The Effects of Homophily on the Arbitrariness of Peer Review

Aliaksandr Birukou **Elise S. Brezis**

PEERE International Conference on Peer Review

07-09 March 2018

1. Motivation of the paper

- The NIPS experiment
  - ➢2014 PC chairs split PC in two, forming 2 committees.
  - ➢Acceptance rate was pre-defined: (22%) - 37 out of 166 papers.
  - ➢two committees disagreed on 43 papers (26%).
  - ➢Moreover, for accepted papers: disagreed on 21/37 (57%).

NIPS 2014 Call For Papers

Neural Information Processing Systems Conference and Workshops December 8-13, 2014 Montreal Convention Center, Montreal, Canada

1. Motivation of the paper

- The purpose of our work:
  - ➢Analyze the arbitrariness of Peer Review
  - ➢Reproduce and Explain the outcomes of the NIPS experiment
  - ➢Show that the papers with very innovative ideas can suffer from peer review.

2. Main assumptions

A. Homophily: from Ancient Greek ὁμοῦ (homou, “together”) and Greek φιλία (philia, "friendship") is the tendency of individuals to associate and bond with similar others “

The model is based on the concepts of homophily:

- ➢ Reviewers have personal bias
- ➢ Ideas closer to one’s mental model are valued more

- ➢ We assume that reviewers are different in their taste for innovation, and it influences their grades

2. Main assumptions

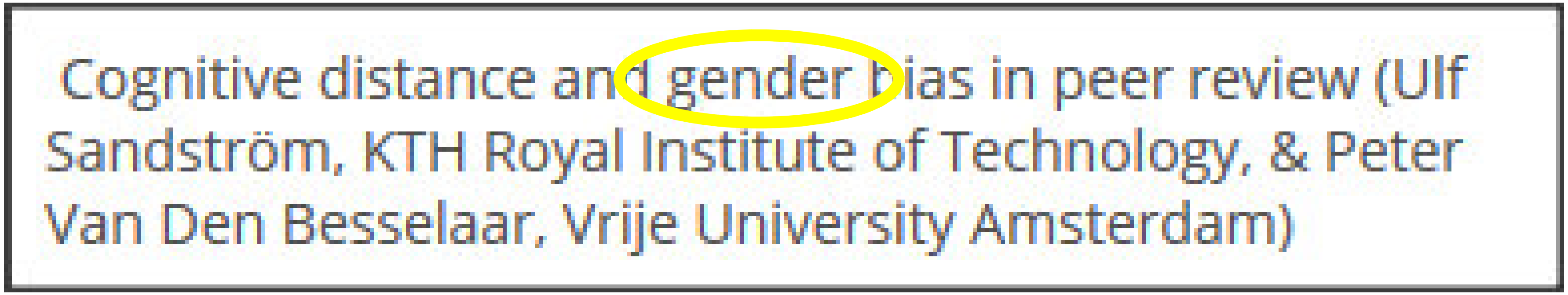
B. Reviewers who do not invest enough time in the review process make mistakes on the “true” value/quality of the project

- ➢ Moreover, referees are not investing the same amount of time to analyze the projects.

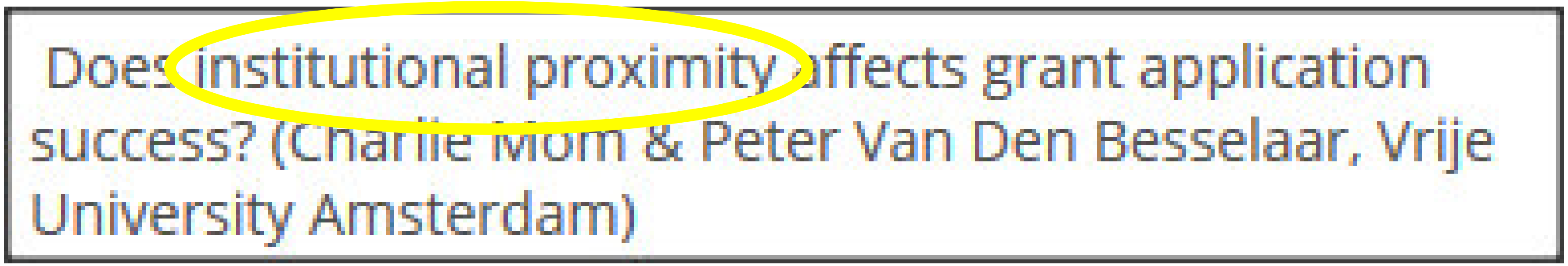
There is heterogeneity between reviewers

