Evaluation of the Generalized PeerRank Method for Peer Assessment

Mert Can ÇIKLA, Emre BEKTAŞ

Dec 30, 2014

Abstract

Massive open-access online courses serve the purpose of providing free, high level education in vast scales. Peer assessment is the most practical solution in case the participants reach tens of thousands, assignments are open to interpretation and can not be graded in an automated manner. However peer assessment is still far from perfect getting on par with instructor assessment. Bias of the peers and their accuracy in assessing another peer are matters that require work to make peer assessment more usable in practice. In this study we compare Generalized PeerRank Method with some highly popular peer assessment methods. We identify and differentiate them considering their way of approaching the problem. We experiment on various forms of synthetically generated data in order to measure their performance in a simulated environment. We propose three ideas to improve accuracy and reduce error in peer assessment.

1 Introduction

In recent years, massive open-access online courses (MOOCs) have gotten immensely popular. Coursera.org and edX.org are a couple of the online platforms where these courses are being distributed to anyone willing to take them, from around the globe, for free, meaning having an internet connection is enough to have access to these courses who are taught by professors in some of the top universities in the world. Aside from being free, being easily accessible, they also provide the flexibility to study at one's own pace.

Although online education can be provided in huge scales, there is a problem that arises if the students need to be examined. Even though most of the exams held are made up of multiple choice answers or questions that can have their answers evaluated in an automated manner, In cases where the exam can not be assessed by a computer whether it is due to its type —such as an essay— or the nature of the field of study, the issue of assessment and evaluation of thousands of peers arise.

Peer grading or peer assessment is one of most popular solutions to this problem where every peer grades certain amount of other peer's exams which in the end will build up to be that peer's grade for that exam. This concept, in general, tries to utilize the amplitude of the number of peers in a course and attempts to create a fair evaluation of peers where otherwise the evaluation of peers by experts would be impossible. As Sadler and Good have stated in [10] this method not only saves time and

has the logistical upper hand(i.e quicker feedback for peers)but also helps students deepen their understanding of the topic at hand and gives the peers the opportunity to identify their strength and weaknesses regarding the subject.

Peer assessment is the most efficient way to evaluate when the peer count reaches thousands but it has it's flaws. Peers with no incentive to spend time on assessment has a big impact on error. Various studies [2, 7, 13]have tried to incentivize this effort by giving a portion of the final grade to the grader depending on how accurately they have graded compared to others. Due to the vast scale of the peers and their different backgrounds, this is not always sufficient as a mean of error reduction. It might be that one does not want to spend time on assessment and grades randomly or one might not be knowledgeable about a certain subject and may not grade others correctly. There are a lot more similar challenges to be overcome in this case where you have next to no information about how skilled a peer is in evaluating so there are a lot of unknown variables in the equation.

In this study, we gather data regarding peer assessment by reviewing some of the popular peer assessment methods proposed in the literature. We take a closer look into PeerRank method proposed by [13]. Our main contribution is that we devise a methodology to compare ordinal and cardinal methods in the same setting and report GeneralizedPeerRank to be the overall best in most cases. In Section 6 we explain our methodology and our algorithms we used for testing. In Section 7 we give charts of the experimentation we held and analyze the methods on various forms of synthetic data. Finally we identify some of the issues with peer assessment and try to point out possible improvements that are to be implemented and tested in the future.

2 Formal Background

In this section we define some of the scientific terminology used throughout this paper.

Normal Distribution

Normal or Gaussian distribution is a continuous probability distribution that occurs very often in nature and it is defined by the equation below where μ refers to the mean value of the distribution and σ is the standard deviation.

$$\mathcal{N}(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{1}$$

Binomial Distribution

Binomial Distribution is another frequently occurring distribution in nature and it is discrete. Binomial Distribution describes the behavior of Yes/No events such as a coin flip.

$$\mathcal{B}(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{2}$$

where n is the number of trials (coin flips), k is the number of successes with each trial having probability p of success and $\binom{n}{k}$ is defined by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{3}$$

Gamma Distribution

The Gamma distribution is another very important distribution, it is the parent of the exponential distribution and can explain many others.

$$\mathcal{G}(x; k, \theta) = \frac{x^{k-1} e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)} \tag{4}$$

where $\Gamma(p) = \int_0^\infty t^{p-1} e^{-t}\, dt$ is the Gamma function, k is the shape and σ is the scale parameter.

