Evaluation of the GeneralizedPeerRank Method for Peer Assessment

Mert Can ÇIKLA, Emre BEKTAŞ

Advisor: Prof.Dr.Brahim Hnich

January 13, 2015

Overview

What is Peer Assesment

Peer Assessment

Is a concept where peers/students assess each other's assignments

What is Peer Assesment

Peer Assessment

Is a concept where peers/students assess each other's assignments

How does Peer Assessment work?

- Peers grade a subset of each other
- A consensus grade is constructed from the subset of grades given by others

What is Peer Assesment

Peer Assessment

Is a concept where peers/students assess each other's assignments

How does Peer Assessment work?

- Peers grade a subset of each other
- A consensus grade is constructed from the subset of grades given by others

Students	A	B	C	D
A		60	70	
B	70			50
C		50		40
D	100		80	

Application of Peer Assessment to MOOCs

Massive Open-Access Online courses

Free university level education via some websites

Application of Peer Assessment to MOOCs

Massive Open-Access Online courses

Free university level education via some websites

- The number of students to be evaluated can go upto tens of thousands
- Automated grading can't be used due to the nature of the program(ie art desing projects, essays) or isn't preferred

Peer Assessment Methods

- **Cardinal/Grading Based**
  Peers assign a numerical value
  - Average & Median
  - CrowdGrader
  - $\mathbf{PG}_1$ Model
  - PeerRank & GeneralizedPeerRank

Peer Assessment Methods

- **Cardinal/Grading Based**
  Peers assign a numerical value
  - Average & Median
  - CrowdGrader
  - $\mathbf{PG}_1$ Model
  - PeerRank & GeneralizedPeerRank
- **Ordinal/Ranking Based**
  Peers make piecewise comparisons and report the better one
  - Mallow's Model
  - Bradley-Terry Model
  - Thurstone Model
  - Plackett-Luce Model

Illustration of AVG and MEDIAN

Students	A	B	C	D	E	F
A			70		80	65
B	70		65			90
C		50		80	55	
D	100		80			80
E	90	80		45		
F		85		45	65	
AVG	86.6	71.6	71.6	56.6	66.6	78.3
MEDIAN	90	80	70	45	65	80

What is PeerRank?

- PeerRank is PageRank adapted to Peer Assessment

Toby Walsh (2014)
The PeerRank Method for Peer Assessment
ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic

