Systemic Challenges and Solutions on Bias and Unfairness in Peer Review

Nihar B. Shah

Machine Learning and Computer Science Departments

Carnegie Mellon University

CHECK IT OUT, I'VE BEEN ASKED TO BE A PAPER REVIEWER FOR NEXT YEAR'S CONFERENCE.
WHOA.
I GUESS THEY CONSIDER ME AN EXPERT IN THE FIELD.
REALLY? WHAT WAS YOUR QUALIFICATION?
JORGE CHAM © 2007
UH, I HAD A PAPER PUBLISHED IN THE SAME CONFERENCE THE YEAR BEFORE.
BUT THAT MEANS YOUR PAPER WAS REVIEWED...
...BY PEOPLE WHOSE ONLY QUALIFICATION WAS TO BE REVIEWED BY THE PREVIOUS YEAR'S PEOPLE WHOSE ONLY QUALIFICATION WAS...
STOP, YOU'RE MAKING MY EGO HURT!
WWW.PHDCOMICS.COM

"Piled Higher and Deeper" by Jorge Cham

Logistics

- On all slides, references are clickable and link to the paper
- Overview article with references: **bit.ly/PeerReviewOverview**
- Please ask questions. Feel free to interrupt! ☺

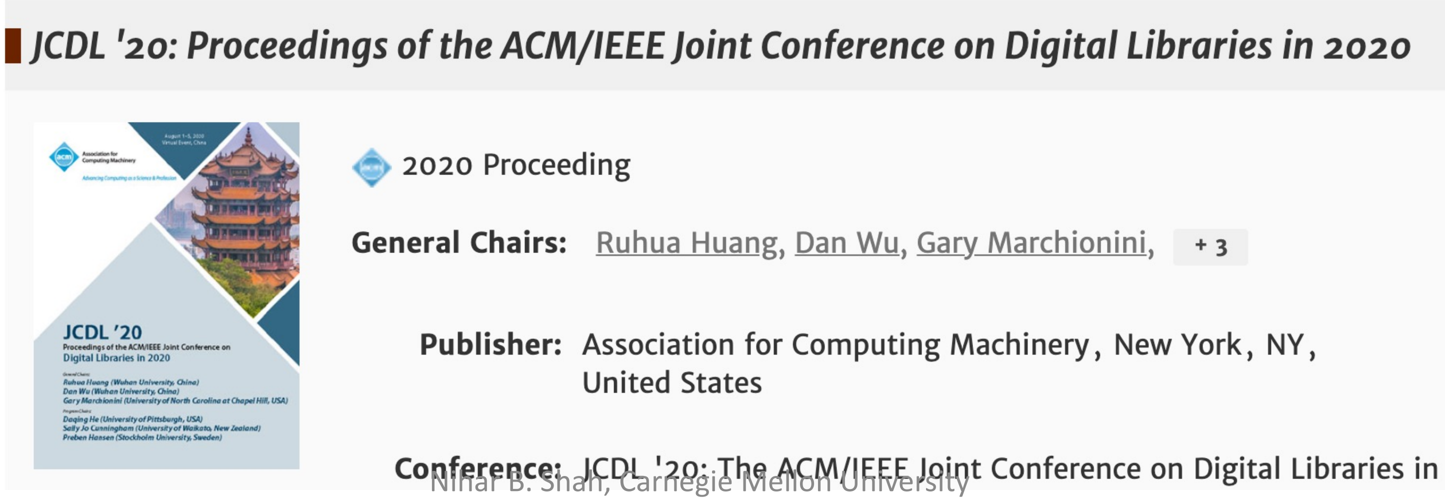
Peer Review

↓

Scientific Digital Libraries

ACM DL DIGITAL LIBRARY

IEEE COMPUTER SOCIETY DIGITAL LIBRARY



JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020

2020 Proceeding

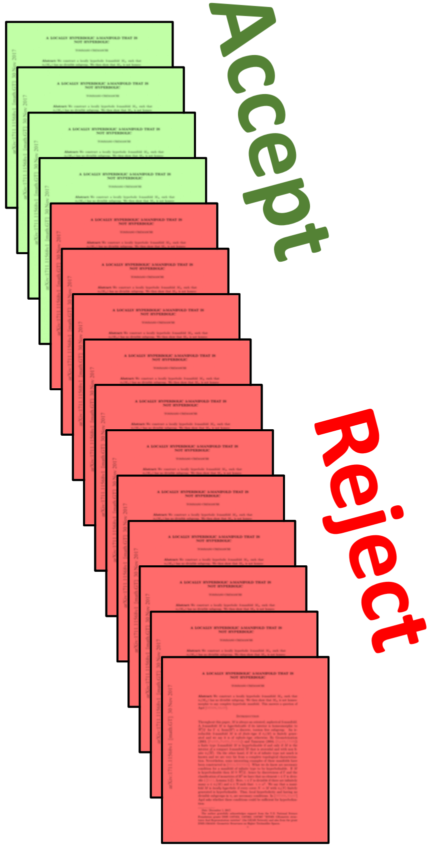
General Chairs: Ruhua Huang, Dan Wu, Gary Marchionini, + 3

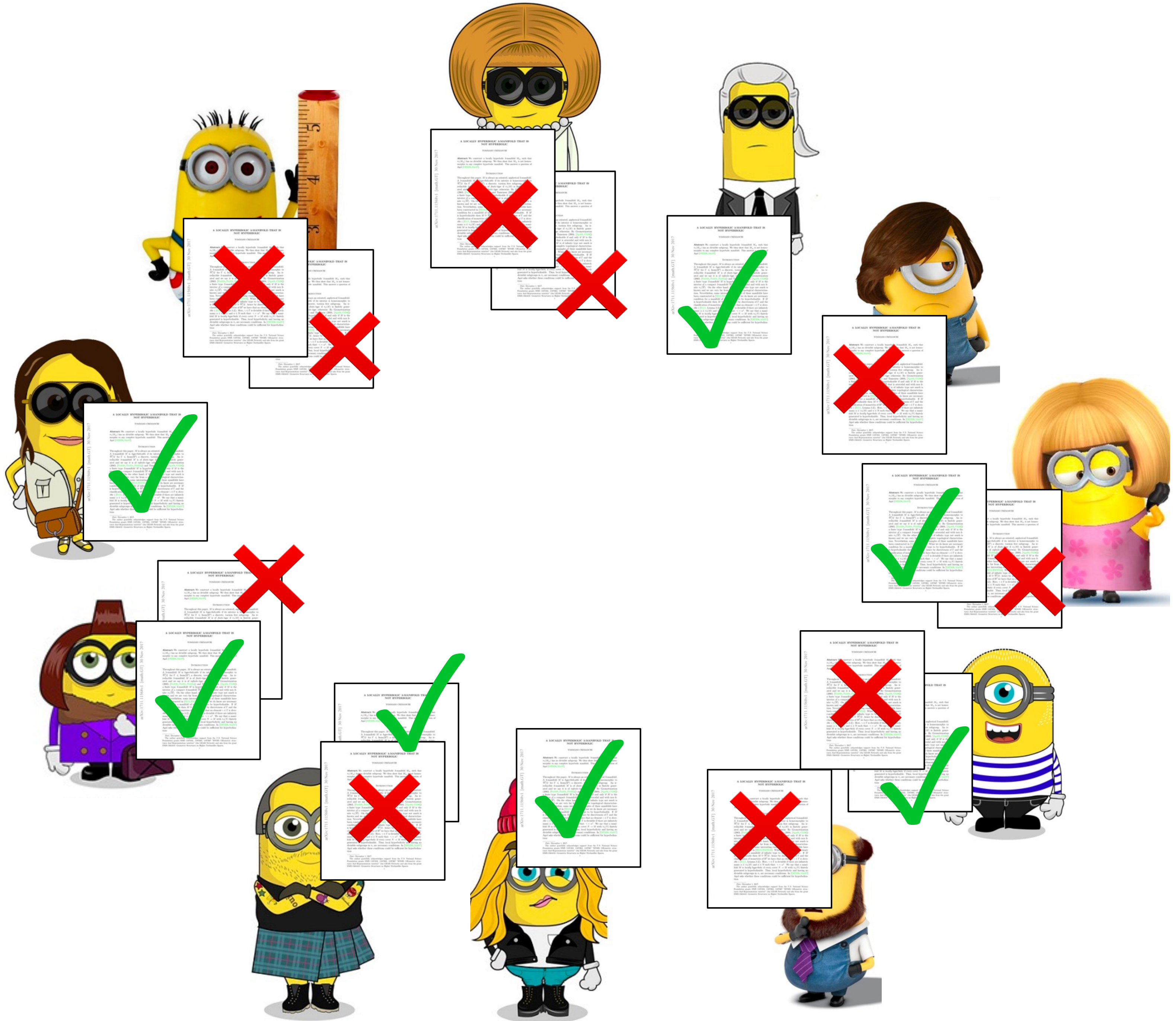
Publisher: Association for Computing Machinery, New York, NY, United States

Conference: JCDL '20: The ACM/IEEE Joint Conference on Digital Libraries in

Peer-review

JCDL 2021



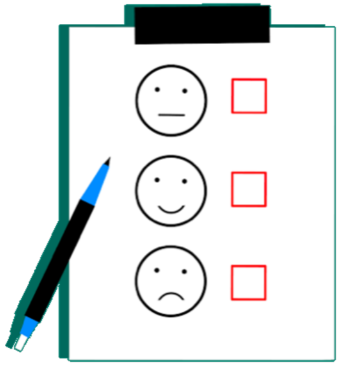


Challenge across many research fields

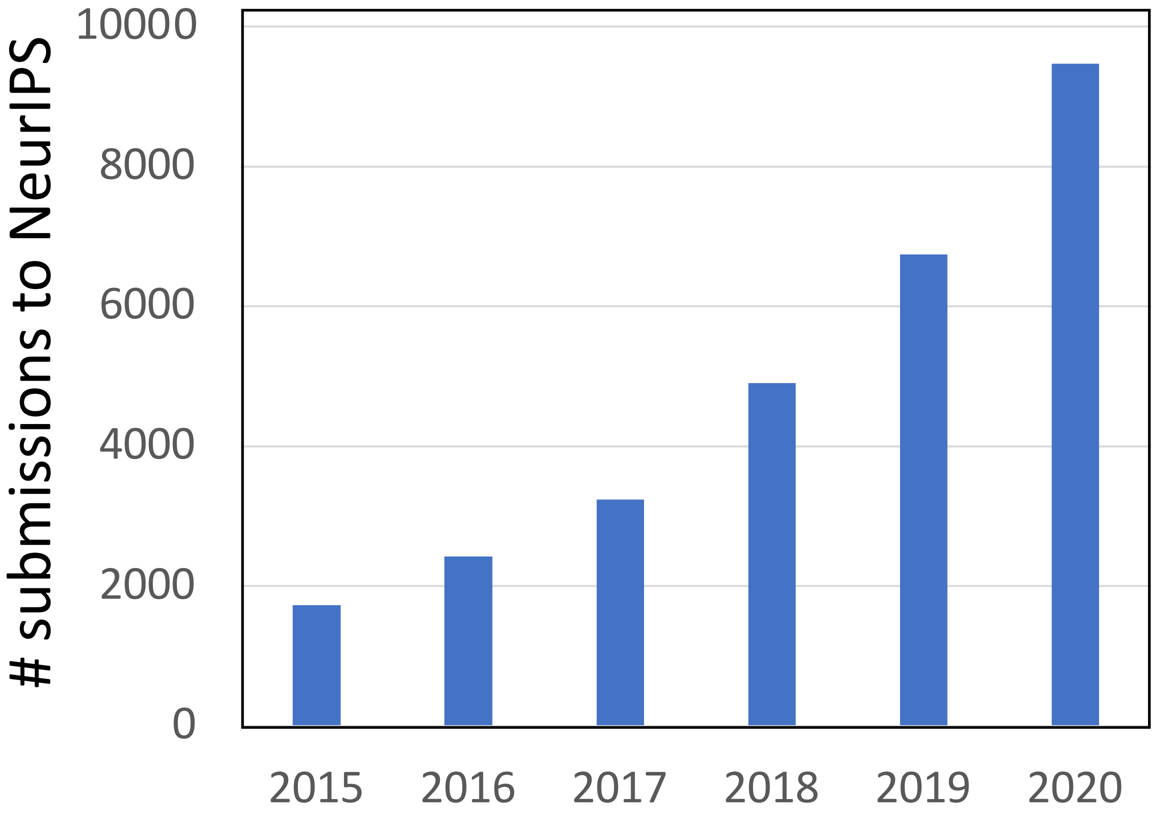
- **“Let's make peer review scientific”** [Rennie, Nature 2016]

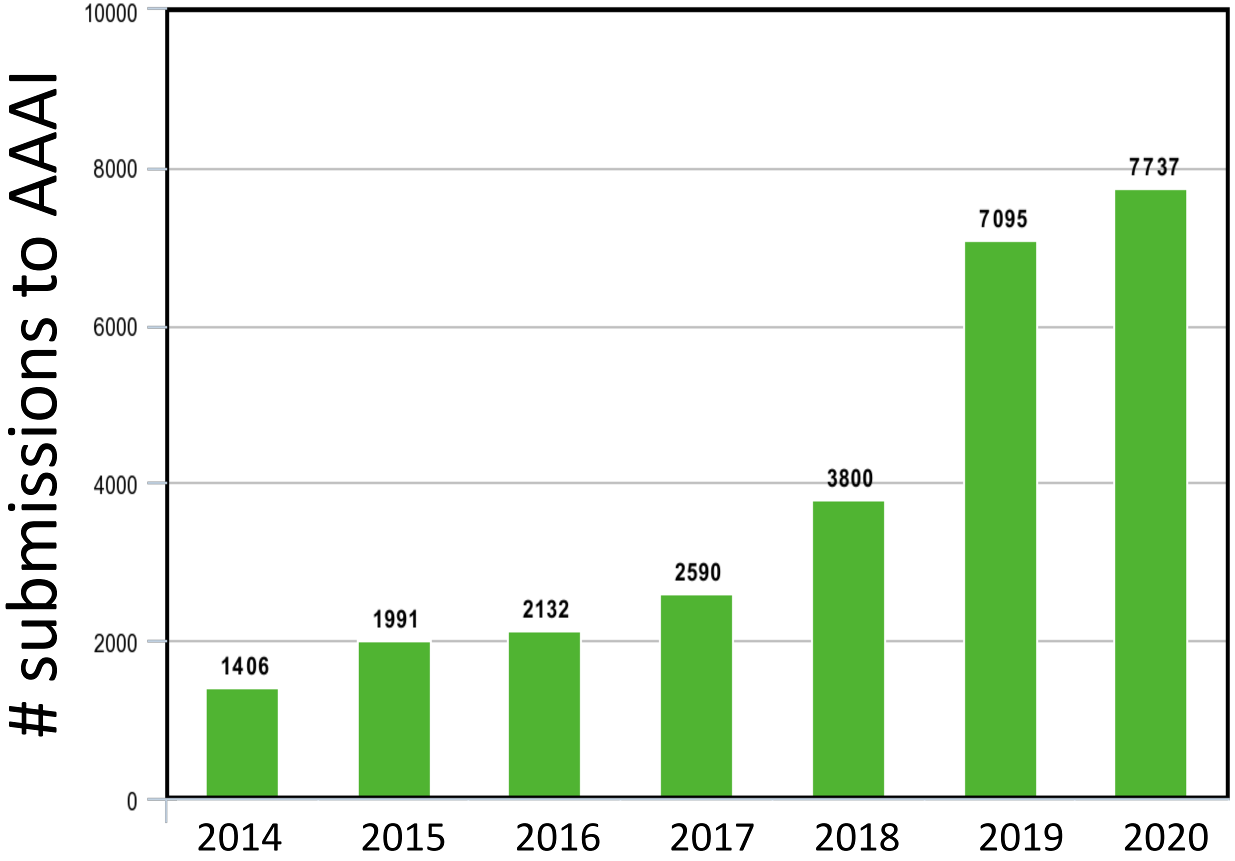
  “Peer review ... is a human system. Everybody involved brings ***prejudices****,* ***misunderstandings*** *and gaps in knowledge, so no one should be surprised that peer review is often* ***biased*** *and* ***inefficient****. It is occasionally* ***corrupt****, sometimes a charade, an open temptation 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]



Several thousand submissions, exponential growth

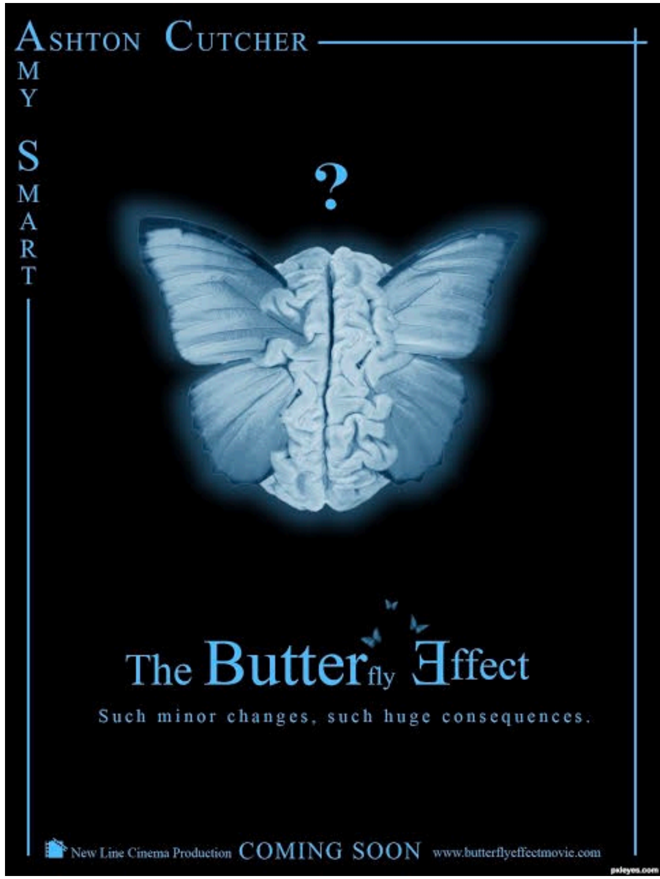




Hurts scientific progress and careers of researchers

“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]


ASHTON CUTCHER
AMY SMART
?
The Butterfly Effect
Such minor changes, such huge consequences.
New Line Cinema Production COMING SOON www.butterflyeffectmovie.com

Peer-review for grant proposals

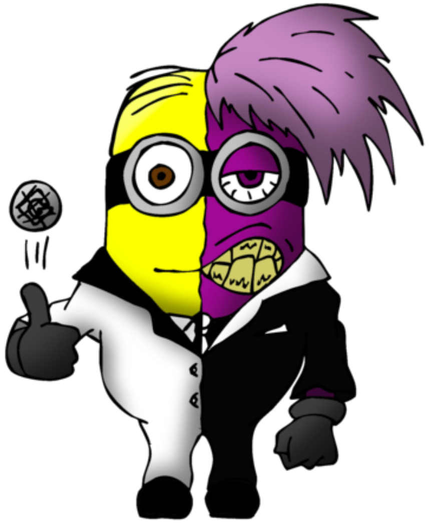
Budget of several billions of $

Peer-evaluation of employees at companies

npr

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

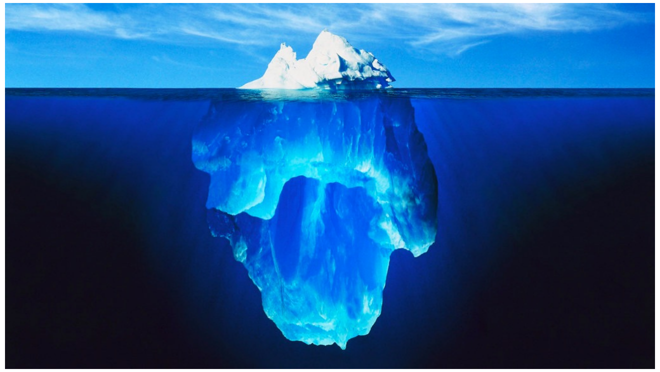
Can make or break careers



“In public, scientists and scientific institutions celebrate truth and innovation. In private, they perpetuate peer review biases that thwart these goals… **what can be done about it?**”
[Lee 2015]

Tutorial Objectives

1) **Challenges** of systemic bias and unfairness in data from people, with a running application of peer review



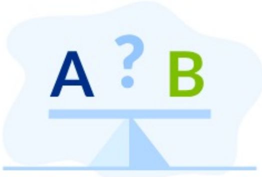
2) **Current research** addressing these challenges

3) How can we **contribute to address** these important open problems?

