Bayesian model compare to the accuracy of the orderings estimated via maximum likelihood estimation (MLE)? To address this question, we compare the following techniques:

- **MLE:** Maximum-Likelihood Estimator of the Mallows model [23]. This is a single point estimate, and it is also used to initialize the Markov Chain.

- **Mode-MAL:** (One of the) Modes of the posterior of the Mallows distribution. Ties are broken randomly.

- **Mode-MAL+G:** (One of the) Modes of the posterior of the Mallows distribution with grader reliability estimates. Ties are broken randomly.

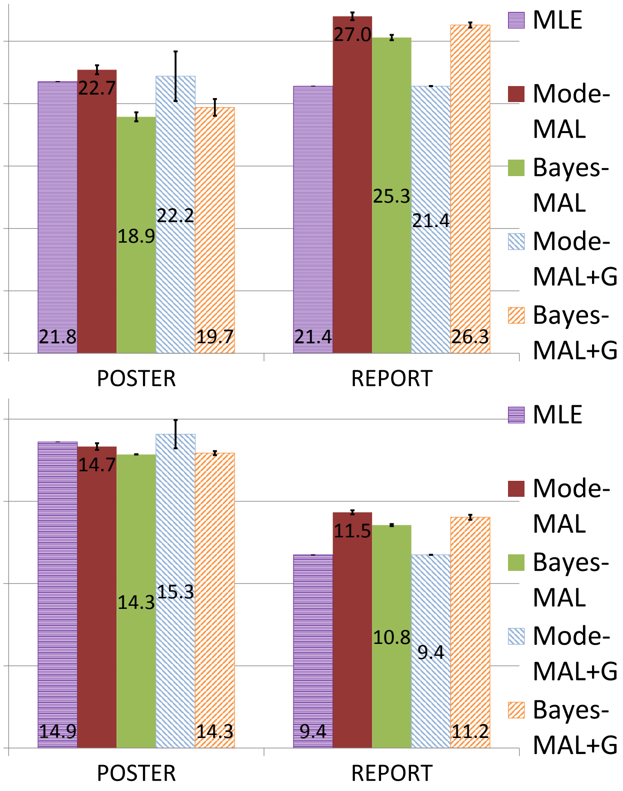

**Figure 2. (*Top*) Normalized Kendall-Tau performance of all methods against the instructor grades for both datasets: Poster (Left) and Report (Right). Figure on the *Bottom* is similar but reports a *weighted* version of the Kendall-Tau error. Note: Performance of a *random baseline* would be 50% for both metrics. For both figures, the lower value is better.**

- **Bayes-MAL**: This is the Bayes estimate minimizing posterior expected $\delta_K$ over the posterior learned by Alg 1. Formally, the predicted ordering is

$$\hat{\sigma} = \texttt{argmin}_{\sigma} \sum_{\sigma'} \delta_K(\sigma', \sigma) P(\sigma'|D),$$

  where $P(\sigma'|D)$ represents the estimated posterior distribution (as output by the Bayesian MCMC method).

- **Bayes-MAL+G**: This is the Bayes estimate minimizing posterior expected $\tau_{KT}$ over the posterior of the Mallows model with grader reliability estimates.

While computing the Bayes-MAL and Bayes-MAL+G predictions is an NP-hard problem, as it requires computing the Kemeny-optimal aggregate [10], we can approximate the optimal solution of the minimization problem efficiently. In particular, we used the simple and efficient Borda-Count technique, which is known to be a 5-approximation [9]. In our case, this also carries a nice semantic meaning as it amounts to simply ordering the assignments by their posterior mean ranks.

The results are shown in Figure 2. As the measure of prediction accuracy, we use the Kendall-Tau error with regards to the instructor rankings. We also compute a *weighted* version of the Kendall-Tau error, where misordering items with

