

# An Overview of Challenges, Experiments, and Computational Solutions in Peer Review

**Nihar B. Shah**

Machine Learning and Computer Science Departments

**Carnegie Mellon University**



"Piled Higher and Deeper" by Jorge Cham

WWW.PHDCOMICS.COM

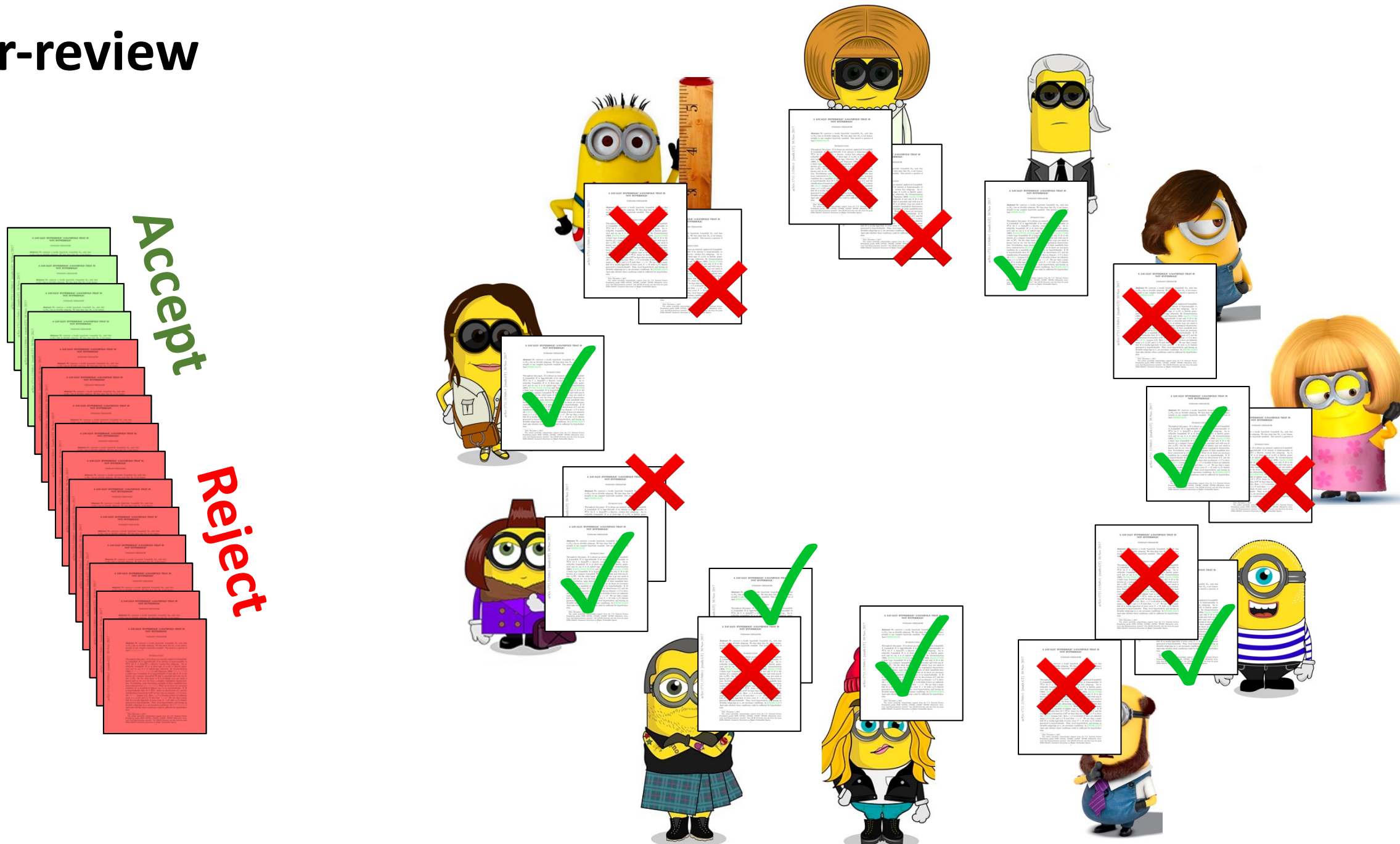
# Logistics

- Overview article with references:

<https://www.cs.cmu.edu/~nihars/preprints/SurveyPeerReview.pdf>

- On all slides, references are clickable and link to the paper

# Peer-review



# Peer-review for grant proposals



Budget of several billions of \$

# Peer-evaluation of employees at companies



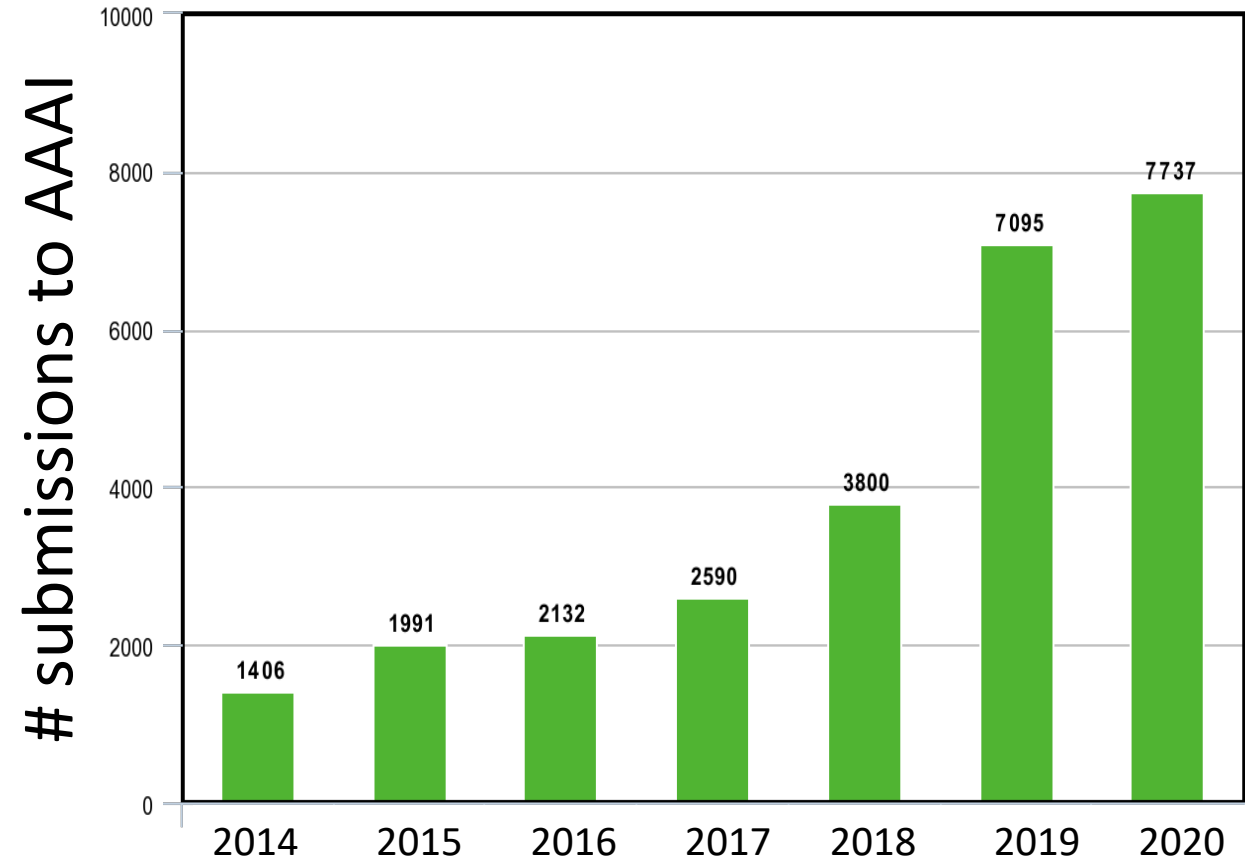
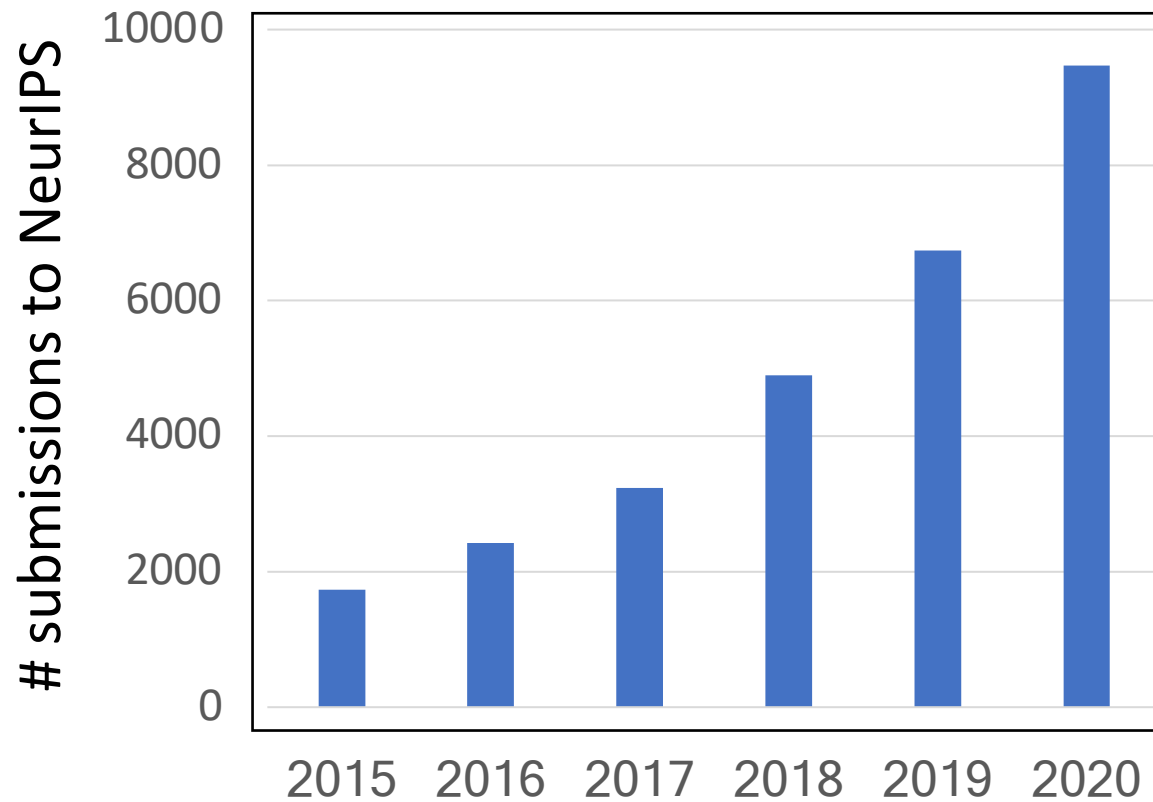
Peer Review Feedback: The Good,  
Bad, The Really Ugly



Can make or break careers



# Several thousand submissions, exponential growth



**Increase in #submissions in many other fields:**

*“Submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint”* [McCook 2006]

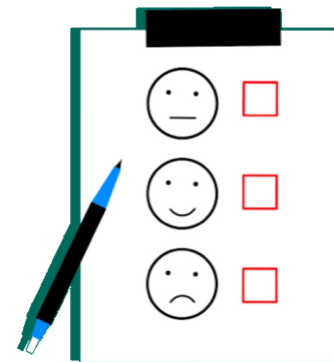
# Challenge across many research fields

- “Let's make peer review scientific” [Rennie, Nature 2016]

*“Peer review ... is a human system. Everybody involved brings **prejudices**, **misunderstandings** and gaps in knowledge, so no one should be surprised that peer review is often **biased** and **inefficient**. It is occasionally **corrupt**, sometimes a charade, an open temptation to plagiarists. Even with the best of intentions, how and whether peer review identifies high-quality science is unknown. It is, in short, **unscientific**.”*

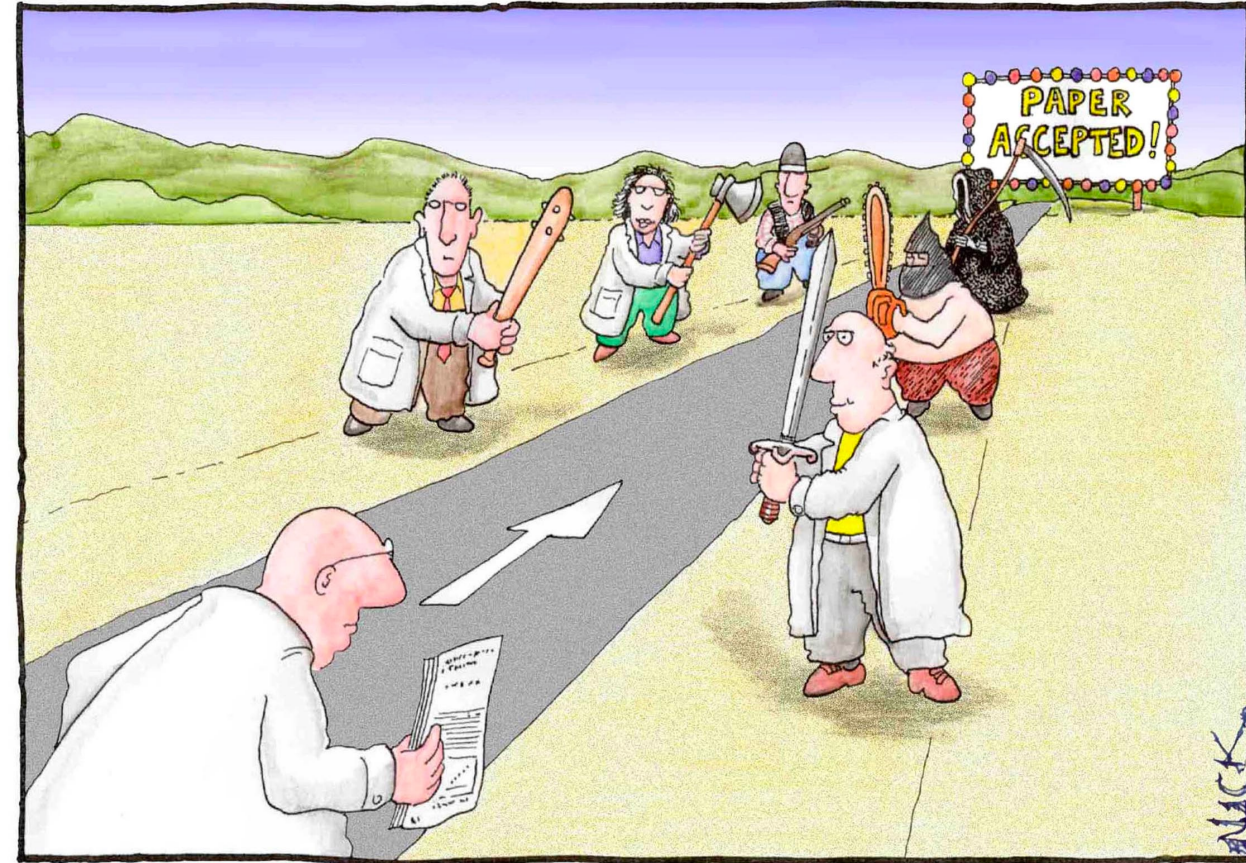
- Overwhelming desire for improvement

[surveys by Smith 2006, Ware 2008, Mulligan et al. 2013]



# Hurts scientific progress

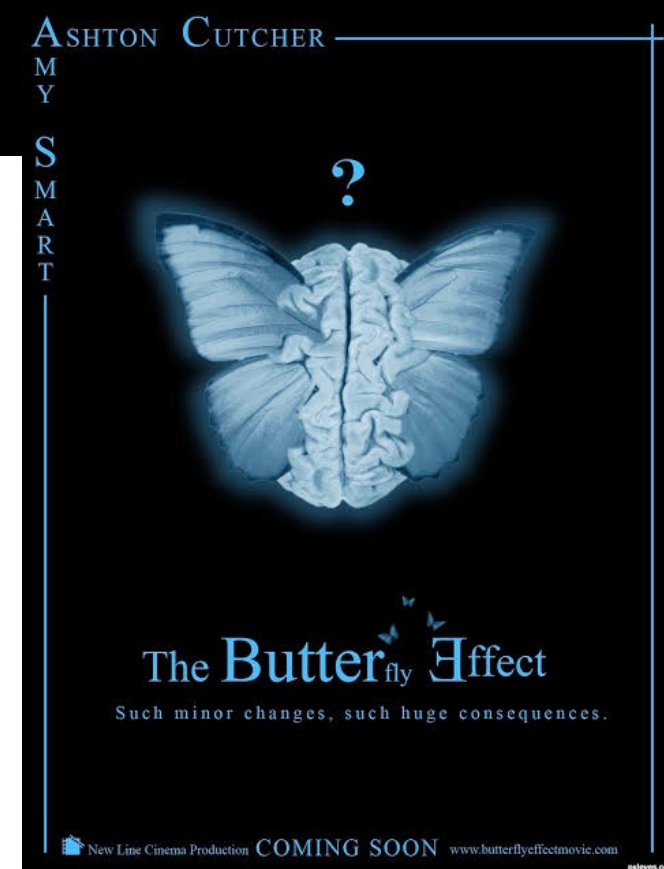
“interdisciplinary research, frontier science, areas of controversy, and risky new departures are all more likely to **suffer from cognitive cronyism** than is mainstream research” [Travis et al. 1991]



# Hurts careers

“an incompetent review may lead to the rejection of the submitted paper, or of the grant application, and the ultimate **failure of the career of the author.**” [Triggle et al. 2007]

“These long term effects arise due to the widespread prevalence of the Matthew effect (‘**rich get richer**’) in academia” [Merton 1968]





# Harms public perception of science



**Donald J. Trump** ✓  
@realDonaldTrump



Follow

Massive combined inoculations to small children is the cause for big increase in autism....