PageRank

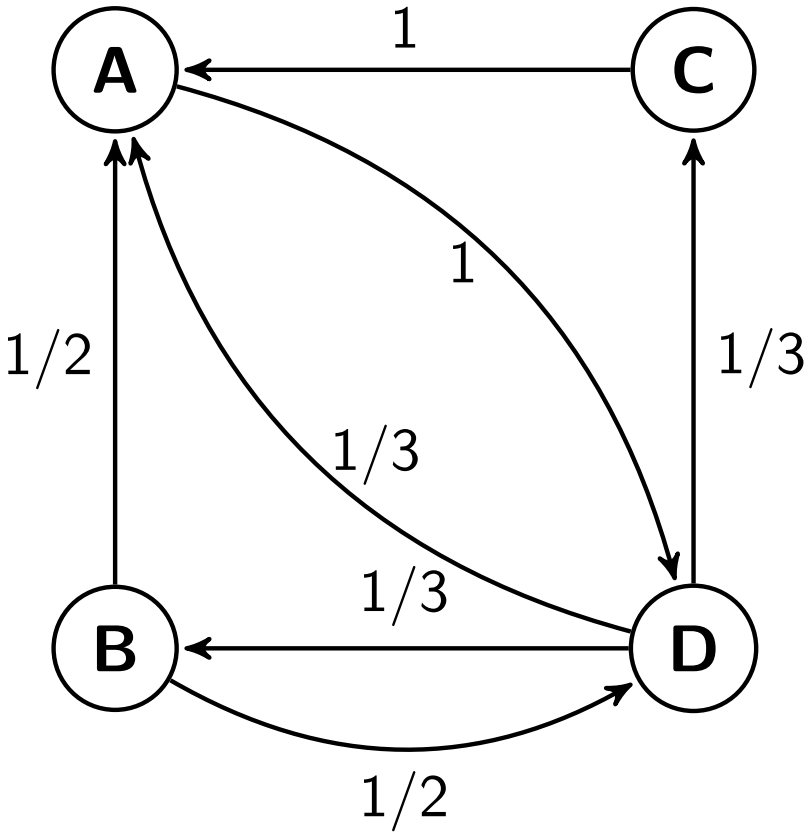


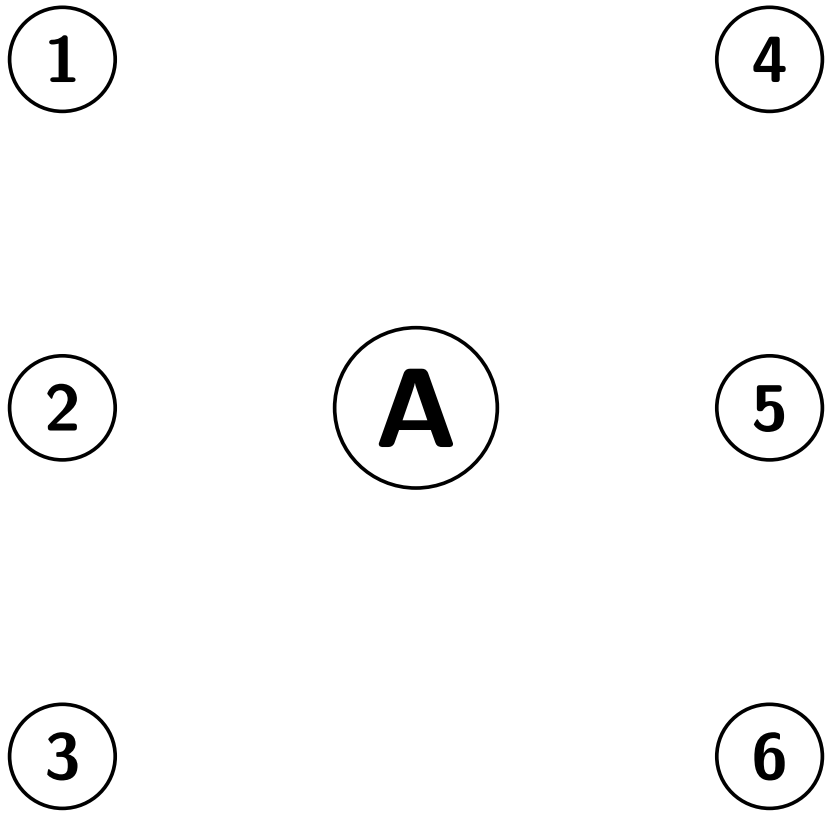
Figure: Hyperlink Graph

0	0	0	1
0,5	0	0	0,5
1	0	0	0
0,33	0,33	0,33	0

Figure: Hyperlink Matrix

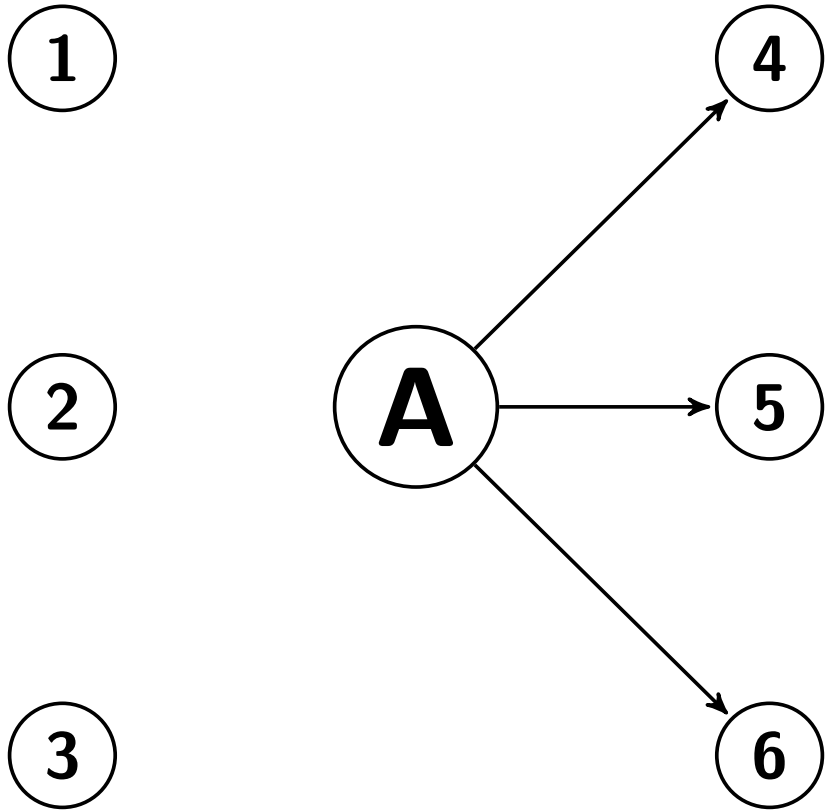
What is PeerRank

- Students grades are assumed to be their ability to grade correctly



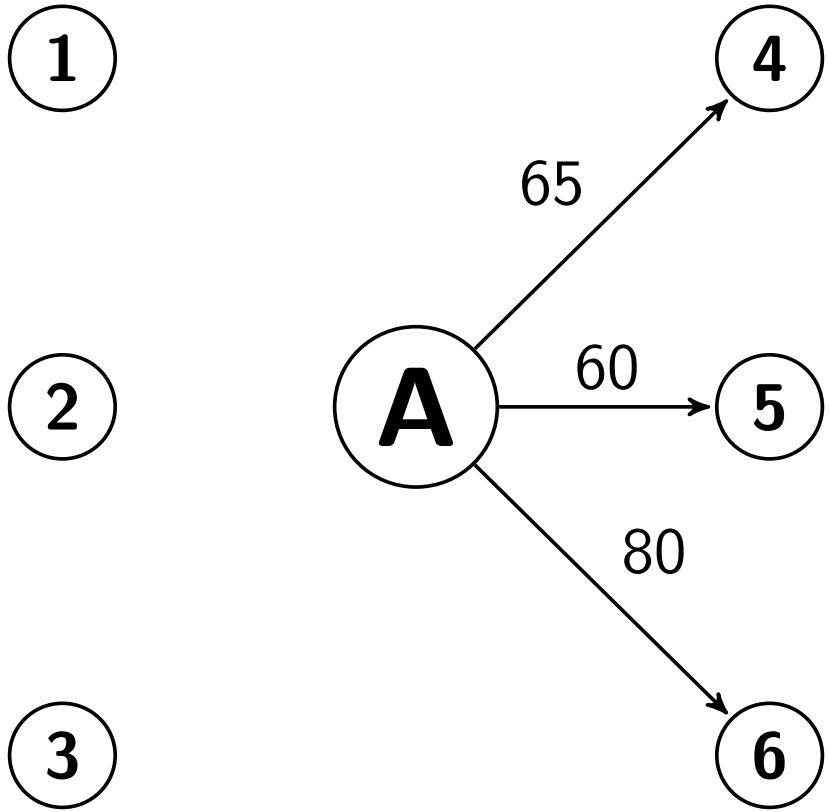
What is PeerRank