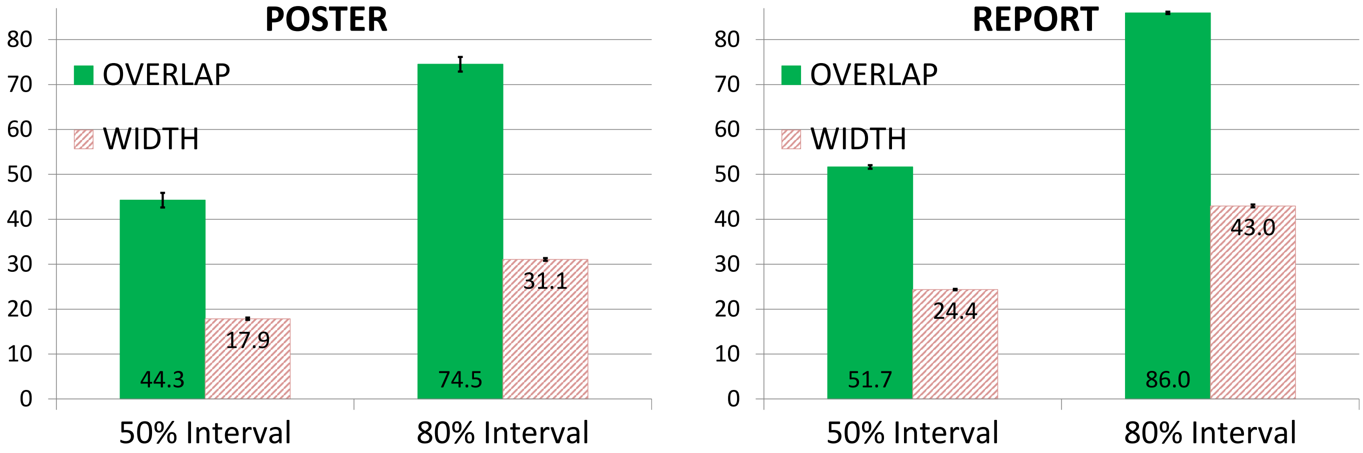

**Figure 3. Average Overlap (solid green bars) of the 50% and 80% Bayesian credible intervals with the instructor rank distribution, for the intervals produced by the Bayesian Mallows MCMC method for the POSTER (*Left*) and REPORT (*Right*) datasets. Along with the overlap is the average size (width) of the interval (as a percentage) of the overall ranking length (the red striped bars).**

a larger (instructor) score difference leads to a worse performance measure. Note that both of these measures are normalized to lie between 0 (indicating perfect agreement with instructors) and 100 (indicating a complete reversal of the instructor ranking). On both datasets, the performance of the proposed Bayesian methods are not substantially different from that of the MLE. There appears to be no clear trend that one method is superior to the others, and the differences are probably due to fact that the instructor grades used as a gold standard are themselves subject to uncertainty. One issue to note is that the "Mode" techniques tend to have larger variance, as performance can vary with the mode that was selected (as the distribution tends to be multi-modal).

Lastly, we also note that the performance does not vary much with adding grader reliability estimation. This observation agrees with a similar finding made in [23] (for both ordinal and cardinal grading techniques). The most likely reason for observing this behavior is the explicit incentive in terms of grade credit that the students were given for doing a thorough job with the peer reviews, such that the number of truly substandard reviews in the data may be low.

### How good are the estimated confidence intervals?

While the previous experiment indicated that the overall quality of the orderings tends to be quite good (with regards to instructor grades), it does not tell us how accurately the Bayesian approach models the uncertainty of the predicted ranks. To address this question, we now evaluate how good the Bayesian confidence intervals (*i.e.,* credible intervals) of the inferred posterior marginal distributions (over position in the overall ranking) for individual assignments are. 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. We find that the intervals produced by the Bayesian MCMC based Mallows technique have are well calibrated. In particular, for both the posters and the reports, the 50% and 80% interval cover roughly that percentage of the instructor grades as desired (as indicated by the overlap values). The observed overlap is far greater than the size of the interval, which indicates predictive performance that is far better than random. These results show that the estimated intervals are meaningful and convey accurate uncertainty information. The results when incorporating grader reliability information are similar and hence left out to avoid redundancy.

### How peaked are the posterior distributions?

The results in Figure 3 show that the confidence interval for the reports have larger width than those for the posters *i.e.,* there is more uncertainty in the marginals of the reports than the posters. This suggests that the posterior distributions are more peaked around the mode for the posters as compared to the reports. To verify this, we computed the *expected* values of the Kendall-Tau error (and the *weighted* Kendall-Tau error) under the posterior distribution:

$$\sum_{\sigma} \delta_K(\sigma^*, \sigma) P(\sigma|D)$$

Note that $\sigma^*$ refers to the instructor ranking and $P(\sigma|D)$ is the learned posterior. We refer to these values as *EXP-MAL*