Broad applicability

Hiring

Admissions



A/B testing

Crowdsourcing

tripadvisor yelp

Product ratings

Healthcare

Peer grading

JCDL 2021

Peer review

…

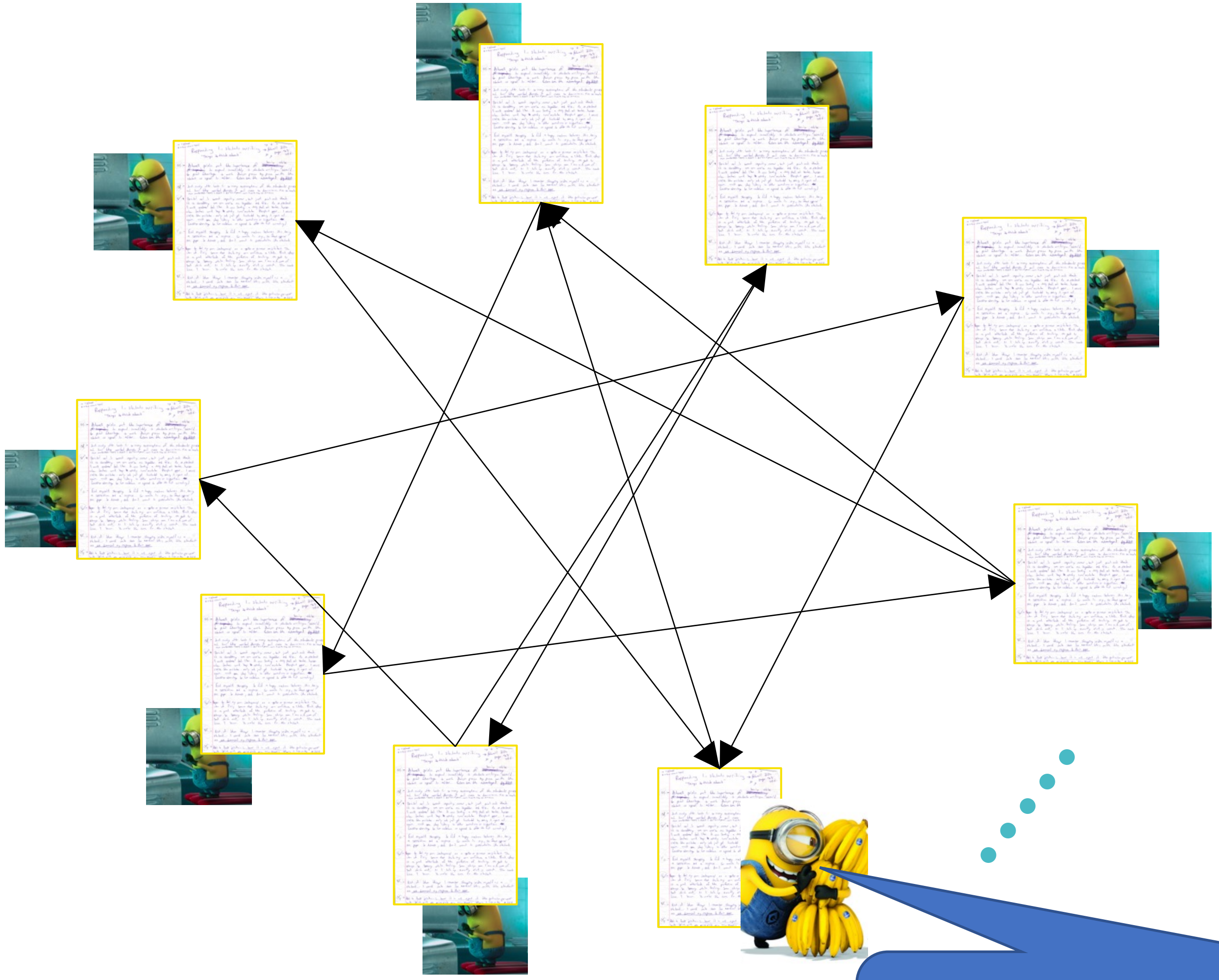
Distributed human evaluations:

Each item evaluated by few people, each person evaluates few items

- Challenges of systemic bias and unfairness
- Problems amplify when this data is used to train AI/ML systems

- Noise
- Fraud
- Miscalibration
- Subjectivity
- Bias regarding author identities
- Norms and policies

Noise



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

[Chapter 3 of bit.ly/PeerReviewOverview]

Noise and reviewer assignment

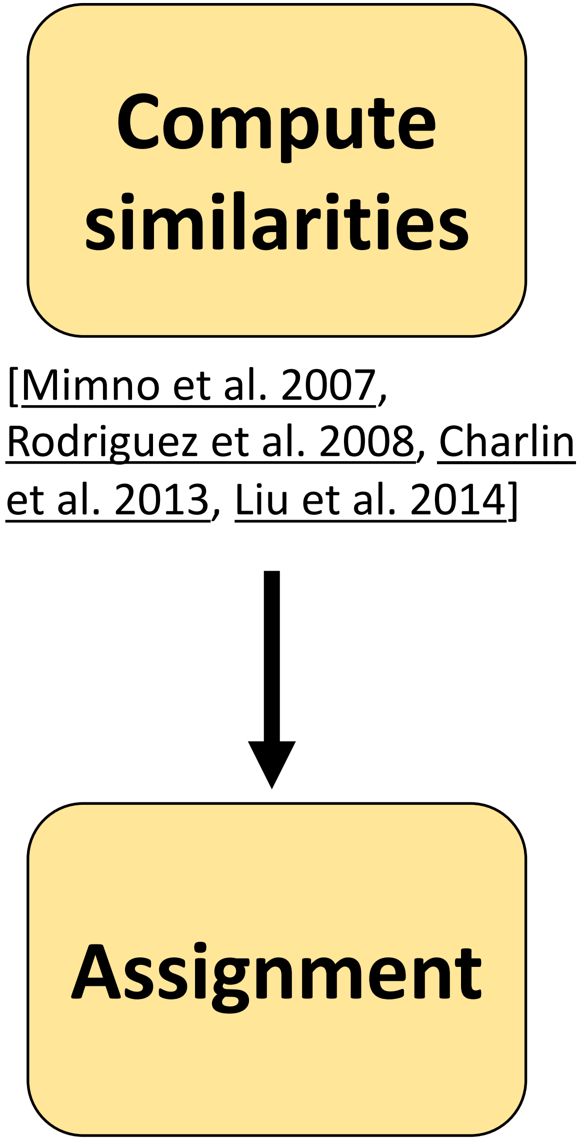
Poor reviews due to inappropriate choice of reviewers

“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,…)



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

- For every pair (paper p, reviewer r), similarity score $\boldsymbol{s_{pr} \in [0,1]}$
- Higher similarity score ⇒ 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,…)

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

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, 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

[Stelmakh et al. 2018]

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

[Stelmakh et al. 2018]

Common approach: Maximize total similarity

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

- **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

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

subject to

Every paper gets at least certain #reviewers

Every reviewer gets at most certain #papers

No paper is assigned to conflicted reviewer

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

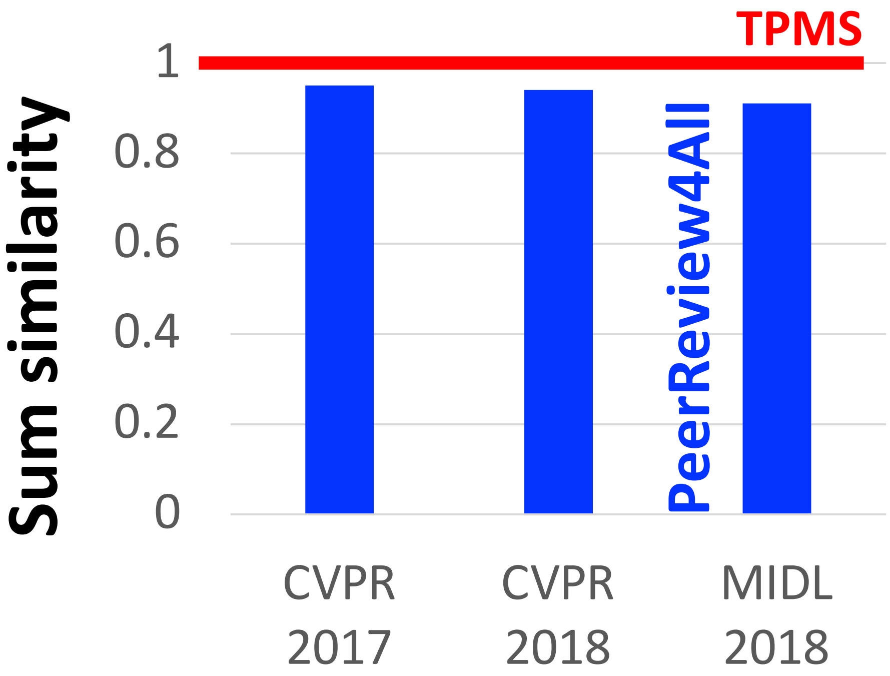
Repeat for remaining papers

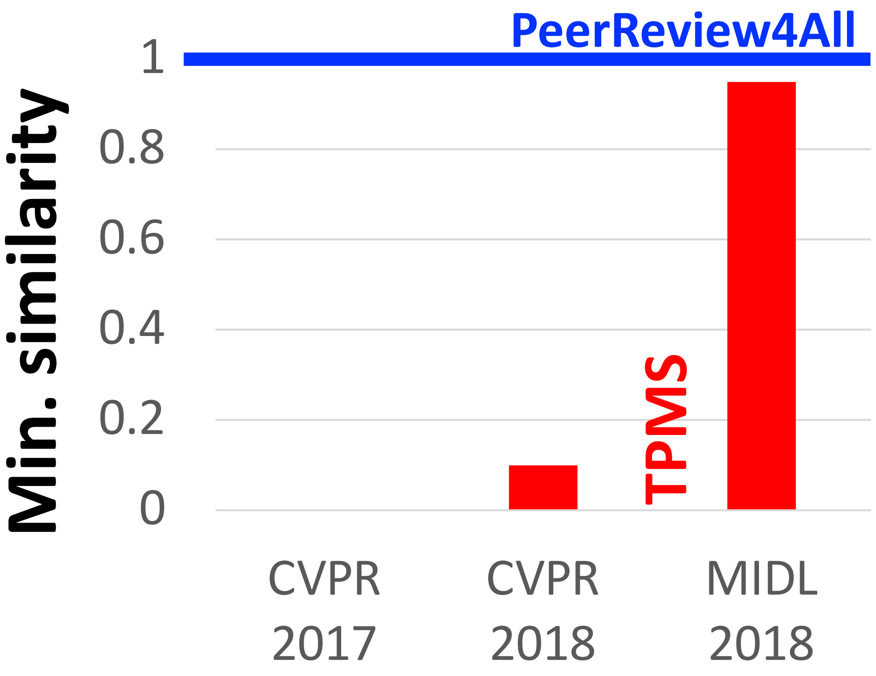
- NP Hard [Garg et al. 2010]
- Approximation algorithm ("PeerReview4All")
- Statistical guarantees on overall top-K selection

[Stelmakh et al. 2018]

Evaluation

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





[Evaluations by Kobren et al. 2019]

Noise: Open problems

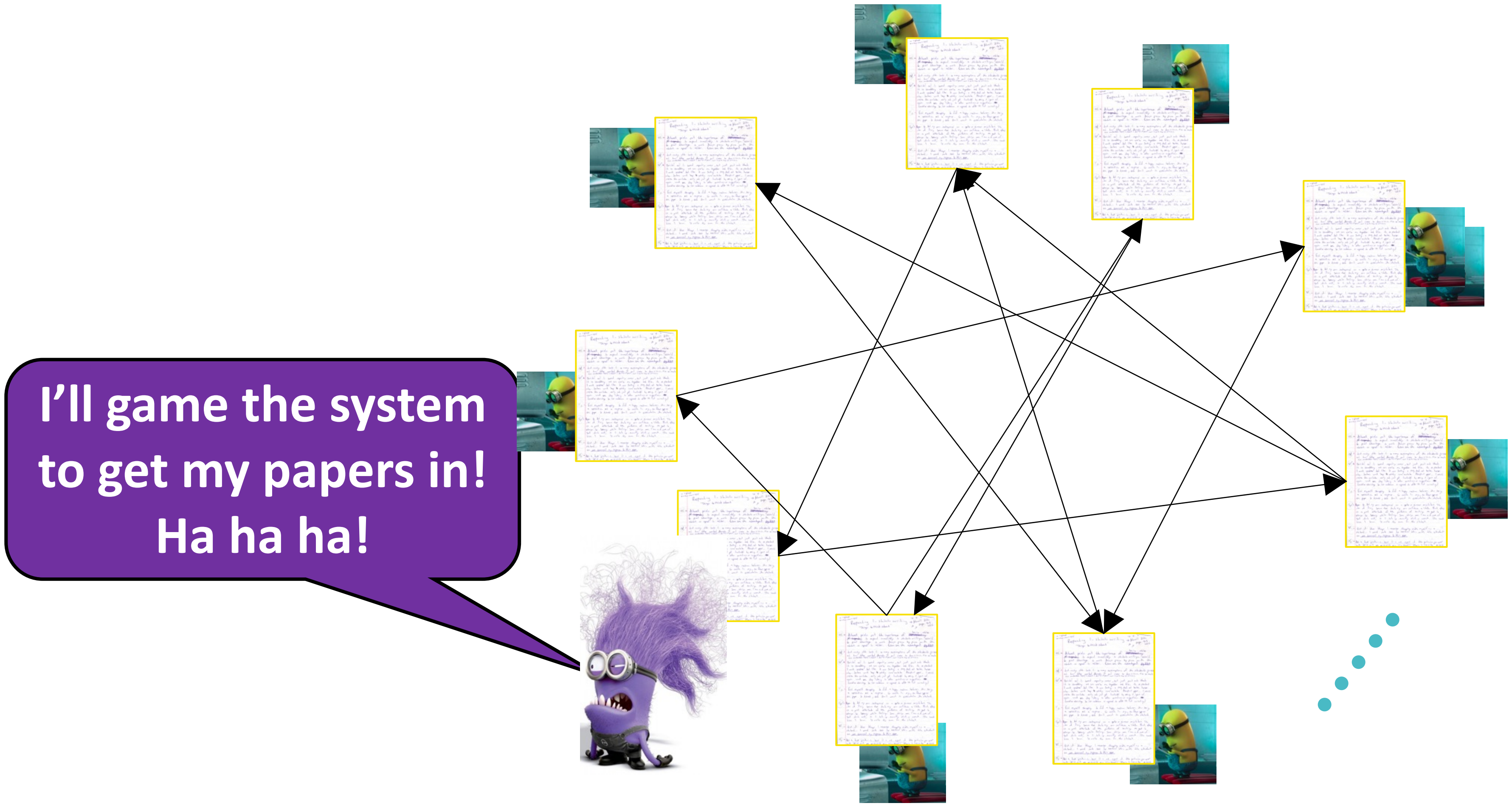
- Better computation of similarities
  - Interdisciplinary papers
  - Joint similarity computation and assignment

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

- Fair and improved bidding process [Fiez et al. 2019, Meir et al. 2020]

- How to combine various sources of data to form similarities?
  - Currently use an arbitrary pre-defined formula
  - NeurIPS 2016: Similarity = 2^{bid} (text-match + subject-match) [Shah et al. 2018]