- Grader's ability to grade is measured by his/her own grade



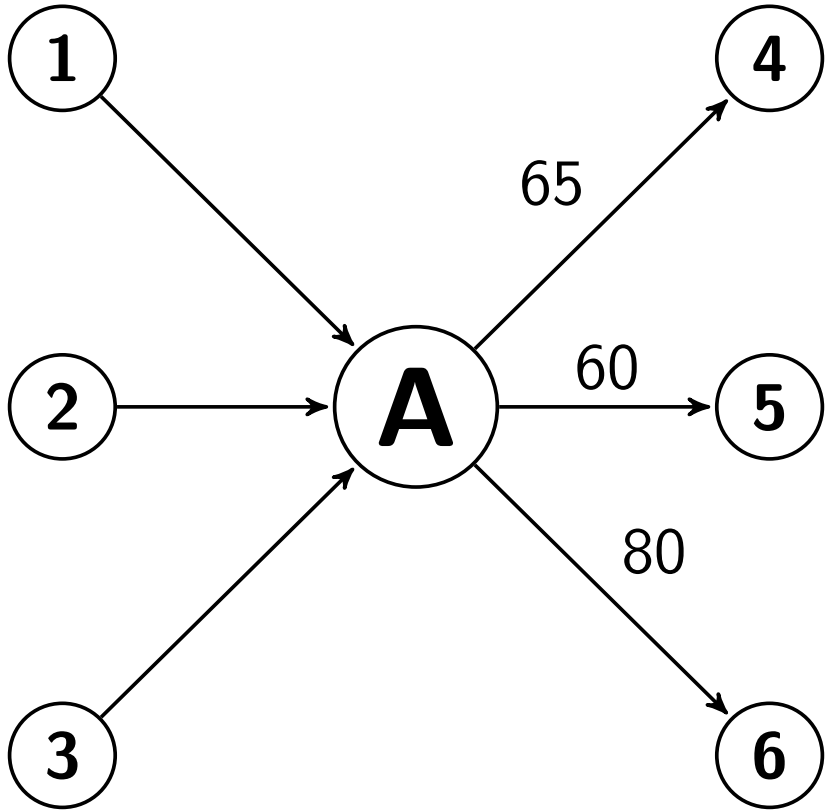
What is PeerRank

- Grader's ability to grade is measured by his/her own grade



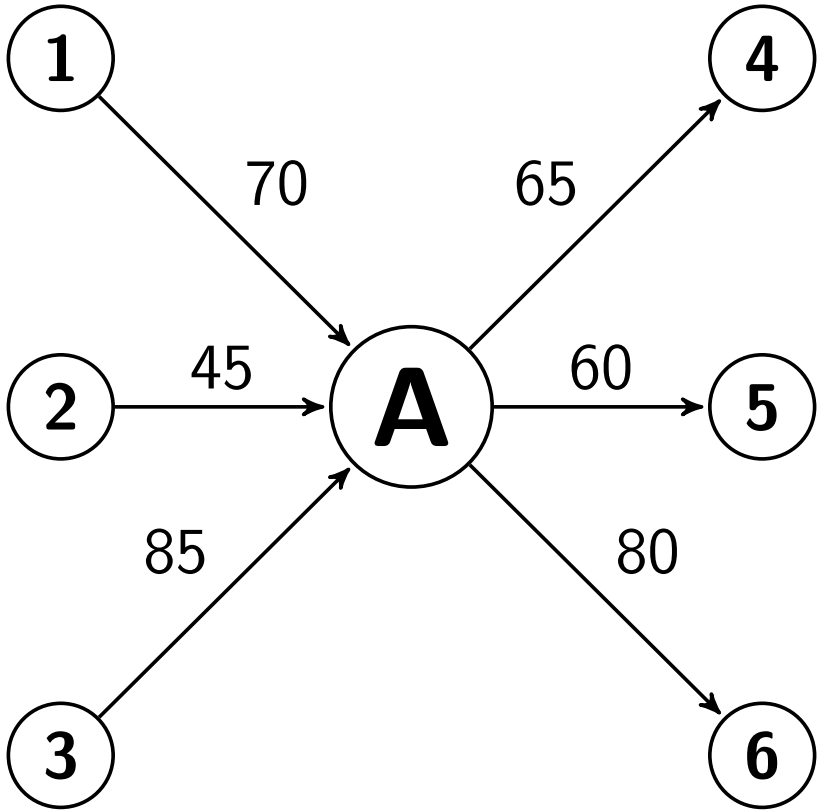
What is PeerRank

- Grader's ability to grade is measured by his/her own grade



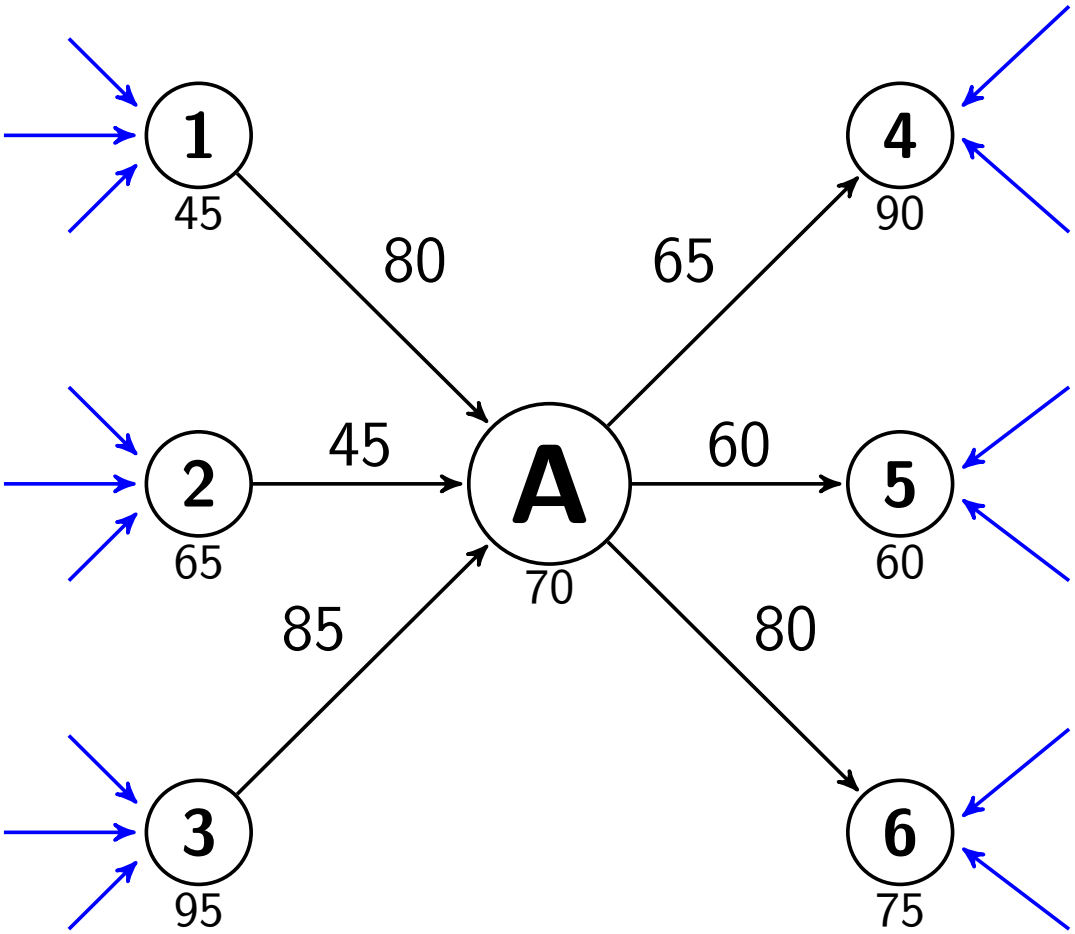
What is PeerRank