Kendall-τ Distance

Kendall-τ is a metric that counts the number of pairwise disagreements between two rankings σ and π.

$$\mathcal{K}(\sigma) = \sum_{i=1}^{\infty} [\sigma_n(i) < \pi_n(i)] \tag{5}$$

Spearman's Footrule Distance

As opposed to Kendall-τ, Spearman's Footrule measures the total displacement between two elements in a ranking which takes into account the degree of disagreement.

$$\mathcal{F}(\sigma) = \sum_{i=1}^{\infty} |\sigma_n(i) - \pi_n(i)| \tag{6}$$

3 The Peer Assessment Problem

Given S a set of students with cardinality n with each student assessing m students including themselves and $m = n$, let g_{ij} be the grade given to the s_j by s_i, rg_i be the grade that an instructor would have given and reliability r_i be a measure of how reliable student s_i is as a grader. g_{ij} is a simple mark of the student between 0 and 100 where 100 is the perfect, score. $r = 0$ means that the student grades at random and $r = 1$ is a perfect reliability score that is equivalent to an Instructor's. The Peer Assessment Problem is to minimize

$$\epsilon = \sum_i \left(rg_i - \sum_j g_{ij} w_j \right) \tag{7}$$

where $\sum_j w_j = 1$. Note that when simply averaging the grades, $w_j = \frac{1}{n}$ for all j.
In order to reduce ϵ the challenge is to find r_i such that

$$w_i = r_i \sum_{j=0}^{n} \frac{1}{r_j} \tag{8}$$

In real world r is never 1 even though it is required to get $\epsilon = 0$ by setting $r_i = w_i = 1$ and all others to 0 however this is usually not the case and is the core of the problem.

Example 1. Lets assume a classroom with $n = 6$ students and their real grades given by an instructor as $rg = \{75, 85, 50, 45, 10, 25\}$. Let G be a matrix where $G_{ij} = g_{ij}$ i.e the grade given by student s_i to s_j.

$$G = \begin{bmatrix} 80 & 85 & 45 & 45 & 5 & 25 \\ 30 & 100 & 95 & 85 & 55 & 0 \\ 70 & 95 & 35 & 55 & 20 & 5 \\ 30 & 60 & 20 & 75 & 35 & 0 \\ 85 & 75 & 50 & 50 & 5 & 35 \\ 100 & 100 & 85 & 95 & 45 & 70 \end{bmatrix} \tag{9}$$

ID	Performance as Student	Performance as Grader
1	Good	Good
2	Good	Bad
3	Average	Average
4	Average	Bad
5	Bad	Good
6	Bad	Bad

Table 1: Identification of student performances that was used as a basis to generate the grades

From the grades in G we find the error caused in student assessment by $E_{ij} = |rg_j - g_{ij}|$

$$E = \begin{bmatrix} 5 & 0 & 5 & 0 & 5 & 0 \\ 45 & 15 & 45 & 40 & 45 & 25 \\ 5 & 10 & 15 & 10 & 10 & 20 \\ 45 & 25 & 30 & 30 & 25 & 25 \\ 10 & 10 & 0 & 5 & 5 & 10 \\ 25 & 15 & 35 & 50 & 35 & 45 \end{bmatrix} \tag{10}$$

Summing each row i.e all the grades given by a particular student and normalizing to unity yields student reliabilities.

$$\begin{aligned} r_i =& 1 - \frac{(E\vec{1})_i}{\sum_i r_i} \\ r =& \{0.02, 0.30, 0.10, 0.25, 0.05, 0.28\} \end{aligned}$$

This now can be used to reduce assessment error ϵ to 0 by assigning $w_i = 1 - a_i$.
Solving this example by simply averaging yields the grades

$$g = \{66, 86, 55, 68, 28, 23\}$$

subtracting these from instructor grades yields

$$\bar{\epsilon} = \{9, -1, -5, -23, -18, 2\}$$

which is on average 10 grade points deviated per student with respect to real grades and maximum error of 23 points for student with ID 4.

4 Literature Review

There have been various studies to increase the use of Peer assessment by reducing the error induced in evaluation. CrowdGrader by [2] combines the grades provided by the students into a consensus grade for each submission students make by utilizing an algorithm that relies on a reputation system. They introduce the Vancouver algorithm which computes a resulting grade for each submission students make by weighing the students input grade by their accuracy which are then used to update each students estimated accuracy of grading.This algorithm outperforms median and average reliability computation techniques but shows mixed results over real life data. The authors also questioned whether it is best to use solely ranking or grading. In their initial queries students preferred to grade rather than rank because it felt not as accurate as grading since merely ranking was unable to differ between students with assignments which are very close to each other in quality and a pair of submissions one of which is fine but the other one not quite so.

Piech et al. [7] defined 3 statistical models for peer grading. The first one PG1 attempts to detect grader's bias and compensate accordingly. Authors try to make use of any possible coherence between a peer's performance on two different assignments at different times with PG2. PG3 relates a peer's evaluation performance and grade like the method in [13]. They have experimented with these methods on a dataset obtained from a sizable MOOC and report significant improvements. In [5], Mi and Yeung propose 2 extensive models they referred to as PG4 and PG5 which respectively utilize Gamma and Gaussian distributions in order to represent the true scores. What these two models have in difference to the first three models proposed are the redefined correlation between a grader's reliability and her true score. Instead of the linear deterministic behavior that has been followed in PG1-3 they propose two probabilistic relationships. Walsh's [13] adapts PageRank —the website ranking algorithm used by Google— [6] to peer assessment, we explain the method in detail in Section 5.