Fraud

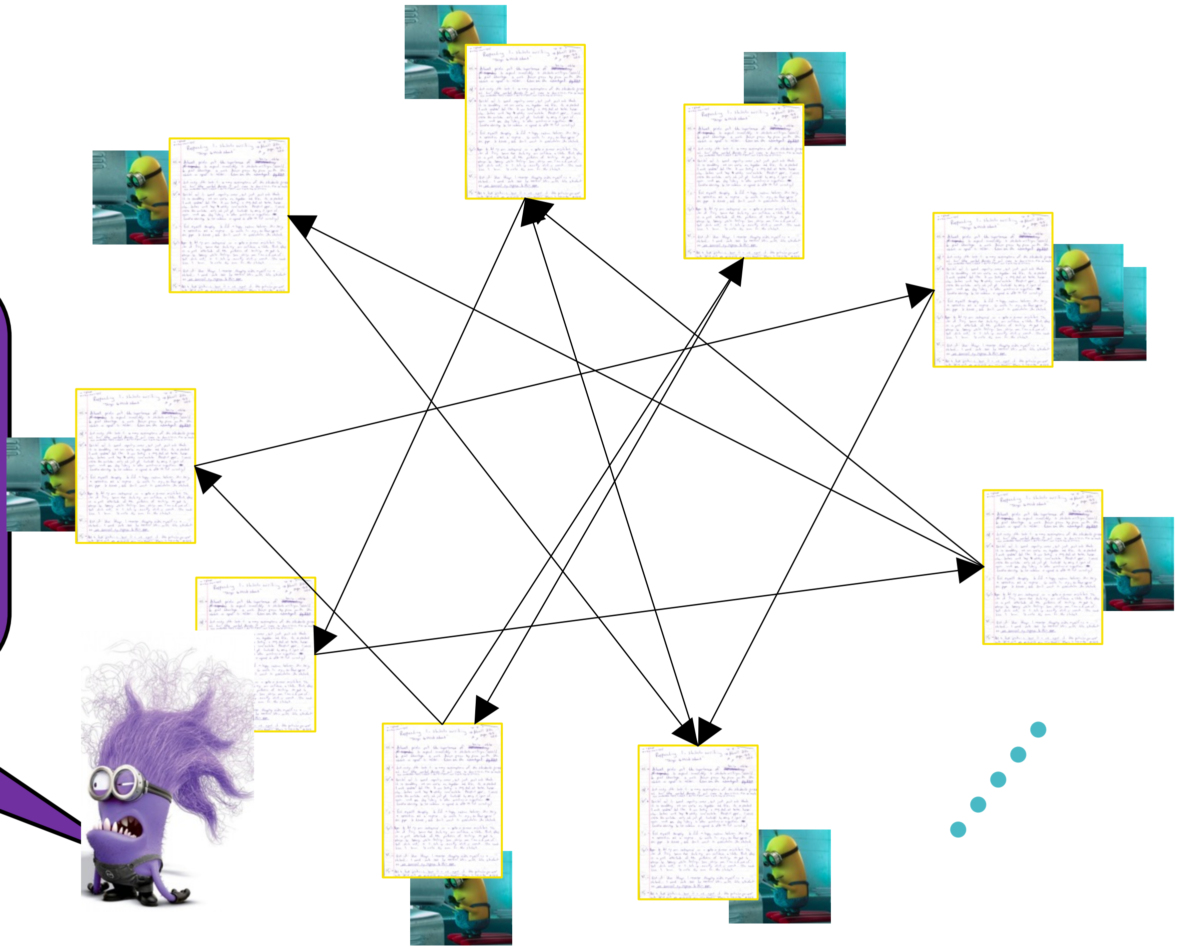


[Chapter 4 of bit.ly/PeerReviewOverview]

Fraud

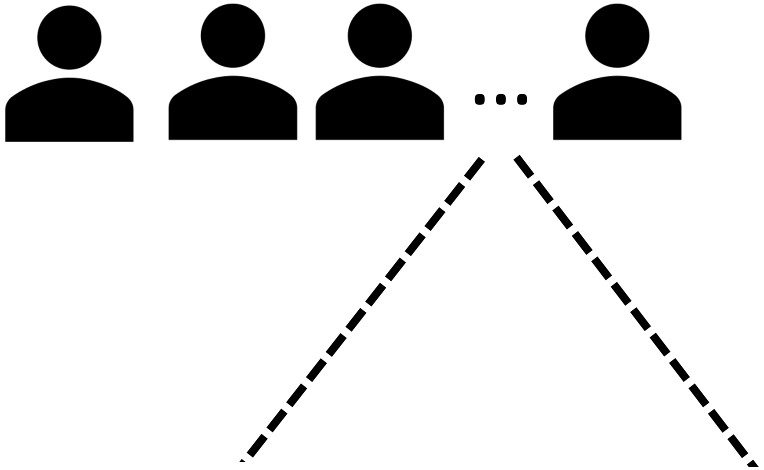
1. Lone wolf
2. Coalition

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

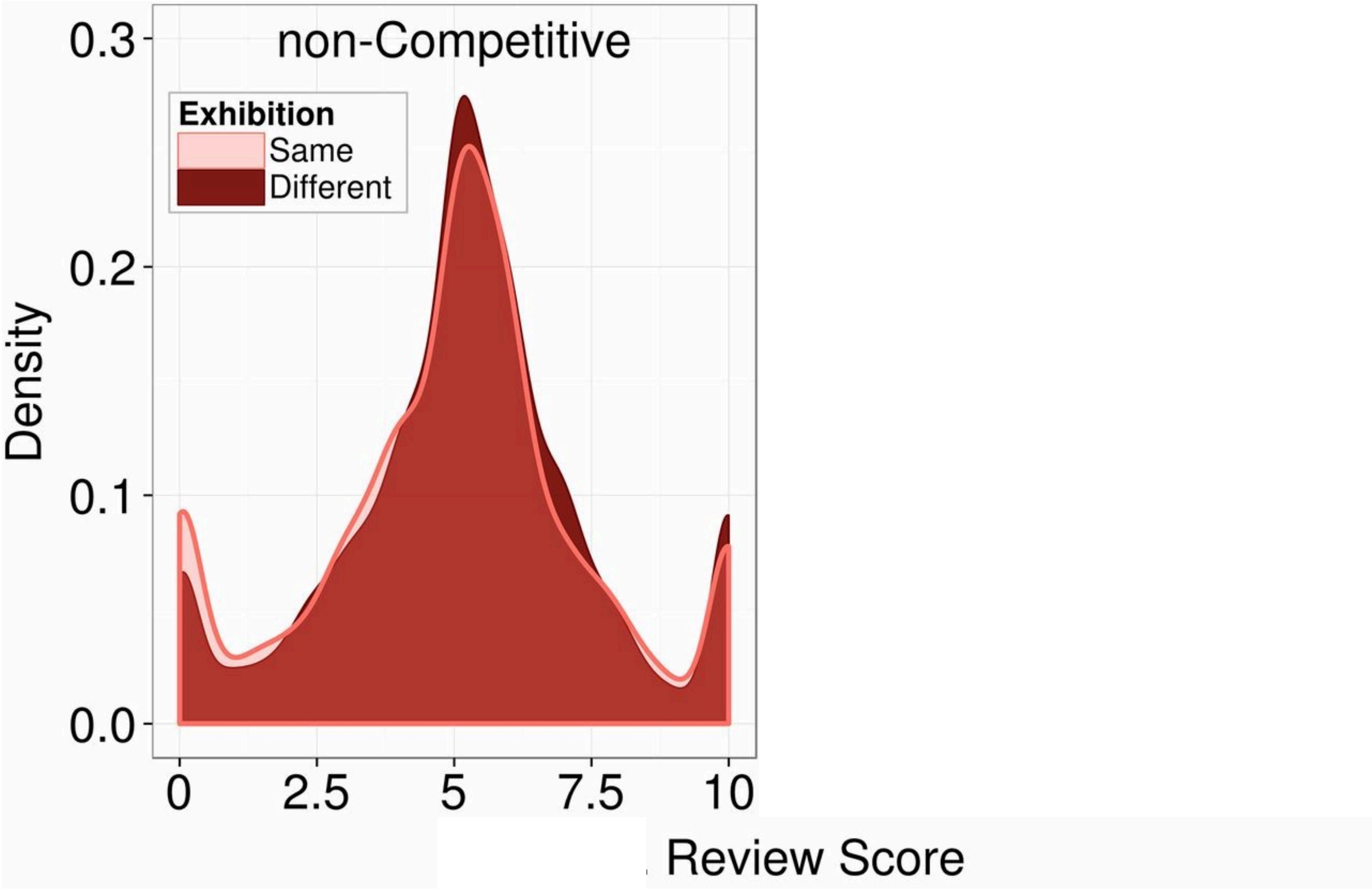
All above certain threshold get award

Competitive

Top certain fraction in each exhibition win award

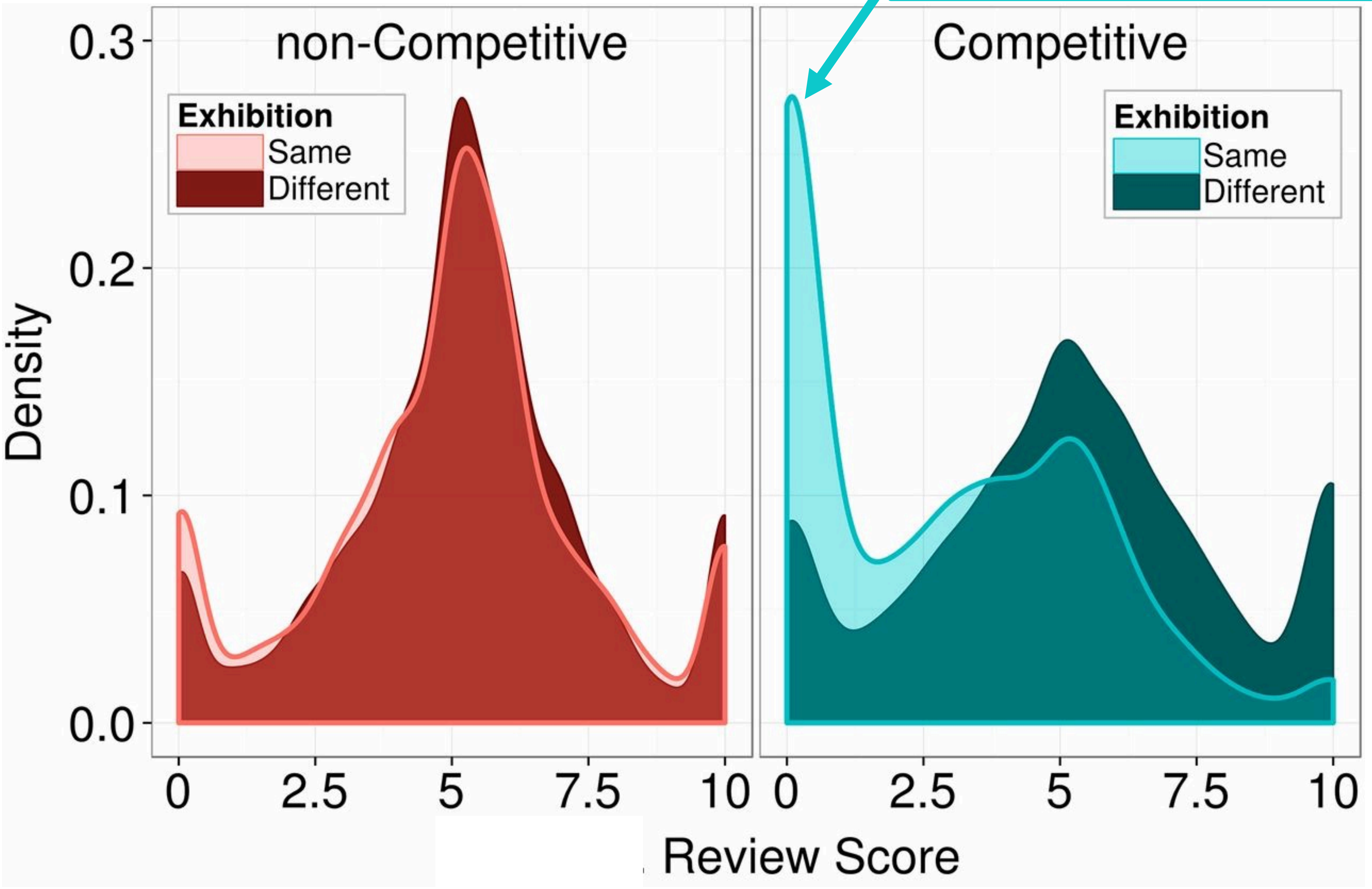
- Each participant knows which exhibition their drawing belongs, and if it is competitive or not
- Each participant also told the exhibition to which the drawings they are reviewing belong

[Balietti et al., 2016]

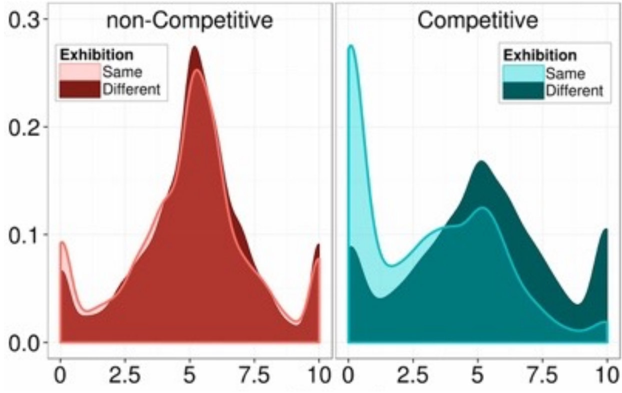


[Balietti et al., 2016]

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



[Balietti et al., 2016]



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

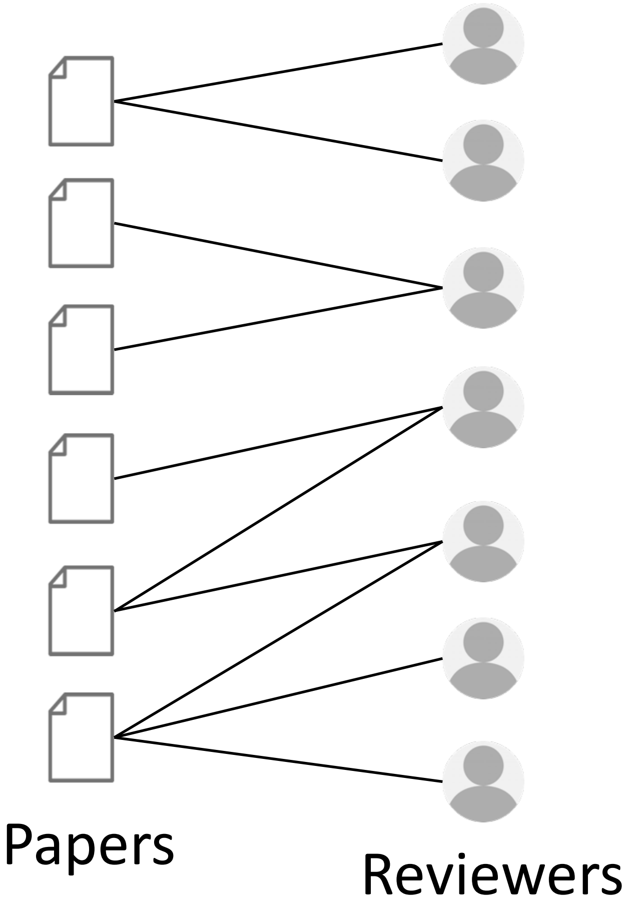
“This result provides further evidence that a substantial amount of gaming of the review system is taking place… competition incentivizes reviewers to behave strategically, which reduces the fairness of evaluations”

[Balietti et al., 2016]

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

How to make peer review strategyproof?

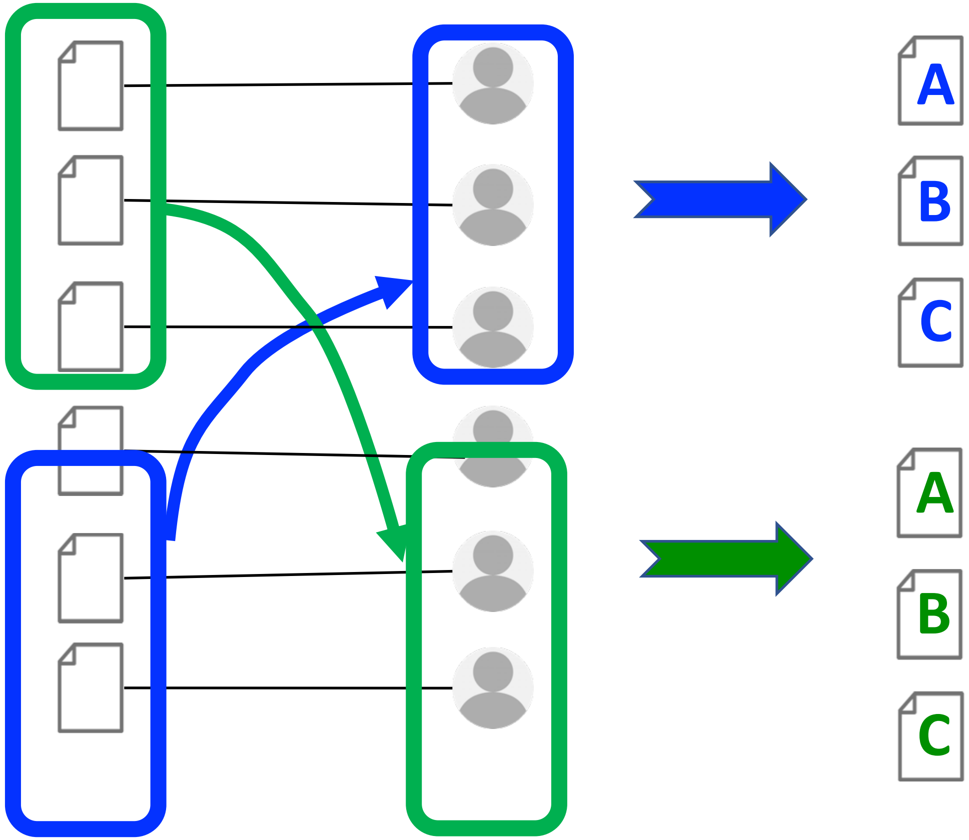
Given: Authorship graph



How to ensure that no reviewer can influence decision of their own paper?