RETWEETS  
**605**

LIKES  
**264**





# Broad applicability

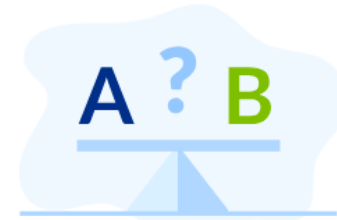
## Distributed human evaluations



Hiring



Admissions



A/B testing



Crowdsourcing



Product ratings



Healthcare



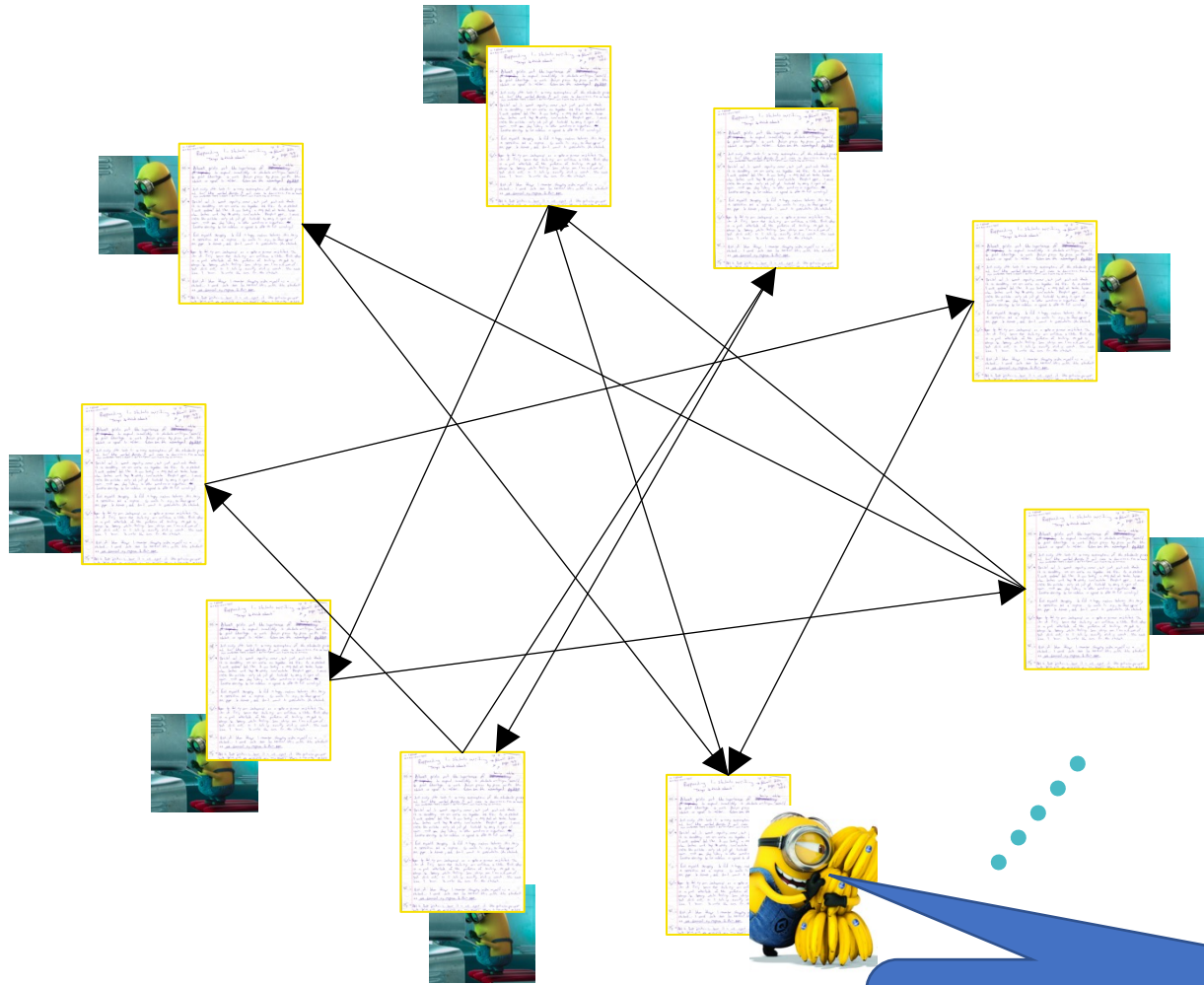
Peer grading

...

Problems amplify when this data is used to train AI/ML systems!

- **Noise**
- **Fraud**
- **Bias**
- **Miscalibration**
- **Subjectivity**
- **Norms and policies**

# Noise



I don't know much about this area.  
Weak reject I guess...

# Noise and reviewer assignment

## Poor reviews due to **inappropriate choice of reviewers**

“one of the first and **potentially most important** stages is the one that attempts to distribute submitted manuscripts to competent referees.” [[Rodriguez et al. 2007](#)]

**Top reason for dissatisfaction:** “Reviewers or panelists not expert in the field, poorly chosen, or poorly qualified” [[McCullough 1989](#)]



# Automated assignment

(Used in AAAI, NeurIPS, ICML,...)

**Compute  
similarities**

[[Mimno et al. 2007](#),  
[Rodriguez et al. 2008](#), [Charlin  
et al. 2013](#), [Liu et al. 2014](#)]



**Assignment**

- For every pair (paper  $p$ , reviewer  $r$ ), similarity score  $s_{pr} \in [0, 1]$
- Higher similarity score  $\Rightarrow$  Better envisaged quality of review
- Based on
  - Match text of submitted paper with reviewer's past papers
  - Match chosen subject areas
  - Reviewer bids
- Use similarity scores to assign reviewers to papers...



# Assignment: Maximize total similarity

(Used in AAAI, NeurIPS, ICML,...)

$$\text{maximize}_{\text{assignment}} \sum_{p \in \text{Papers}} \sum_{r \in \text{Reviewers}} s_{pr} \mathbb{I}\{\text{paper } p \text{ assigned to reviewer } r\}$$

subject to

Every paper gets at least certain #reviewers

Every reviewer gets at most certain #papers

No paper is assigned to conflicted reviewer

[Conference management systems: TPMS (Charlin and Zemel 2013), EasyChair, HotCRP]

[Goldsmith et al. 2007, Taylor 2008, Tang et al. 2010, Charlin et al. 2012, Long et al. 2013]

# Toy example

- One reviewer per paper
- One paper per reviewer

	Paper A	Paper B	Paper C
Reviewer 1	1	0	0.5
Reviewer 2	0.7	1	0
Reviewer 3	0	0.7	0

**Assignment is unfair to paper C**

**There exists another more balanced assignment**

# Another example

- Two reviewers per paper
- One paper per reviewer

	Paper A	Paper B	Paper C
Reviewer 1	0.9	0	0.5
Reviewer 2	0.6	0	0.5
Reviewer 3	0	0.9	0.5
Reviewer 4	0	0.6	0.5
Reviewer 5	0	0	0
Reviewer 6	0	0	0

**Assignment is unfair to (inter-disciplinary) paper C**

**There exists another more balanced assignment**

# Common approach: Maximize total similarity

$$\underset{\text{assignment}}{\text{maximize}} \sum_{p \in \text{Papers}} \sum_{r \in \text{Reviewers}} s_{pr} \mathbb{I}\{\text{paper } i \text{ assigned to reviewer } j\}$$

- **Unbalanced:** Can assign all relevant reviewers to some papers and all irrelevant reviewers to others [Stelmakh et al. 2018]
- **Can be particularly unfair** to interdisciplinary papers
- On CVPR 2017 data, assigns at least one paper **all reviewers with 0 similarity** (there are other assignments that do much better) [Kobren et al. 2019]

# More balanced assignment

$$\begin{array}{ll} \text{maximize} & \text{minimum} \\ \text{assignment} & p \in \text{Papers} \end{array} \sum_{r \in \text{Reviewers}} s_{pr} \mathbb{I}\{\text{paper } i \text{ assigned to reviewer } j\}$$

subject to

Every paper gets at least certain #reviewers

Every reviewer gets at most certain #papers

No paper is assigned to conflicted reviewer

Fix assignment for the worst-off paper  $\operatorname{argmin}_{p \in \text{Papers}}$

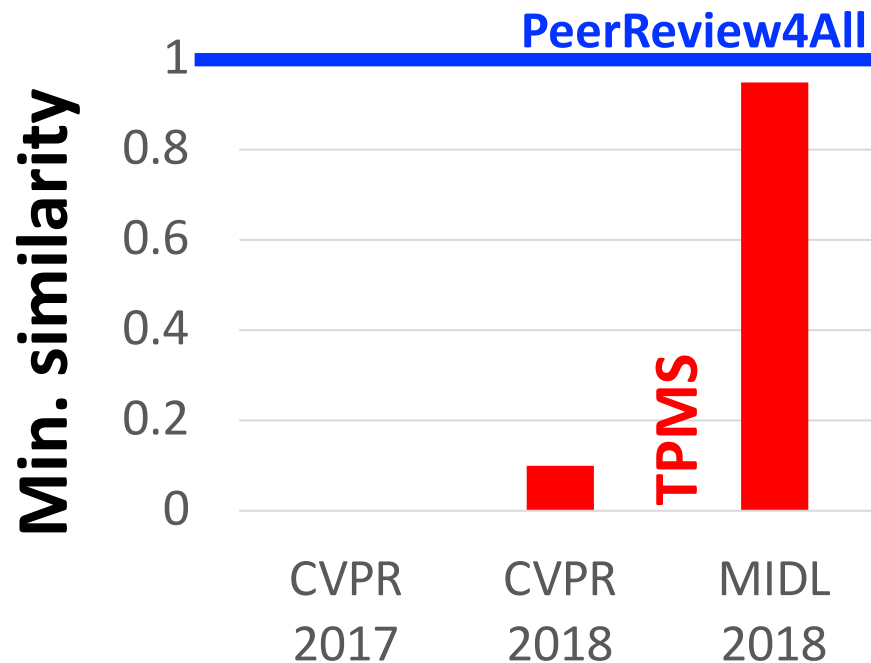
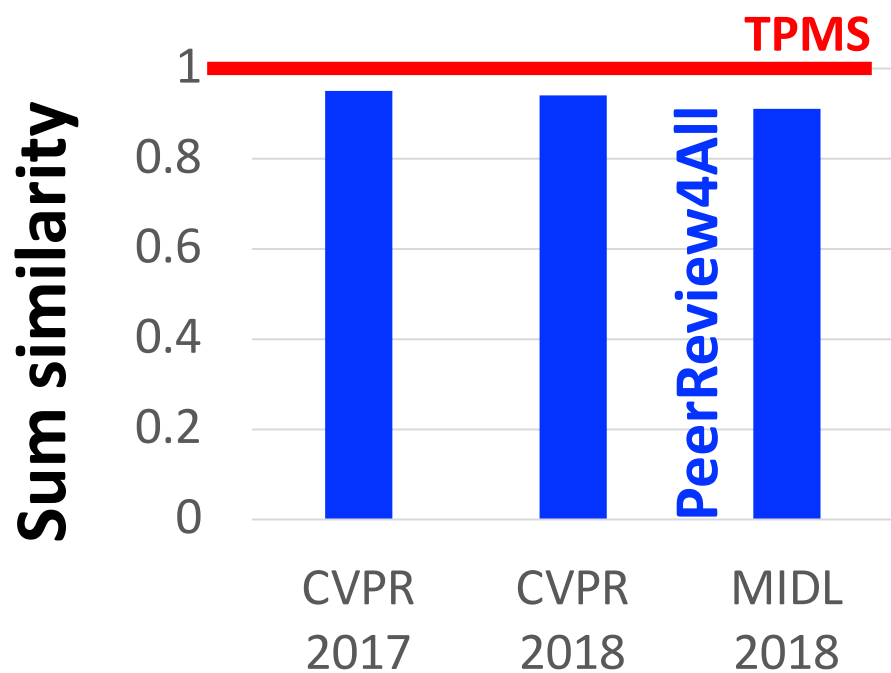
Repeat for remaining papers

- NP Hard [[Garg et al. 2010](#)]
- Approximation algorithm (“PeerReview4All”)
- Statistical guarantees on overall top-K selection



# Evaluation

- **TPMS algorithm** optimizes **sum similarity**
- **PeerReview4all algorithm** [Stelmakh et al. 2018] optimizes **minimum similarity**



[Evaluations by Kobren et al. 2019]

- **PeerReview4All used in ICML 2020: Outcome similar to above**

# Noise: Open problems

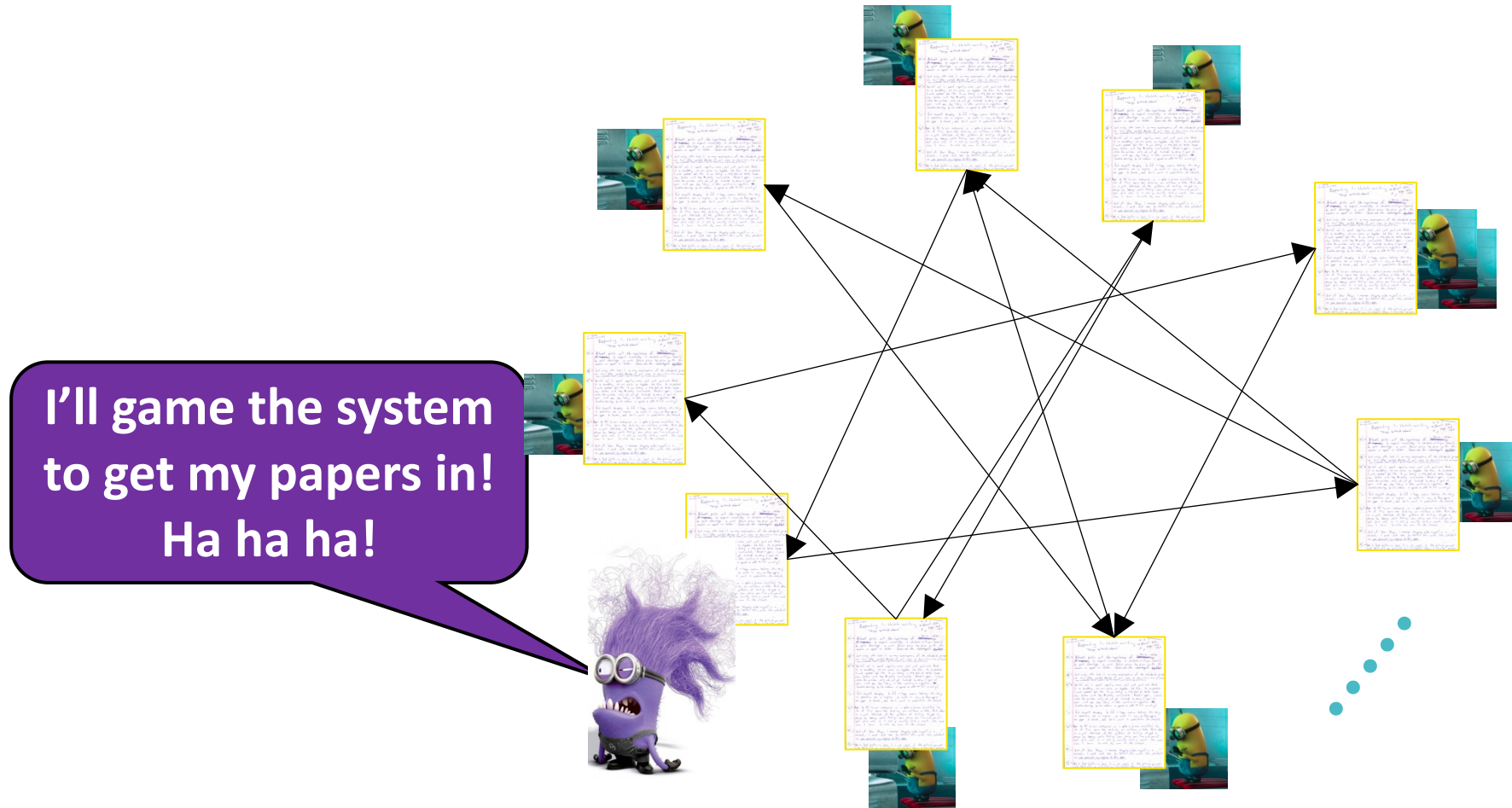


- Better computation of similarities
  - Interdisciplinary papers
  - Joint similarity computation and assignment

[[Mimno et al. 2007](#), [Rodriguez et al. 2008](#), [Charlin et al. 2013](#), [Liu et al. 2014](#), [Tran et al. 2017](#)]

- Denoise using text of reviews [[Fromm et al. 2020](#), [Cheng et al. 2020](#)]
- Computationally faster fair assignment with guarantees [[Stelmakh et al. 2018](#), [Kobren et al. 2019](#)]
- Fair and improved bidding process [[Fiez et al. 2019](#), [Meir et al. 2020](#)]