Footnote 1

- Examples of reviewer heterogeneity in Day 1 talks:



Cognitive distance and gender bias in peer review (Ulf Sandström, KTH Royal Institute of Technology, & Peter Van Den Besselaar, Vrije University Amsterdam)



Does institutional proximity affects grant application success? (Charlie Mom & Peter Van Den Besselaar, Vrije University Amsterdam)

Footnote 2

- Peer Review are used for selecting:
  - best projects, e.g. Horizon 2020.
  - best papers for a conference, e.g., NIPS
  - best papers for a journal, e.g., Nature.

  - Rankings based on peer review are used only for the first two.

3. Main results

A.

Analyzing the arbitrariness of peer review
and
Reproducing the results of the NIPS experiment

B.

Policy results

B. Policy results

1. More referees do not improve the peer review process
   - in fact, with more referees, worse projects were accepted
2. More specific guidance/criteria to referees does not improve the peer review process

3. Lower acceptance rate disadvantages innovation and does not improve the peer review process

A. Reproducing the results of the NIPS experiment

- In our model:

	Our model	NIPS	Horizon 2020
Projects reviewers disagreed on	40%	26%	
Accepted projects – disagreed	70%	57%	
Acceptance rate	30%*	22%	1.8%

- 2/3 of chosen projects were not the best and not most innovative.

* this acceptance rate is still higher than in some H2020 calls (1.8%) or <10% acceptance rate fashionable in computer science conferences.

B. Policy results

- More specific guidance/criteria to referees does not improve the peer review process
  - ➢We analyze review criteria in CS conferences
  - ➢There are 12 in total
    - some conferences use up to 6
  - ➢We show that we can group them in 3 categories
    - Soundness / Presentation
    - Contribution / Validity
    - Innovation