Partitioning method

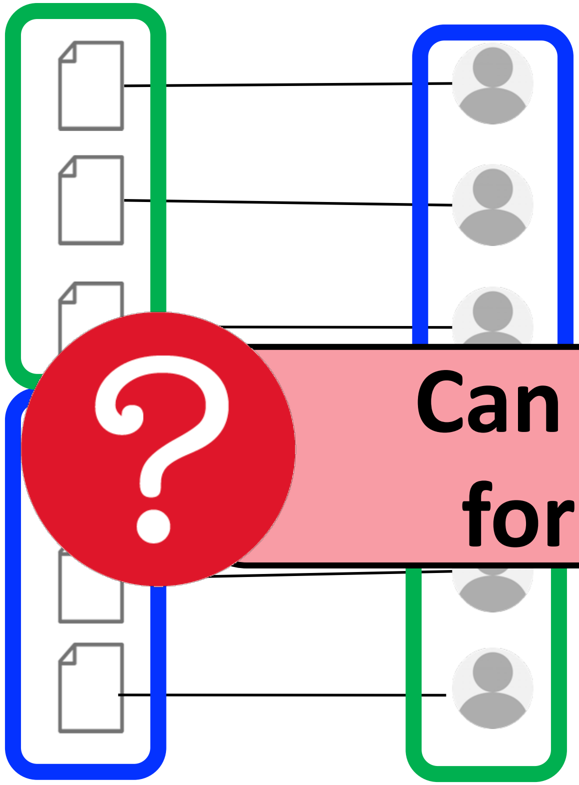
Primarily studied for peer grading



[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]

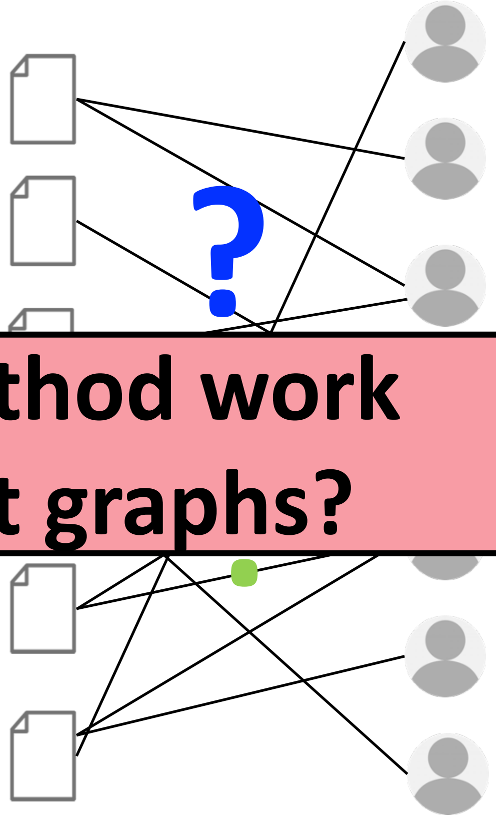
Peer grading

1-1 conflict graphs



Conference peer review

More **complex** conflict graphs



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

[Xu et al. 2018]

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.

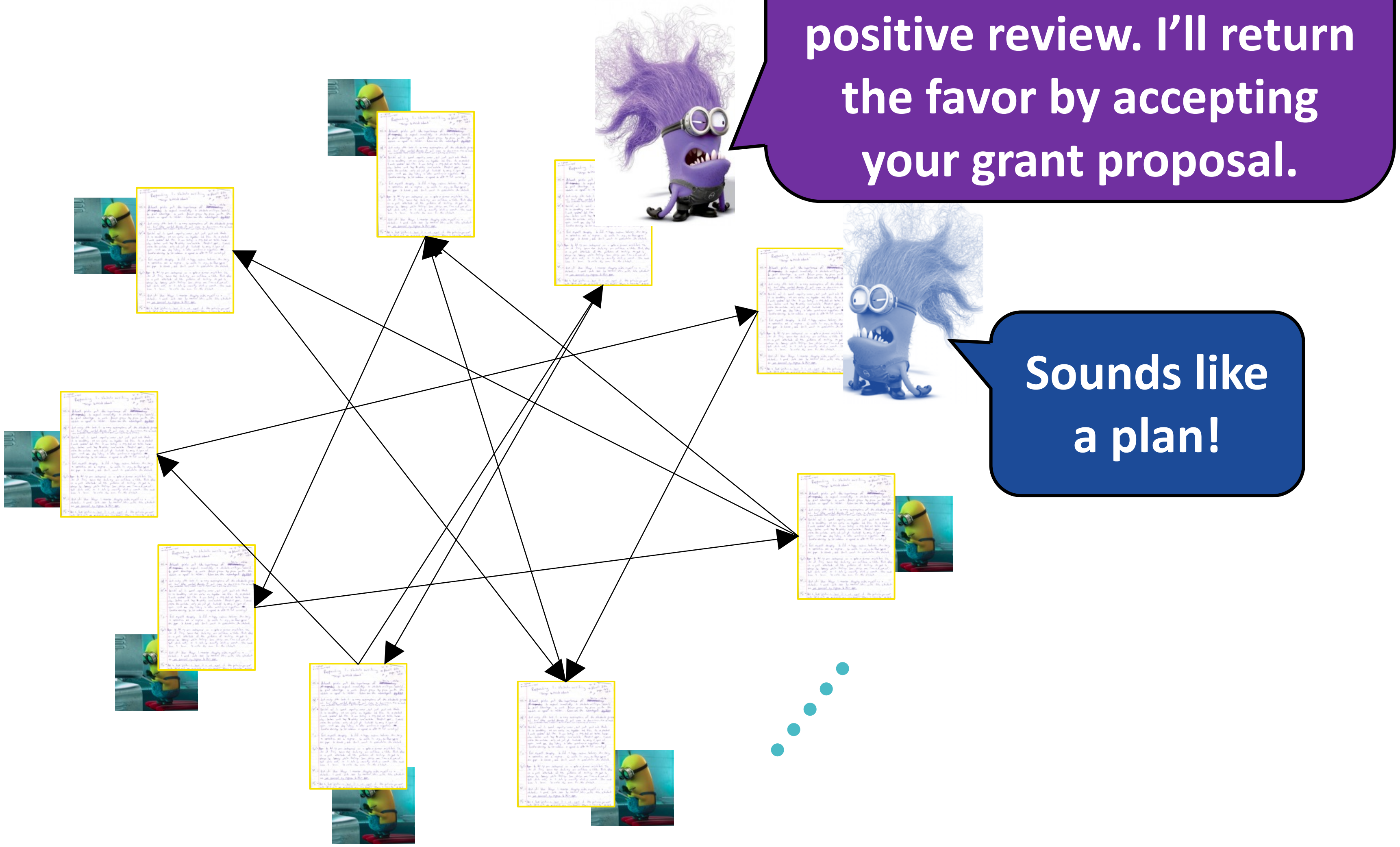
- Heuristics for more flexibility: Removing 3.5% of reviewers from the reviewer pool reduces size of the largest component by 86%

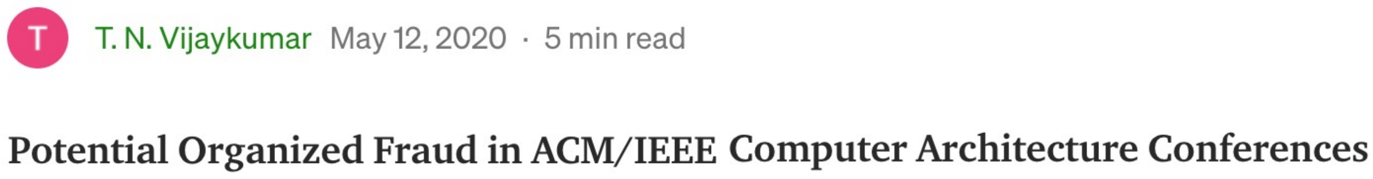
[Xu et al. 2018]

Fraud (Lone wolf): Open problems

- Maximum sum similarity under partitioning-based method? [Xu et al. 2018]
- Is strategyproofing possible when conflict graph cannot be partitioned? [Aziz et al. 2019]
- Detecting such fraud [Stelmakh et al. 2021]

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]

[https://medium.com/@tnvijayk/potential-organized-fraud-in-acm-ieee-computer-architecture-conferences-ccd61169370d]

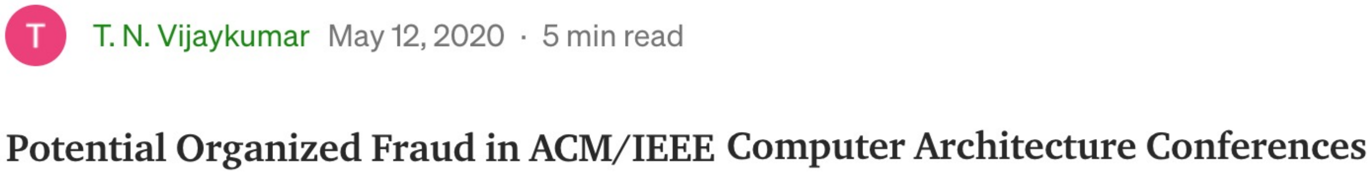
Defense 1: Conflicts of Interest

CoI

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

Challenges:

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



T 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).”

Rings

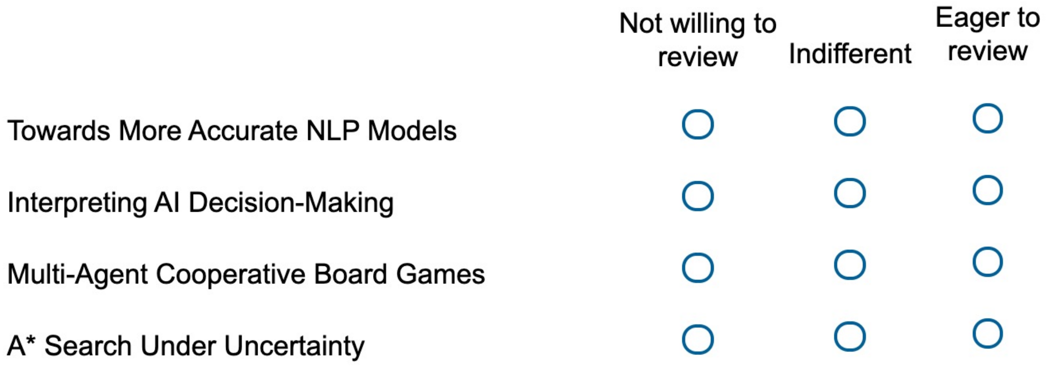
Defense 2: Detect or Remove Rings

[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	○	○	○
Interpreting AI Decision-Making	○	○	○
Multi-Agent Cooperative Board Games	○	○	○
A* Search Under Uncertainty	○	○	○

- 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.

Bids

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]

Bids

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.

→

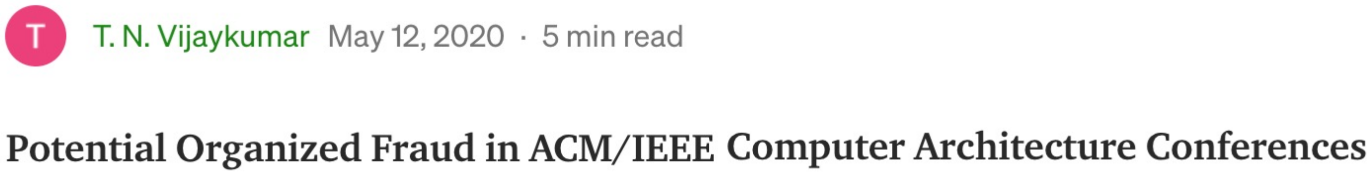
Visible to humans:

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

[Jecmen et al. 2020]

Recall: Automated assignment

$$\underset{\text{assignment}}{\text{maximize}} \sum_{p\,\in\,\text{Papers}} \; \sum_{r\,\in\text{Reviewers}} s_{pr}\, \mathbb{1}\{\text{paper } p \text{ assigned to reviewer } r\}$$

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 $\boldsymbol{Q} \in [\mathbf{0}, \mathbf{1}]^{\#\mathbf{papers} \times \#\mathbf{reviewers}}$ such that

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

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

[Jecmen et al. 2020]

Randomized assignment

Example: $\mathrm{Q_{ij}} = 0.5 \; \forall \; \boldsymbol{i}, \boldsymbol{j}$

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

subject to

Every paper gets **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

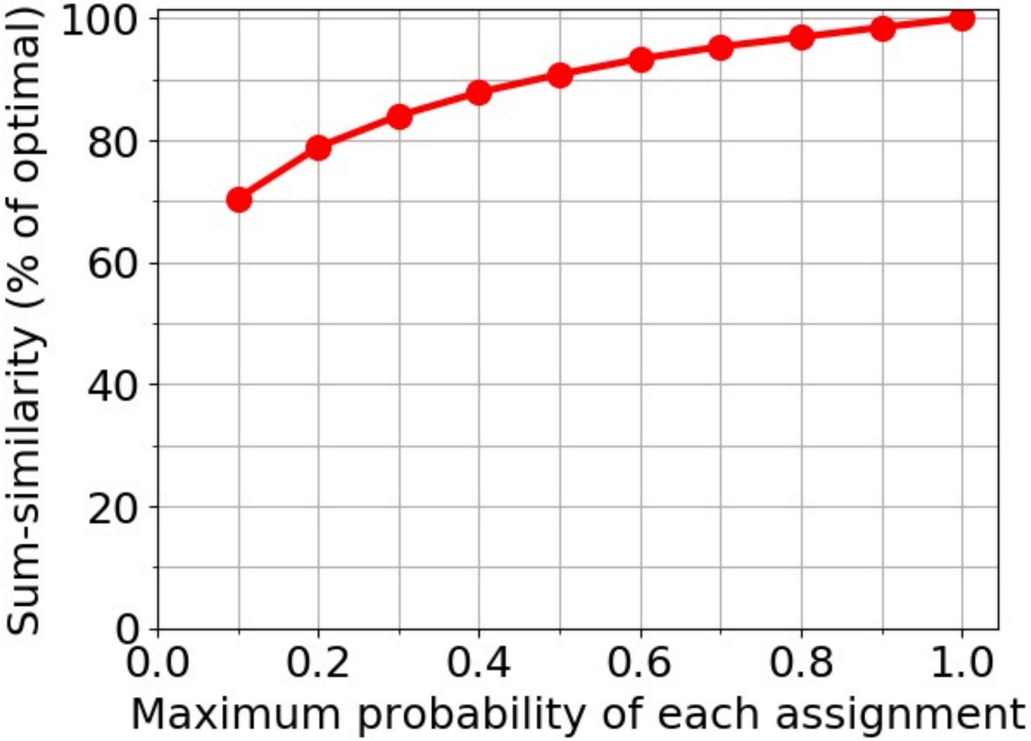
Every reviewer gets at most **3** papers

P(any reviewer assigned to any paper) ≤ 0.5

[Jecmen et al. 2020]

How about expertise?

ICLR 2018



Any reviewer has at best a 50% chance of getting a paper

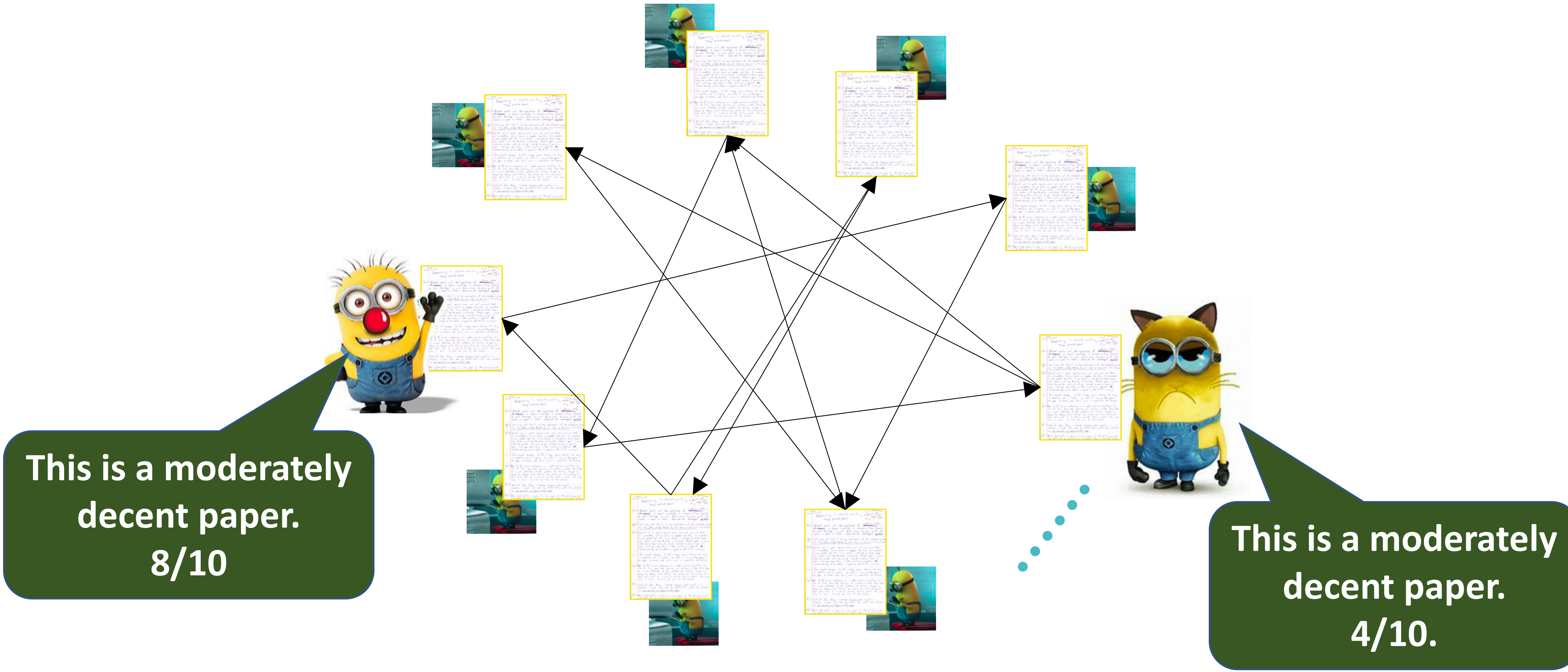
Sum similarity is 90% of original

[Jecmen et al. 2020]

Fraud (Coalitions): Open problems

- Detect such fraud [Wu et al. 2021]
- (Game-theoretic) equilibria?
  - Game between program chairs and colluders
- Other kinds of dishonest behavior [Ferguson et al. 2014, Gao et al. 2017, Lauer et al. 2019]

Miscalibration


This is a moderately decent paper. 8/10
This is a moderately decent paper. 4/10.