# Fraud



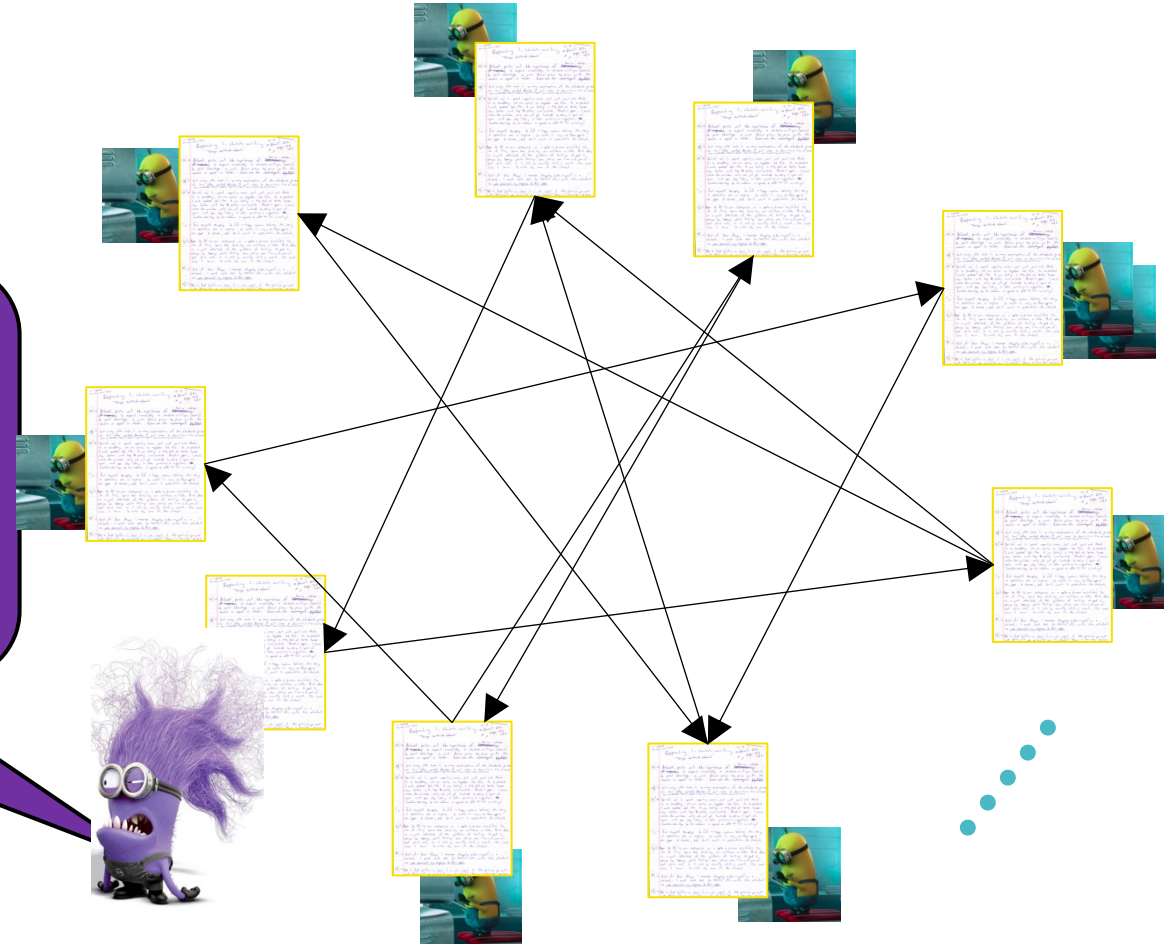
# **Fraud**

**1. Lone wolf**

**2. Coalition**

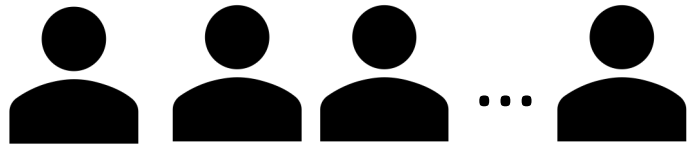
# Fraud: Lone wolf

Giving lower scores to other papers will increase chances of my own paper getting accepted! Ha ha ha ha!





# An experiment



1. Make a drawing
2. Enter one of 3 “exhibitions”
3. Peer review others’ drawings
4. Possibly win an award

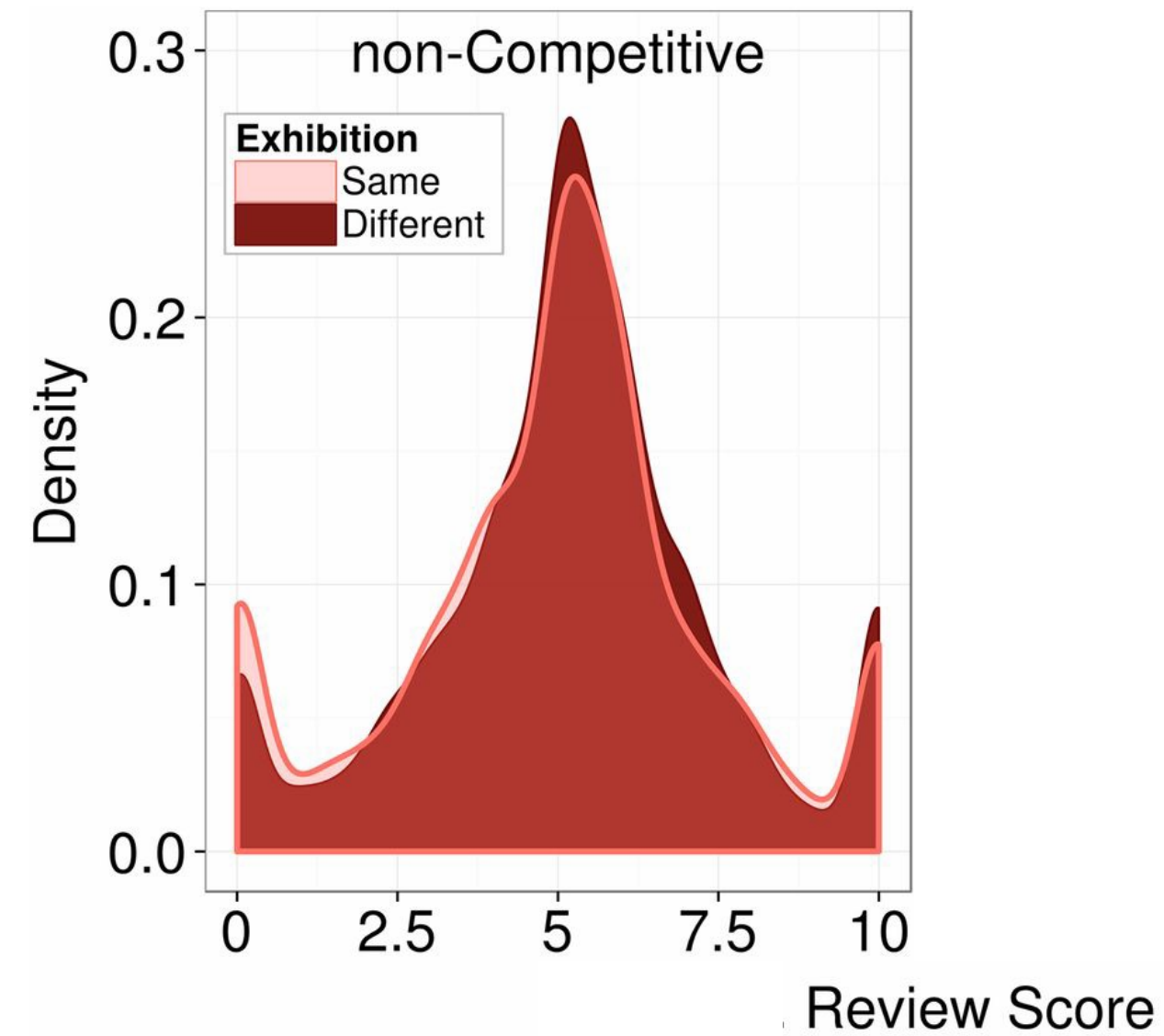
## Non-competitive

All above certain  
threshold get award

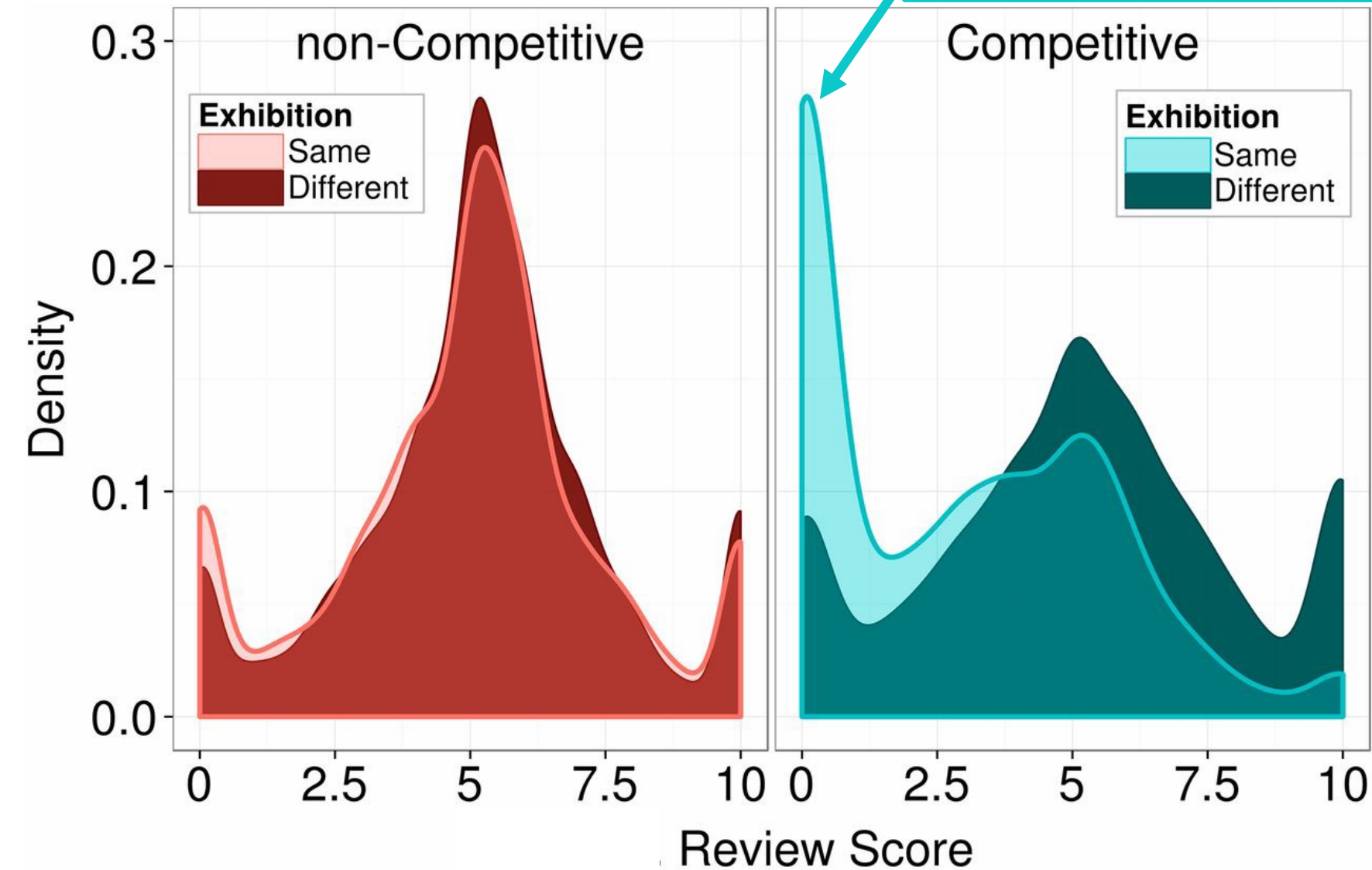
## Competitive

Top certain fraction in  
each exhibition win award

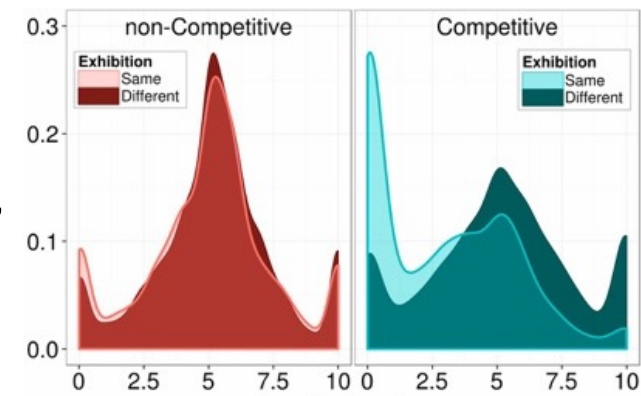
- Each participant knows which exhibition their drawing belongs, and if it is competitive or not
- Each participant also told the exhibition to which the drawings they are reviewing belong



Giving a lower score increases chances of their drawing getting an award



- “competitive sessions produce considerably more [strategic] reviews”
- “the number of [strategic] reviews increases over time”



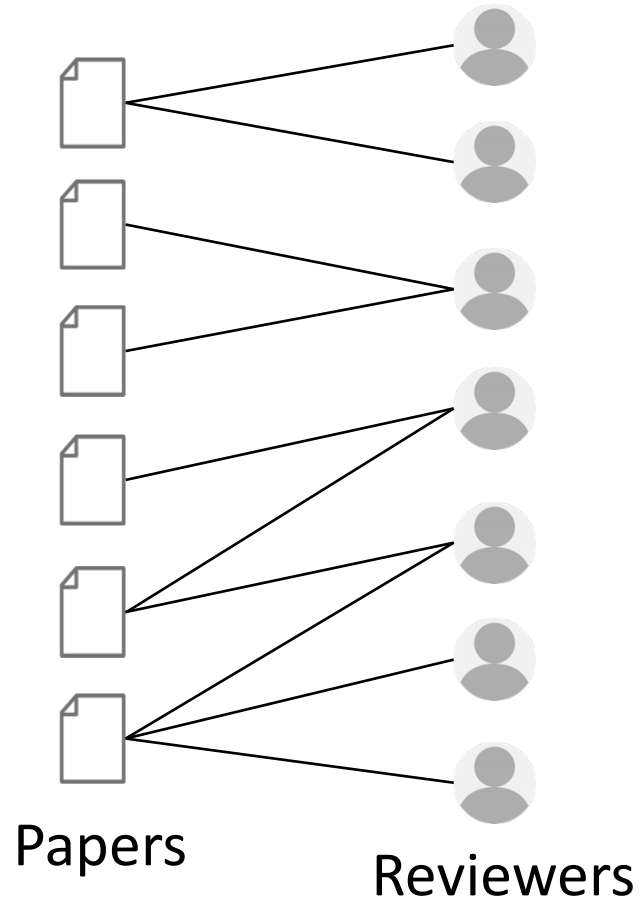
**“This result provides further evidence that a substantial amount of gaming of the review system is taking place... competition incentivizes reviewers to behave strategically, which reduces the fairness of evaluations”**

[Balietti et al., 2016]

Also [[Anderson et al. 2007](#), [Langford 2008 \(blog\)](#), [Akst 2010](#), [Turner and Hanel 2011](#)]

# How to make peer review strategyproof?

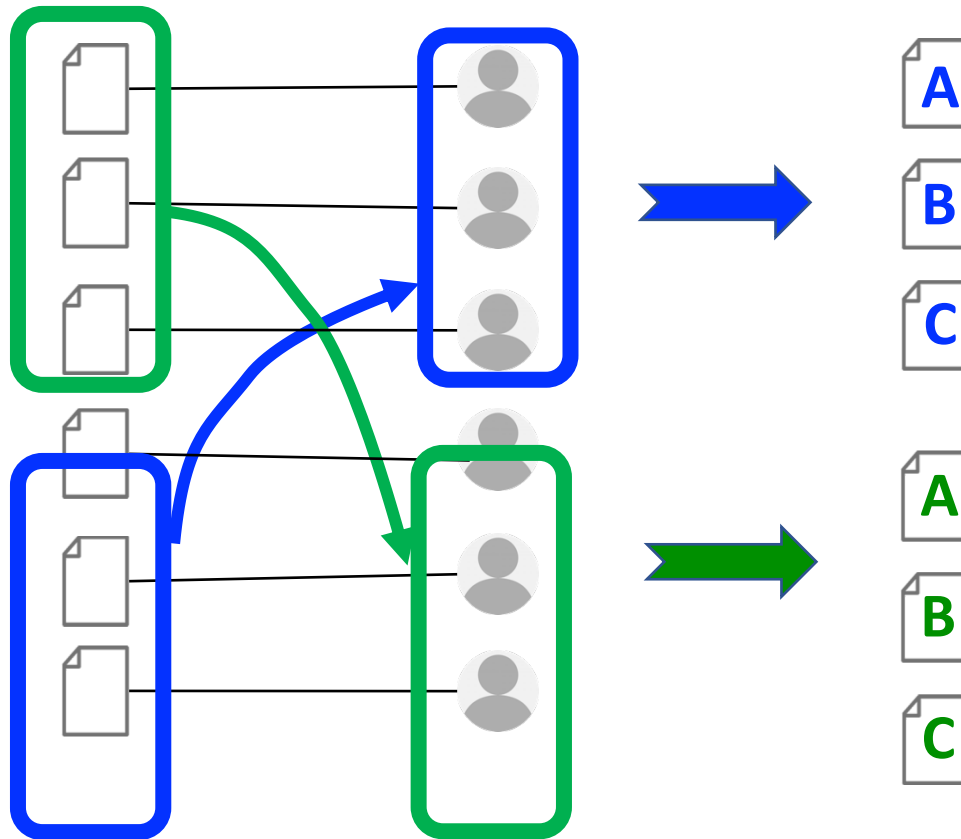
**Given:** Conflict graph  
(e.g., authorship graph)



