Aggregating Partial Rankings with Applications to Peer Grading in MOOCs

Ioannis Caragiannis, George A. Krimpas and Alexandros A. Voudouris
University of Patras and CTI "Diophantus"

Introduction

MOOCs:

- **Massive:** available to a large number of people (e.g. 16–18 million in 2014)
- **Online:** through the Internet/Web
- **Open:** no cost for the students
- **Courses:** series of lectures on a subject

The challenge: Assess students' performance in open type questions (critical thinking, ability in mathematical proofs, etc.)

The bottleneck: Limited available qualified human resources (professional graders, TAs)

The solution: peer grading

Ordinal Peer Grading

- The students **order** the assignments they are given from best to worst
- The partial rankings are **aggregated** into a global ranking that represents the students' relative performance

Important Questions

- How **many** assignments should we give to each student?
- How can we **distribute** the assignments?
- How can we **merge** the partial rankings?
- Objective: to **learn the ground truth** ranking

Model

- n students/graders
- k assignments per grader
- k graders per assignment

Grading scheme:

- Distribute assignments to graders
- **Bundle graph:** a bipartite k-regular graph that contains bundle and assignment nodes; an edge represents the fact that an assignment belongs to a bundle
- **Constraint:** graders cannot grade their own assignments

Grading Scenarios

Perfect grading:

- After all students have submitted their assignments, the instructor **announces** indicative solutions and grading instructions
- The students use this info when grading

Imperfect grading:

- **No info** by the instructor
- Students' grading performance is similar to their performance in the exam

More Info

I. Caragiannis, G. A. Krimpas, and A. A. Voudouris. Aggregating partial rankings with applications to peer grading in massive online open courses. AAMAS 2015, pp. 675–683.

Theorem

Under perfect grading, the expected fraction of pairwise relations in the ground truth that are correctly recovered by Borda is at least $1-\mathcal{O}(1/k)$ when the bundle graph is square-free, and at least $1-\mathcal{O}\left(1/\sqrt{k}\right)$ in general.

Proof Idea

- Two assignments with true ranks $r < q$
- $B_{r,q}$ = difference in their Borda scores
- $\mathbb{E}[B_{r,q}]$ is proportional to $q-r$
- **Martingales** + **Azuma inequality** $\Rightarrow B_{r,q}$ sharply concentrated around $\mathbb{E}[B_{r,q}]$
- $\Pr[\text{pairwise relation correctly recovered}] = \Pr[B_{r,q} > 0]$
- Sum over all pairs of assignments

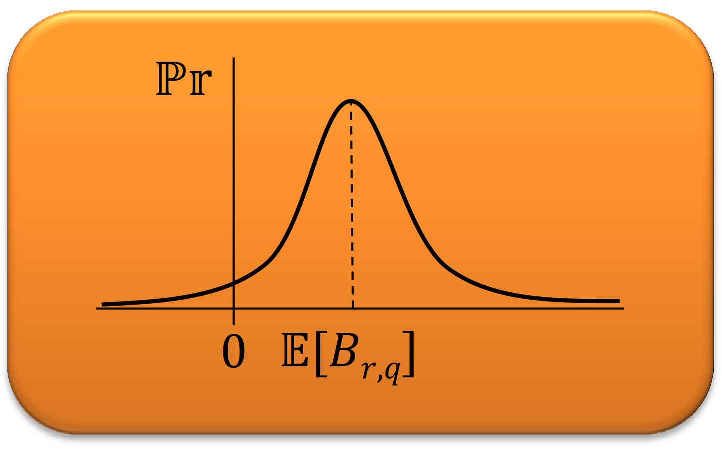

Figure 1: Concentration of $B_{r,q}$ around $\mathbb{E}[B_{r,q}]$

RSD

- Initially, the global ranking is empty
- Serial phase:
  - Randomly permute all partial rankings
  - Traverse the partial rankings and copy all pairwise relations that do not contradict relations copied earlier
  - Augment with pairwise relations implied due to transitivity
- Random completion phase:
  - Iteratively, pick randomly an undecided pair of elements
  - Decide randomly and update all relations implied due to transitivity