[Chapter 5 of bit.ly/PeerReviewOverview]

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]

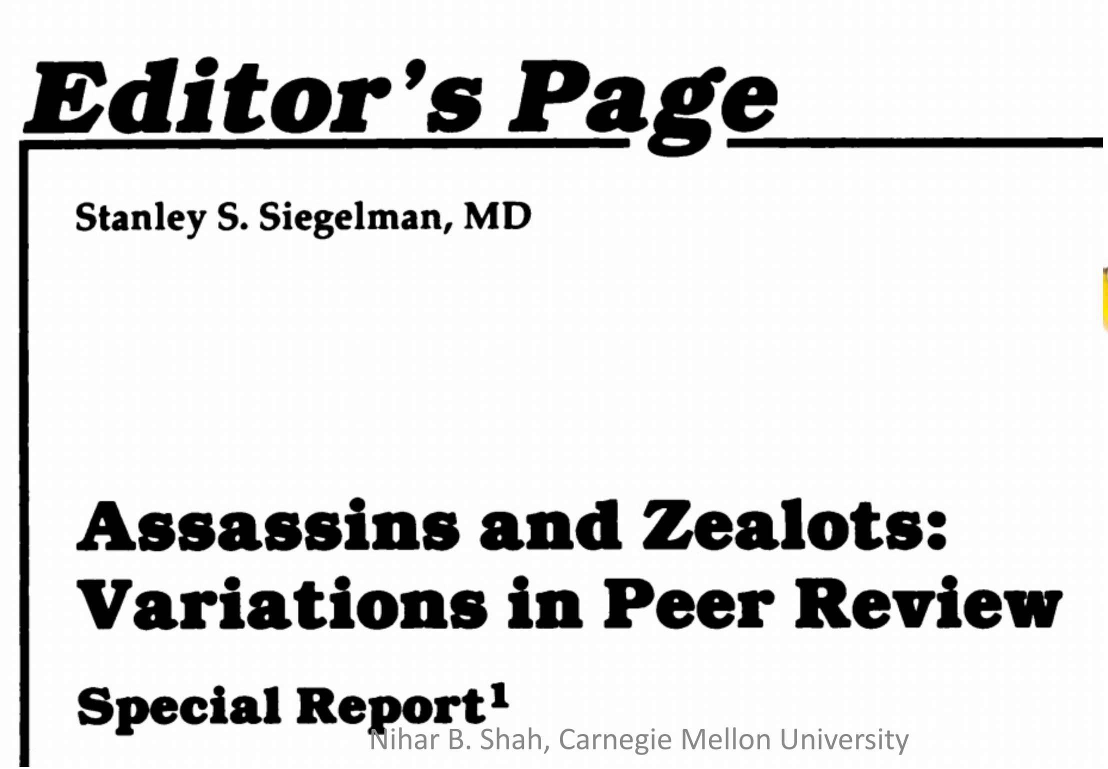
“The rating scale as well as the individual ratings are often arbitrary and may not be consistent from one user to another.” [Ammar et al. 2012]

“[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."

[Siegelman 1991]



Editor's Page

Stanley S. Siegelman, MD

Assassins and Zealots: Variations in Peer Review

Special Report[1]

Two approaches in the literature

1 Assume simplified (affine) models for calibration

- Every paper i has some "true" quality Θ_i
- Every reviewer j has two implicit parameters α_j and β_j

- Model assumes that score given by reviewer j to a paper i:

$$\alpha_j \Theta_i + \beta_j + \text{noise}$$

- Algorithms estimate Θ_i's (as well as α_j's and β_j's) from observed scores

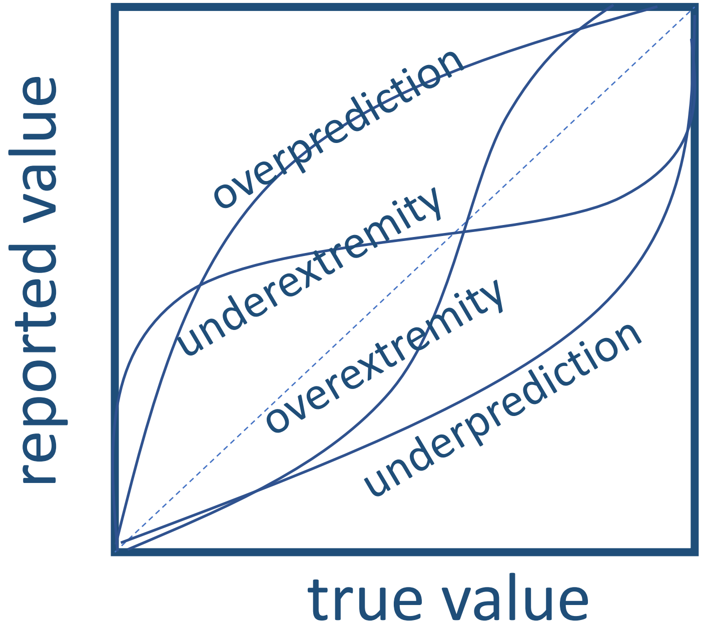
[Paul 1981, Flach et al. 2010, Roos et al. 2011, Baba et al. 2013, Ge et al. 2013, Mackay et al. 2017]

1 Assume simplified (affine) models for calibration

- 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]



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

?

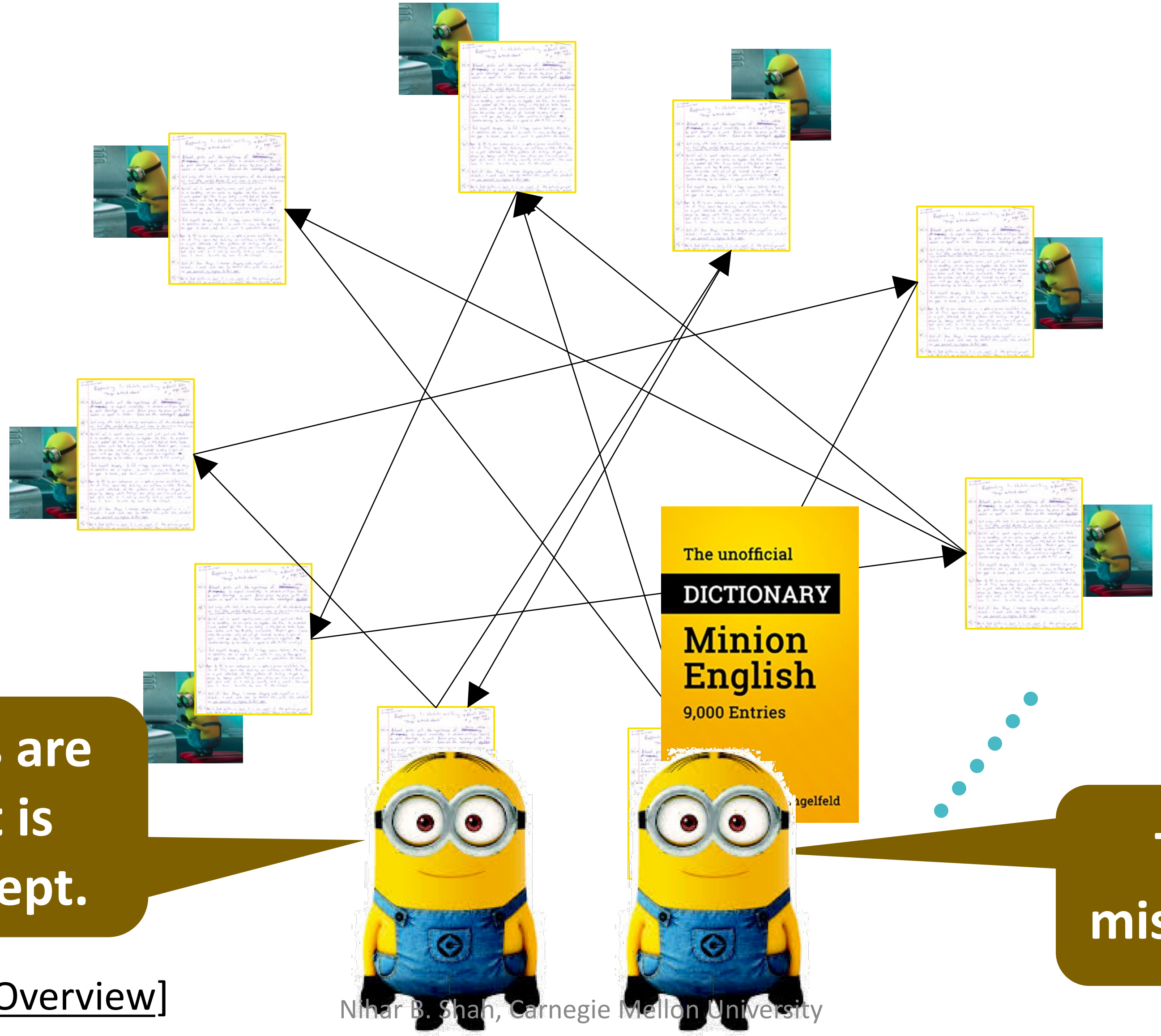
Is it possible to do better using ratings than rankings, with essentially no assumptions on the miscalibration?

Yes! But decisions need to be randomized. [Wang et al. 2018]

Miscalibration: Open problems

- Very small sample size per reviewer (especially in conferences)
  - Approach: Privacy-preserving sharing of some data [Ding et al. 2021]
  - Estimate of reviewer calibration from past conferences
  - Should not hurt reviewer confidentiality
- Better models and calibration algorithms
  - Capture more complexity than affine
- Use rankings and ratings together
  - E.g., use rankings to break ties in ratings
  - About 40% of ratings given by a reviewer to a pair of papers are tied in NeurIPS 2016 [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.

[Chapter 6 of bit.ly/PeerReviewOverview]

Subjectivity

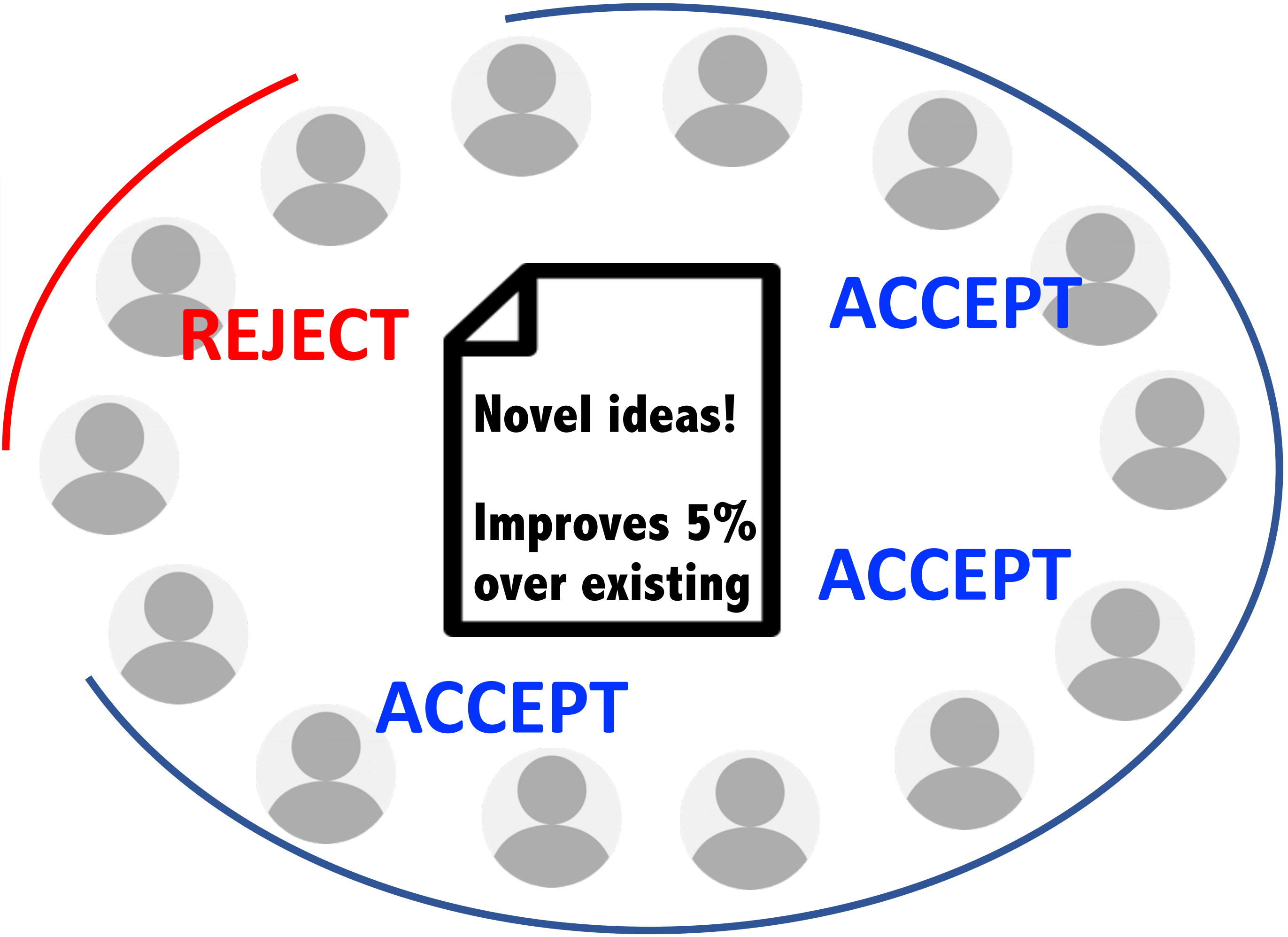
1. Commensuration bias
2. Confirmation bias
3. Dr. Fox effect
4. Hindering novelty

Subjectivity

1. Commensuration bias
2. Confirmation bias
3. Dr. Fox effect
4. Hindering novelty

Differing opinions about relative importance of criteria

Novelty is not useful unless improvement by at least 10%



REJECT

Novel ideas!

Improves 5% over existing

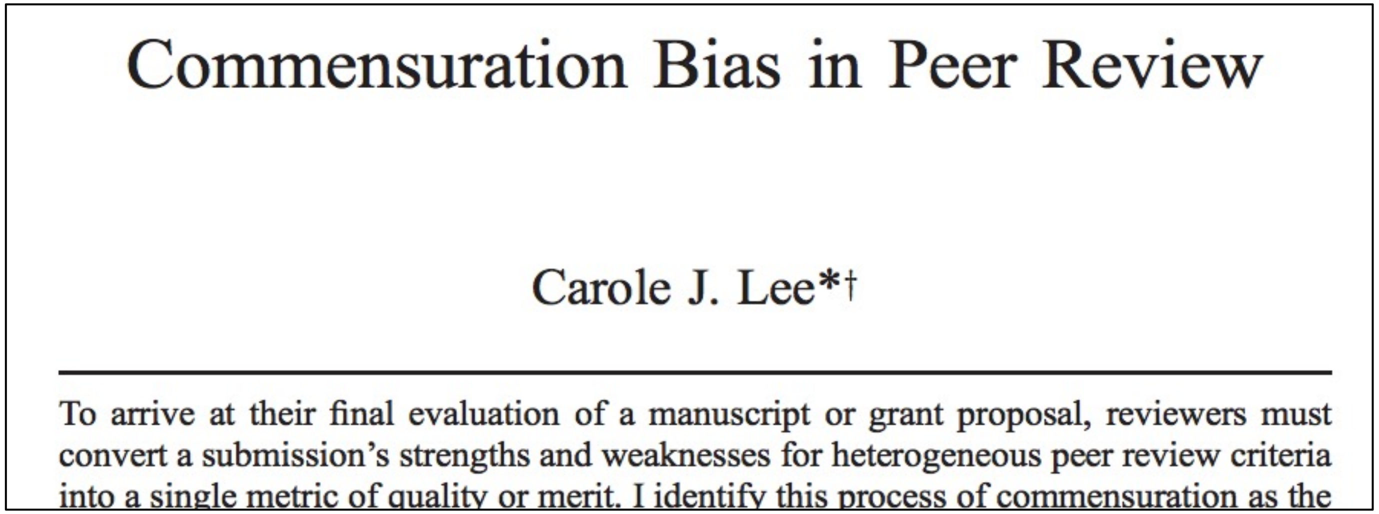
ACCEPT

ACCEPT

ACCEPT

Novelty is extremely important

[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?

[Lee 2015]

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 reviews**

[Noothigattu et al. 2018]

Handcrafted design?

AAAI-13
July 14-18, 2013
IAAI-13
July 16-18, 2013
BELLEVUE, WASHINGTON, USA

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 admirable, but handcrafted design did not work well
  - 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 strong accept when 5 or 6 in clarity, but 2 in every other criterion
- Challenging to manually specify an 8-dimensional function

Data-driven approach: Learn a mapping

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

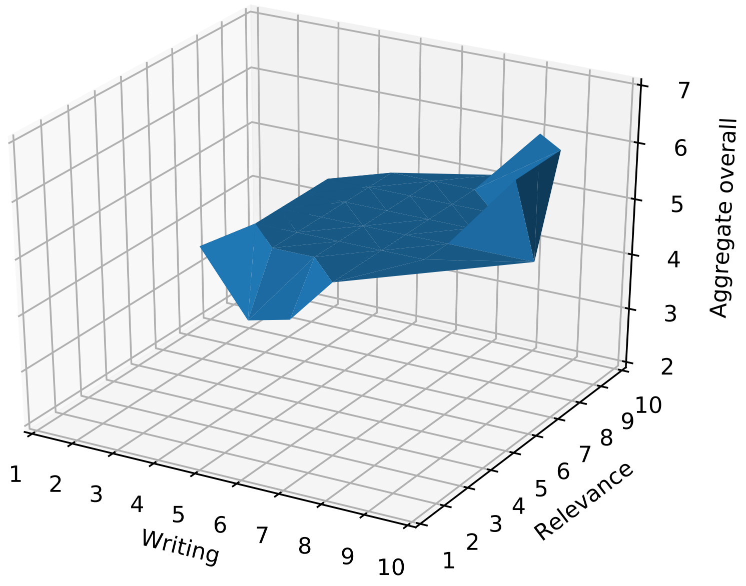
👍 **Using ML and social choice theory**

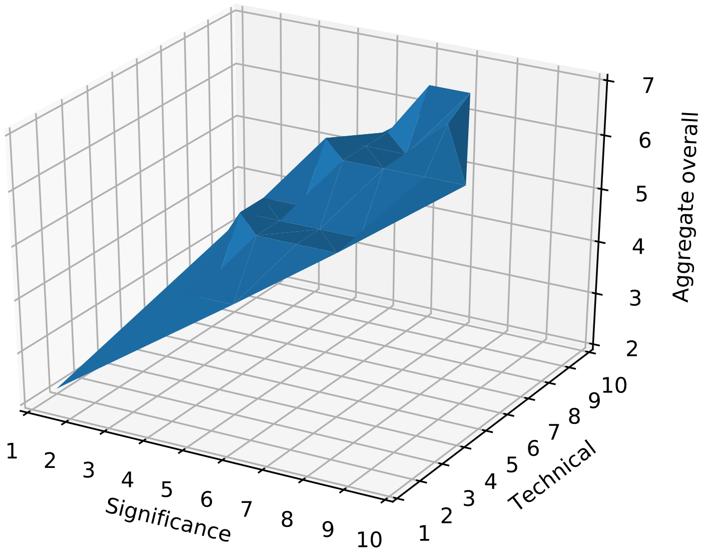
Theorem (informal): The only mapping that satisfies three natural requirements is:

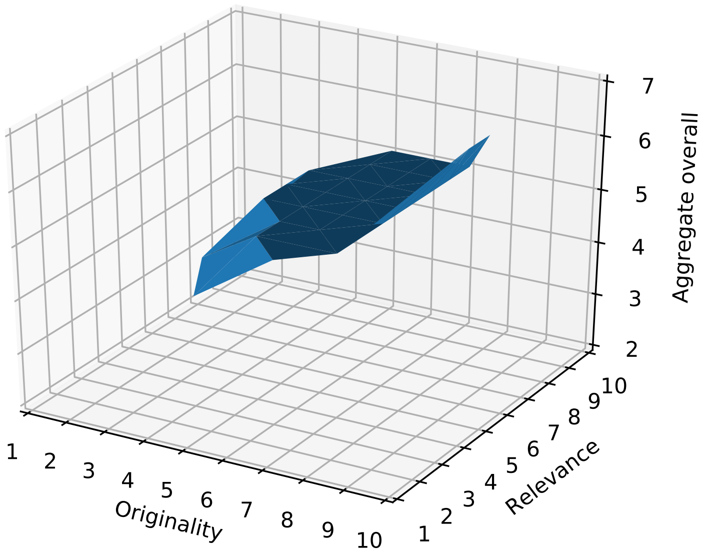
$$\hat{f} \in \operatorname*{argmin}_{\substack{f:[0,1]^k \rightarrow [0,1],\\ \text{non-decreasing}}} \sum_{\text{reviews}} |f(\text{criteria scores given by reviewer to paper}) - \text{overall score given by reviewer to paper}|$$

[Noothigattu et al. 2018]

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.