**How to ensure that no reviewer  
can influence decision of any  
conflicted paper?**

# Partitioning method

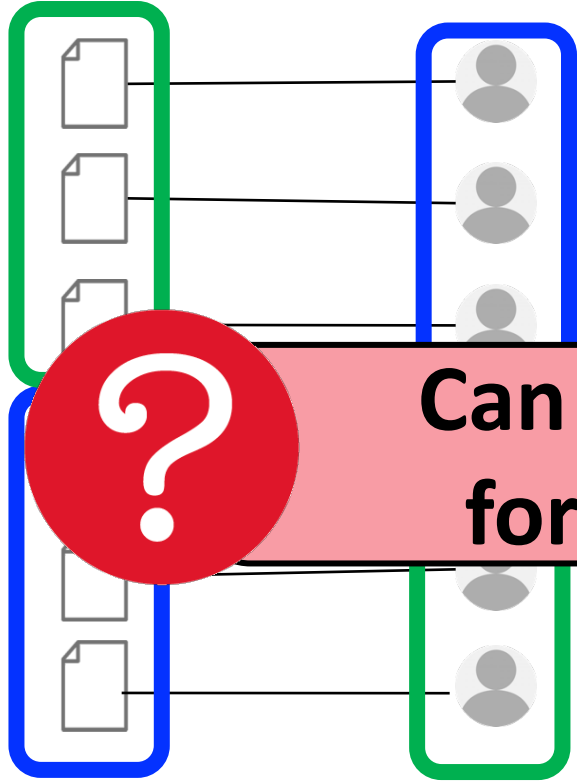
Primarily studied for peer grading



[Alon et al. 2011, Holzman et al. 2013, Bousquet et al. 2014, Fischer et al. 2015, Kurokawa et al. 2015, Kahng et al. 2017; see also Aziz et al. 2019, Mattei et al. 2020]

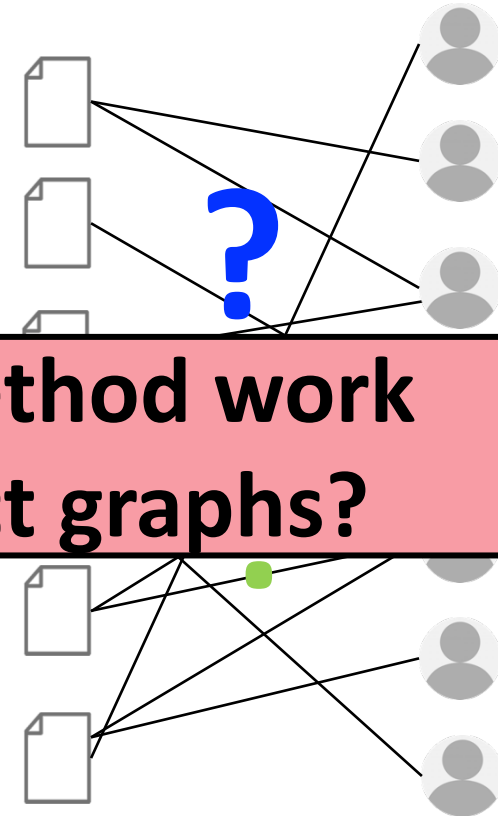
## Peer grading

1-1 conflict graphs



## Conference peer review

More **complex** conflict graphs



Can the partitioning method work  
for peer-review conflict graphs?

# ICLR empirical evaluation (authorship conflicts)

## Q1. Is partitioning of conflict graph feasible?

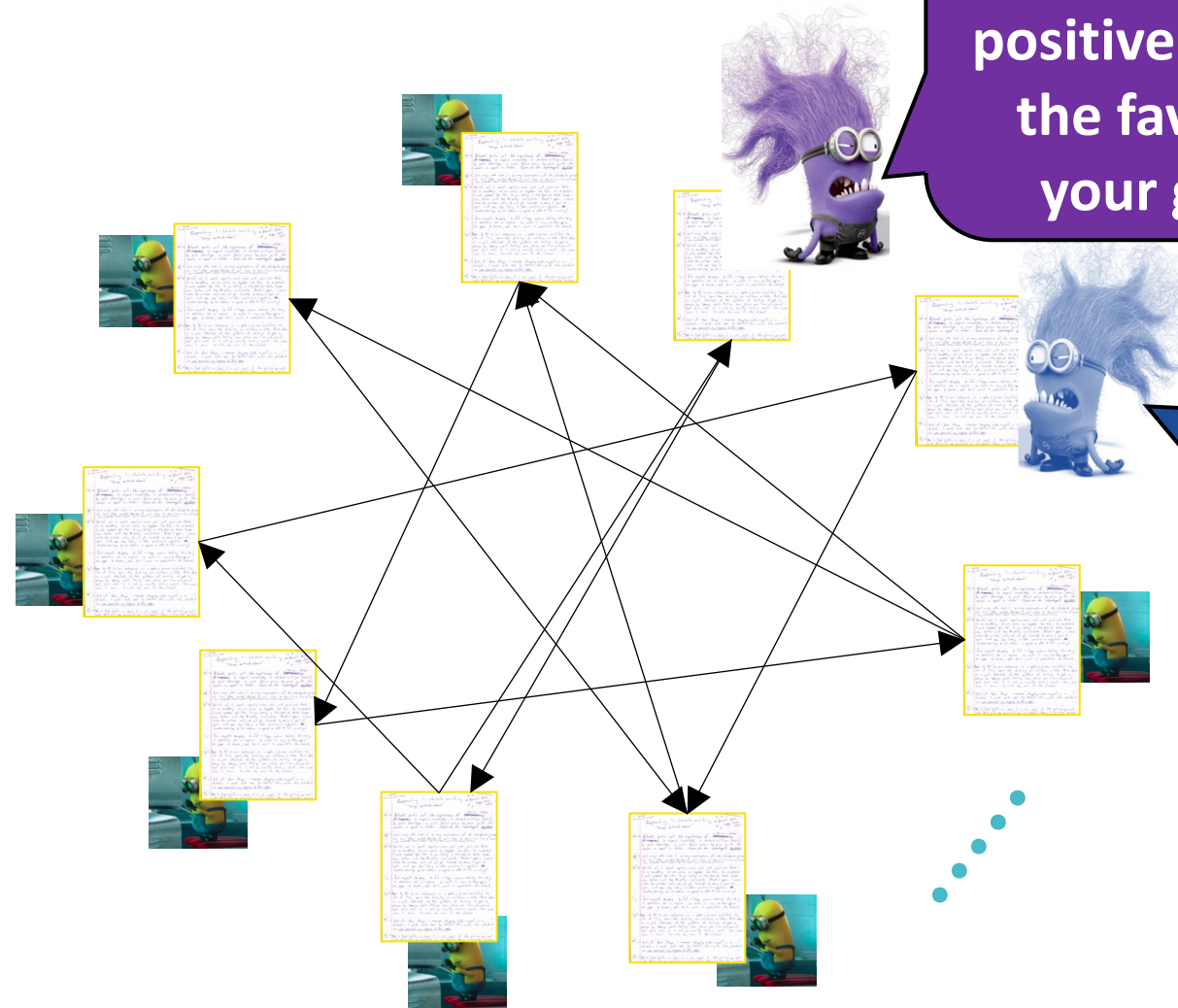
Yes! 253 disjoint components

## Q2. How does assignment quality fare under strategyproofness?

- 372 reviewers and 133 papers in largest connected component  
∴ Assigned reviewers may lack expertise.
- Heuristics for more flexibility: Removing 3.5% of reviewers from the reviewer pool reduces size of the largest component by 86%



# Fraud: Coalition



Why don't you bid on my paper and give it a positive review. I'll return the favor by accepting your grant proposal.

Sounds like a plan!



T. N. Vijaykumar May 12, 2020 · 5 min read

## Potential Organized Fraud in ACM/IEEE Computer Architecture Conferences

*“investigators found that a group of PC members and authors colluded to bid and push for each other’s papers. They give high scores to the papers. Our process is not set up to combat such collusion.”*

**Such collusions also uncovered in conferences in other research areas and in grant reviews**

[[Lauer 2020](#), [Littman 2021](#)]




# Defense 1: Conflicts of Interest

- Don't assign papers to collaborators/colleagues of authors

## Challenges:

- Colluders may not be collaborators/colleagues
- Colluders skirt conflicts-of-interest detectors

 **T. N. Vijaykumar** May 12, 2020 · 5 min read

Potential Organized Fraud in ACM/IEEE Computer Architecture Conferences

*“There is a chat group of a few dozen authors who in subsets work on common topics and carefully ensure not to co-author any papers with each other so as to keep out of each other’s conflict lists (to the extent that even if there is collaboration they voluntarily give up authorship on one paper to prevent conflicts on many future papers).”*



# Defense 2: Detect or Remove Rings

[[Guo et al. 2018](#)]

## Challenges:

- A reviewer may target an author's paper, and author may offer quid pro quo elsewhere.



# Defense 3: Detect Malicious Bids / Disable Bids

	Not willing to review	Indifferent	Eager to review
Towards More Accurate NLP Models	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Interpreting AI Decision-Making	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Multi-Agent Cooperative Board Games	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
A* Search Under Uncertainty	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

- Bidding is easily gameable [[Jecmen et al. 2020](#), [Wu et al. 2021](#)]
  - Via strategic bidding, reviewers can increase chances of getting assigned a paper from ~10% to ~90% [[Jecmen et al. 2020](#)]
- Remove outlier bids [[Wu et al. 2021](#)]
  - Use bids from all reviewers as labels to train a machine learning model which predicts bids based on the other sources of data.
  - Use this predictive model as the similarities for making the assignment.
  - Mitigates dishonest behavior by de-emphasizing bids that are significantly different from the other data sources.



# Defense 3: Detect Malicious Bids / Disable Bids

## Challenges:

- Other aspects of automated assignment systems, like subject area choices or reviewer profiles, can also be gamed

*“TPMS can be gamed through rare keywords”* [[Ailamaki et al. 2019](#)]



# Defense 3: Detect Malicious Bids / Disable Bids

## Challenges:

PDF embedding attacks on text-matching [Markwood et al. 2017; Tran and Jaiswal 2019]

- Most frequent word in colluding paper: “review”
- Most frequent word in colluding reviewer’s previous papers: “minion”
- PDF allows authors to define their own fonts:

Font 0: Default

Font 1: m → r, i → e, n → v

Font 2: o → e, n → w

- Appropriately choose fonts for rendering text in submitted paper

**Visible to an automated plain-text parser:**

Each minion in peer minion will undergo minion



**Visible to humans:**

Each review in peer review will undergo review



# Defense 3: Detect Malicious Bids / Disable Bids

## Challenges:

Colluding reviewers may already have expertise for that paper, and can be assigned even without bids

 T. N. Vijaykumar May 12, 2020 · 5 min read

Potential Organized Fraud in ACM/IEEE Computer Architecture Conferences

*"They exchange papers before submissions and then either bid or get assigned to review each other's papers by virtue of having expertise on the topic of the papers. "*



# Defense 4: Mitigating strategy

Miti-  
gate



Idea!

Assign reviewers to papers  
uniformly at random!

**Problem: Assigned reviewers  
may not have expertise**

# Defense 4: Mitigating strategy



Idea 2.0!

Trade off between  
randomness and expertise  
via controlled randomness  
in the assignment

# Recall: Automated assignment

$$\begin{array}{ll} \text{maximize} & \\ \text{assignment} & \sum_{p \in \text{Papers}} \sum_{r \in \text{Reviewers}} s_{pr} \mathbb{I}\{\text{paper } p \text{ assigned to reviewer } r\} \end{array}$$

subject to

Every paper gets **3** reviewers

Every reviewer gets at most **3** papers

No paper is assigned to conflicted reviewer

# Randomized assignment

Program chairs specify matrix  $Q \in [0, 1]^{\#papers \times \#reviewers}$  such that

$$P(\text{reviewer } r \text{ is assigned to paper } p) \leq Q_{pr} \quad \forall p, r$$

- Can choose a constant matrix (e.g., all entries 0.5)
- Or can choose  $Q$  based on other information/requirements

# Randomized assignment

**Example:**  $Q_{ij} = 0.5 \forall i, j$

$$\begin{array}{ll} \text{maximize} & \\ \text{assignment} & \sum_{p \in \text{Papers}} \sum_{r \in \text{Reviewers}} s_{pr} \mathbb{I}\{\text{paper } p \text{ assigned to reviewer } r\} \end{array}$$

subject to

Every paper gets **6** reviewers

Every reviewer gets at most **6** papers

No paper is assigned to conflicted reviewer

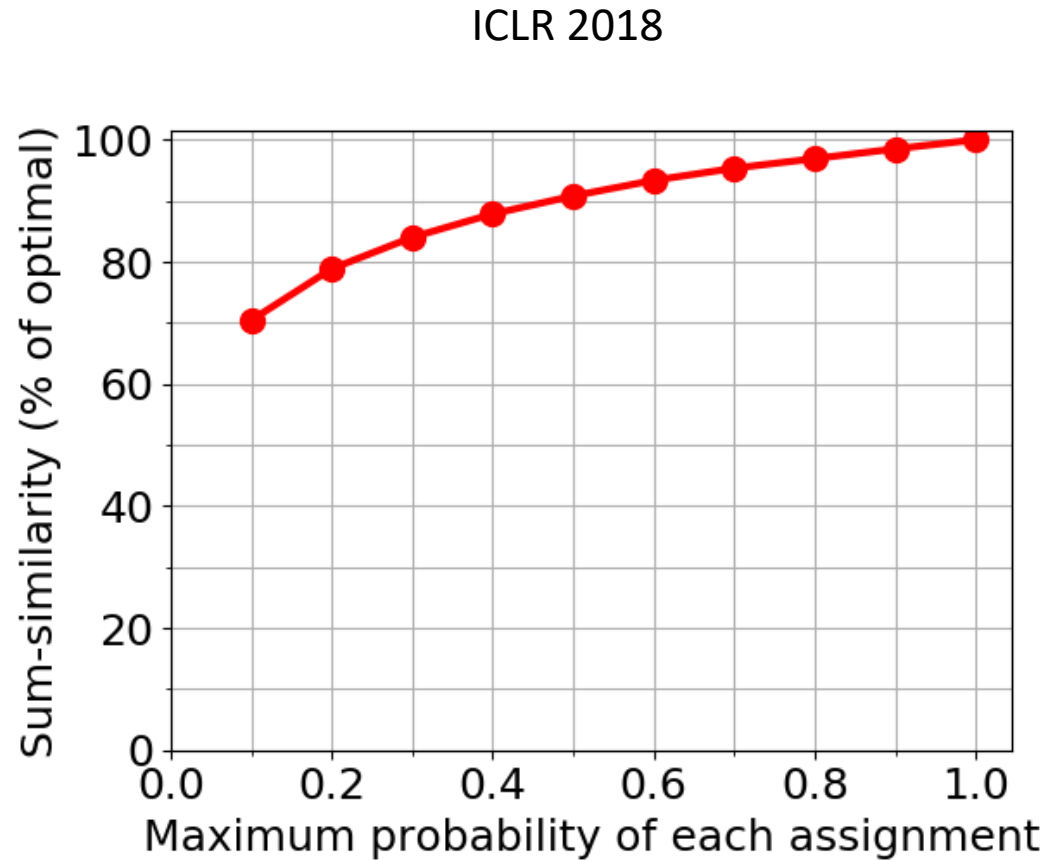
**Sample** an assignment at random so that