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

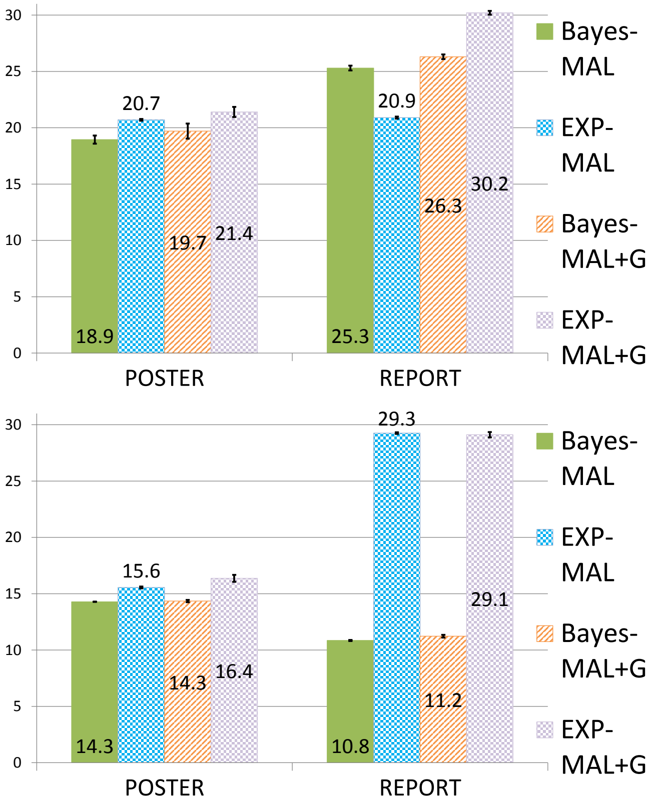

**Figure 4. Kendall-Tau (*Top*) and Weighted Kendall-Tau (*Bottom*) performance of the Bayesian point estimate rankings versus the expected performance of the posterior ranking distribution. For both figures, the lower value is better.**

(without grader reliabilities) and *EXP-MAL+G* (with grader reliability estimation). The results are shown in Figure 4.

We find that the difference in performance between the Bayes estimate (Bayes) and the expected value (EXP) of the full posterior is typically larger for the reports than for the posters. For the posters, it appears that the posterior is so narrow that almost any sample from the posterior is close to the Bayes estimate. For the reports, the posterior is less peaked. One explanation is the larger number of reviews available for the posters.

Finally, we would like to investigate which assignments the Bayesian peer grading method is most uncertain about, and how this uncertainty relates to the scores given by the instructors. To provide some insight, we compute the posterior marginal entropies of all assignments, and then average the entropies for all assignments with the same cardinal instructor grade. The result is visualized in Fig. 5. The assignments that receive the highest and the lowest instructor scores tend to be the assignments with the lowest posterior marginal entropy. The assignment in the middle tend to have higher entropies, indicating that the method is less certain about their position in the ranking. Based on these findings, our conjecture is that it is "easy" to for both students and instructors to identify very good and very bad assignments. The assignments in the middle are more difficult to grade, since they require careful tradeoff between different types of errors. It may also be the case that some of the assignments in the middle are difficult to compare, since they are so different in topic that a meaningful comparison is difficult.

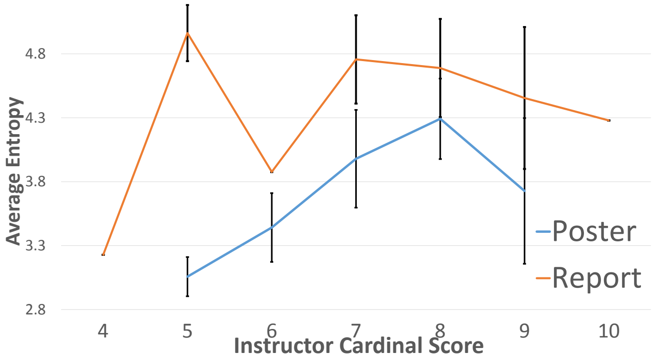

**Figure 5. Distribution of the *average entropies* of the marginals when aggregated by the cardinal instructor grades. Points with no error bars indicate just a single assignment with that score.**

## CONCLUSIONS AND FUTURE WORK

In this work we proposed the use of Bayesian techniques for the problem of ordinal peer grading so as to provide instructors richer information that communicates uncertainty in addition to the predicted ordinal grades. Our proposed method utilizes a Metropolis-Hastings based MCMC sampler for the peer grading Mallows model. We empirically validated the proposed techniques and show the inferred posteriors to agree with instructor grades and to convey an accurate amount of uncertainty.