- Grader's ability to grade is measured by his/her own grade



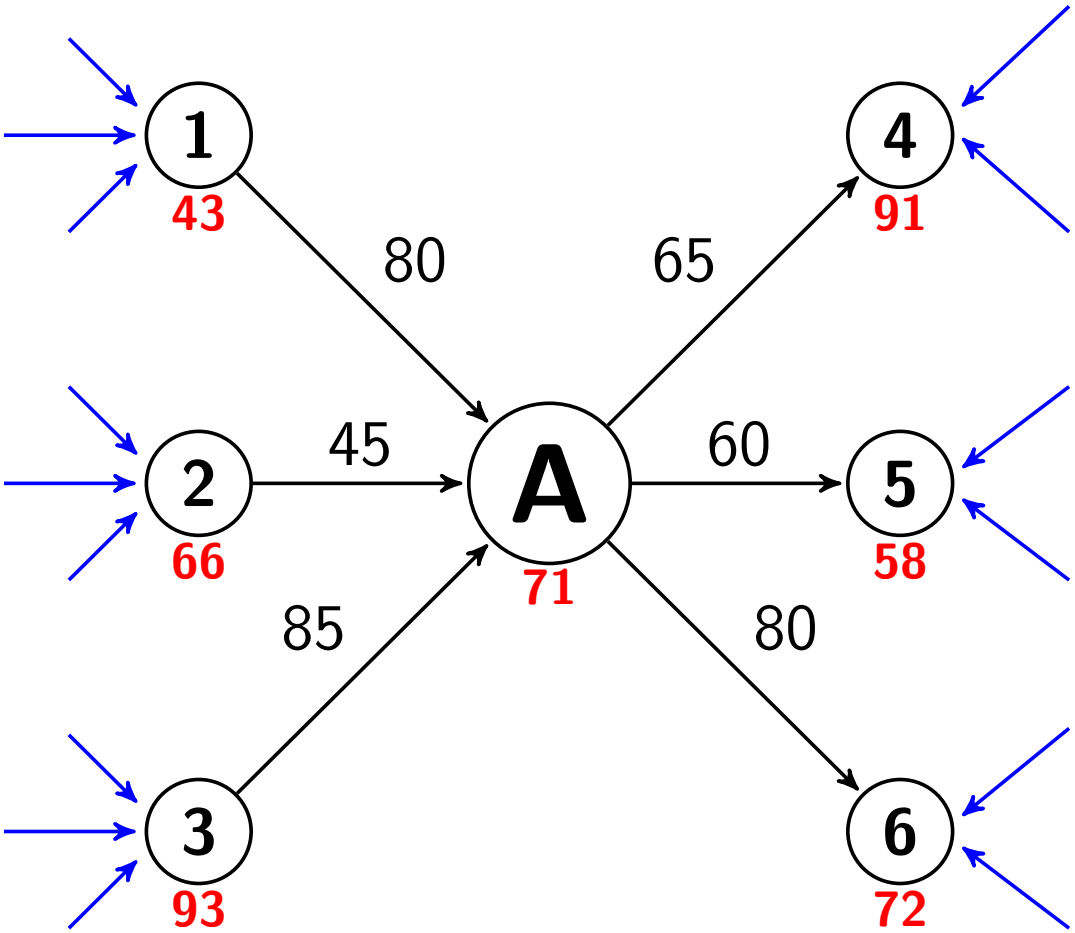
What is PeerRank

- Grader's ability to grade is measured by his/her own grade



What is PeerRank

- Grader's ability to grade is measured by his/her own grade



Mathematical Representation

$X_i^0 =$ Average of grades given to peer i

$\vdots$ $\vdots$

$X_i^n = X_i^{n-1} + \mathcal{A} + \mathcal{B}$