Every paper gets **3** reviewers

Every reviewer gets at most **3** papers

**$P(\text{any reviewer assigned to any paper}) \leq 0.5$**

# How about expertise?



**Any reviewer has at best a 50% chance of getting a paper**  
**Sum similarity is 90% of original**

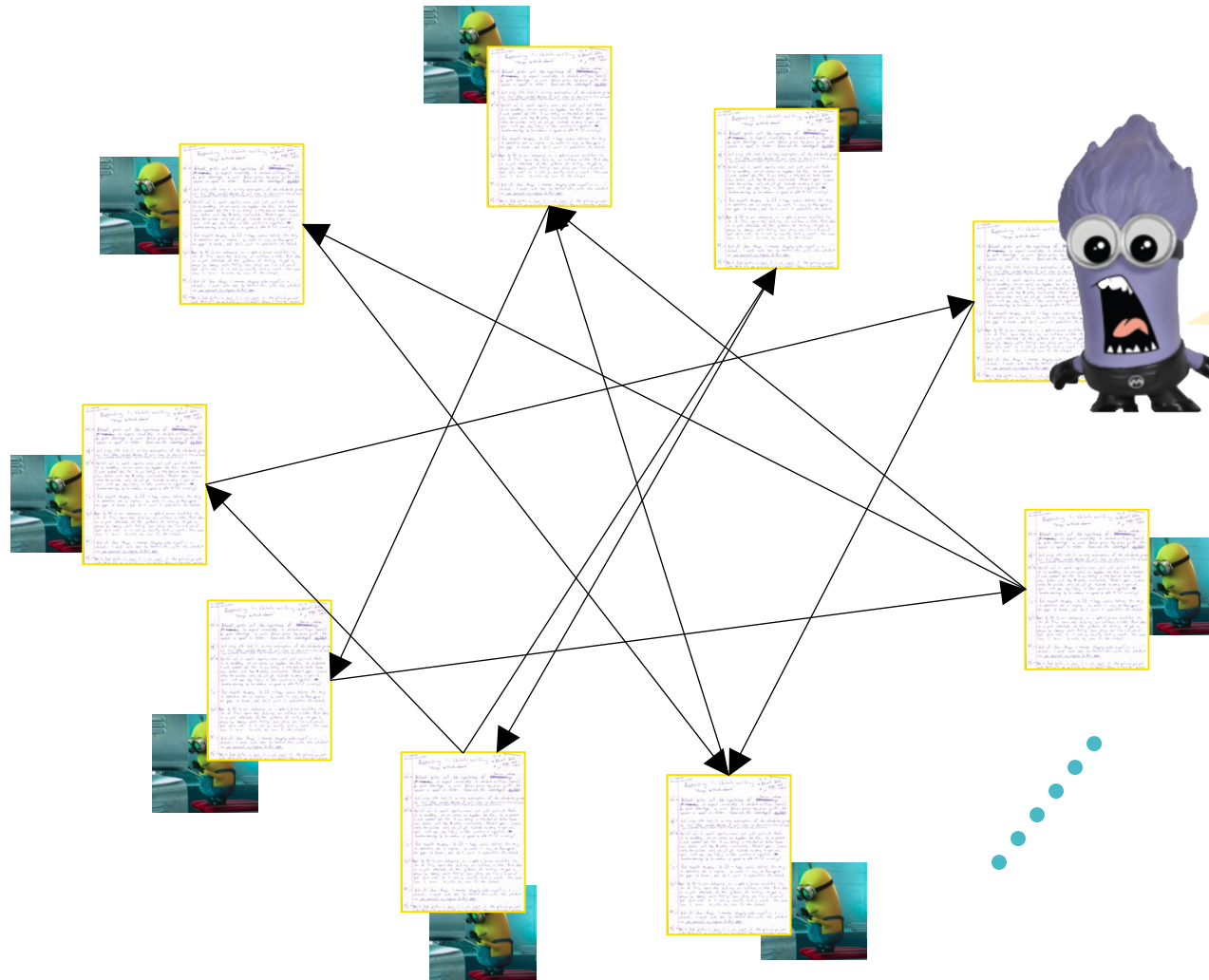


# Fraud: Open problems



- Lone wolf
  - Maximum sum similarity under partitioning-based method? [[Xu et al. 2018](#)]
  - Is strategyproofing possible when conflict graph cannot be partitioned? [[Aziz et al. 2019](#)]
- Coalitions
  - How to make use of various meta data?
- Detect such fraud [[Stelmakh et al. 2021](#), [Wu et al. 2021](#)]
- Other kinds of dishonest behavior [[Ferguson et al. 2014](#), [Gao et al. 2017](#), [Lauer et al. 2019](#)]

# Bias



**It would probably be beneficial  
to find one or two male  
researchers to work with**

True story

Review in PLOS ONE, 2015

Authors: Fiona Ingleby, Megan Head

# Single blind versus double blind

## A Principled Interpretation of Minion Speak

S. Overkill and F. Gru  
Cartoony Minion University

In this paper we present a new understanding of...

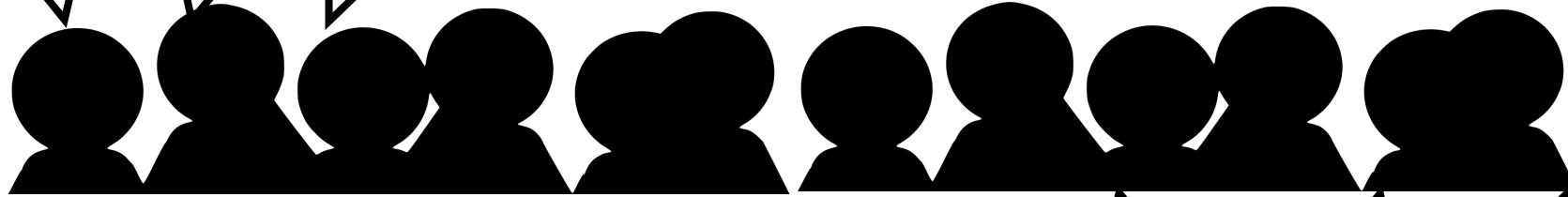
## A Principled Interpretation of Minion Speak

Anonymous Authors  
Anonymous Affiliation

In this paper we present a new understanding of...

# Lot of debate!

Single blind can lead to gender/fame/race/... biases



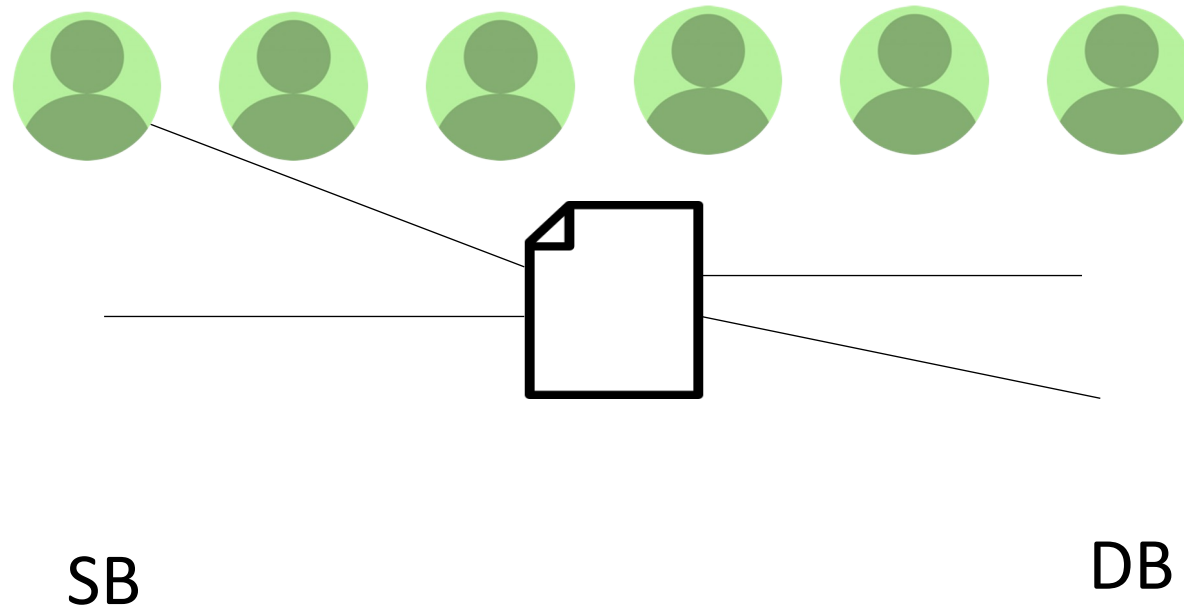
Where is the evidence of bias in my research community?



**How to rigorously test for biases in peer review?**

# WSDM'17 experiment: Setup

A remarkable experiment!



- Reviewers randomly split into single blind (SB) and double blind (DB) conditions
- Each paper assigned 2 SB reviewers and 2 DB reviewers

# WSDM'17 experiment: Tests for bias regarding...

- Gender
- Famous author
- Top university
- Top company
- From USA
- Academic institution
- Reviewer same country as author



# WSDM'17 experiment: Testing procedure

- For any paper  $p$ , let  $q_p$  = “intrinsic” value of paper  $p$
- **Logistic model:**  $P(\text{single blind reviewer accepts paper } p)$   
$$= \frac{1}{1 + \exp(-[\beta_0 + \beta_1 q_p + \sum_{\text{attributes } a} \beta_a \mathbb{I}\{\text{Paper } p \text{ has author attribute } a\}])}$$
- **Use DB reviewers** to estimate  $q_p$  for each paper  $p$
- **Fit decisions of SB reviewers** into logistic model to estimate  $\beta$ 's

Test:  $\beta_a = 0$  vs.  $\beta_a \neq 0$   
(no bias) (bias)

# WSDM'17 experiment: Findings

- Famous author
  - Top university
  - Top company
- } Significant bias
- At least one woman author
- } Not statistically significant; high effect size  
Meta analysis is statistically significant
- From USA
  - Academic institution
  - Reviewer same country as author
- } No evidence of bias

WSDM moved to double blind from the following year.



# Peculiar characteristics of peer review

# Statistical testing preliminaries

**False alarm (Type I error)** Claiming **presence** of bias when the bias is **absent**

**Detection (1 - Type-II error)** Claiming **presence** of bias when the bias is **present**

For a given  $\alpha$ , must ensure  
 $P(\text{false alarm}) \leq \alpha$

Typical choice:  $\alpha = 0.05$

Want high detection subject to false alarm control



## **Characteristic 0:** Correlations between quality of papers and certain attributes

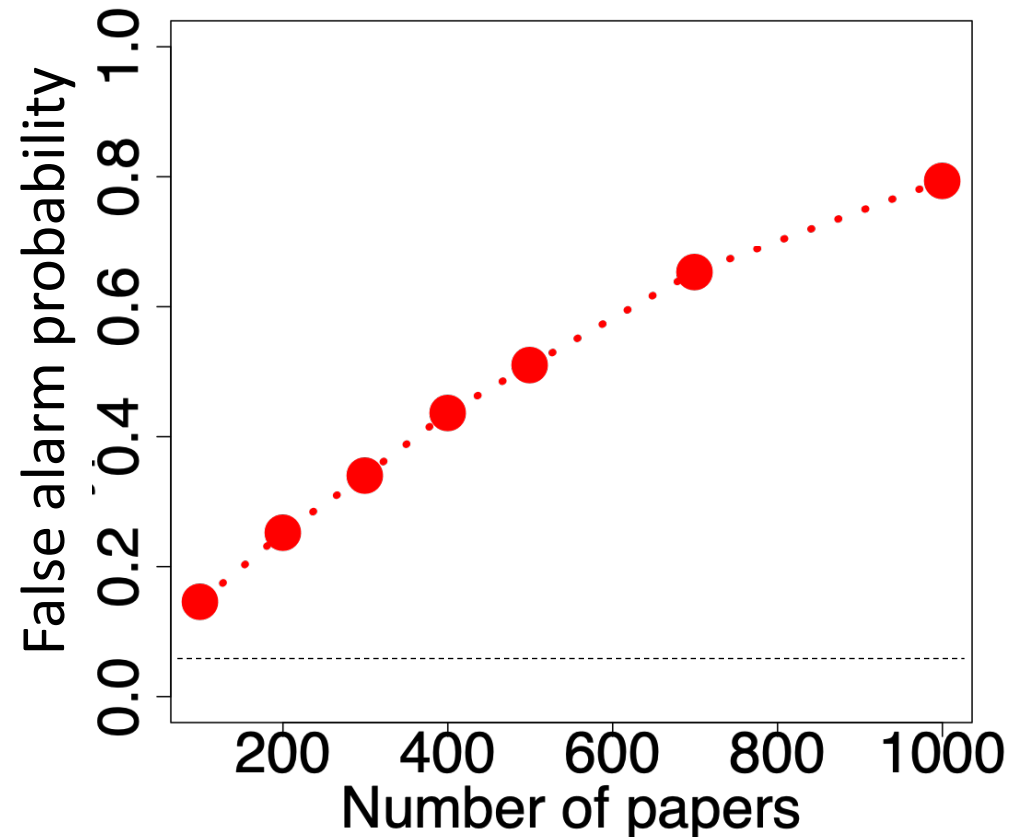
- Famous author
- Top university
- Top company

Combined with other characteristics...

# Characteristic 1: Reviews are noisy

Reviewers are imperfect (noisy)

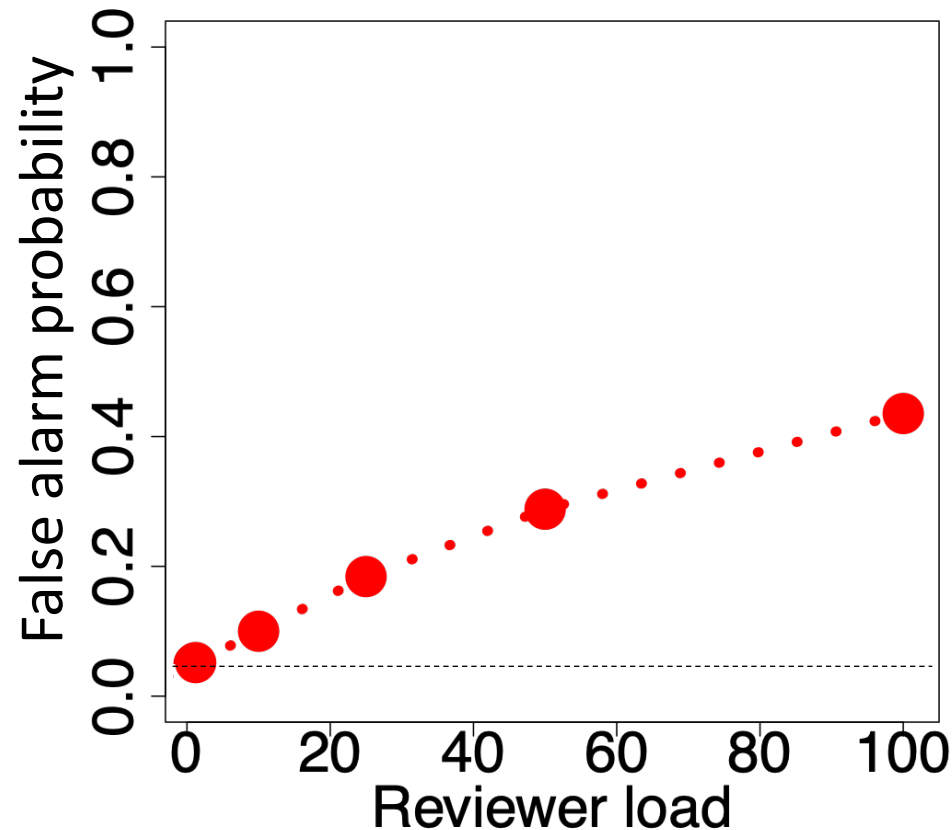
Must ensure:  $P(\text{declare bias when no bias}) \leq 0.05$



# Characteristic 2: Intra-reviewer dependency

Reviews of different papers by the same reviewer are dependent, e.g., a reviewer may be lenient or strict

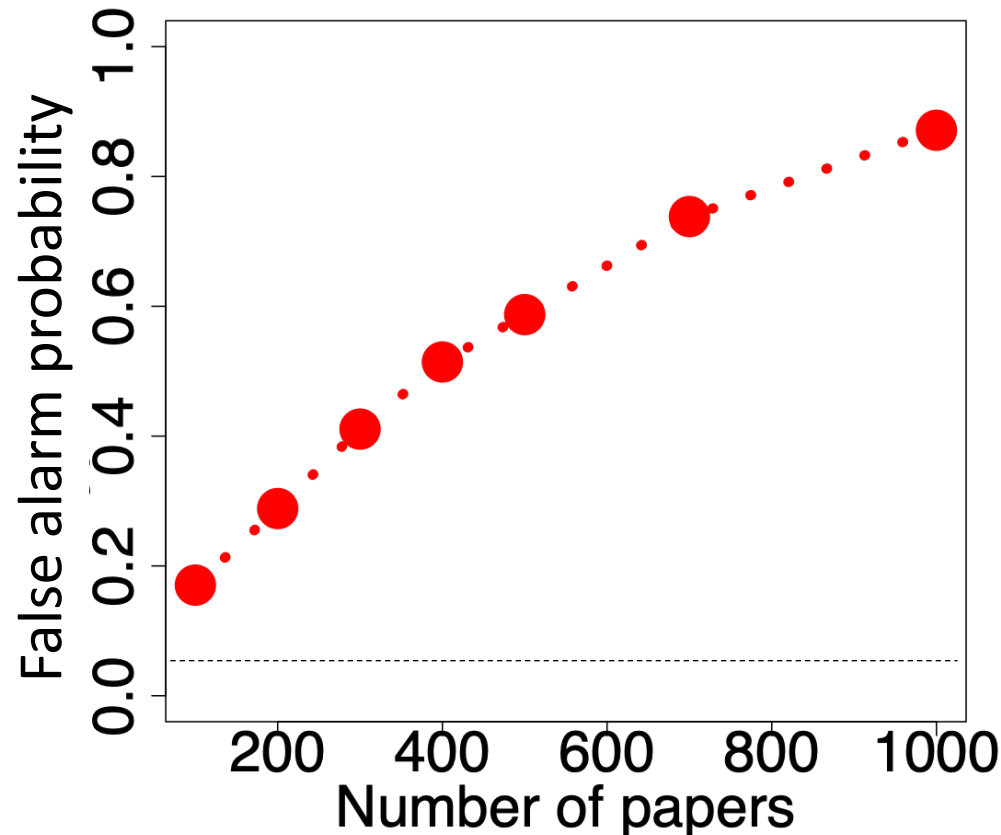
Must ensure:  $P(\text{declare bias when no bias}) \leq 0.05$



# Characteristic 3: Model complexity

Human evaluations may be more complex than simple parametric/logistic models

Must ensure:  $P(\text{declare bias when no bias}) \leq 0.05$

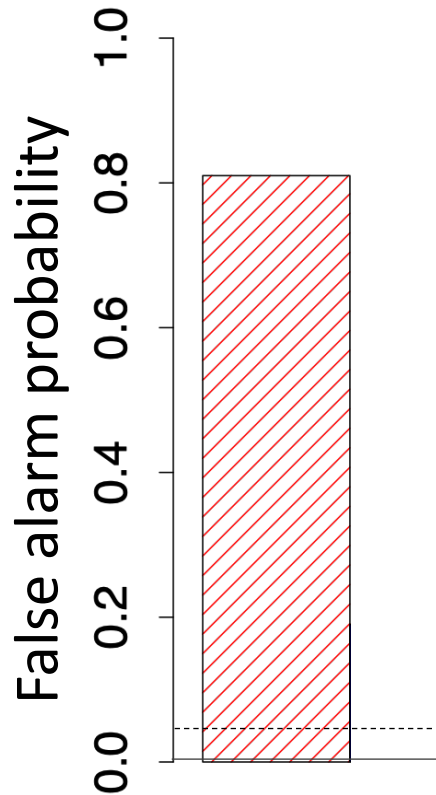




# Characteristic 4: Non-random assignment

Assignment of reviewers to papers is NOT random

Must ensure:  $P(\text{declare bias when no bias}) \leq 0.05$



# A solution

## Step 1: Experimental setup (Reviewer assignment)

- (1a) **Initial assignment:** Each paper assigned 2 reviewers; at most 1 paper per reviewer
- (1b) **Randomization:** For each paper, send 1 reviewer to SB and 1 to DB uniformly at random
- (1c) **Final assignment:** Assign remaining reviewers in any manner desired

