Peer Review, Biases, and Statistical Learning

Nihar B. Shah

Machine Learning and Computer Science Departments

Carnegie Mellon University

I'm in fabulous France!
Je suis très heureux ☺
I'll accept this paper!

Peer Review

Papers

Grant proposals

anr
agence nationale de la recherche
AU SERVICE DE LA SCIENCE

NSF

Promotions

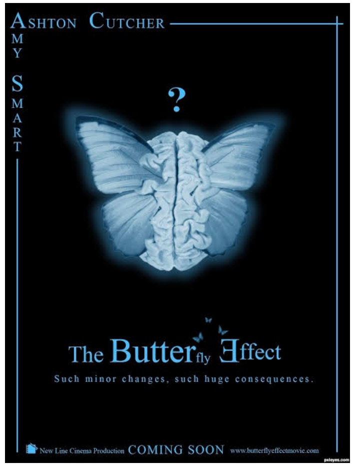
npr

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

Problems in peer review hurt…

Scientific progress

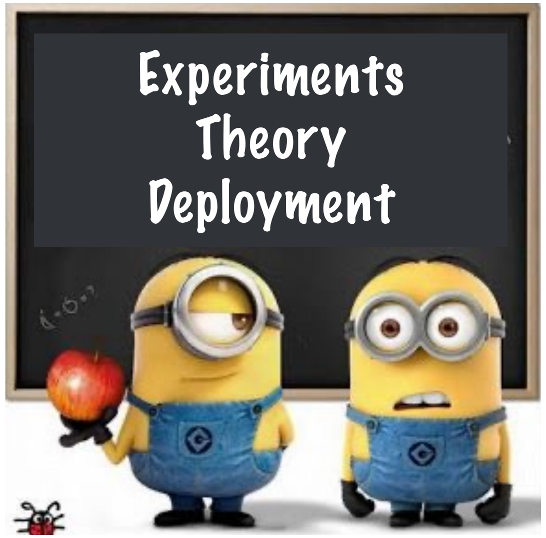
Careers



Public perception of science

Tackle systemic problems in peer review

via principled and practical approaches



Overview article on peer review:

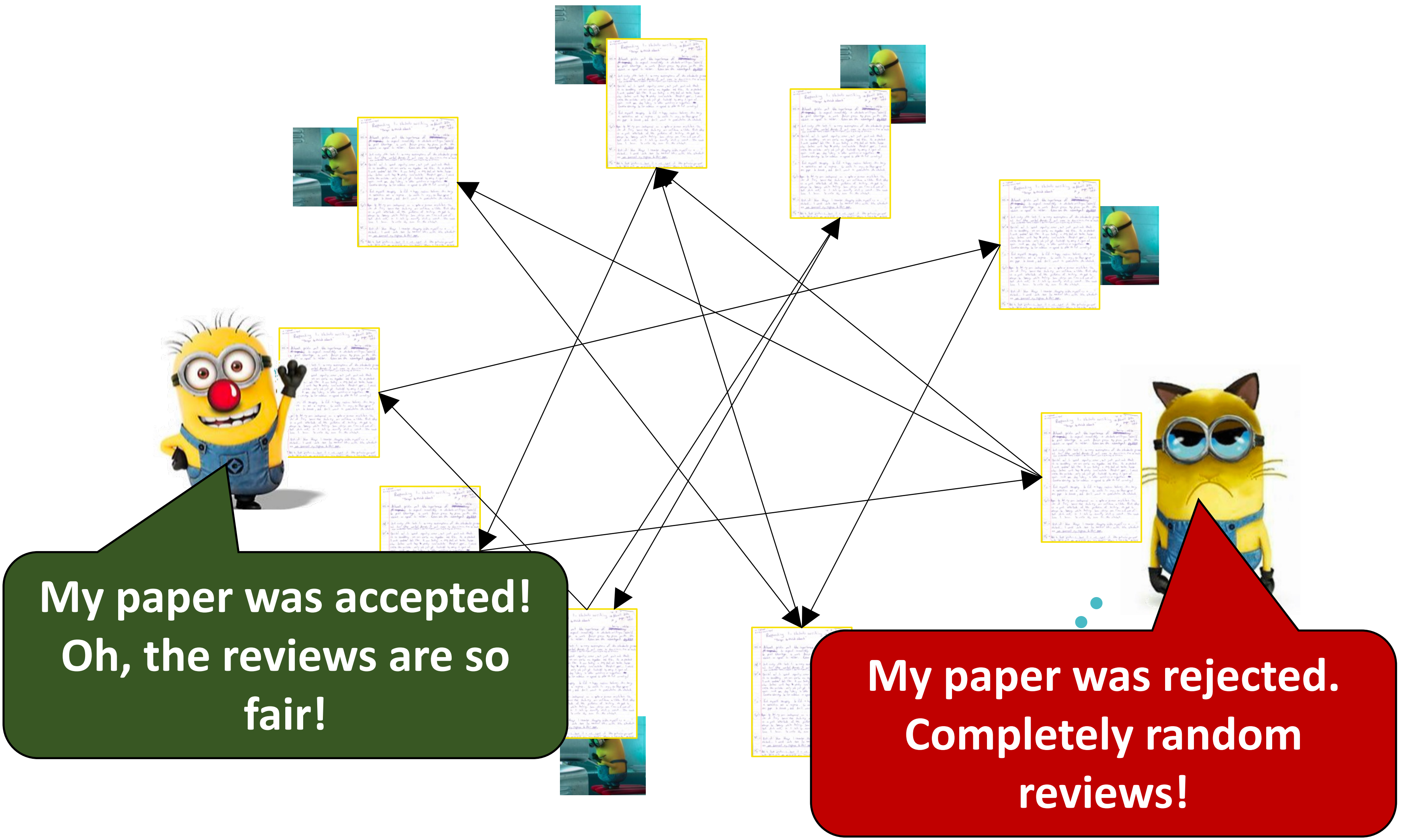
bit.ly/PeerReviewOverview

Outline for this talk

Feedback bias

Author identity bias

Feedback bias



Joint work with:

Jingyan Wang
Ivan Stelmakh
Yuting Wei

Feedback Loop Crucial for any System

R&D, testing, deploy

How satisfied are you with this system?

Feedback

Improve the system,
create better incentives,
etc.

How to obtain feedback?

- How to evaluate the peer-review process or specific review(er)s?
- Quite common opinion: Authors know their papers best, so ask them to rate the reviews

JSys Journal of Systems Research

"The three reviews will be graded A/B/C by the authors in terms of helpfulness... Reviewers with a history of poor reviews will be removed from the editorial board."

But…

Authors are biased by the outcomes of their papers

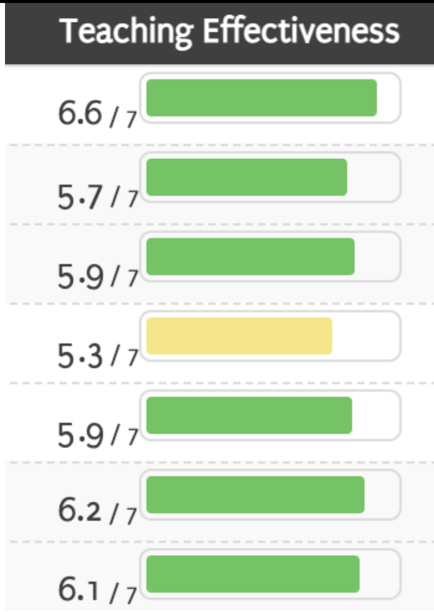
“Satisfaction [of the author with the review] had a strong, positive association with acceptance of the manuscript for publication… Quality of the review of the manuscript was not associated with author satisfaction.”