Another approach used in the literature is to work on rankings instead of a cardinal grade. This approach eases the load of students by asking for pairwise comparisons or a set of rankings but lacks in the value of information retrieved. The method in [11] compares ranking methods to those who only grade and show cases where making pairwise comparisons reduce error in comparison to cardinal evaluation. Their work extends the Bradley-Terry-Luce Model [1, 3] to peer grading and report ordinal evaluation to be more robust to lack of grader expertise.

The work of [9] tackles the Peer assessment as a rank aggregation and extends works of [1, 4, 8, 12]. They evaluated those methods using Kendall-τ on data gathered from a University course and report that ordinal methods are in some cases superior to the cardinal method PG1 of [7].

5 The PeerRank Method

PeerRank method constructs a matrix composed of grades and computes peer grades depending on their own grade and their bias as a grader. The method supposes that the grade of a peer is a measure of their ability to grade correctly and grade of each peer is constructed from the grades of the peers who are grading that peer. As those grades are being weighted by the grading peers own grade which are the peers' grade whom evaluated that peer in the first place hence building up a system of equations that converge to a fixed point.

$$\begin{aligned} X_i^0 &= \frac{1}{m}\sum_j A_{i,j} \\ X_i^{n+1} &= (1-\alpha)X_i^n + \frac{\alpha}{\sum_j X_j^n}\sum_j X_j^n A_{i,j} \end{aligned}$$

Breaking down the equation, the part

$$(1-\alpha).X_i^n$$

corresponds to how much the grader's grade in the previous iteration is affecting the current one.

$$\frac{\alpha}{\sum_j X_j^n}\sum_j X_j^n A_{i,j}$$

is the average grade of the peers' that graded s_i which is weighted by their own grades which is the basis of PeerRank. Our experiments show the method converges to the fixed point without any hassle and the performance of it is of no issue even with particularly large grade matrices.

Some propositions made by Walsh are as follows

- The resulting fixed point grades are the eigenvalues of the grade matrix
- PeerRank always return grades with the interval set
- Lets assume every grade given in the grade matrix is θ, then the PeerRank always assigns θ as the final grade for all the peers
- There is no set of grades for which it is impossible to get it as output of the matrix
- All the grades in the grade matrix contribute to the final grade one way or another i.e there is no dummy.
- Swapping grades of two peers result in their final grade being swapped hence PeerRank is symmetrical

The GeneralizedPeerRank adds the notion of reliability that couples grader's accuracy as grader with their own grade to incentivize correct grading and reduce error. Reliability is added by merely adding a factor into the equations which mean that the grades are being partially weighted by that notion. This method performs quite well especially in a setup where the average real grades of the students are relatively higher as we will observe in Section 7.

$$X_i^{n+1} = (1-\alpha-\beta).X_i^n + \frac{\alpha}{\sum_j X_j^n}.\sum_j X_j^n.A_{i,j} + \frac{\beta}{m}.\sum_j 1 - |A_{j,i} - X_j^n|$$

The last term $\frac{\beta}{m}.\sum_j 1 - |A_{j,i} - X_j^n|$ is added in order to incentivize the peers towards assessing accurately since it ties their accuracy to their final grade. All the propositions made for PeerRank are also true for GeneralizedPeerRank and for the detailed proofs refer to the original paper [13].

Walsh moves on to the definition of Sum of Binomials Marking Model which generates synthetic data for testing.

Algorithm 1 Sum of Binomials Marking Model

review(s) ▷ s: size of class, w: workload
for all $G_{i,j}$ **do** ▷ jth grader of the student i
  $\alpha = \mathcal{B}(rg_i, rg_{G_{i,j}})$ ▷ $\mathcal{B}(n,p)$ n trials and p probability
  $\beta = \mathcal{B}(100 - rg_i, 1 - rg_{G_{i,j}})$
  $g_{i,j} = \alpha + \beta$;
end for

The model has a very intuitive way of assigning grades. Aside from the α component which is straight-forward and represents a student grading correctly a question that is correct. The β component of the mark is added by getting a peer's wrong solution incorrectly graded as correct. However an issue arises with this model that, when a student with a real grader lower than 50 grades a similar grade student they grade themselves much higher than possibly anticipated. For instance, assume the grade of every student in the classroom is 20, this yields on average a grade of $bin(20, 0.2) + bin(80, 0.8) = 80$ even though one would expect an average grade of 20.

A unique feature of PeerRank that sets itself apart from other similar methods is that it incorporates the incentivization grade into the equation, which makes it more accurate when the grades are more deviated.

Walsh suggests three possible extensions to the method, first one is for peers to grade only a subset of each other which is in fact the only suitable solution in practice as it would be unreasonable to ask thousand students to grade thousand other. Similar works have suggested a workload between 4 and 10. Another extension mentioned is to return a group of groups that reflect uncertainty instead of just one final grade.

6 Implementation Details

We have written a Java program that enables us to run experiments with different parameters. In order to generate synthetic data to test we have used the following algorithm.