## Step 2: Statistical test (after getting reviews)

- Condition on triples from (1a) where reviewers disagree on their decisions
- Run permutation test at the level  $\alpha$

- No assumption of existence of any “true scores”
- Non-parametric model
- Guaranteed false alarm control

# Biases: Open problems

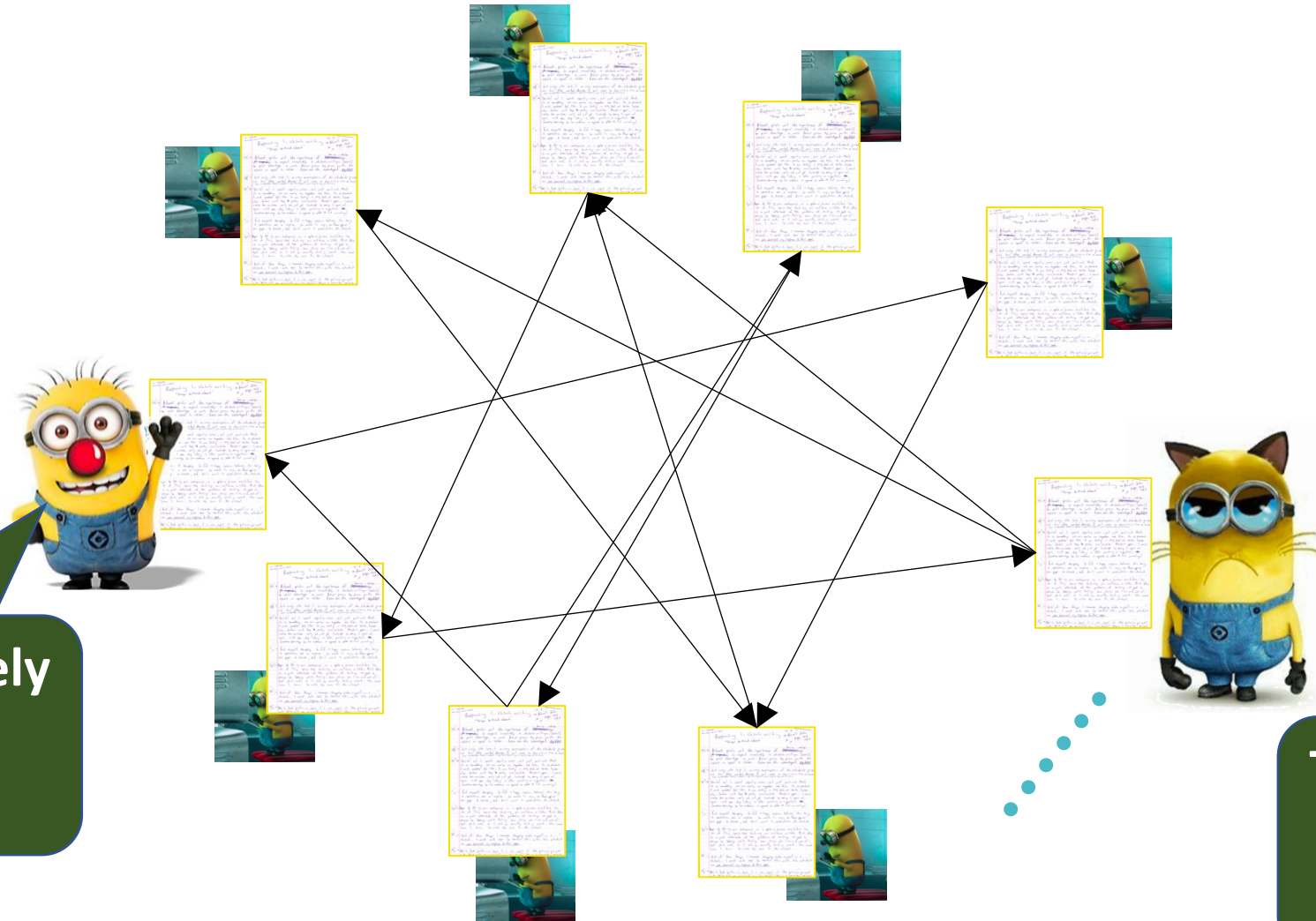


- Optimal detection for given false alarm control
- Tests on observational peer-review data [[Thelwall et al. 2019](#), [Tran et al. 2020](#), [Shah et al. 2018](#)]
- Biases in other review components such as program committee meetings and discussions
- Biases in text [[Manzoor et al. 2021](#)]



Observational; uses the fact that ICLR switched from SB to DB

# Miscalibration



This is a moderately  
decent paper.  
8/10

This is a moderately  
decent paper.  
4/10.

# Miscalibration in ratings

“A raw rating of 7 out of 10 in the absence of any other information is **potentially useless.**” [Mitliagkas et al. 2011]

“The rating scale as well as the individual ratings are often **arbitrary** and may not be consistent from one user to another.” [Ammar et al. 2012]

“[Using rankings instead of ratings] becomes very important when we combine the rankings of many viewers who often use **completely different ranges of scores** to express identical preferences.” [Freund et al. 2003]

# Unfairness in peer review

“the existence of disparate categories of reviewers creates the potential for **unfair treatment of authors**. Those whose papers are sent by chance to assassins/demoters are at an unfair disadvantage, while zealots/pushovers give authors an unfair advantage.”



## ***Editor's Page***

Stanley S. Siegelman, MD

### **Assassins and Zealots: Variations in Peer Review**

#### **Special Report<sup>1</sup>**

Nihar B. Shah, Carnegie Mellon University



[Siegelman 1991]

# NeurIPS 2016

	1 (low or very low)	2 (sub-standard)	3 (poster level: top 30%)	4 (oral level: top 3%)	5 (award level: top 0.1%)
Impact	6.5%	36.1%	45.7%	10.5%	1.1%
Quality	6.7%	38.0%	44.7%	9.5%	1.1%
Novelty	6.4%	34.8%	48.1%	9.7%	1.1%
Clarity	7.1%	28.0%	48.6%	14.6%	1.8%

**≥ 3: 57%** instead of intended 30%

**≥ 4: 10%** instead of intended 3%

**≥ 5: 1%** instead of intended 0.1%

# Two approaches in the literature

1

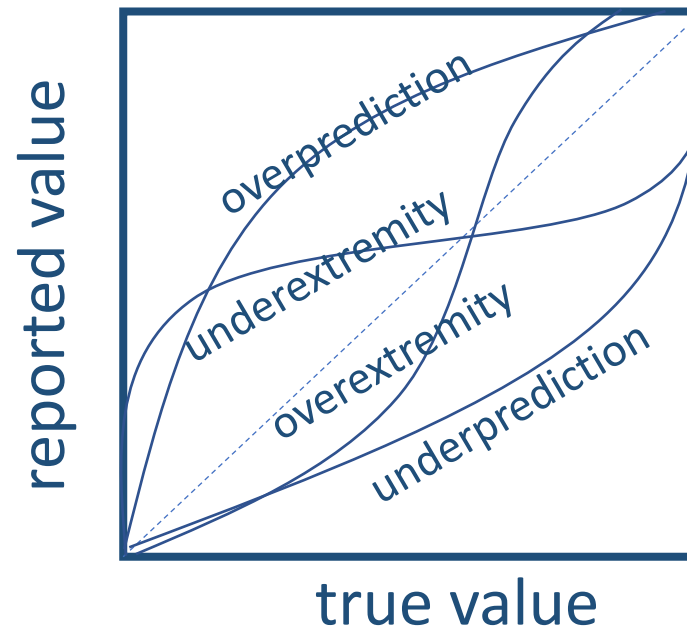
## Assume simplified (affine) models for calibration

[[Paul 1981](#), [Flach et al. 2010](#), [Roos et al. 2011](#), [Baba et al. 2013](#), [Ge et al. 2013](#), [Mackay et al. 2017](#)]

- Did not work well [NeurIPS 2016 program chairs; personal communication]
- *“We experimented with reviewer normalization and generally found it significantly harmful.”* [[Langford](#) (ICML 2012 program co-chair)]

## Miscalibration is quite complex:

[[Brenner et al. 2005](#)]





# Two approaches in the literature

## 2 Use rankings

[[Rokeach 1968](#), [Freund et al. 2003](#), [Harzing et al. 2009](#),  
[Mitliagkas et al. 2011](#), [Ammar et al. 2012](#), [Negahban et al. 2012](#)]

- Use rankings induced by ratings or directly collect rankings
- Commonly believed to be the best option if no assumptions on miscalibration



**Is it possible to do better using ratings than rankings, with essentially no assumptions on the miscalibration?**

# Canonical 2x2 setting



$$z_A^* \neq z_B^* \in [0,1]$$



Miscalibration function:  $f_1 : [0,1] \rightarrow [0,1]$

Given paper  $i \in \{A, B\}$ , outputs  $f_1(z_i^*)$



Miscalibration function  $f_2 : [0,1] \rightarrow [0,1]$

Given paper  $i \in \{A, B\}$ , outputs  $f_2(z_i^*)$

- Adversary chooses  $z_A^*, z_B^*$  and strictly monotonic  $f_1, f_2$
- One paper assigned to each reviewer at random
- **Goal: Given (assignment, score given by each reviewer)**  
**estimate if  $z_A^* > z_B^*$  or  $z_B^* > z_A^*$** 
  - Eliciting rankings is vacuous; amounts to random guessing

# Impossibility on deterministic estimators

## Theorem

**No deterministic estimator has a success probability better than ranking.**

# A randomized estimator

## Theorem

**There is a randomized estimator that strictly outperforms ranking.**

With probability  $(1 + |\text{difference between the two scores}|)/2$ ,  
pick paper which received higher score

	Reviewer 1: $f_1(x) = x/2$	Reviewer 2: $f_2(x) = (3+x)/4$
Paper A: $z_A^* = 0.2$	$f_1(0.2) = 0.1$	$f_2(0.2) = 0.8$
Paper B: $z_B^* = 0.6$	$f_1(0.6) = 0.3$	$f_2(0.6) = 0.9$

- Under **blue** assignment, pick paper **B** with probability

$$\frac{1 + |0.1 - 0.9|}{2} = 0.9$$

(output is correct)

- Under **red** assignment, pick paper **A** with probability

$$\frac{1 + |0.3 - 0.8|}{2} = 0.75$$

(output is wrong)

- On average, correct with probability

$$\frac{1}{2}(0.9) + \frac{1}{2}(1 - 0.75) = 0.575 > 0.5$$

# Miscalibration: Open problems



Strong assumptions:  
parametric, affine

**Sweet spot**

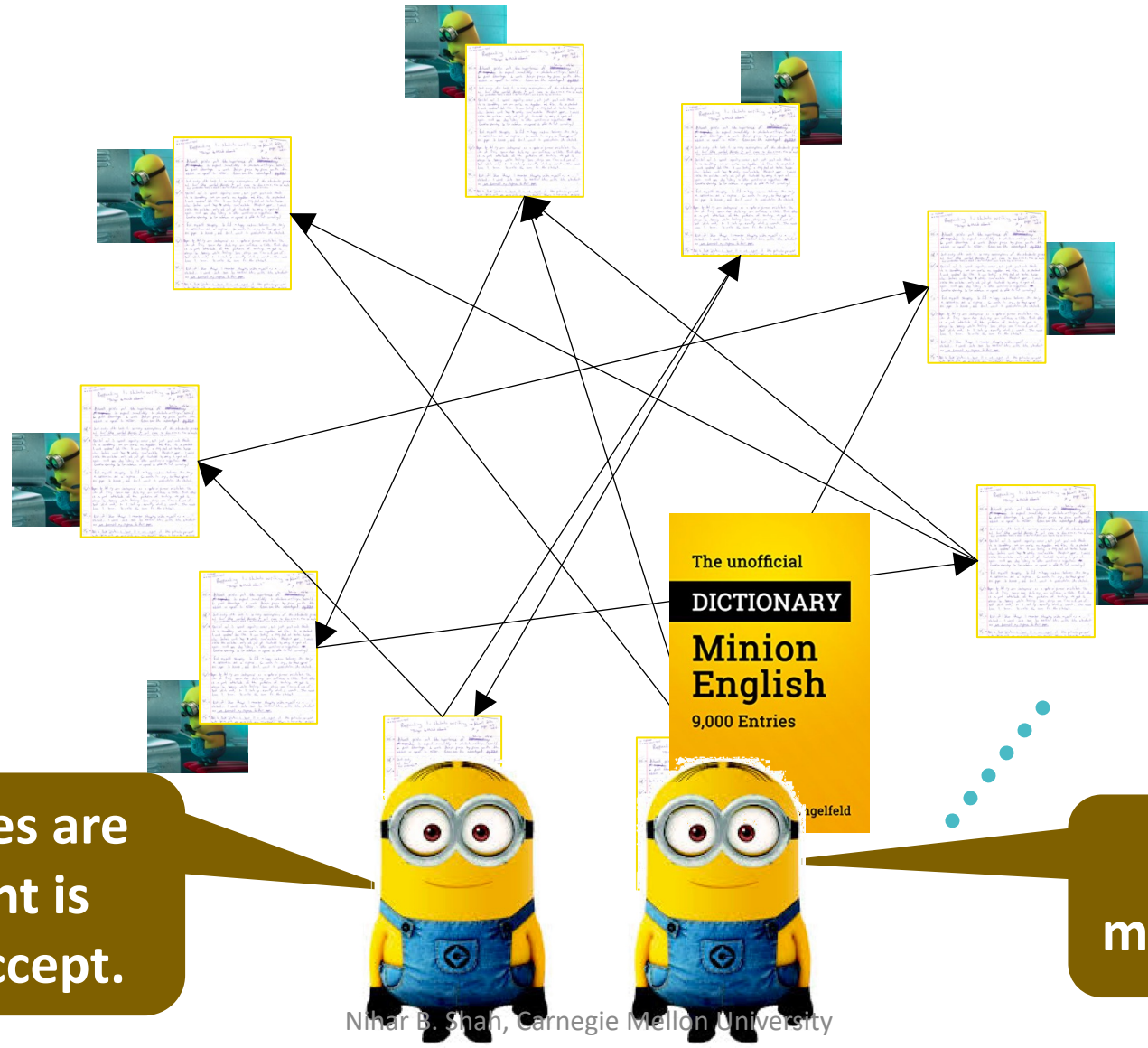
Arbitrary/adversarial  
miscalibration

Ranking

?

- Weaker assumptions: non-parametric, non linear (e.g., permutation-based models [[Shah 2017 part 1](#)])
- Amenable to small sample sizes: Avoid overkill
- Use rankings and ratings together
  - About 40% of ratings given by a reviewer to a pair of papers are tied [[Shah et al. 2018](#) Section 3.8.1]

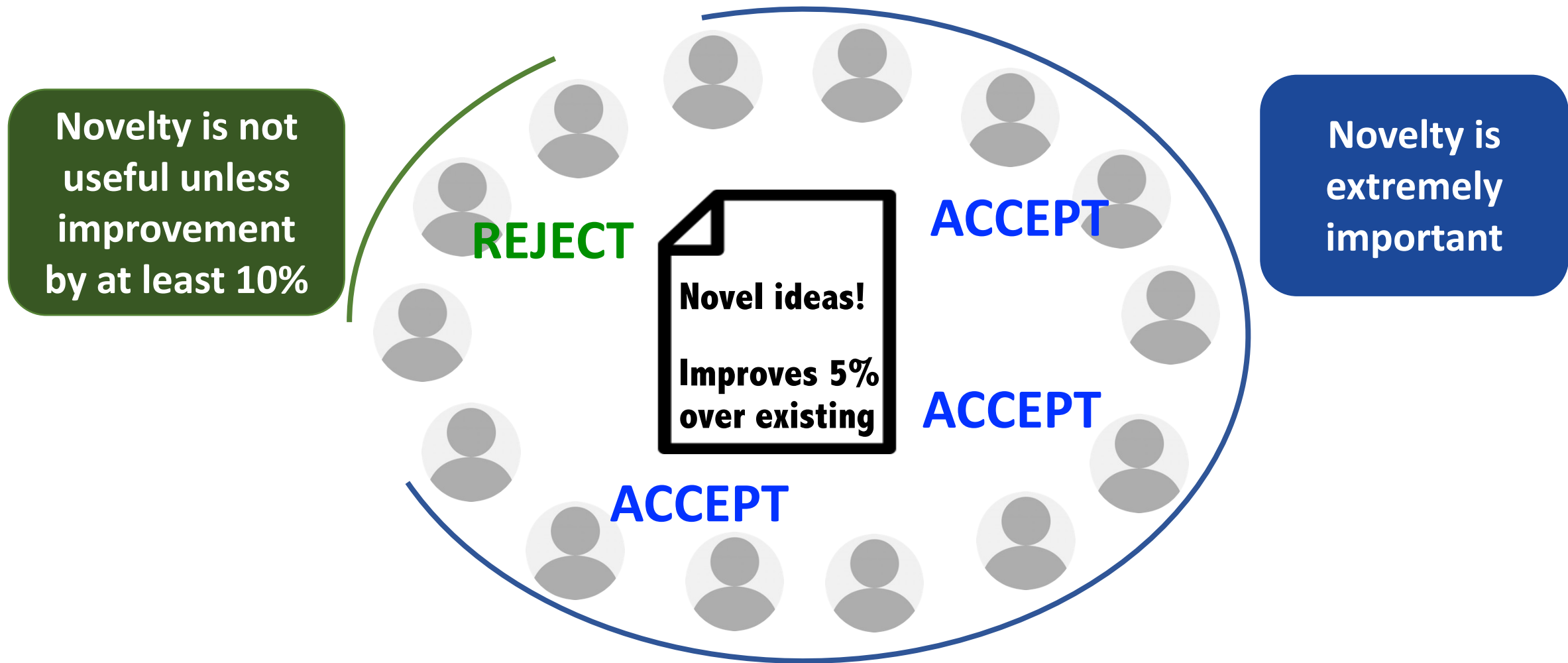
# Subjectivity



Spelling mistakes are  
ok. The content is  
great. Strong accept.