3 dimensions capture it all

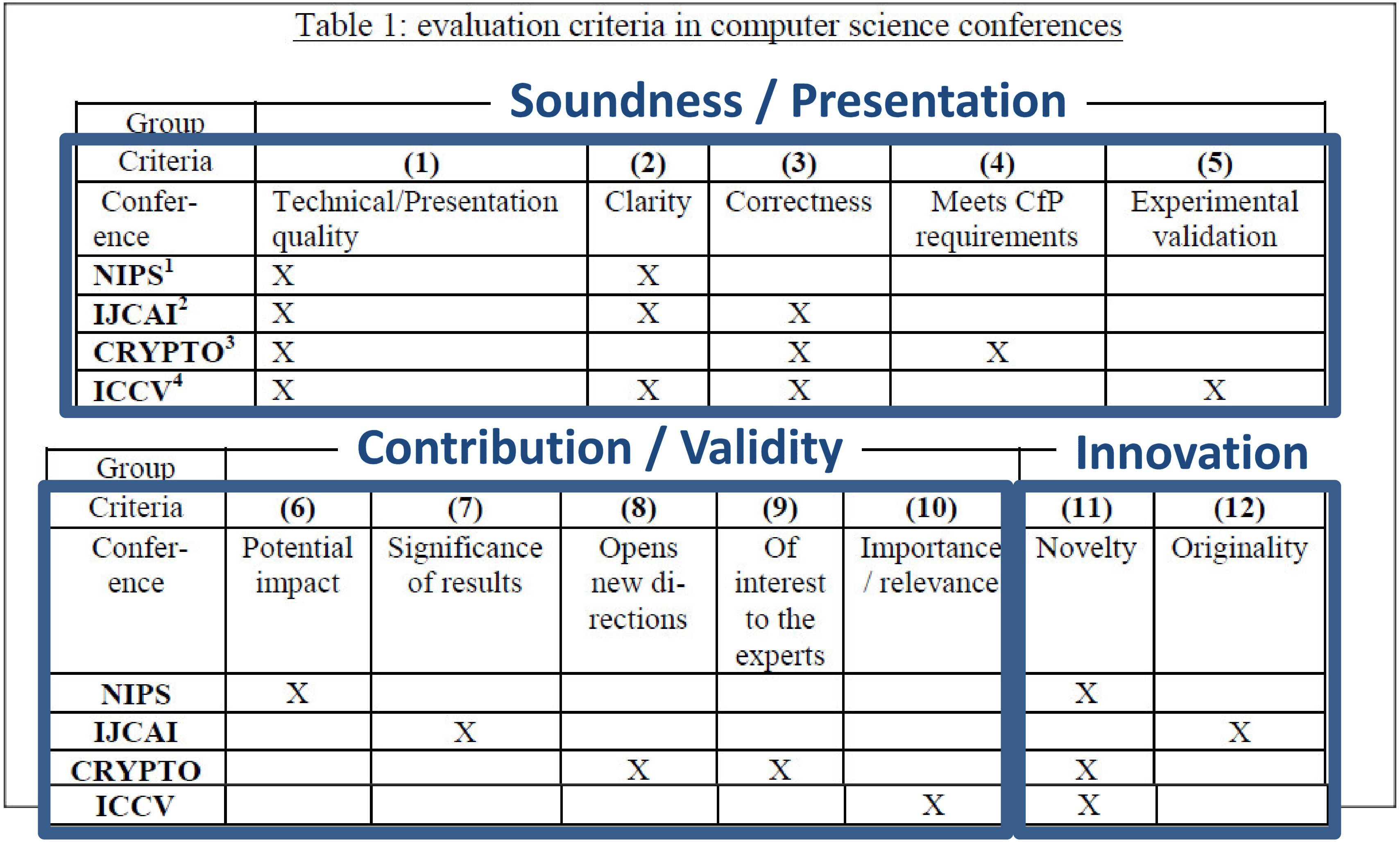


Table 1: evaluation criteria in computer science conferences

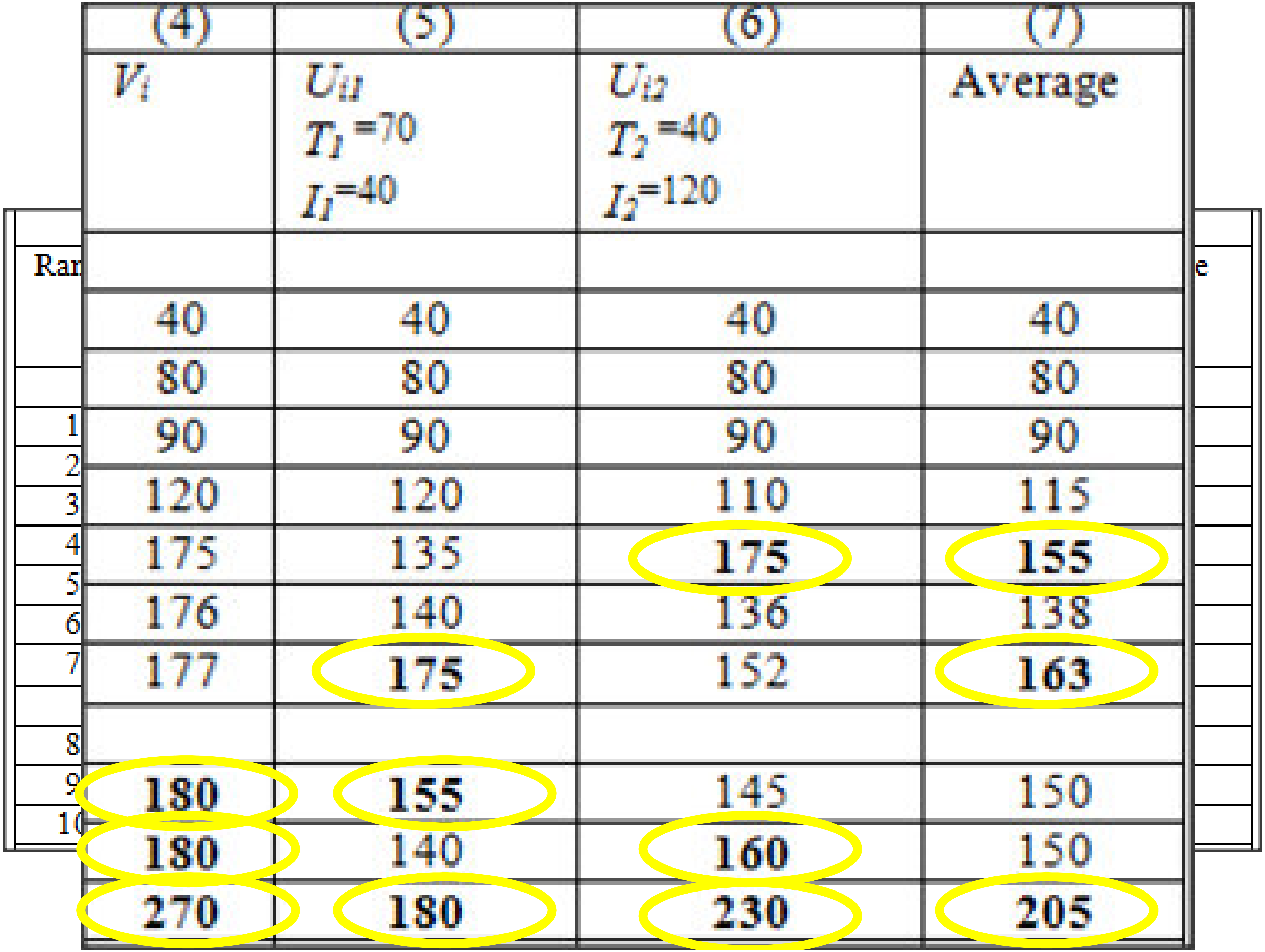
Group	Soundness / Presentation				
Criteria	(1)	(2)	(3)	(4)	(5)
Conference	Technical/Presentation quality	Clarity	Correctness	Meets CfP requirements	Experimental validation
NIPS[1]	X	X			
IJCAI[2]	X	X	X		
CRYPTO[3]	X		X	X	
ICCV[4]	X	X	X		X

Group	Contribution / Validity					Innovation	
Criteria	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Conference	Potential impact	Significance of results	Opens new directions	Of interest to the experts	Importance / relevance	Novelty	Originality
NIPS	X					X	
IJCAI		X					X
CRYPTO			X	X		X	
ICCV					X	X	

More about our results

How do we get our results?

We model the decision making of referees given their distribution on homophily as well as time devoted to peer review.



(4)	(5)	(6)	(7)
V_i	U_{i1} T_1 =70 I_1=40	U_{i2} T_2 =40 I_2=120	Average
40	40	40	40
80	80	80	80
90	90	90	90
120	120	110	115
175	135	**175**	**155**
176	140	136	138
177	**175**	152	**163**
180	**155**	145	150
180	140	**160**	150
270	**180**	**230**	**205**

Conclusions

1. This paper shows that picking papers/projects based on peer review is quite arbitrary, due to **heterogeneity of reviewers.**

- The arbitrariness is of almost 50%.

2. Our policy results:

- Adopting more criteria, or asking for more referees is not improving the results (quite counter-intuitive!)

3. Less tightness of acceptance leads to accept on average, better projects/papers.

What do we learn from this?

Ratings are not robust!

In peer review: less is more!!