Algorithm 2 Grade Assignment Algorithm

```
gen(s)                                   ▷ s: size of class, w: workload
S_i ← G_{i+1}..G_{i+5}                   ▷ students i + 1..i + 5 grade student i
for i ← 1, s do
    rg_i ← N(70, 30)                     ▷ real grade of the student
    if rg_i < 0 then
        g_i ← 0
    else if rg_i ≥ 100 then
        g_i ← 100
    else
        g_i ← rg_i
    end if
end for
```

Grade Assignment Algorithm creates a classroom of s students and assigns them w graders and a grade sampled from a Normal distribution $\mathcal{N}$ with $\mu = 70$ and $\sigma = 30$. Also note that we have also experimented with various other distributions.

We have implemented Sum of Binomials Marking Model and Grade Assignment Algorithm to measure performance of different peer assessment algorithms in the literature. The PeerRank and GeneralizedPeerRank methods [13] were implemented by us. To measure the algorithm in [2] we used the implementation of authors' and the ranking methods described in [9] were also measured using the authors' implementation after verification.

7 Experiments

Throughout this paper PR refers to PeerRank [13] and GPR is the GeneralizedPeerRank from the same paper. CG refers to CrowdGrader of [2]. MAL refers to Mallow's model originally proposed by Mallow [4] and was extended to peer assessment by [9]. MALS is the score-weighted Mallow's model and MALBC is Mallow's model with Borda count approximation also proposed by Raman and Joachims [9]. BT which stands for Bradley-Terry Model, THUR, the Thurstone model and PL the Plackett-Luce Model are also explained in [9]. PG1 is the PG1 model proposed by Piech et al. [7] but with a maximum likelihood estimator instead of Gibbs Sampler.

Our first set of experiments are based on the experiments conducted by Walsh in [13] where we measure the Root Mean Square Error which we will refer to as RMSE to evaluate the performances of peer assessment methods. We have generated a synthetic dataset that has 500 students with each student assigned to grade 5 other. Note that the original experiment by Walsh used 10 students all grading each other, with that said, our results verify those of Walsh in [13] and show exacting behaviour.

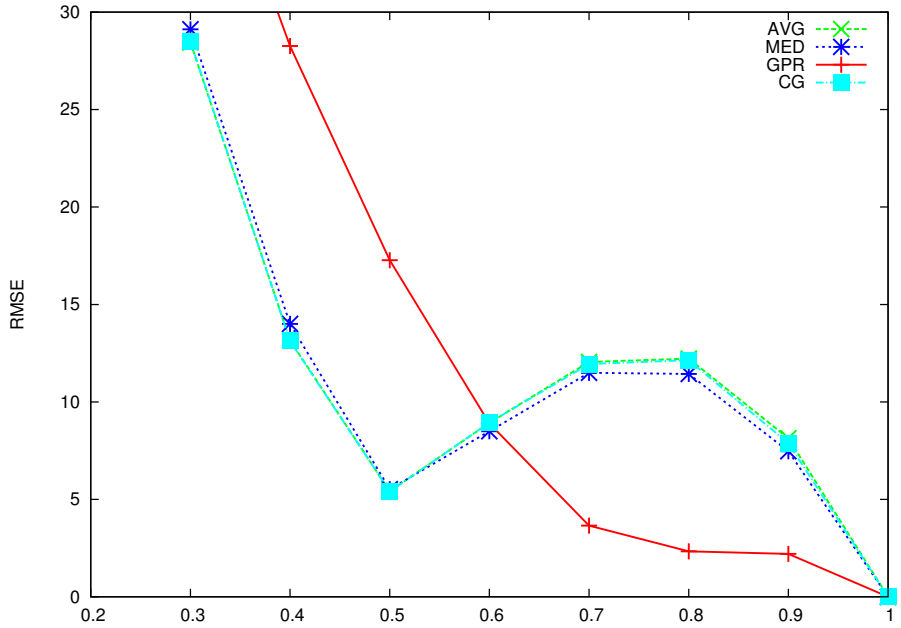

Figure 1: Binomial distribution with varying probability, lower is better

GeneralizedPeerRank outperforms Median and AVG methods in our recreation of binomial model with varying probability, verifying tests conducted by [13]. Additionally the CrowdGrader performs quite similarly to non GPR method which might be due to Vancouver Algorithm's nature utilized by CrowdGrader of [2].