$X_i^{n+1} = X_i^n + \mathcal{A} + \mathcal{B}$

$\vdots$ $\vdots$

$X_i^k =$ GPR grade of peer i

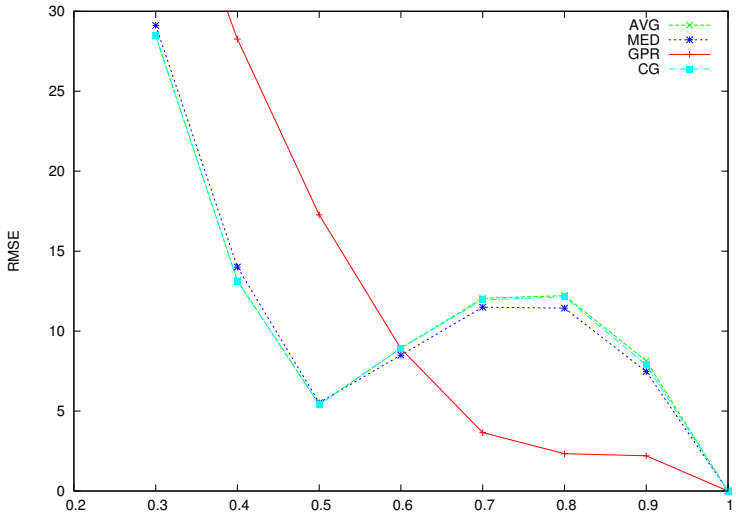
Experiments

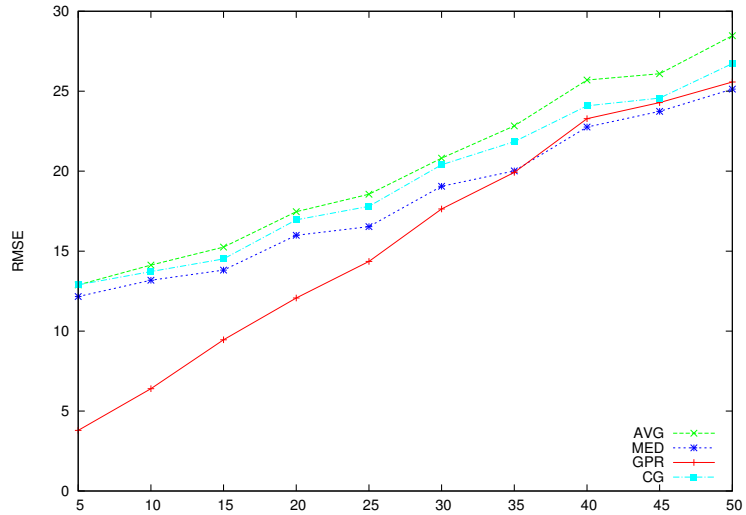
- Purpose of the experiments
  - Validation of the original GPR experiments
  - Comparing GPR to various other cardinal and ordinal peer assessment methods over synthetic data.