Too many spelling  
mistakes. Strong reject.

# Differing opinions about relative importance of criteria



[Kerr et al. 1977, Bakanic et al. 1987, Hojat et al. 2003, Church 2005, Lamont 2009, Lee 2015]



# Commensuration Bias in Peer Review

Carole J. Lee\*†

---

To arrive at their final evaluation of a manuscript or grant proposal, reviewers must convert a submission's strengths and weaknesses for heterogeneous peer review criteria into a single metric of quality or merit. I identify this process of commensuration as the

“Illuminates how intellectual priorities in individual peer review judgments can collectively subvert the attainment of community-wide goals”



**How to ensure that every paper is  
judged by the same yardstick?**

# Problem setting

- Reviewers asked to judge papers on **k criteria**
  - E.g. (IJCAI 17): Originality, Relevance, Significance, Writing, Technical
  - Give **criteria scores** in  $[0,1]^k$
- And an **overall score** in  $[0,1]$
- Each reviewer has a coordinate-wise non-decreasing **(subjective) mapping** from criteria scores in  $[0,1]^k$  to overall score in  $[0,1]$

**Need a common mapping** (from criteria to overall scores) **for all papers**

# Handcrafted design?



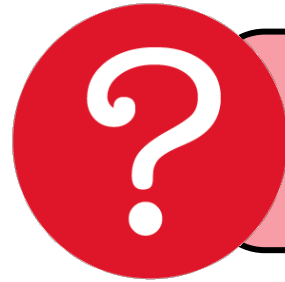
## AAAI 2013

- Similar goal
- Reviewers asked to score papers according to 8 criteria
- Program Chairs provided detailed instructions on how to map criteria to an overall recommendation
- The goal was really admirable, but handcrafted design did not work well
- Quite challenging to manually specify an 8-dimensional function
- For example, strong accept when paper gets a score of 5 or 6 (out of 6) for some criterion, and does not get a 1 for any criteria.
- Implies to strongly accept paper that has a 5 or 6 in clarity, but can be below average in every other criterion

# Data-driven approach: Learn a mapping

- Obtain (criteria scores, overall score)  $\in [0,1]^k \times [0,1]$  for every review
- Learn a mapping  $\hat{f}: [0,1]^k \rightarrow [0,1]$  from this data
- For every review, augment overall score with  $\hat{f}(\text{criteria scores})$

# Framework



Which loss function to use?

$$\hat{f} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \ell \left( \begin{matrix} & \text{Papers} \\ \text{Reviewers} & \begin{bmatrix} f([.8 \ .9 \ .9]) & f([.8 \ .6 \ .1]) \\ f([.9 \ .1 \ .4]) & f([.2 \ .7 \ .4]) \\ f([.1 \ .3 \ .2]) & f([.8 \ .9 \ .2]) \\ f([.9 \ .5 \ .4]) & f([.7 \ .8 \ .2]) \\ f([.3 \ .2 \ .1]) & f([.4 \ .7 \ .9]) \end{bmatrix} \end{matrix} \right), \begin{matrix} & \text{Papers} \\ \text{Reviewers} & \begin{bmatrix} .9 & .6 \\ .2 & .4 \\ .6 & .6 \\ .4 & .9 \\ .2 & .3 \end{bmatrix} \end{matrix} \right)$$

Criteria scores                      Overall scores

$\mathcal{F}$  = set of all coordinate-wise non-decreasing functions

# An axiomatic approach

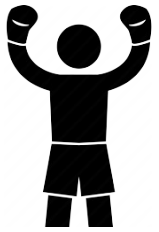
- **Challenge:** There is no ground truth!
- **Axiomatic approach**
  - Approach is popular in economics and social choice theory
  - Identify specific scenarios that is easy to reason about
  - Establish necessary conditions (or “axioms”)
  - Use these to narrow down the possible choices at hand

# Three natural axioms



## Axiom 1: Consensus

For some criteria score vector  $x$  and some overall score  $y$ , if all reviewers map  $x$  to  $y$  then the learnt mapping must also map  $x$  to  $y$ .



## Axiom 2: Dominance

If a paper  $a$  is “*at least as good as*” paper  $b$ , then the learnt mapping should reflect that.



## Axiom 3: Strategyproofness

No reviewer can bring the learnt mapping closer to their own opinion by strategic manipulation.



# Theorem

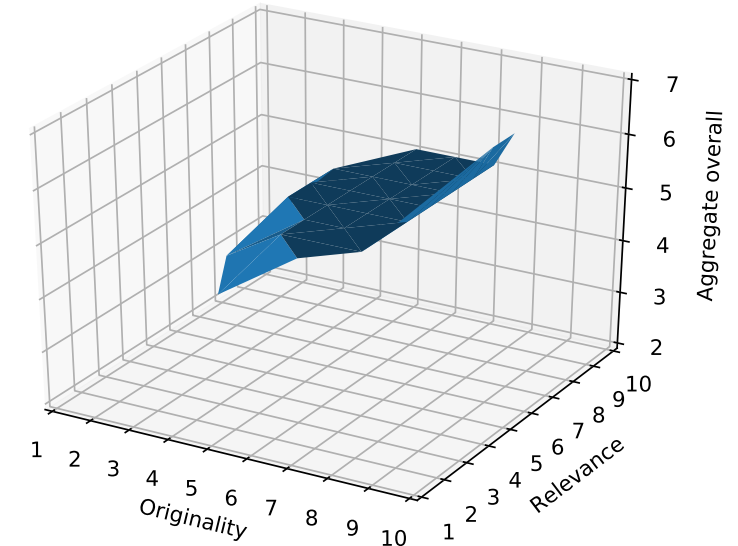
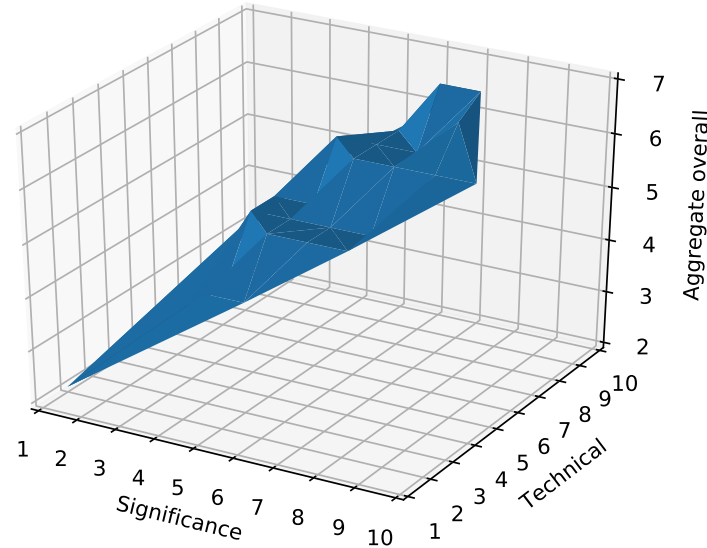
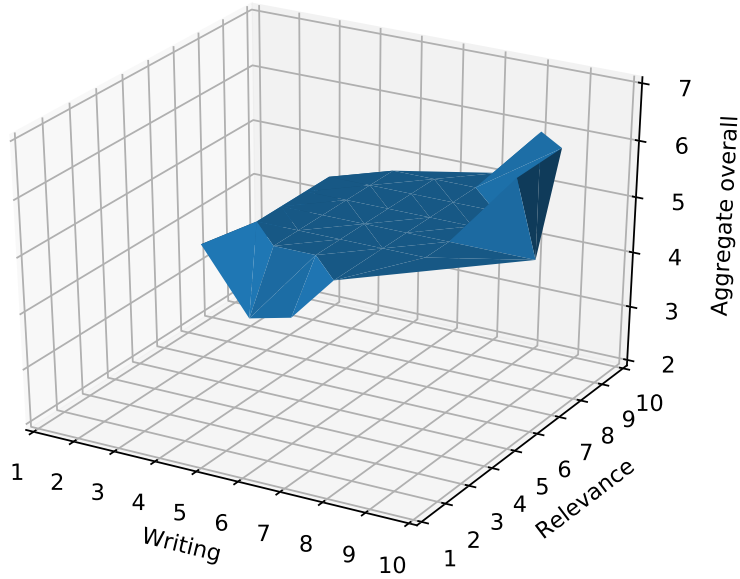
Under these three axioms, there is **exactly one possible loss**.

“**L(1,1) loss**, i.e., sum of absolute differences of all entries”

$$\hat{f} \in \operatorname{argmin}_{f \in \mathcal{F}} \sum_{\text{all entries}} \left| \begin{pmatrix} f([.8 \ .9 \ .9]) & & f([.8 \ .6 \ .1]) \\ f([.9 \ .1 \ .4]) & f([.2 \ .7 \ .4]) & \\ & f([.1 \ .3 \ .2]) & f([.8 \ .9 \ .2]) \\ f([.9 \ .5 \ .4]) & & f([.7 \ .8 \ .2]) \\ f([.3 \ .2 \ .1]) & f([.4 \ .7 \ .9]) & \end{pmatrix} - \begin{pmatrix} .9 & & .6 \\ .2 & .4 & \\ & .6 & .6 \\ .4 & & .9 \\ .2 & .3 & \end{pmatrix} \right|$$

$\mathcal{F}$  = set of all coordinate-wise non-decreasing functions

# IJCAI 2017



- **Writing** and **Relevance**: Really bad - significant downside, really good - appreciated, in between - irrelevant.
- **Technical** quality and **Significance**: high influence; the influence is approximately linear.
- **Originality**: moderate influence.

# Subjectivity: Open problems



- Mixture models
- How much do program-chair-specified criteria explain overall scores?
  - In NeurIPS 2016, 55 cases of a reviewer rating a paper strictly higher than another for all criteria but inverting the relative ranking of the two papers in the overall ordering [Shah et al. 2018 Section 3.9]
- Homophily [Lamont 2009, Brezis et al. 2020] & preconceptions [Ernst et al. 1994]
- Novelty
  - “Reviewers love safe (boring) papers, ideally on a topic that has been discussed before (ad nauseam). Precedents are good; novelty is bad” [Church 2005]
  - “Today reviewing is like grading: When grading exams, zero credit goes for thinking of the question. When grading exams, zero credit goes for a novel approach to solution. (Good) reviewing: acknowledges that the question can be the major contribution. (Good) reviewing: acknowledges that a novel approach can be more important than the existence of the solution.” [Naughton 2010]
- Multiple problems together

# Norms and Policies

Alright, so here's  
what everyone  
must do...



# Norms and Policies

1. Resubmission bias
2. Novice reviewers
3. Herding in discussions
4. Alphabetical author-ordering bias
5. Gender distribution of paper awards

# Resubmission Bias

Many conferences ask authors to declare previous rejections of submitted paper



“authors must declare the resubmission by including a cover letter with their submission... should summarize the main reasons for rejection and should describe the changes the authors have made to address the reviewers’ comments. **The cover letter should be inserted at the beginning of the submitted PDF, along with the previous reviews and previous anonymized rejected submission, before the 6+1 pages of the paper...** A paper rejected from these conferences and omitting to declare resubmission will be directly rejected without further review.”

# Question



Do reviewers get biased when they know that the paper they are reviewing was previously rejected from a similar venue?



# A controlled experiment

- Auxiliary conference review process associated to ICML 2020
- 134 junior reviewers each reviewing 1 paper
- Randomly divided into:

---

## A SUPER\* Algorithm to Optimize Paper Bidding in Peer Review

---

### Author checklist:

- If applicable, will you make the code and data publicly available upon acceptance?  
Answer: **Yes**

### Abstract

A number of applications involve the sequential arrival of users, and require showing each user a set of items. It is well known that the order in which the items are presented to a user can have a

In typical peer review process, when the bidding phase begins, reviewers enter the system in an arbitrary sequential order. Upon entering, a list of papers is shown and the reviewer places bids on papers they would prefer to review.

It is known that the order of papers presented to reviewers

*Control condition*

---

## A SUPER\* Algorithm to Optimize Paper Bidding in Peer Review

---

### Author checklist:

- If applicable, will you make the code and data publicly available upon acceptance?  
Answer: **Yes**
- Was this paper submitted to NeurIPS'19?  
Answer: **Yes, the paper was rejected from NeurIPS**

### Abstract

A number of applications involve the sequential arrival of users, and require showing each user a set of items. It is well known that the order in which the items are presented to a user can have a

In typical peer review process, when the bidding phase begins, reviewers enter the system in an arbitrary sequential order. Upon entering, a list of papers is shown and the reviewer places bids on papers they would prefer to review.

It is known that the order of papers presented to reviewers

*Test condition*



# Key findings



- Reviewers give almost **one point lower score** on a 10-point Likert item for the overall evaluation of a paper when they are told that a paper is a resubmission.
- In terms of narrower review criteria, reviewers tend to **underrate “Paper Quality” the most.**

## Implications.

- Informs debate on whether and how to use resubmission information.
- Consider revealing resubmission information after the initial reviews are submitted.
- Consider whether reviews of rejected papers should be publicly available on systems like openreview.net and others.

# Herding in Discussions

ML/AI conferences have a discussion (via typed comments in a forum) between reviewers of a paper after reviews are submitted.

There is no specified policy on who initiates the discussion.

Past research on human decision making shows that decision of a group can be **biased towards the opinion of the group member who initiates the discussion.**



**Problematic in peer review: Final decisions depends on who initiated discussion**

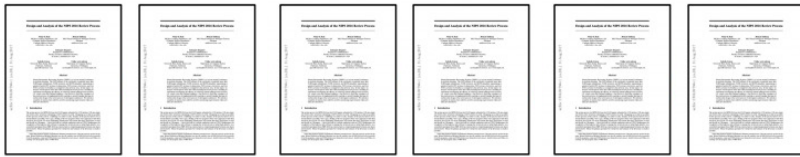
# Question



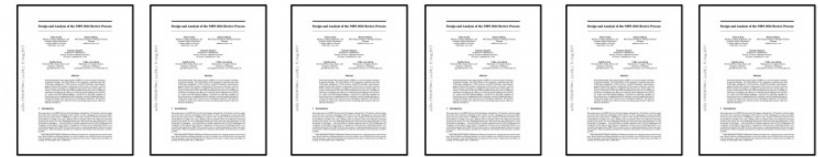
Conditioned on a set of reviewers who actively participate in a discussion of a paper, does the final decision of the paper depend on the order in which reviewers join the discussion?

# A controlled experiment

- Discussions in ICML 2020
- 1500 papers, 2000 reviewers
- Split papers uniformly at random into two groups



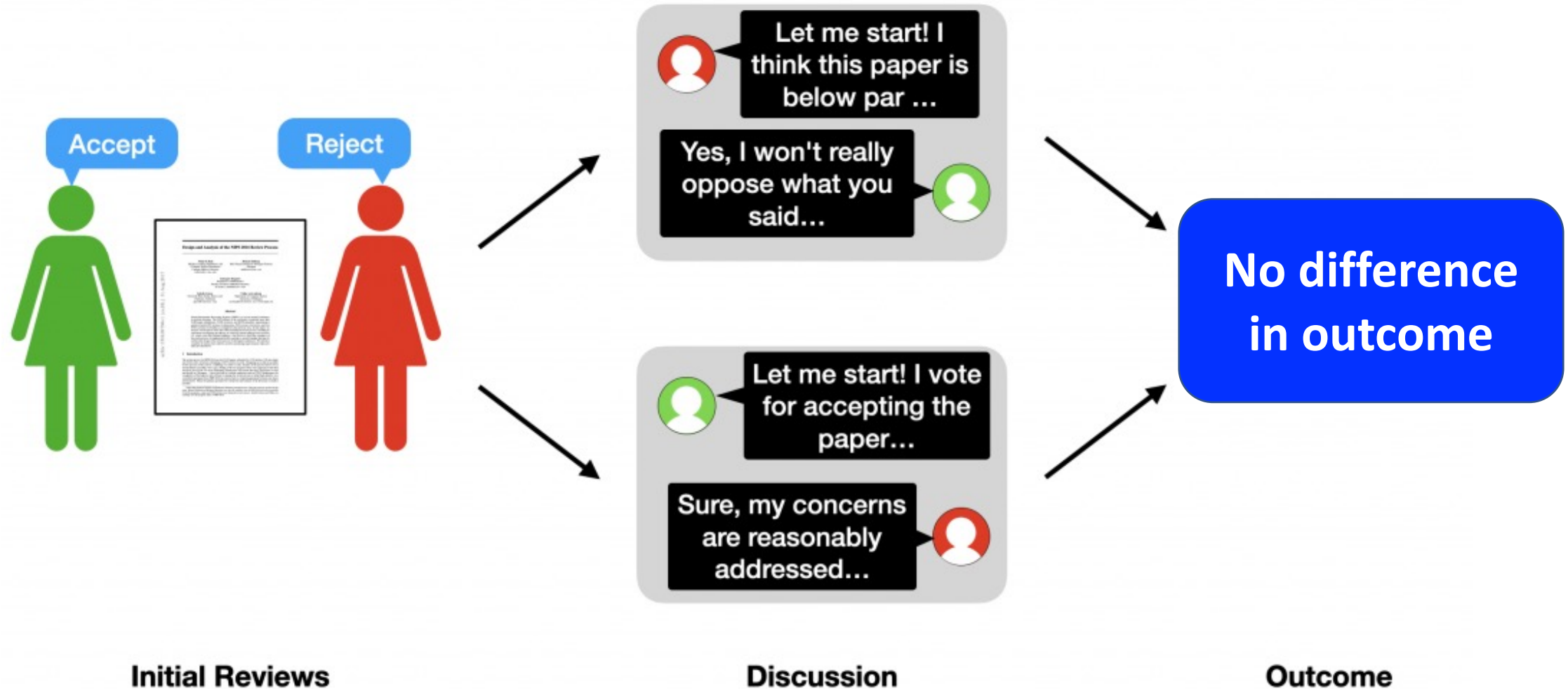
First ask most **positive** reviewer to start the discussion, then later ask the most **negative** reviewer to contribute to the discussion.