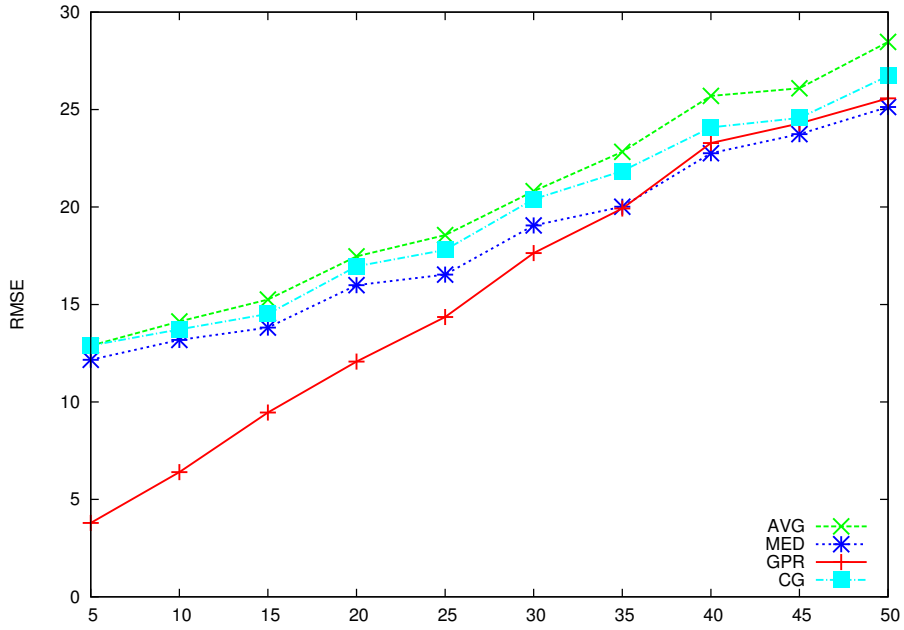

Figure 2: Normal distribution with increasing deviation, lower is better

Similarly to the Binomial case, GeneralizedPeerRank performs better within a low deviation range with Normal Distributions. Here we observe that this is simply due to the fact that GeneralizedPeerRank is heavily dependent on the knowledge amongst peers. CrowdGrader, the method we newly add to this experimentation again displays a behaviour similar to Median and AVG methods.

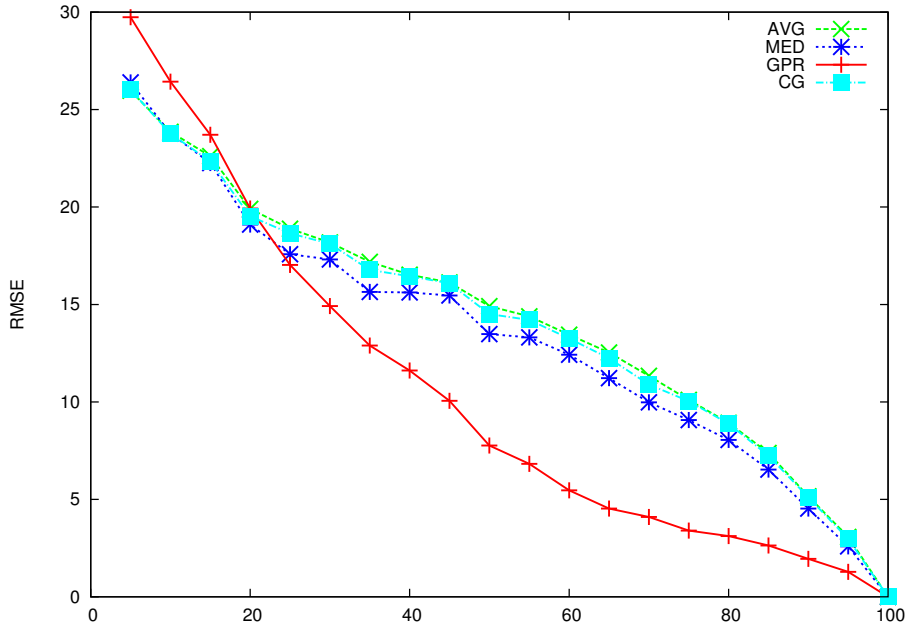

Figure 3: Uniform distribution with increasing lower bound.

GeneralizedPeerRank outperforms averaging when the lower bound is greater than 20. In a similar fashion to previous experiments, the dependence on the knowledge of peers continue as we observe the drop in RMSE as lower bound increases.