[Noothigattu et al. 2018]

Subjectivity

1. Commensuration bias
2. Confirmation bias
3. Dr. Fox effect
4. Hindering novelty

Confirmation bias

Reviewers are favorable to those manuscripts whose results agree with the reviewer’s own views.

Review: This is a wonderful paper with rigorous methods. Accept!

Methods:...

Conclusion: Coffee is bad for health

Review: This is a poor paper fatally flawed methods. Reject!

Papers that agreed with reviewer’s views:

- rated as methodologically better
- as having better data presentation
- reviewer was less likely to catch mistakes in the paper.

[Mahoney 1977, Travis et al. 1991, Ernst et al. 1994]

Subjectivity

1. Commensuration bias
2. Confirmation bias
3. Dr. Fox effect
4. Hindering novelty
5. Interdisciplinary research

Dr. Fox effect

Can complex presentation influence reviewers positively?

“acceptance via obfuscation”

Dr. Fox effect

Too simple. Reject.

We add up the two values and…

Denote the two values as α and β and compute (α + β)

What an advanced paper. Accept

[Armstrong 1980]

Subjectivity

1. Commensuration bias
2. Confirmation bias
3. Dr. Fox effect
4. Hindering novelty

Hinders novelty

"Reviewers love safe (boring) papers, ideally on a topic that has been discussed before (ad nauseam)...The process discourages growth" [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]

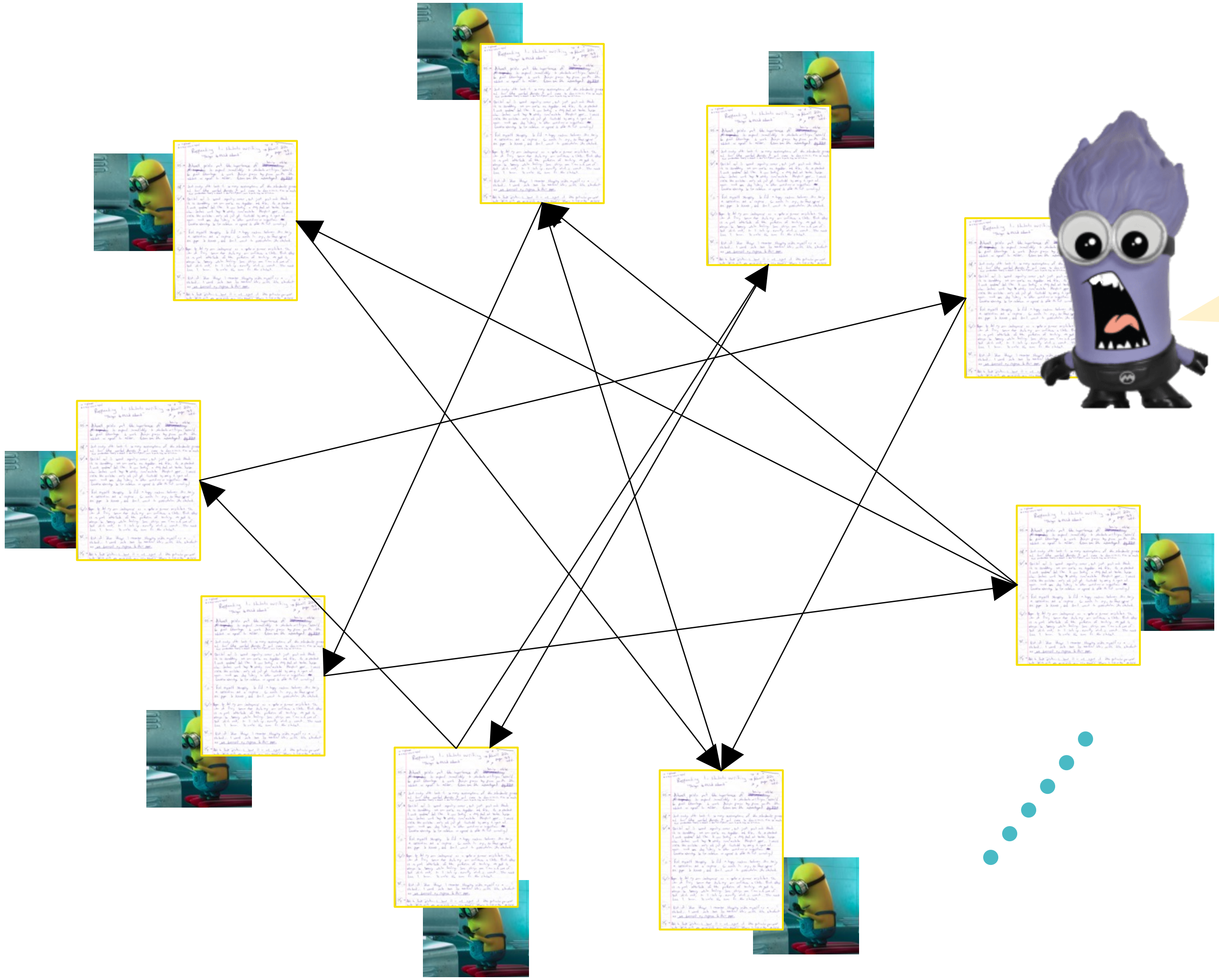
Subjectivity

1. Commensuration bias
2. Confirmation bias
3. Dr. Fox effect
4. Hindering novelty

Subjectivity: Open problems

- Statistical techniques to address commensuration bias
- 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]
- Computational or policy-based methods to address other issues of subjectivity
  - Automated finding of reviewer/paper characteristics
  - Debiasing algorithms/policies

Bias regarding author identities



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

[Chapter 7 of bit.ly/PeerReviewOverview]

Single blind versus double blind

A Principled Interpretation of Minion Speak

S. Overkill and F. Gru
Cartoony Minion University

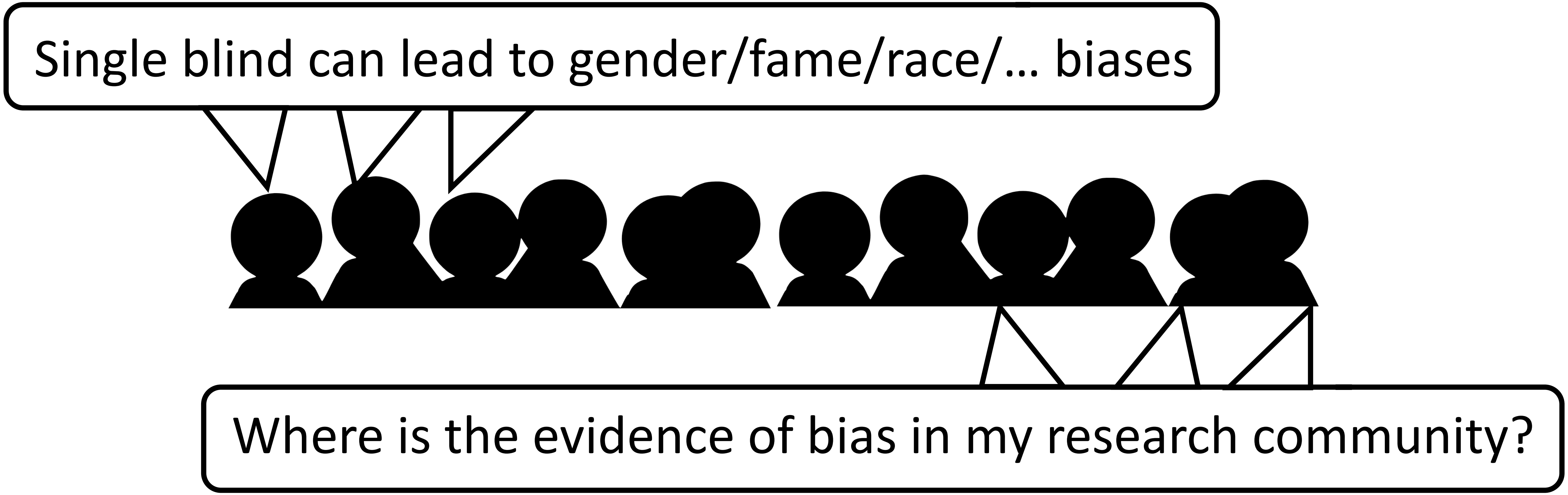
In this paper we present a new understanding of…

A Principled Interpretation of Minion Speak

Anonymous Authors
Anonymous Affiliation

In this paper we present a new understanding of…

Lot of debate!



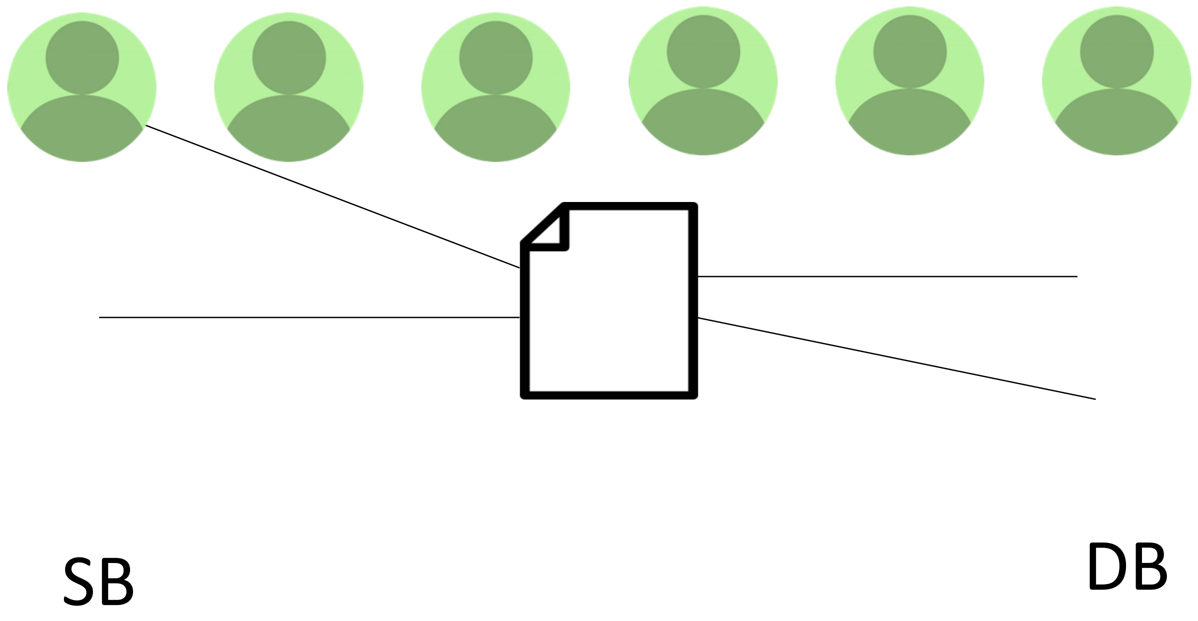
Single blind can lead to gender/fame/race/… biases

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

[Tomkins et al. 2018]

WSDM’17 experiment: Tests for bias regarding…

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

[Tomkins et al. 2018]

WSDM’17 experiment: Testing procedure

- For any paper p, let q_p = “intrinsic” value of paper p

- Logistic model: $P(\text{single blind reviewer accepts paper } p)$

$$= \frac{1}{1+\exp(-\boxed{(\beta_0+\beta_1 q_p + \sum_{\text{attributes } a} \beta_a \mathbb{I}\{\text{Paper } p \text{ has author attribute } a\}})))}$$

- Use DB reviewers to estimate q_p for each paper p

- Fit decisions of SB reviewers into logistic model to estimate β’s

Test: $\beta_a = 0$ vs. $\beta_a \neq 0$

(no bias) (bias)

[Tomkins et al. 2018]

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.

[Tomkins et al. 2018]

Peculiar characteristics of peer review

[Stelmakh et al., 2019]

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 α, must ensure
P(false alarm) $\leq \alpha$

Typical choice: $\alpha = 0.05$

Want high detection subject to false alarm control

[Stelmakh et al., 2019]

Characteristic 0: Correlations between quality of papers and certain attributes

- Famous author
- Top university
- Top company

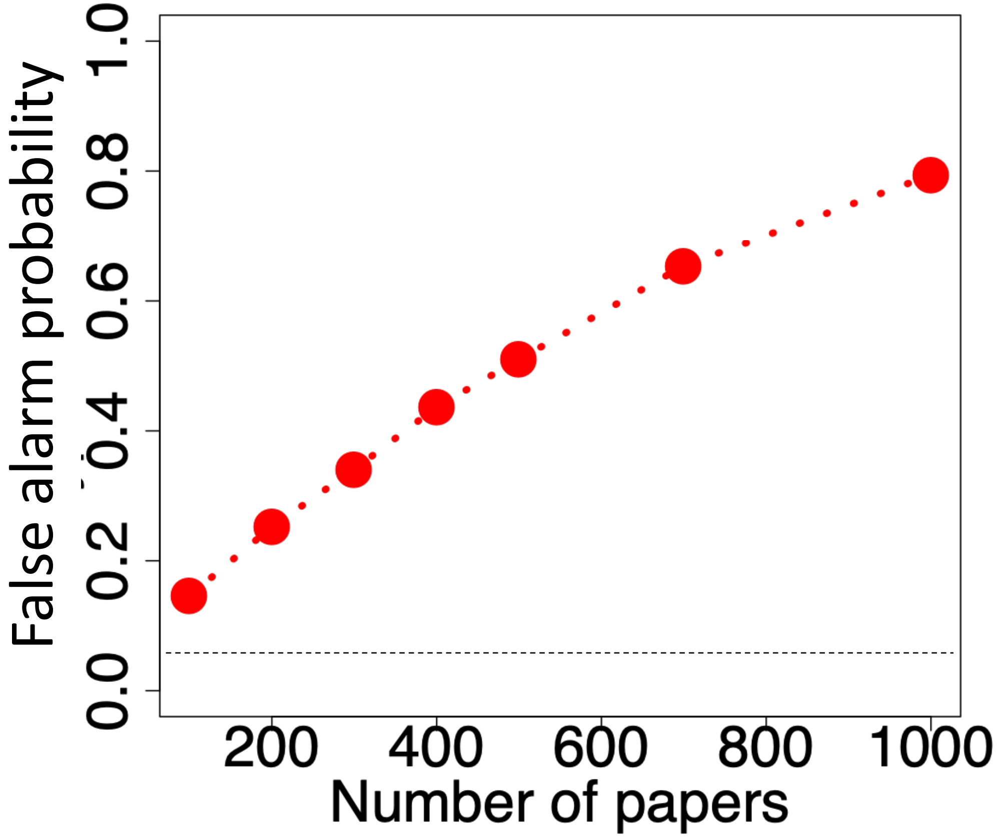
Combined with other characteristics…

[Stelmakh et al., 2019]

Characteristic 1: Reviews are noisy

Reviewers are imperfect (noisy)