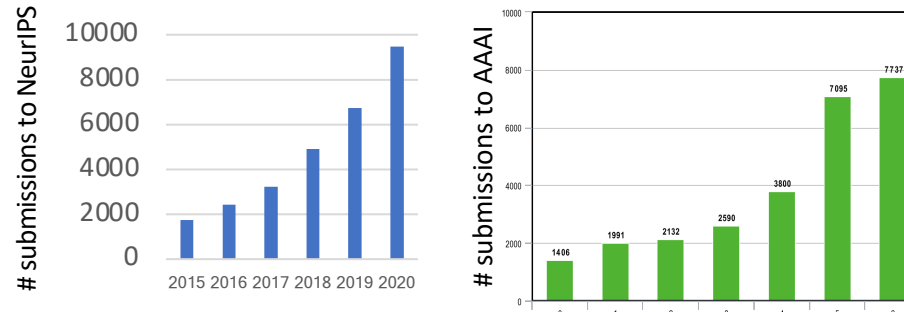
First ask the most **negative** reviewer to start the discussion, then later ask the most **positive** reviewer to contribute to the discussion.

Measure difference in outcomes

# Key findings



# Novice Reviewers



“There is significant evidence that the process of reviewing papers in machine learning is creaking under several years of exponentiating growth.” [Langford 2018]

“Submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint” [McCook 2006]

**Challenge 1.** To avoid overloading reviewers, need to find new sources of reviewers.

**Challenge 2.** Ensure newly added reviewers can write reviews of good quality.

# Common policy

Relax experience or seniority bar for reviewers

- Researchers with limited publication history
- 70% of reviewers in NeurIPS 2016 are PhD students

- Challenge 1 (more reviewers) ✓

- Challenge 2 (quality) ?

- “graduate students seem to be unable to provide very useful comments” [[Patat et al. 2019](#)]
- Junior reviewers are more critical than their senior counterparts [[Mogul 2013](#)]

# Question



Can researchers with limited or no publication history be recruited and guided such that they enlarge the reviewer pool of leading ML and AI conferences without compromising the quality of the process?



# An experiment

## Supplement expansion of reviewer pool with:

### Selection

- Auxiliary conference review process involving 134 junior reviewers.
- Reviews evaluated by authors of papers used in the experiment (authors happy to do so since they get good feedback on their paper)
- Invited 52 best reviewers for ICML 2020

### Mentoring

- In the actual conference, additional mentoring of selected reviewers by a senior researcher
- Additional guidelines
- There to answer questions
- Examples on how to review or participate in discussions etc.
- Point out common issues in reviews

*Amount of additional work for organizers: Comparable to work of one area chair*

# Key findings



- Reviews by experimental reviewers are comparable and/or of higher-rated quality as compared to conventional reviews
- 30% of reviews written by experimental reviewers received highest ratings by area chairs, compared to 14% for the main pool
- Experimental reviewers more engaged
- Experimental reviewers are junior but no more or less critical than experienced reviewers
- Positive feedback from participants who appreciated the opportunity to become a reviewer in ICML 2020

# Biases due to alphabetical ordering

In Economics, norm is to order authors in alphabetical order of last names.

**Faculty with last name starting with an earlier alphabet are:**

- Significantly more likely to receive tenure
- Significantly more likely to become fellows of the Econometric Society
- More likely to receive the Clark Medal and the Nobel Prize

The (related) field of Psychology, which does not order by alphabet, does not show any of these biases.



# What causes these biases?

## In papers

Implicit bias – Primacy effects

Explicit bias – “*First author et al.*”

Conference	#Total papers	#Papers using “ <i>First author et al.</i> ” in its text
STOC 2017	99	70
STOC 2016	79	59
FOCS 2017	79	48
FOCS 2016	73	43
EC 2017	75	48
EC 2016	99	87

## On websites

Serial position effects



### AAAI-19 Program Committee

Hussein Abbass  
Sherief Abdallah  
Abbas Abdolmaleki  
Naoki Abe  
David Abel  
Ayan Acharya  
Maribel Acosta  
Shuchin Aeron  
Maedeh Aghaei

# Let's fix this!

## In papers

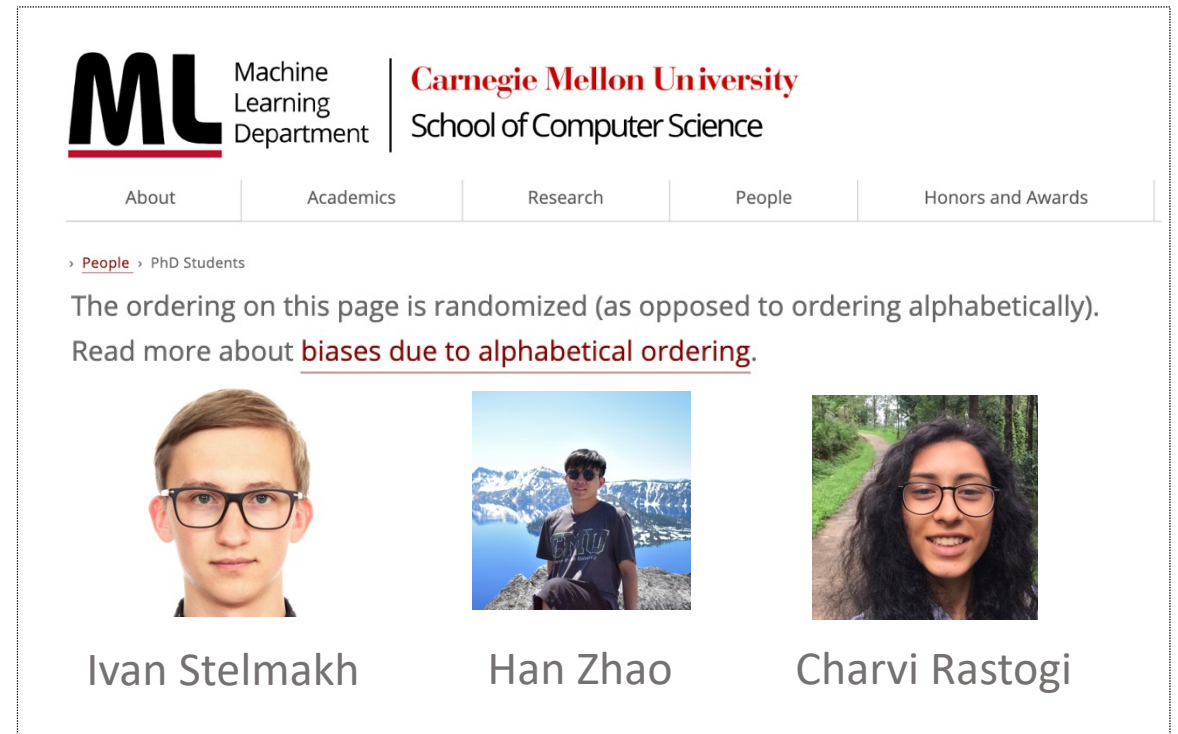
ACM EC conference now uses numbering instead of “first author et al.” citation style

Can randomize author ordering

## On websites

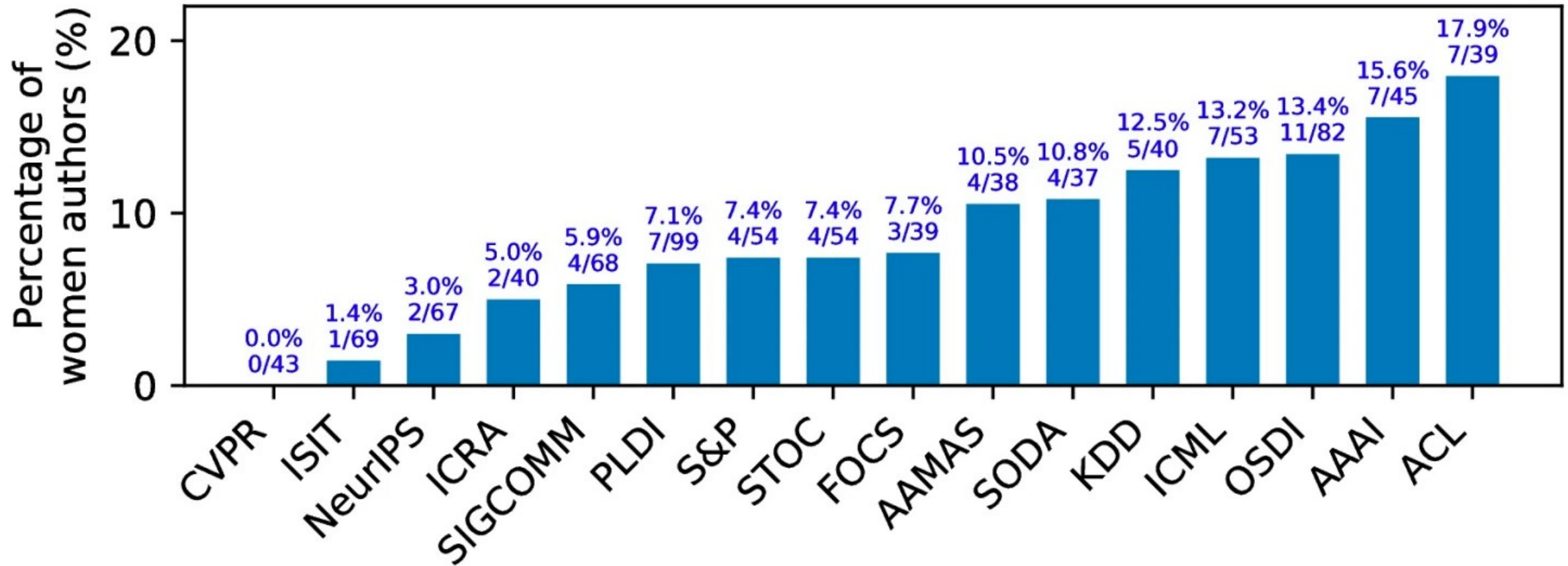
CMU Machine Learning Department website now uses dynamic randomization for ordering people

[www.ml.cmu.edu/people/phd-students.html](http://www.ml.cmu.edu/people/phd-students.html)



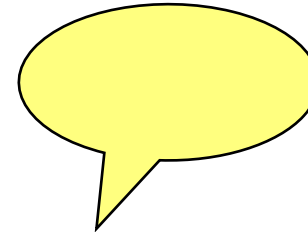
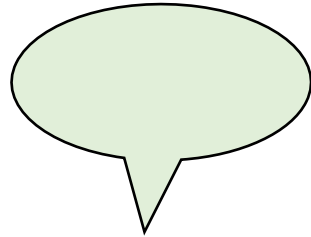
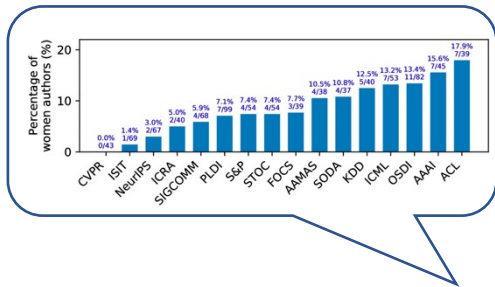
# Gender distribution in paper awards

(2010–2018)

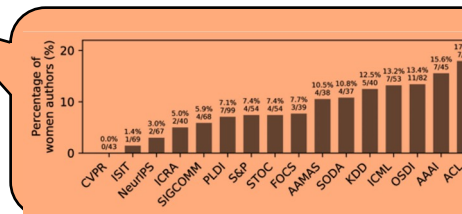
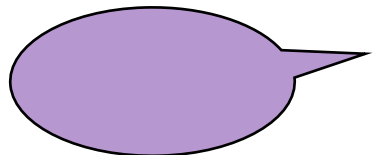


# Need for transparency

- Are author identities visible to the award committee?
- How is the committee determined?
- What criteria are used?



**Started conversations in information theory society,  
NLP community, ML community, vision community,...**





# Norms and Policies: Open problems

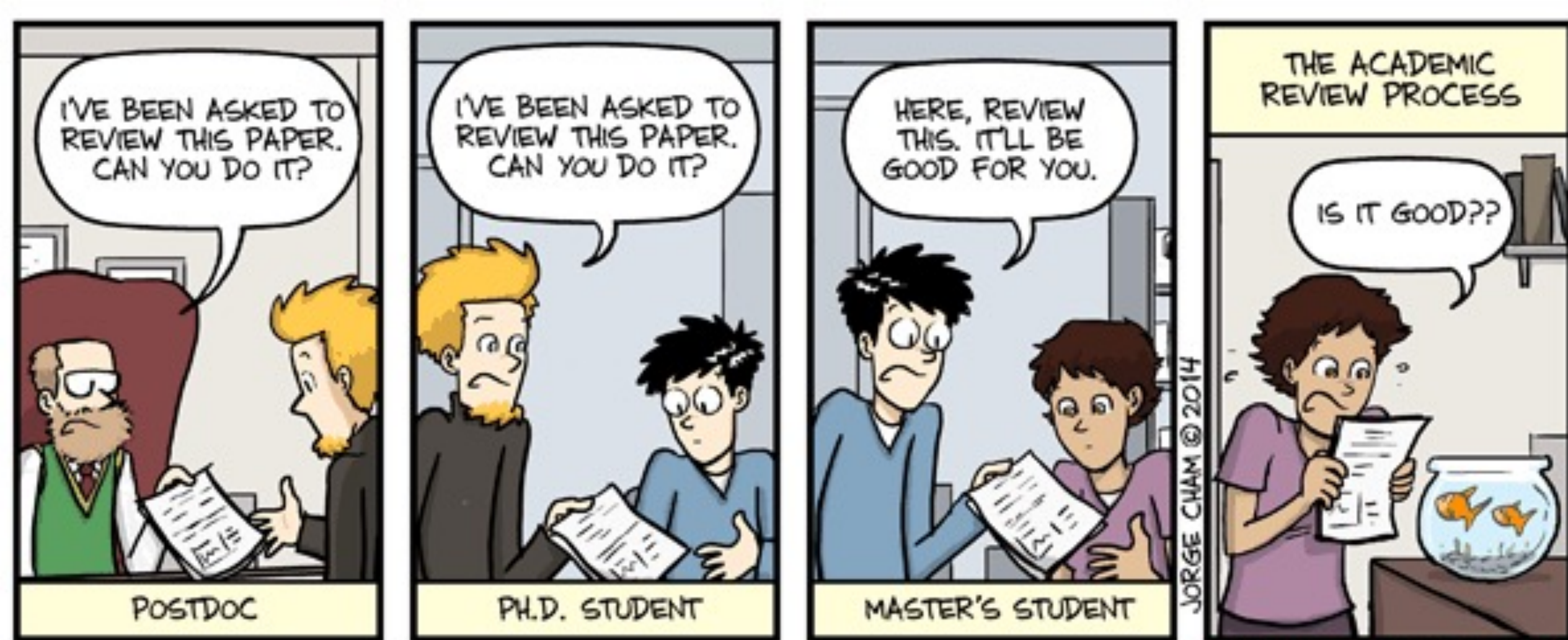


- More experiments: Science for science!
- Privacy-preserving techniques for researchers to use peer-review data [\[Ding et al. 2021\]](#)
- Evaluation metrics for peer-review algorithms and policies [\[Wang et al. 2021\]](#)



# Conclusions

- **Many sources of biases and unfairness in peer review**
- **Urgent need to systematically address challenges in peer review, at scale**
  - Lot at stake: Careers, Scientific progress
- **Lots of open problems!**
  - Exciting
  - Theoretical / Applied / Conceptual
  - Challenging
  - **Impactful**



"Piled Higher and Deeper" by Jorge Cham

# Thank you! Questions?

<http://cs.cmu.edu/~nihars>

[nihars@cs.cmu.edu](mailto:nihars@cs.cmu.edu)

Nihar B. Shah, Carnegie Mellon University