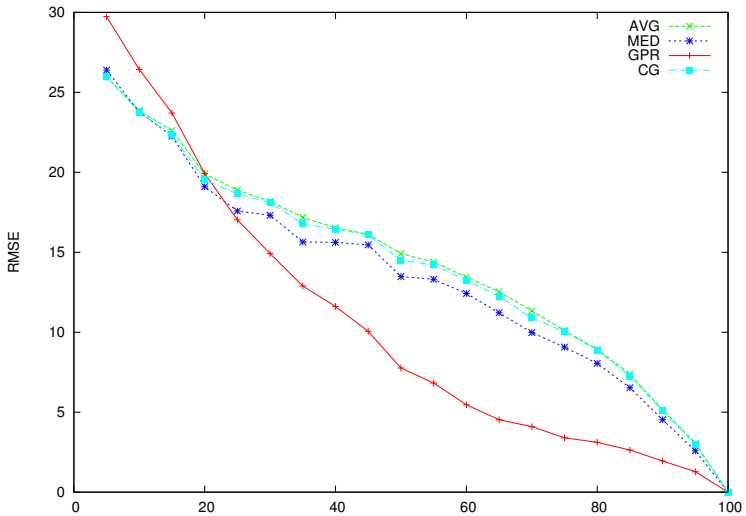
Experiment Setup

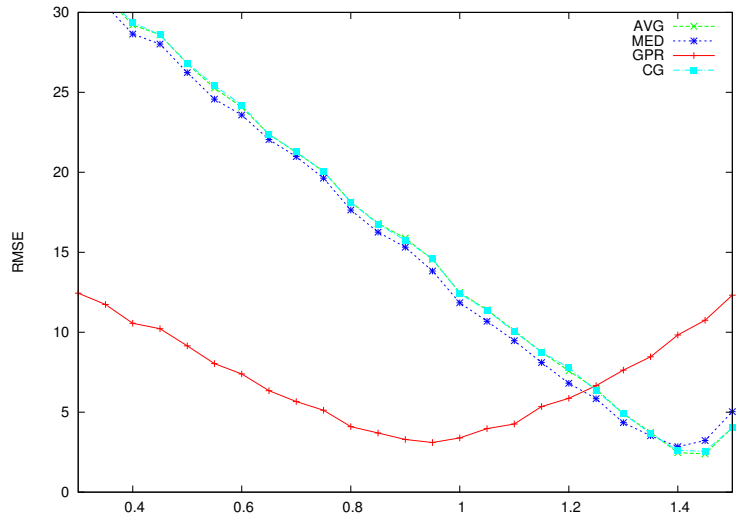
- 100 students each grading 5 others
- Methods are measured using RMSE which is simply the absolute value difference in grades

Validation of the Original Experiments









Results from the Original Experiments

GeneralizedPeerRank's performance is;

1. in most cases more powerful than primitive Average and Median

Results from the Original Experiments

GeneralizedPeerRank's performance is;

1. in most cases more powerful than primitive Average and Median
2. heavily dependent on the knowledge amongst peers ie. the avg grade

Results from the Original Experiments

GeneralizedPeerRank's performance is;

1. in most cases more powerful than primitive Average and Median
2. heavily dependent on the knowledge amongst peers ie. the avg grade
3. proportional to the deviation in the knowledge amongst peers

Results from the Original Experiments

GeneralizedPeerRank's performance is;

1. in most cases more powerful than primitive Average and Median
2. heavily dependent on the knowledge amongst peers ie. the avg grade
3. proportional to the deviation in the knowledge amongst peers
4. mostly unaffected by the peers' bias in assessment

Comparing grades with rankings

Assume σ a ranking of 4 items and θ a set of grades assigned to those

	σ	θ
A	2	86
B	3	24
C	4	71
D	1	49

	σ	σ'	θ	θ'
A	2	66	86	1
B	3	33	24	4
C	4	0	71	2
D	1	100	49	3

	σ	θ'
A	2	1
B	3	4
C	4	2
D	1	3

In order to compare a ranking σ and a grading θ format conversion is necessary

- Convert ranking to grading by grading on a curve?
  - Induces extra error on the already disadvantageous ranking methods
- Convert grades to a ranking and compare

Spearman's Footrule & RMSE

	σ	θ'
A	2	1
B	3	4
C	4	2
D	1	3

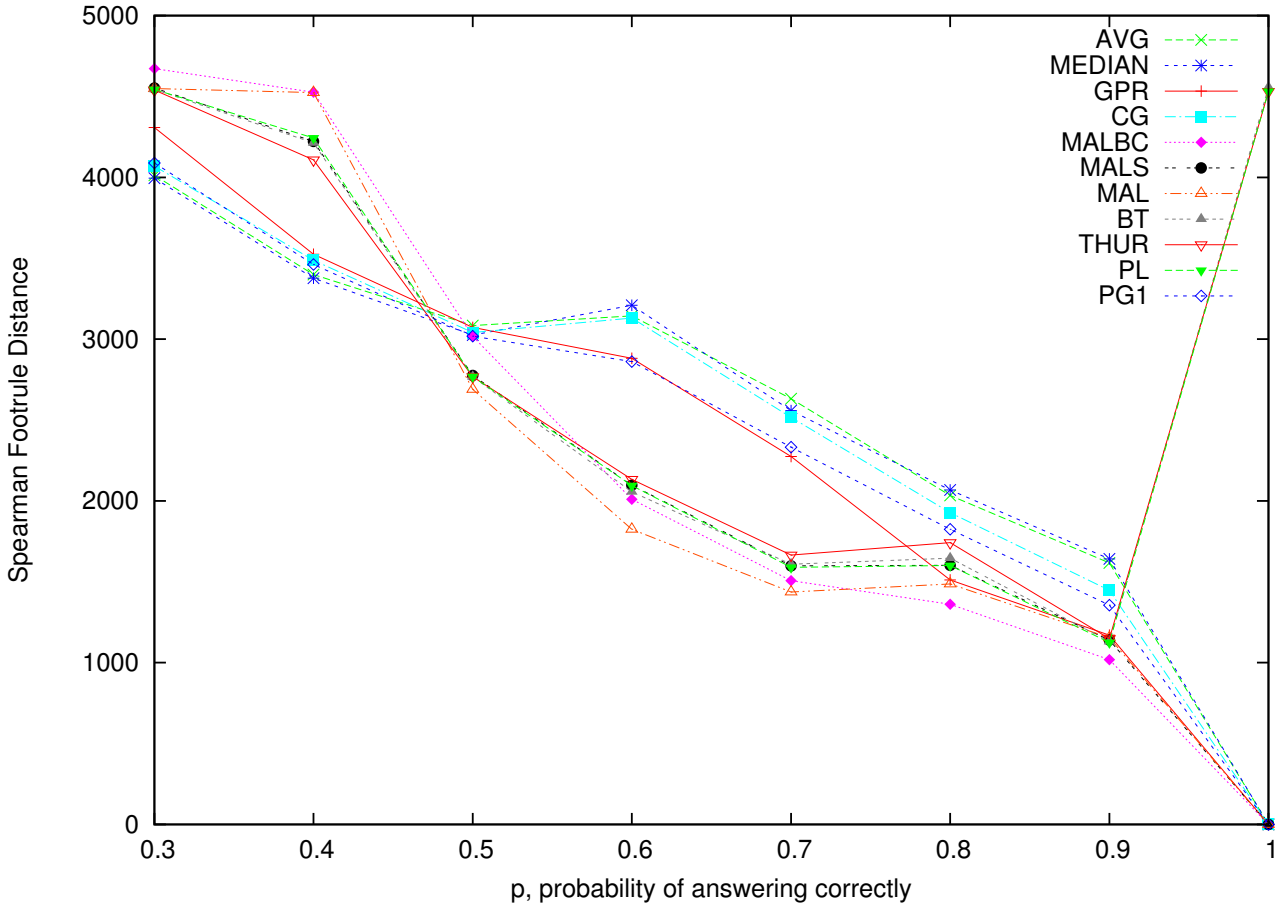
Spearman footrule distance between σ and θ' is