Mallows Noise Model

- Each student i has a quality q_i
- The ground truth $\succ$ is defined as the ranking of the assignments in decreasing order of student qualities.
- Student i orders the assignments in her bundle according to the following procedure:
  - For every pair of assignments x, y such that $x \succ y$, with probability q_i set $x \succ_i y$
  - If a cycle is created, repeat from scratch

Experimental Evaluation – Perfect Grading

graph		random k-regular		square-free		copies of $K_{k,k}$	
k	n	Borda	RSD	Borda	RSD	Borda	RSD
2	1002	73.3	62.7	73.5	60.3	66.8	56.8
3	1001	83.0	77.2	83.2	66.0	73.1	60.2
4	1001	87.5	86.8	87.7	68.7	77.1	62.2
6	1023	92.0	94.6	92.1	72.7	81.6	65.2
8	1026	94.2	97.2	94.1	72.8	84.3	66.5
12	1064	96.3	98.9	96.6	76.0	87.3	68.5

Table 1: Performance of Borda and RSD with perfect grading on different bundle graphs of similar size.

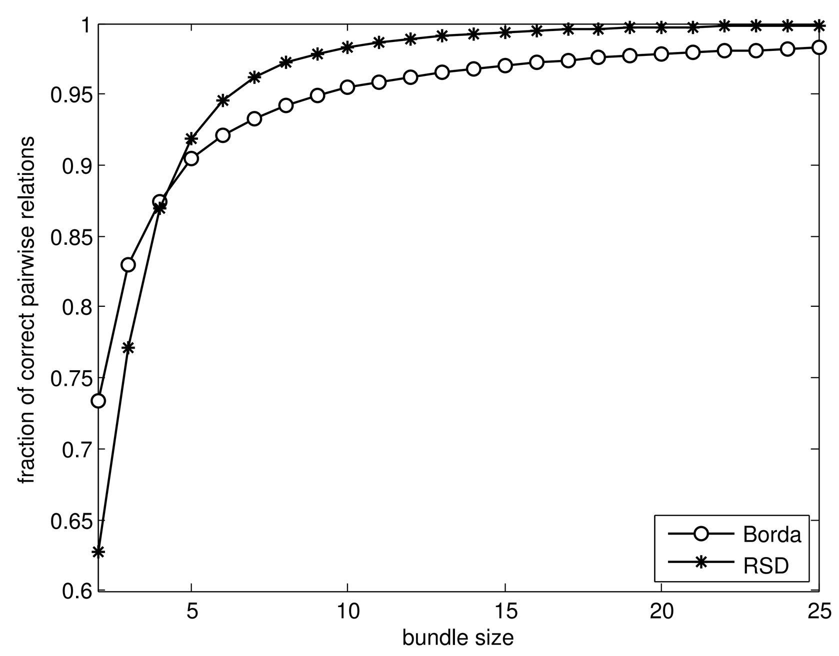

Figure 2: Borda *vs.* RSD – bundle size ranging from 2 to 25.