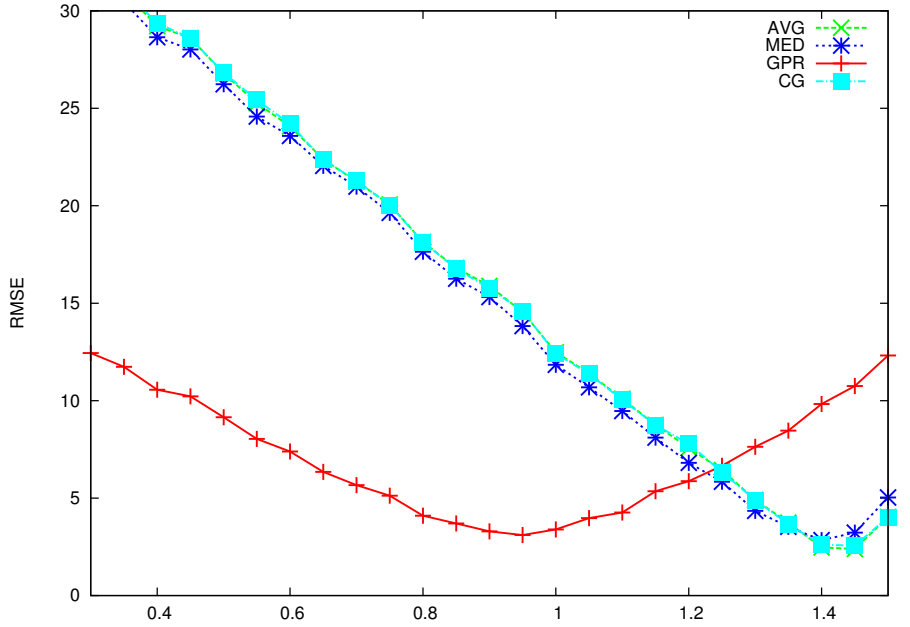

Figure 4: Normal distribution with varying grader bias.

A very interesting result that can be observed from Figure 4 is that GeneralizedPeerRank is mostly unaffected by the peers' biases in assessment. Even in extreme bias levels the method manages to keep the error low and grades considerably much more accurate than the average.

In order to compare ordinal methods' performance with cardinals' we have conducted the following experiments using grades coming from Binomial, Normal and Uniform distributions and we have measured their Spearman's Footrule distances.

The experiments show the AVG, method proposed by [13] outperforms others generally when the grades are Normally or Uniformly distributed. However, the method seems to be average when the grades are Binomial.

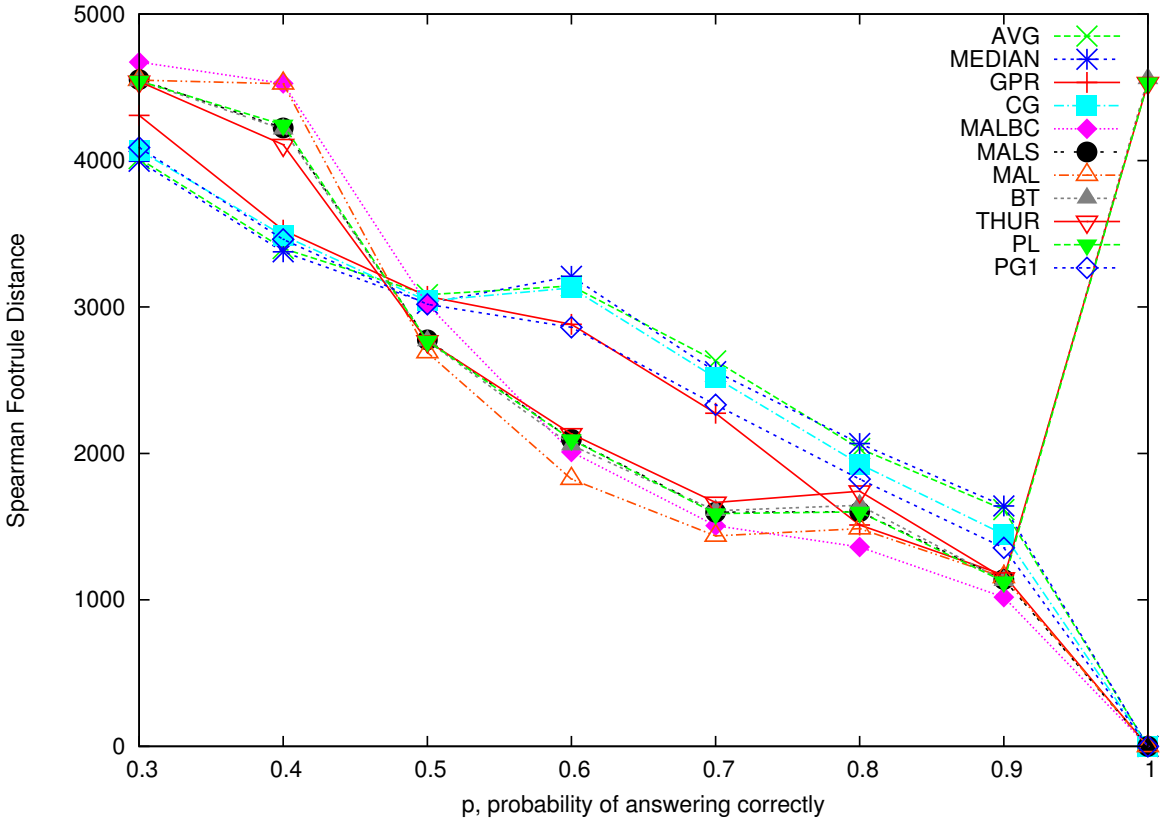

Figure 5: Binomial distribution with varying probability, lower is better

We observe the methods form into to two groups according to the Spearman Distance they incur, the first group is composed of all Cardinal methods AVG, MEDIAN, GPR, CG, PG1 and the second group is all Ordinal with the slight exception MALS which is still ordinal in nature but is score-weighted. Unexpectedly, we also observe the Ordinal methods outperforming Cardinal methods.

