# How Effective Can Simple Ordinal Peer Grading Be?

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

## The problem

Grading in **massive open online courses**

- Assess students' performance in open type questions (critical thinking, ability in mathematical proofs, essay writing)

**The challenge:**

- huge scale (e.g., 1 million students per exam)
- limited availability of qualified professional graders, TAs, etc.

**The solution:** peer grading

## Ordinal peer grading

After the end of an exam:

- **Distribute** $k$ copies of each exam paper to the students so that each of them gets a bundle of $k$ different papers
- **Grading phase:** each student ranks the $k$ papers in his/her bundle
- **Rank aggregation phase:** the partial rankings are combined into a global one

Grades such as "top $17\%$ among $35\,000$ students"

## Important issues

- How large should $k$ be?
- How should we distribute the assignments?
- How should we merge the partial rankings?
- Can we exploit statistical information about the students' grading behaviour in order to select the best possible rank aggregation rule?

**Objective:** To learn the ground truth ranking approximately

## Related material


- Caragiannis, Krimpas, & Voudouris. Aggregating partial rankings with applications to peer grading in massive online open courses. AAMAS 2015
- Caragiannis, Krimpas, & Voudouris. How effective can ordinal peer grading be? EC 2016
- Caragiannis, Krimpas, Panteli, & Voudouris. co-rank: an online tool for collectively deciding efficient rankings among peers. AAAI 2016
- `co-rank.ceid.upatras.gr`


## Grading scenarios

**Perfect grading**

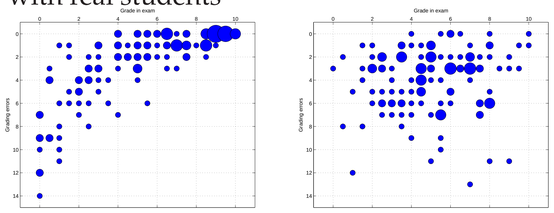
- Students rank the exam papers in their bundle correctly (consistently to the ground truth $\succ$)

**Imperfect grading (Mallows)**

- Every student $i$ has a random quality $q_i$
- Student $i$ ranks the paper exams as follows:
  - For every pair of exam papers $x, y$ such that $x \succ y$, with probability $q_i$ set $x \succ_i y$
  - If a cycle is created, repeat from scratch

**Realistic grading**

- Using data collected through a field experiment with real students




## Aggregation rules

**Borda:**

- Every exam paper gets $k - i$ points when it is ranked $i$-th by a student
- Final outcome is obtained by sorting the exam paper in terms of their Borda score; ties are broken randomly

**Type ordering aggregation rules:**

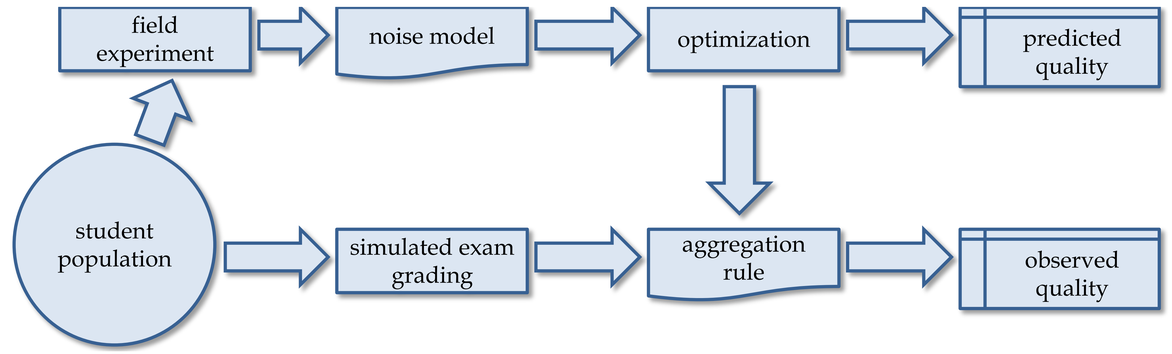
- Every exam paper has a **type**, which is a vector of ranks an exam paper gets in the partial rankings it appears in
- Example: a paper with type $(1, 3, 3, 4, 4, 4)$, is once first, twice third, and three times fourth
- Such a rule defines an ordering over all types
- The final outcome is computed by ranking the exam papers in terms of their types, breaking ties uniformly at random

## Flavour of theoretical results

**Theorem 1.** *Under perfect grading, the expected fraction of pairwise relations in the ground truth that are correctly recovered by Borda is at least* $1 - \mathcal{O}(1/k)$.

**Theorem 2.** *In the perfect grading model, Borda is the optimal type-ordering aggregation rule.*

## The approach




## Computing optimal aggregation rules

**Assessing quality:**

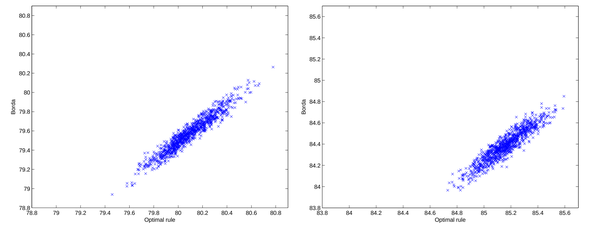
- Quality of a type ordering aggregation rule is defined as the fraction of correctly recovered pairwise relations
- Assuming infinitely many exam papers (the several dependencies among random variables vanish), we get:

$$\begin{aligned} Q(\succ) &= \int_0^1 \int_x^1 \sum_{\sigma,\sigma':\sigma\succ\sigma'} \Pr[x \rhd \sigma \text{ and } y \rhd \sigma']\,\mathrm{d}y\,\mathrm{d}x \\ &\approx \sum_{\sigma,\sigma':\sigma\succ\sigma'} \int_0^1 \int_x^1 \Pr[x \rhd \sigma] \cdot \Pr[y \rhd \sigma']\,\mathrm{d}y\,\mathrm{d}x \\ &= \sum_{\sigma,\sigma':\sigma\succ\sigma'} W(\sigma,\sigma') \end{aligned}$$

**Optimizing:**

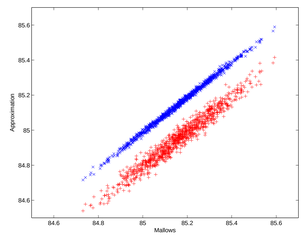
- Compute an ordering $\succ$ among types so that $\sum_{\sigma,\sigma':\sigma\succ\sigma'} W(\sigma,\sigma')$ is maximized
- Equivalent to Feedback Arc Set problem on weighted tournaments
- NP-hard in general but fairly easy in our setting

## Experimental evaluation




Quality of Borda vs optimal rule in realistic and mallows noise scenarios

## Effect of inaccuracies




Blue/red points: comparison of the optimal rule for the 1000/100-sample Mallows approximation vs the Mallows-optimal rule

## Predicted vs observed quality

| perfect grading | | realistic grading | | | | mallows grading | | | |
|---|---|---|---|---|---|---|---|---|---|
| theory | $n = 10^4$ | theory | | $n = 10^4$ | | theory | | $n = 10^4$ | |
| borda | borda | opt | borda | opt | borda | opt | borda | opt | borda |
| 92.01 | 92.02 | 80.01 | 79.57 | 80.09 | 79.57 | 85.15 | 84.38 | 85.16 | 84.39 |

Values are theoretical predictions (theory) or experimental measurements with $10\,000$ students ($n = 10^4$)