Experimental Evaluation – Imperfect Grading

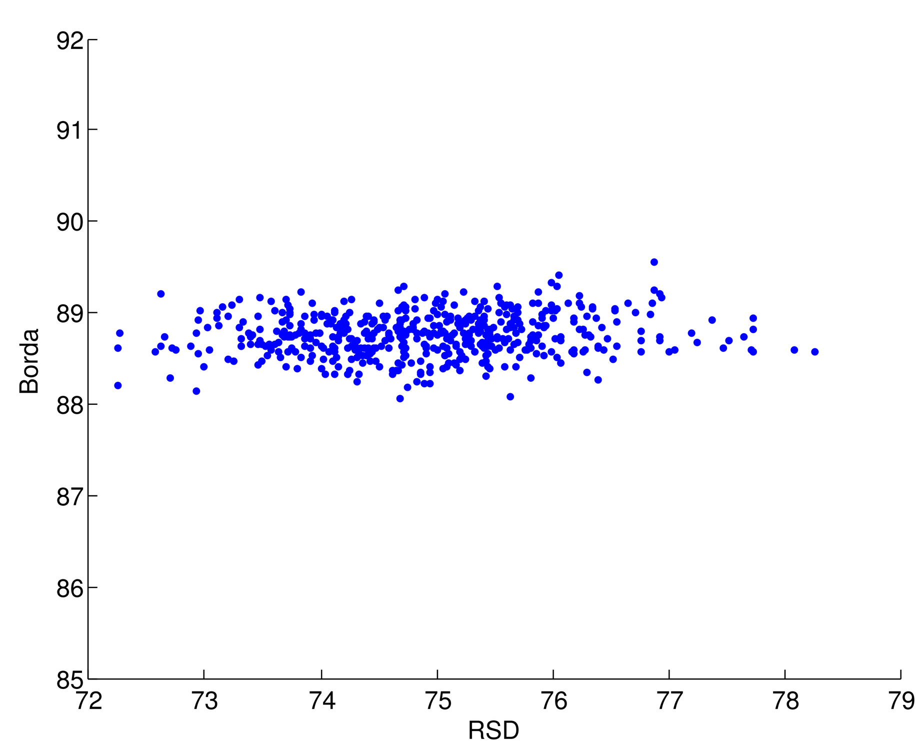

Figure 3: Borda *vs.* RSD – noise level 50%

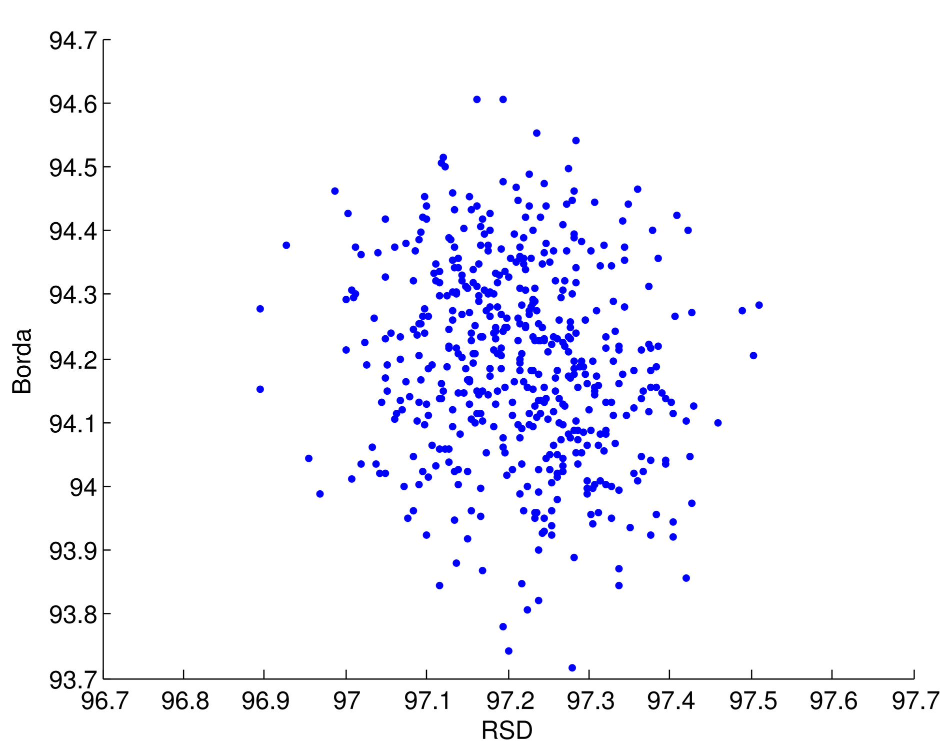

Figure 4: Borda *vs.* RSD – noise level 0%