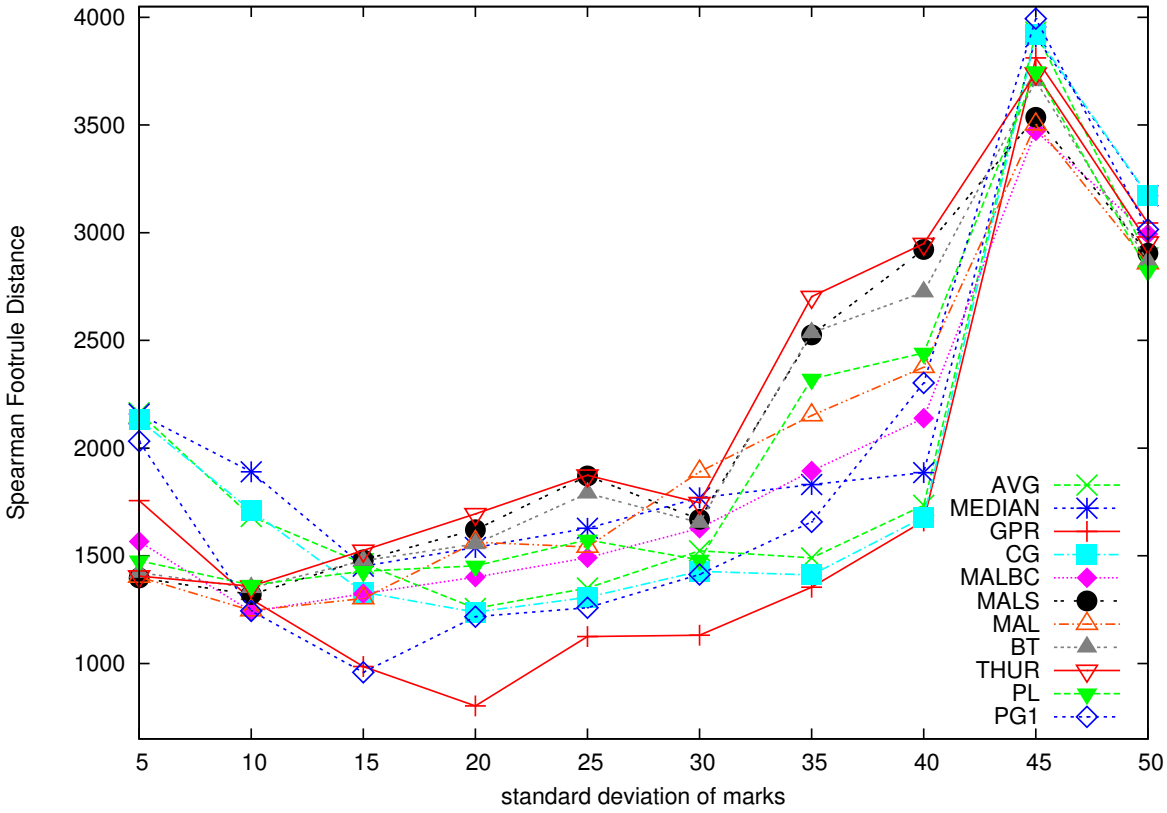

Figure 6: Normal distribution with increasing stdev, lower is better

In the Normal distribution case we observe a much different picture than the Binomial, all the methods perform a lot more similar to one another without any significant formations. We also observe that GeneralizedPeerRank is strictly better when the standard deviation is within the range of 15-40, and is still comparable for the others.