Must ensure: P(declare bias when no bias) ≤ 0.05

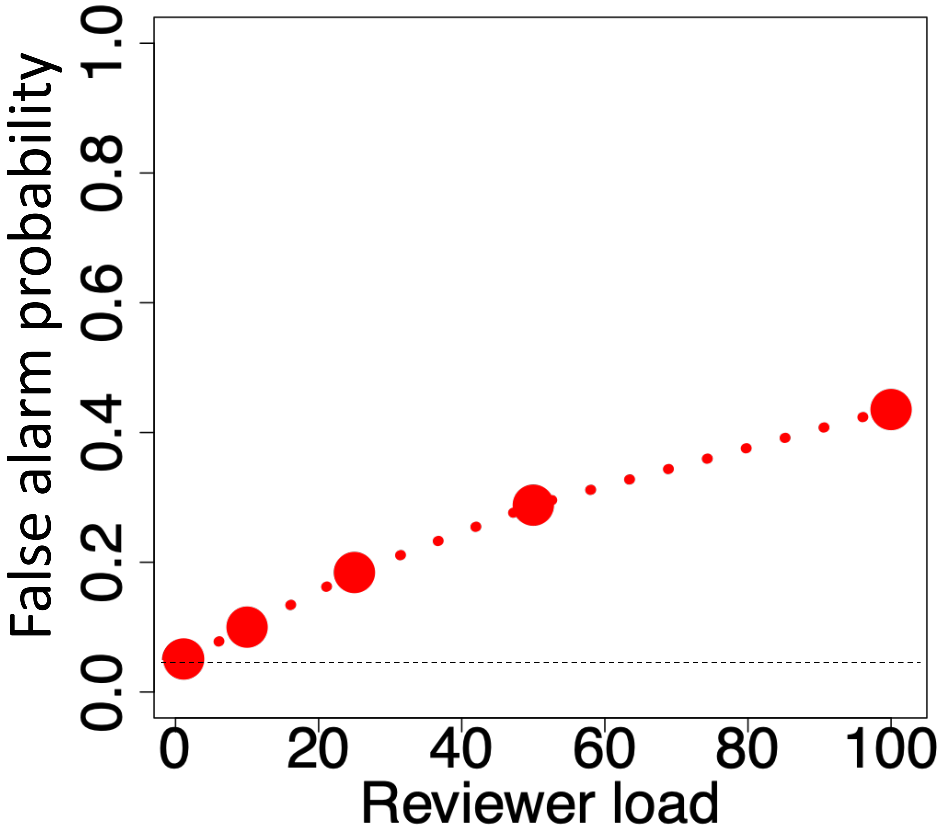


[Stelmakh et al., 2019]

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(declare bias when no bias) $\leq$ 0.05

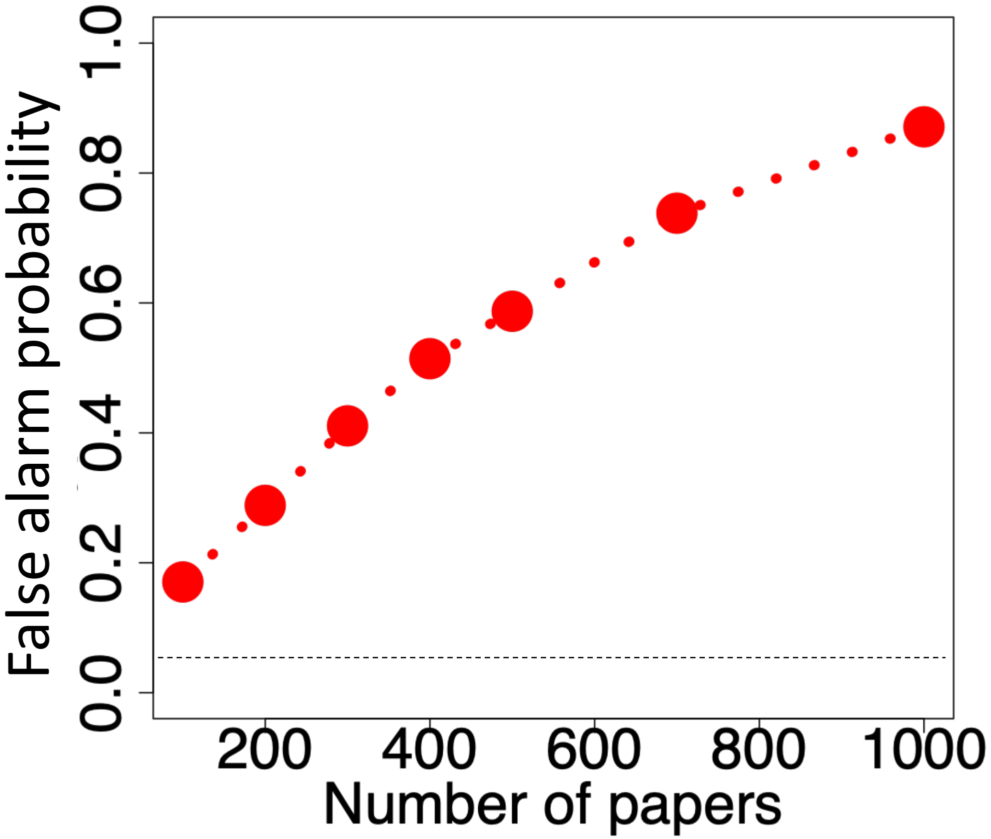


[Stelmakh et al., 2019]

Characteristic 3: Model complexity

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

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

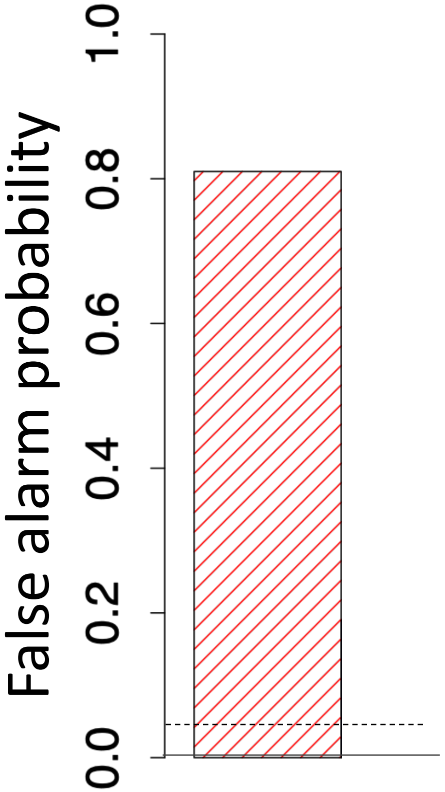


[Stelmakh et al., 2019]

Characteristic 4: Non-random assignment

Assignment of reviewers to papers is NOT random

Must ensure: P(declare bias when no bias) ≤ 0.05



[Stelmakh et al., 2019]

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 α

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

[Stelmakh et al. 2019]

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

Norms and Policies

Alright, so here's what everyone must do…

[Chapter 8 of bit.ly/PeerReviewOverview]

Norms and Policies

1. Author incentives
2. Review quality
3. Author rebuttal
4. Discussions and group dynamics

Norms and Policies

1. Author incentives
2. Review quality
3. Author rebuttal
4. Discussions and group dynamics

Resubmission Bias

Many conferences ask authors to declare previous rejections of submitted paper

IJCAI-PRICAI
YOKOHAMA
2020

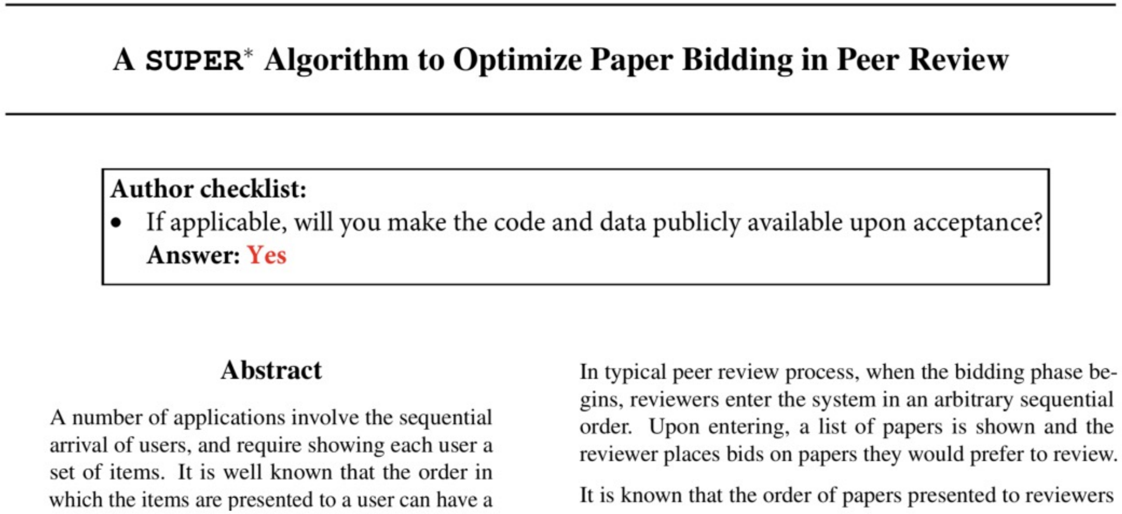
“authors must declare the resubmission by including a cover letter with their submission… should summarize the main reasons for rejection and should describe the changes the authors have made to address the reviewers’ comments. **The cover letter should be inserted at the beginning of the submitted PDF, 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

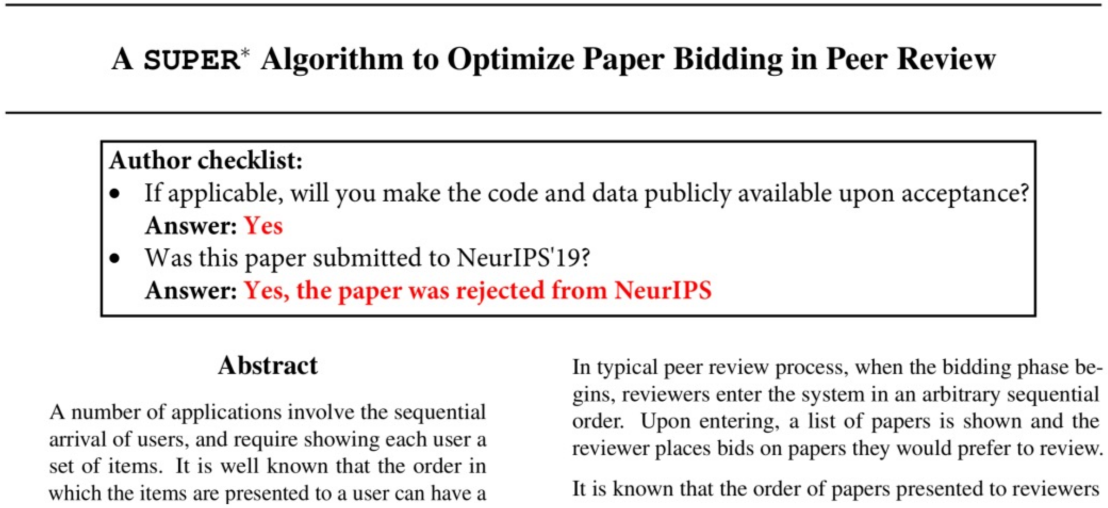
Abstract

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

[Stelmakh et al. 2021]

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.

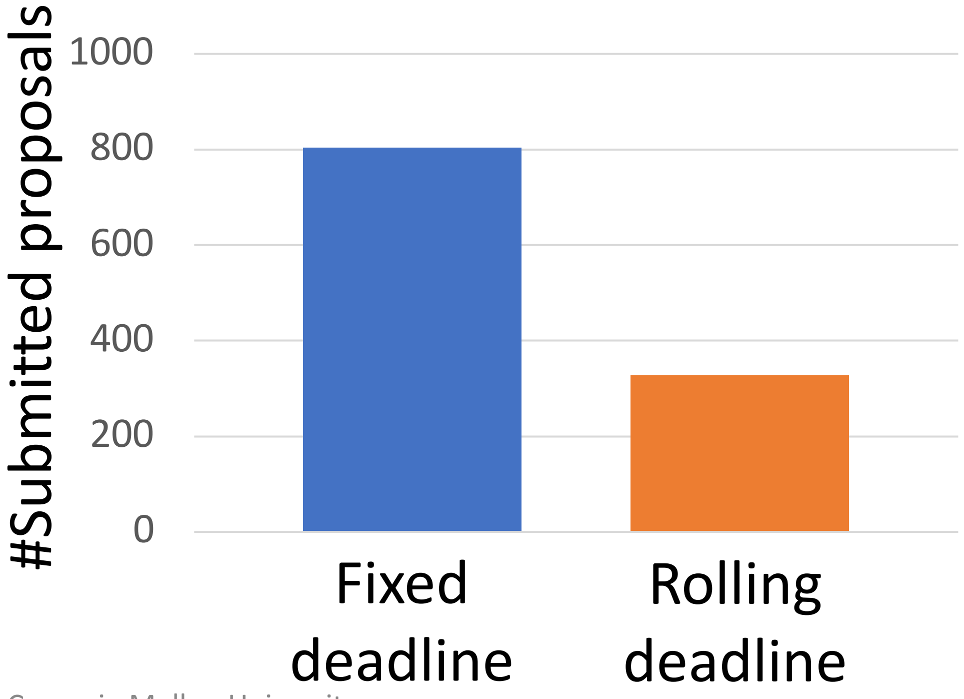
[Stelmakh et al. 2021]

Rolling deadlines

- Majority of papers submitted at or near the deadline [Soergel et al. 2013]

- What if there is no deadline?
  - Authors have time to polish paper
  - Researchers don't all have to use resources (such as compute) at the same time

- NSF experiment [Hand 2016]



Norms and Policies

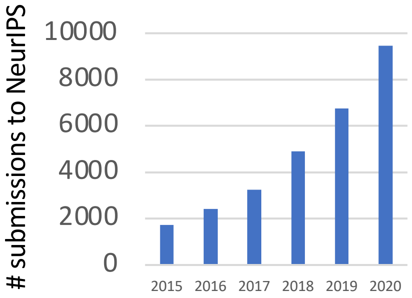
1. Author incentives

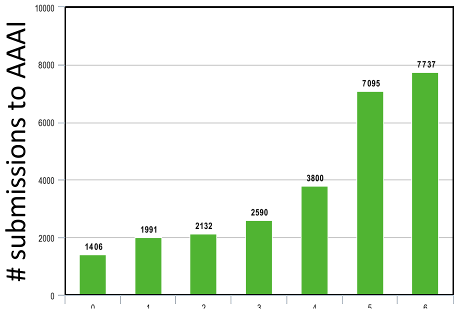
2. Review quality

3. Author rebuttal

4. Discussions and group dynamics

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

[Stelmakh et al. 2021]

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

[Stelmakh et al. 2021]

Reviewer training and progression

Shadow program committee

- SIGCOMM 2005 [Feldmann 2005] and IEEE S&P 2017 [Parno et al. 2017]
- Separate committee of junior researchers to mirror the actual process
- Shadow committee decisions do not influence the actual decisions

What about long term? [Callaham et al. 2011, Joyner et al. 2020]

- Quality of individual reviewers' review quality reduces over time
- Possibly because of increasing time constraints or in reaction to poor-quality reviews they receive

Norms and Policies

1. Author incentives

2. Review quality

3. Author rebuttal

4. Discussions and group dynamics

Do rebuttals change reviews?

NAACL 2015 [Daumé 2015]

- Rebuttal did not alter reviewers’ opinions much
- Most (87%) review scores did not change after the rebuttals
- Among those which did, scores were nearly as likely to go down as up
- Review text did not change for 80% of the reviews
- Does rebuttal lead to more discussions? Not really

NeurIPS 2016 [Shah et al. 2017]

- Fewer than 10% of reviews changed scores after the rebuttal

ACL 2017 [Kan 2017]

- Scores changed after rebuttals in about 15-20% of cases
- Change was positive in twice as many cases as negative

Do authors like rebuttals?

Survey of authors of accepted papers at 56 computer systems conferences [Frachtenberg et al. 2020]:

- About 90% of authors found rebuttal process helpful
- Non-native English speakers found it helpful at a slightly higher rate
- Authors who found the rebuttal process as helpful are only half as experienced (in terms of publication records, career stage, as well as program committee participation) as compared to the set of authors who did not find it helpful

Positive sentiment in other surveys [Parno et al. 2017]

Norms and Policies

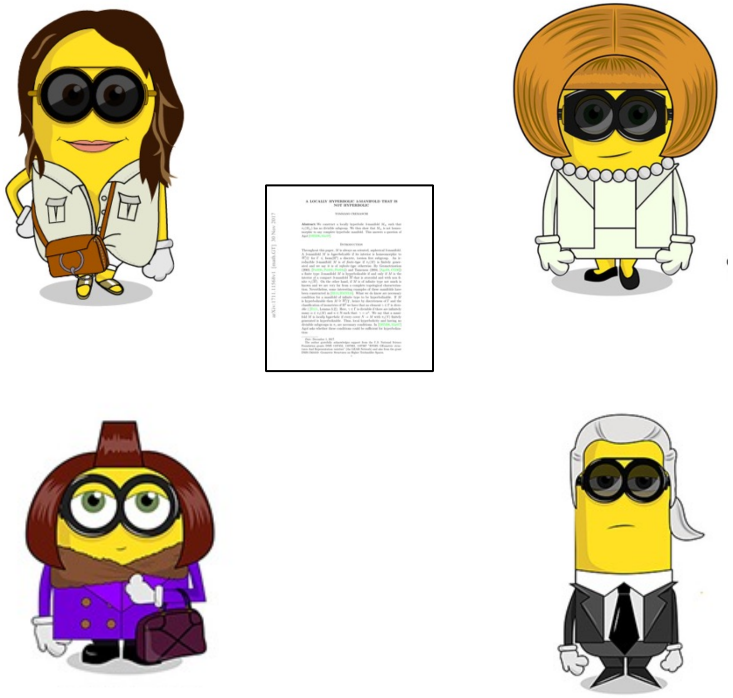
1. Author incentives

2. Review quality

3. Author rebuttal

4. Discussions and group dynamics

Consistency of discussions



- Each paper is independently reviewed by multiple reviewers

- Reviewers then see each others' reviews and discuses the paper

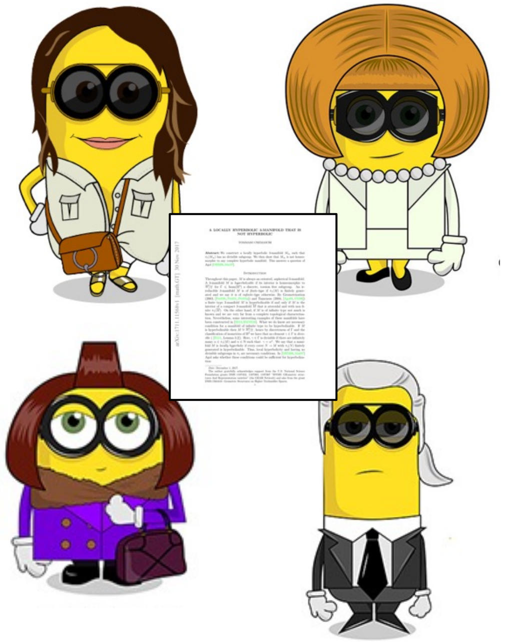
- They arrive at a consensus on the paper

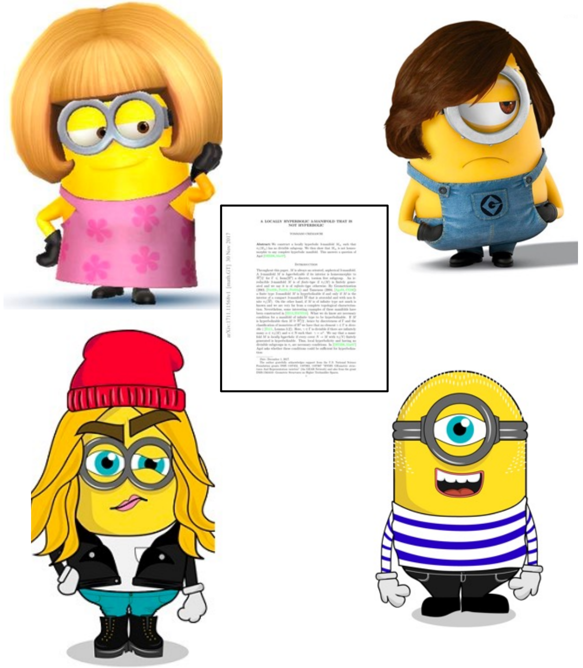
[Pier et al. 2017, Fogelholm et al. 2012, Obrecht 2007]

Do discussions improve consistency?