$$|2-1|+|3-4|+|4-2|+|1-3| = \mathbf{6}$$

Experiment Setup

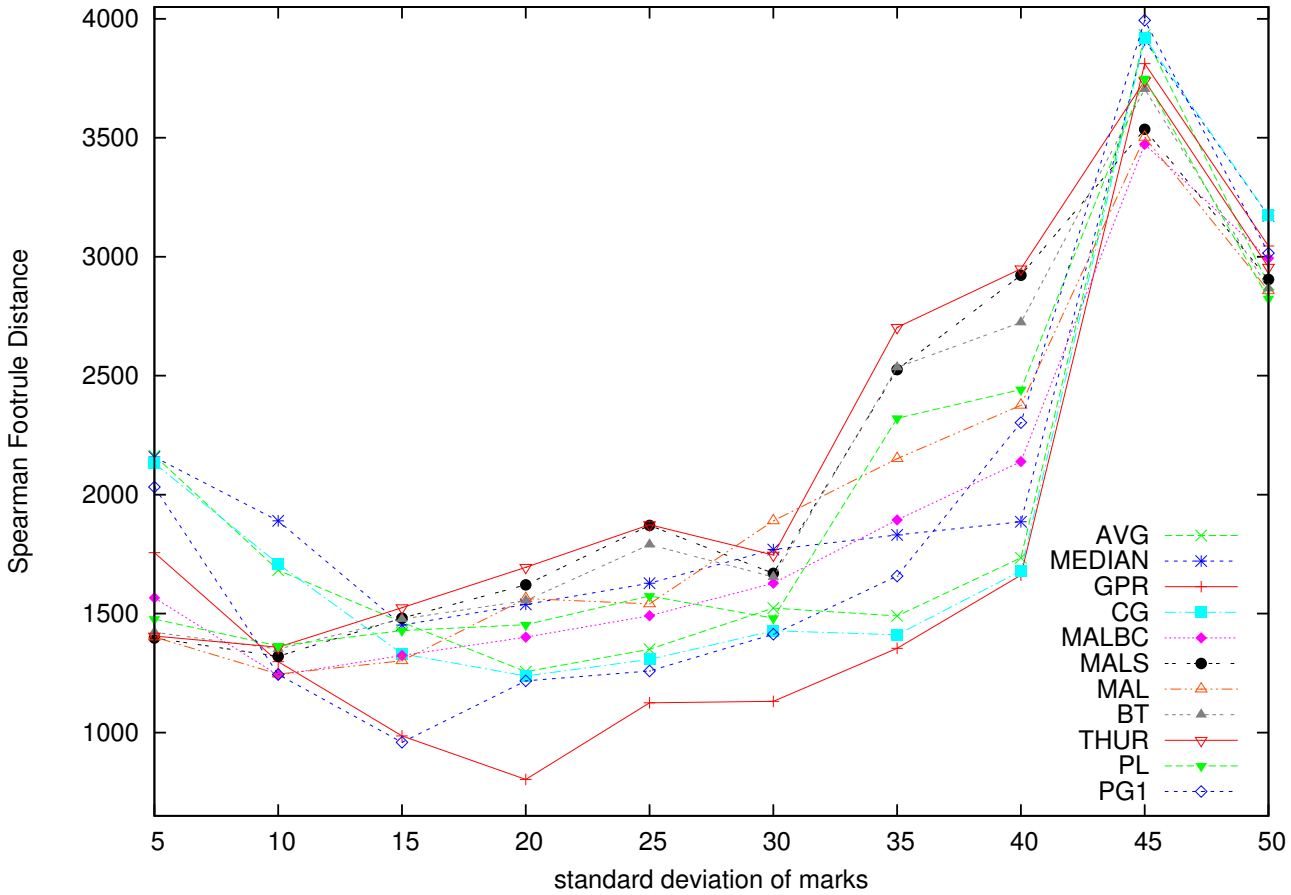
- 500 students each grading 5 others
- Grades are sampled from binomial, uniform and normal distributions in order to simulate a classroom
- Cardinal methods are converted to ranking and measured using Spearman Footrule Distance together with Ordinal methods