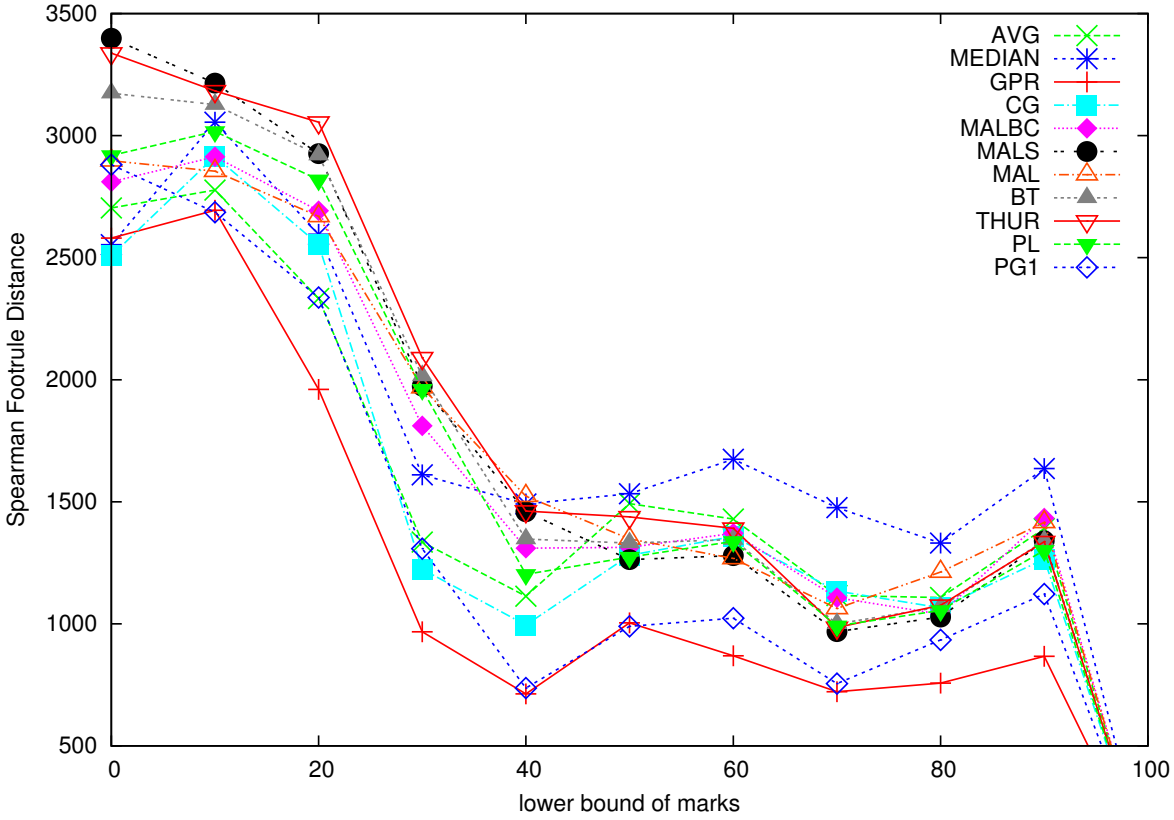

Figure 7: Uniform distribution with increasing lower bound, lower is better

Lower bound of 40 marks as a critical point in behaviour of all the methods. As the average grade increases up to that mark, we can observe a sizable decrease in error. GeneralizedPeerRank stands out as the overall best method.

8 Conclusion

MOOCs are essential to take the education to the next era and Peer assessment has a crucial part in this. However Peer assessment methods still have a long way to go to reduce the error induced due to malicious assessment or lack of assessment performance.

We have experimented using synthetically generated data and distinguished Generalized PeerRank method proposed by [13] to be the method with the least error induced and we are planning to work on ways of improving this particular method.

We also propose an idea that has great potential to improve assessment performance by assigning graders things that they are relatively more knowledged about.

9 Future Work

First possible improvement we have theorized is to improve the Generalized PeerRank by looking at the cause of sub-par performance when the grades are Binomially distributed as it is odd for the results to be vastly different although the Binomial and Normal cases map a very similar distribution of grades overall.

Our second idea for improvement is to gather new information about students by splitting the exam into parts according to a rubric or simply assessing each question separately. Consider the case where a student has gotten a full-mark from the first question but zero marks from the second. In this case the student can be a much more effective as a grader if we can assign the student to only grade the first question for this exam instead of the exam as a whole. This methodology is not limited to exams that can be divided into questions. Any exam with a rubric is an applicable case since for example the grammar used in an English essay can be graded by a grader with grammar expertise and another can grade its cohesion. Also note that this can be used to improve any peer assessment method independent of how the method itself works.

Lastly there is a decent chance of extracting valuable information by finding a coherence between a peer's performance on different times. The likely coherence and identification of this should help better estimate a grader's reliability.

References

[1] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs. *Biometrika*, 39:324–345, 1952.

[2] Luca de Alfaro and Michael Shavlovsky. Crowdgrader: Crowdsourcing the evaluation of homework assignments. *CoRR*, abs/1308.5273, 2013.

[3] R.. Ducan Luce. *Individual Choice Behavior a Theoretical Analysis*. John Wiley, 1959.

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

[5] Fei Mi and Dit-Yan Yeung. Probabilistic graphical models for boosting cardinal and ordinal peer grading in moocs.

[6] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[7] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong B. Do, Andrew Y. Ng, and Daphne Koller. Tuned models of peer assessment in moocs. *CoRR*, abs/1307.2579, 2013.

[8] R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):pp. 193–202, 1975.

[9] Karthik Raman and Thorsten Joachims. Methods for ordinal peer grading. *CoRR*, abs/1404.3656, 2014.

[10] Philip M Sadler and Eddie Good. The impact of self-and peer-grading on student learning. *Educational assessment*, 11(1):1–31, 2006.

[11] Nihar B. Shah, Joseph Bradley, Abhay Parekh, Martin J. Wainwright, and Kannan Ramchandran. A case for ordinal peer-evaluation in moocs. *Neural Information Processing Systems (NIPS): Workshop on Data Driven Education*, 2013.

[12] Leon L Thurstone. The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, 21(4):384, 1927.

[13] Toby Walsh. The peerrank method for peer assessment. *CoRR*, abs/1405.7192, 2014.