[Pier et al. 2017, Fogelholm et al. 2012, Obrecht 2007]

An experiment

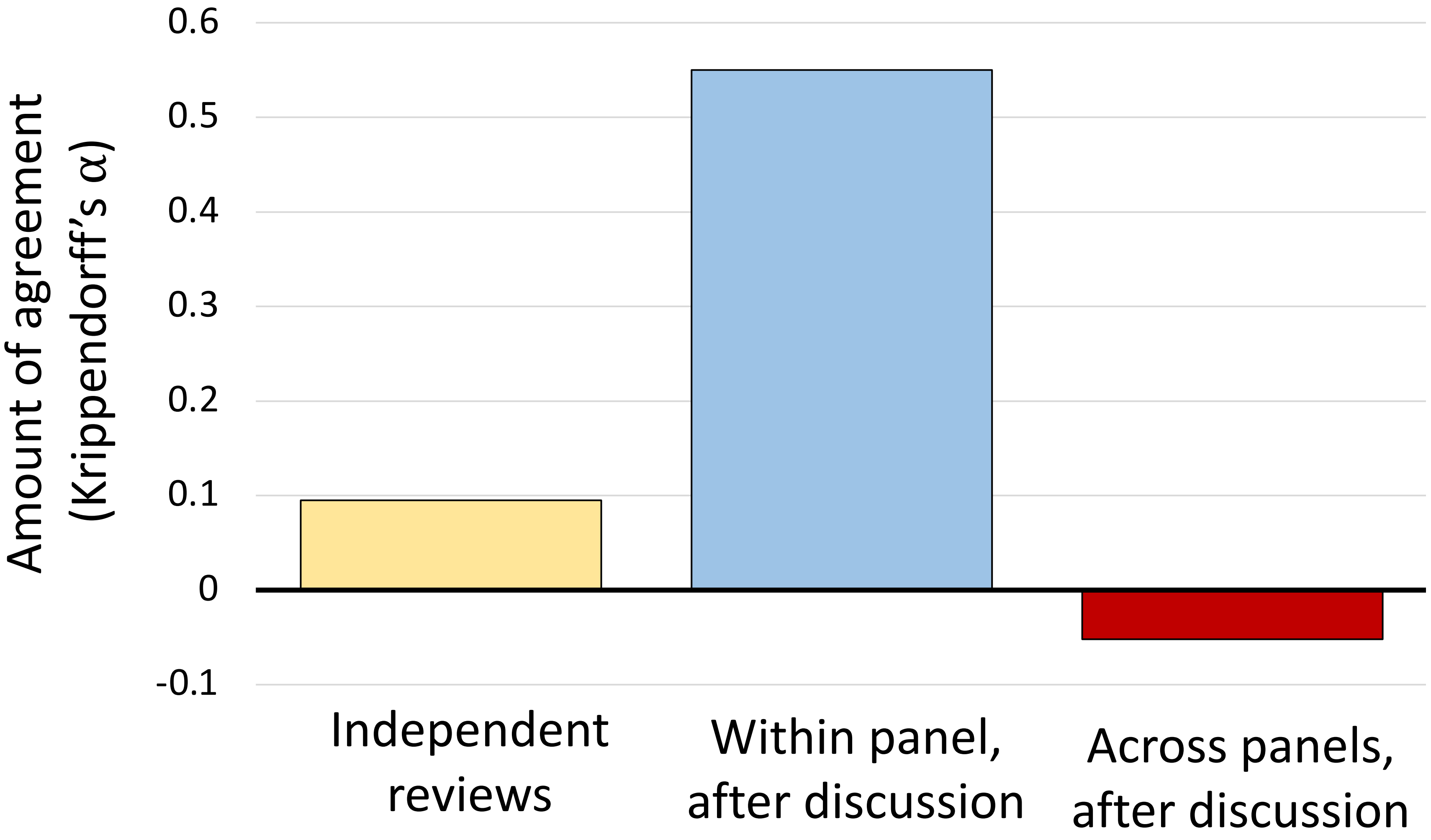




- Multiple independent panels per paper (or proposal)
- Measure agreements

[Pier et al. 2017, Fogelholm et al. 2012, Obrecht 2007]

Key findings



[Pier et al. 2017, Fogelholm et al. 2012, Obrecht 2007]

Influence of other reviews

Do reviewers get unduly influenced by other reviews?

[Teplitskiy et al. 2019]

An experiment

~~6/10~~
7/10

By the way, another reviewer gave 9/10

[Teplitskiy et al. 2019]

Key findings

- Reviewers updated the ratings they had given 47% of the time
- Women reviewers updated the ratings they had given 13% more frequently than men
- Highly-cited reviewers updated 24% less than others
- Review ratings which were originally low were updated 38% less than medium and high ratings. This asymmetry can favor conservative proposals which may have low variance in ratings.

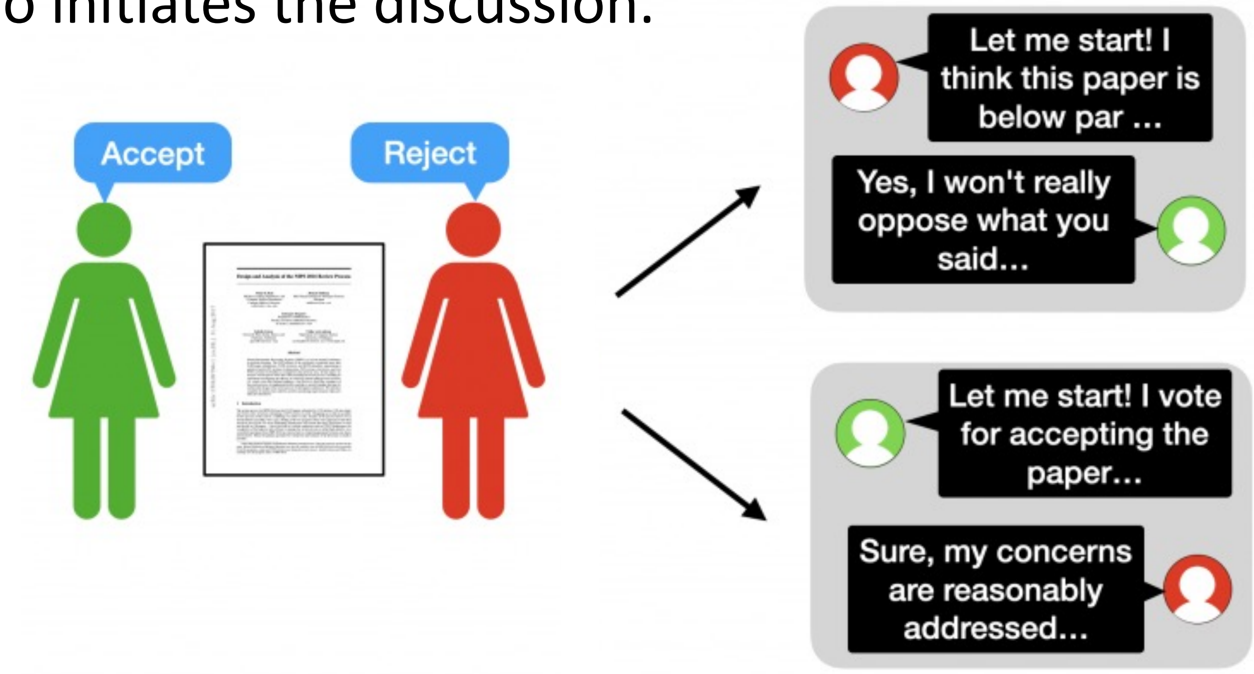
[Teplitskiy et al. 2019]

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**.


Accept
Reject
Let me start! I think this paper is below par ...
Yes, I won't really oppose what you said...
Let me start! I vote for accepting the paper...
Sure, my concerns are reasonably addressed...

Initial review

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

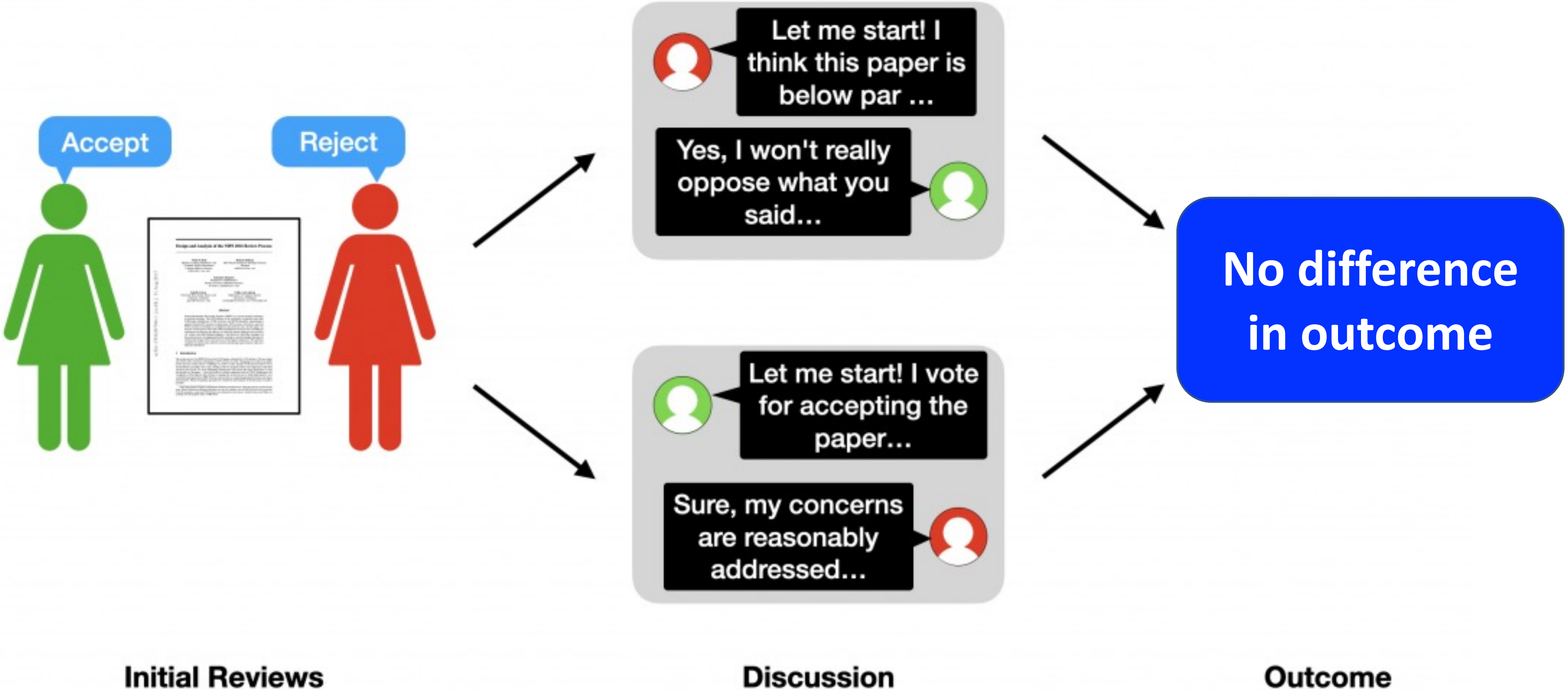
First ask most positive reviewer to start the discussion, then later ask the most negative reviewer to contribute to the discussion.

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

[Stelmakh et al. 2020]

Key findings



Let me start! I think this paper is below par ...

Yes, I won't really oppose what you said...

Let me start! I vote for accepting the paper...

Sure, my concerns are reasonably addressed...

No difference in outcome

[Stelmakh et al. 2020]

Epilogue

AI reviewers?

- Few recent attempts [Huang 2018, Wang et al. 2020, Yuan et al. 2021]
- Not very successful

- Use AI to evaluate specific aspects of the paper [Houle 2016, Foltynek 2019]
  - Do experiments have an appropriate sample size?
  - Do they report relevant metrics?
  - Does it adhere to required format?
  - Plagiarism

Open problems

- Address multiple challenges in tandem

Open problems

- Address multiple challenges in tandem
- Privacy-preserving techniques for researchers to use peer-review data [Ding et al. 2021, Jecmen et al. 2020]

“The main reason behind the lack of empirical studies on peer-review is the difficulty in accessing data.” [Balietti et al., 2016]

“We would prefer to make available the raw data used in our study, but after some effort we have not been able to devise an anonymization scheme that will simultaneously protect the identities of the parties involved and allow accurate aggregate statistical analysis. We are familiar with the literature around privacy preserving dissemination of data for statistical analysis and feel that releasing our data is not possible using current state-of-the-art techniques.” [Tomkins et al. 2018]

Open problems

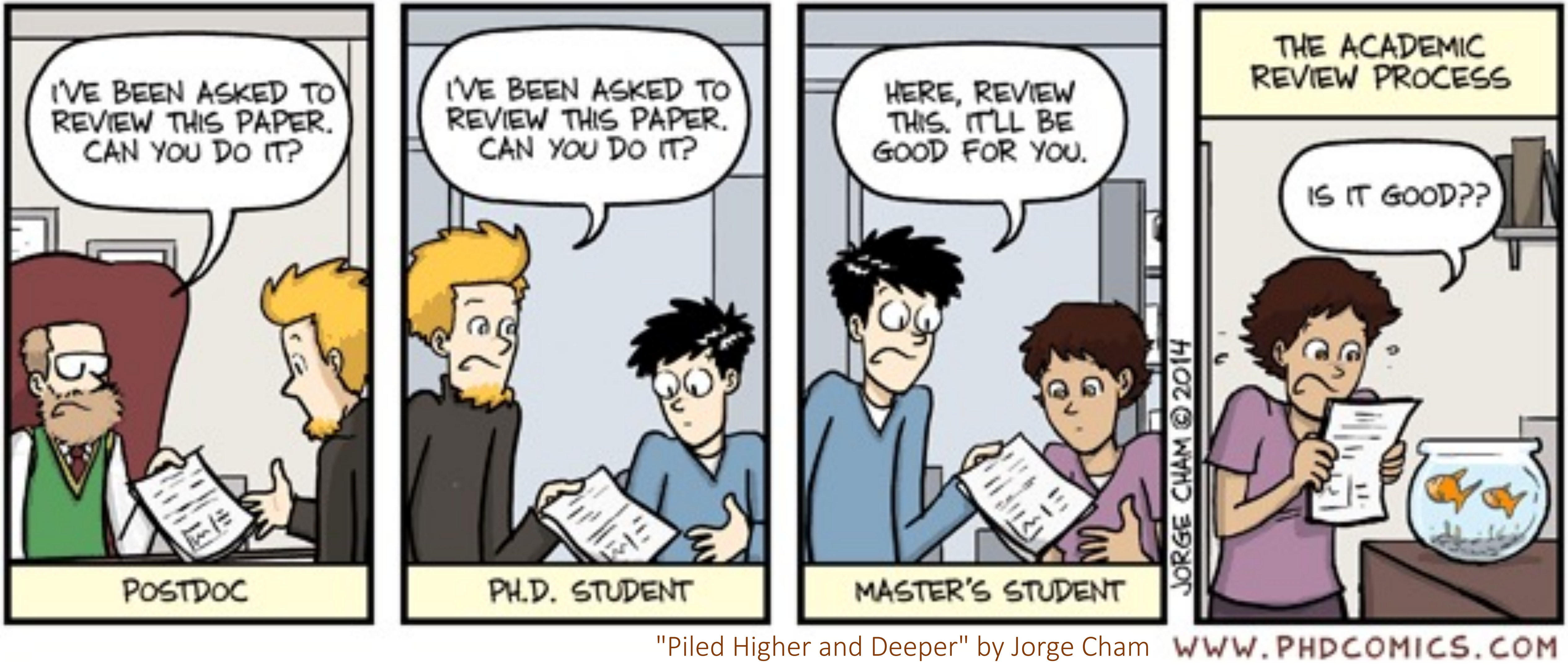
- Address multiple challenges in tandem
- Privacy-preserving techniques for researchers to use peer-review data [Ding et al. 2021, Jecmen et al. 2020]
- Evaluation metrics for peer-review algorithms and policies
  - Meta-reviewers also suffer from similar issues
  - Author feedback biased by decisions
    - "Satisfaction [of the author with the review] had a strong, positive association with acceptance of the manuscript for publication... Quality of the review of the manuscript was not associated with author satisfaction" [Weber et al. 2002]
    - Some initial work on debiasing [Wang et al. 2021]
  - Using citations as metric has other challenges

Open problems

- Address multiple challenges in tandem
- Privacy-preserving techniques for researchers to use peer-review data [Ding et al. 2021, Jecmen et al. 2020]
- Evaluation metrics for peer-review algorithms and policies
- More experiments: Science for science!

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**


I'VE BEEN ASKED TO REVIEW THIS PAPER. CAN YOU DO IT?
POSTDOC
I'VE BEEN ASKED TO REVIEW THIS PAPER. CAN YOU DO IT?
PH.D. STUDENT
HERE, REVIEW THIS. IT'LL BE GOOD FOR YOU.
MASTER'S STUDENT
JORGE CHAM © 2014
THE ACADEMIC REVIEW PROCESS
IS IT GOOD??
WWW.PHDCOMICS.COM

"Piled Higher and Deeper" by Jorge Cham

Thank you! Questions?

bit.ly/PeerReviewOverview

nihars@cs.cmu.edu