In addition to further empirical studies into the quality of the learned posteriors, we are exploring how to develop Bayesian inference methods also for other ordinal peer grading models. An open question regards the quality of the credible intervals of the estimated grader reliabilities. Furthermore, there are many interesting questions regarding how to elicit the feedback from the students. For example, it may be cognitively less demanding on the students to break their ordinal assessment task into pairwise comparisons [15], especially if the number of items to assess is large.

## ACKNOWLEDGMENTS

This research was funded in part by NSF Awards IIS-1217686 and IIS-1247637, the JTCII Cornell-Technion Research Fund and a Google PhD Fellowship.

## REFERENCES

1. Arrow, K. J. *Social Choice and Individual Values*, 2nd ed. Yale University Press, Sept. 1970.
2. Aslam, J. A., and Montague, M. Models for metasearch. In *SIGIR* (2001), 276–284.
3. Barnett, W. The modern theory of consumer behavior: Ordinal or cardinal? *The Quarterly Journal of Austrian Economics 6*, 1 (2003), 41–65.
4. Birenbaum, M., and Tatsuoka, K. K. Open-ended versus multiple-choice response formatsit does make a

difference for diagnostic purposes. *Applied Psychological Measurement 11*, 4 (1987), 385–395.

5. Bradley, R. A., and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika 39*, 3/4 (1952), pp. 324–345.
6. Carterette, B., Bennett, P. N., Chickering, D. M., and Dumais, S. T. Here or there: Preference judgments for relevance. In *ECIR* (2008), 16–27.
7. Chen, X., Bennett, P. N., Collins-Thompson, K., and Horvitz, E. Pairwise ranking aggregation in a crowdsourced setting. In *WSDM* (2013), 193–202.
8. Chib, S., and Greenberg, E. Understanding the Metropolis-Hastings Algorithm. *The American Statistician 49*, 4 (1995), 327–335.
9. Coppersmith, D., Fleischer, L. K., and Rurda, A. Ordering by weighted number of wins gives a good ranking for weighted tournaments. *ACM Trans. Algorithms 6*, 3 (July 2010), 55:1–55:13.
10. Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. Rank aggregation methods for the web. In *WWW* (2001), 613–622.
11. Freeman, S., and Parks, J. W. How accurate is peer grading? *CBE-Life Sciences Education 9*, 4 (2010), 482–488.
12. Haber, J. `http://degreeoffreedom.org/between-two-worlds-moocs-and-assessment`.
13. Haber, J. `http://degreeoffreedom.org/mooc-assignments-screwing/`, Oct. 2013.
14. Kendall, M. *Rank correlation methods*. Griffin, London, 1948.
15. Krosnick, J. A. Survey research. *Annual Review of Psychology 50*, 1 (1999), 537–567. PMID: 15012463.
16. Kulkarni, C., Wei, K., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., and Klemmer, S. Peer and self assessment in massive online classes. *ACM Trans. CHI 20*, 6 (Dec. 2013), 33:1–33:31.
17. Liu, T.-Y. Learning to rank for information retrieval. *Found. Trends Inf. Retr. 3*, 3 (Mar. 2009), 225–331.
18. Lu, T., and Boutilier, C. Learning mallows models with pairwise preferences. In *ICML* (June 2011), 145–152.
19. Lu, T., and Boutilier, C. E. The unavailable candidate model: A decision-theoretic view of social choice. In *EC* (2010), 263–274.
20. Mallows, C. L. Non-null ranking models. *Biometrika 44*, 1/2 (1957), pp. 114–130.
21. Niu, S., Lan, Y., Guo, J., and Cheng, X. Stochastic rank aggregation. *CoRR abs/1309.6852* (2013).
22. Qin, T., Geng, X., and Liu, T.-Y. A new probabilistic model for rank aggregation. In *NIPS* (2010), 1948–1956.
23. Raman, K., and Joachims, T. Methods for ordinal peer grading. In *KDD*, KDD '14, ACM (New York, NY, USA, 2014), 1037–1046.
24. Stewart, N., Brown, G. D. A., and Chater, N. Absolute identification by relative judgment. *Psychological Review 112* (2005), 881–911.
25. Veloski, J. J., Rabinowitz, H. K., Robeson, M. R., and Young, P. R. Patients don't present with five choices: an alternative to multiple-choice tests in assessing physicians' competence. *ACADEMIC MEDICINE-PHILADELPHIA- 74*, 5 (1999), 539–546.
26. Volkovs, M. N., and Zemel, R. S. A flexible generative model for preference aggregation. In *WWW* (2012), 479–488.