Experiment 1 : Binomially Distributed Grades



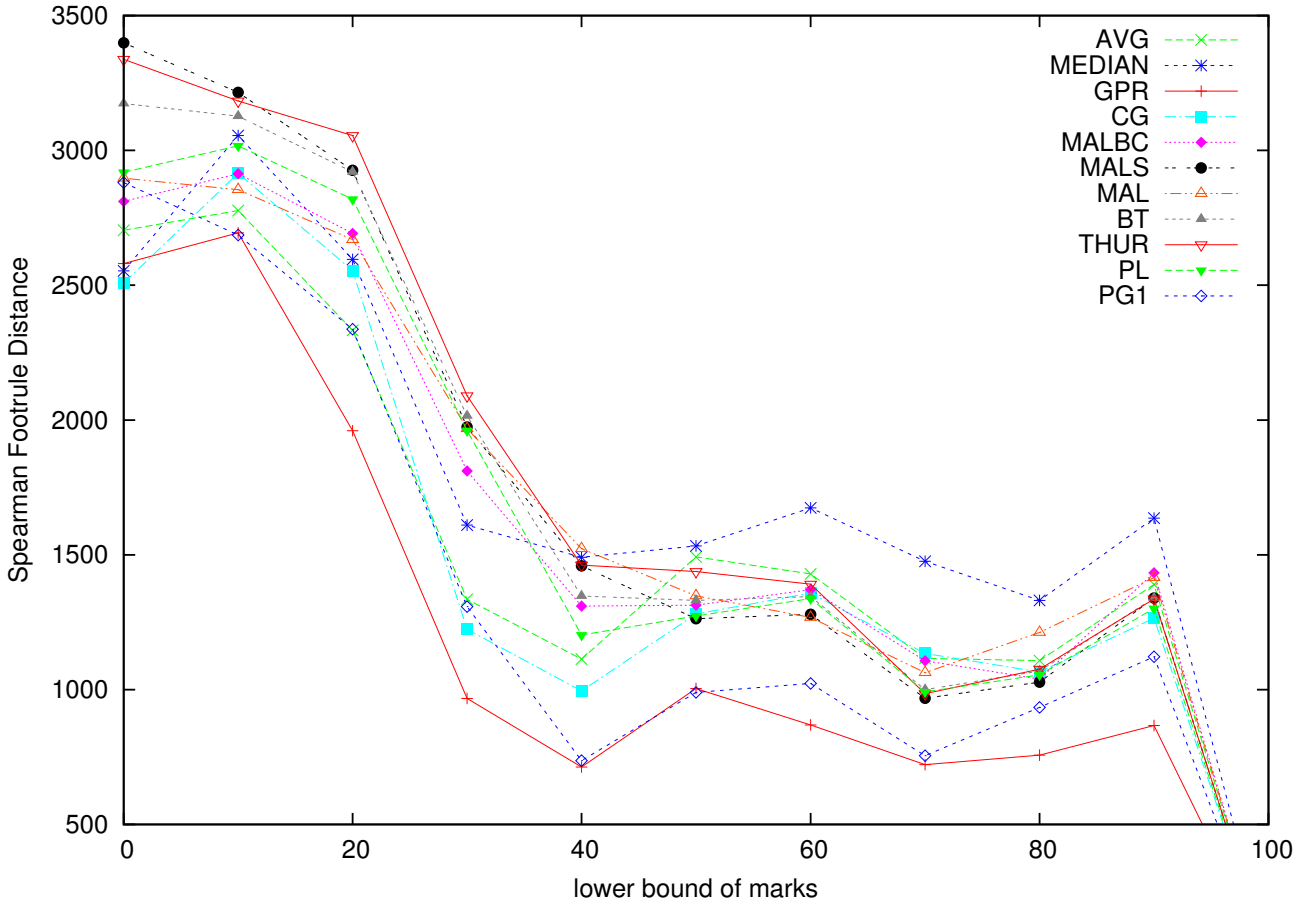
- Methods form into two clearly distinguishible groups as Ordinal and Cardinal without any outperformers
- Surprisingly Ordinal methods perform better

Experiment 2 : Normally Distributed Grades



- No distinguishable separations amongst methods
- GPR is the overall best except with very high and very low deviations

Experiment 3 : Uniformly Distributed Grades



- GPR is better for most of the lower bound values
- Similarly to the original experiments, dependence on the knowledge amongst peers continues

Conclusion

- Peer assessment can assign grades very accurately
  - 5 RMSE on average
  - Or within 2 Spearman Footrule Distance among 500
- Cardinal methods are generally more powerful
- GeneralizedPeerRank is the best method for peer assessment

Future Work

- Experiment with Real Data
- Possible Ways of Extension
  - Temporal Coherence: Correlation between two homeworks given to a student at different times
  - Have peers evaluate each subject separately to gather additional information

Thank you for listening

Any Questions?