[Weber et al., 2002]

[Also: Van Rooyen et al. 1999; Papagiannaki, 2007; Khosla, 2013; Kerzendorf et al. 2020]

Goal: Debias author-provided feedback

Similar Problem in Teaching Evaluations



- Students are asked to rate instructors' teaching effectiveness
- Highly biased by grading leniency:

"...the effects of grades on teacher–course evaluations are both substantively and statistically important..." [Johnson, 2003]

[Also: Carrell & West, 2008; Braga et al., 2014; Boring et al., 2016]

- Introduces incentives for inflating grades

"... instructors can often double their odds of receiving high evaluations from students simply by awarding A's rather than B's or C's." [Johnson, 2003]

Problem formulation and model

- Set of items to evaluate (e.g., review processes or reviewers or courses)
- Unknown true quality $x_i^* \in \mathbb{R}$ for each item i
- Set of evaluators per item (e.g., authors or students)
- If evaluator j rates item i, observed rating $y_{ij} \in \mathbb{R}$ has three components: true quality, feedback bias, and noise. Model:

$$y_{ij} = x_i^* + \text{bias}_{ij} + \text{noise}_{ij}$$

next slide

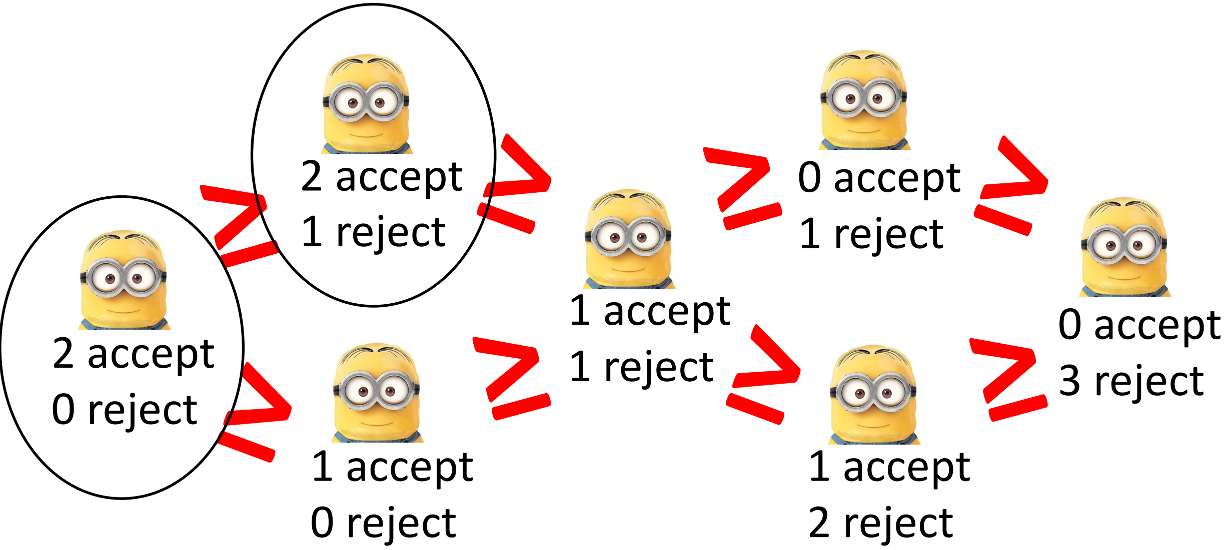
i.i.d. zero-mean Gaussian,
unknown variance

Goal: Estimate x^* minimizing the mean squared error

Model: Bias

$$y_{ij} = x_i^* + \boxed{\text{bias}_{ij}} + \text{noise}_{ij}$$

Peer review



Program chairs know outcomes of evaluators' papers

Courses

University knows outcomes of evaluators' scores

Assume: Biases follow a *known* partial ordering

Model: Bias

$$y_{ij} = x_i^* + b_{ij} + \mathrm{noise}_{ij}$$

- Bias b_{ij}'s
  - Generate i.i.d. zero-mean Gaussian, unknown variance
  - Permuted to align with known partial ordering

Proposed Estimator

$$\hat{x}^{(\lambda)} \in \boxed{\underset{x \in \mathbb{R}^{\#\text{items}}}{\text{argmin}}} \; \boxed{\underset{b_{ij}\text{'s obey partial ordering}}{\min}} \; \underbrace{\sum_{(i,j)} \left(y_{ij} - x_i - b_{ij}\right)^2}_{\text{Noise}} + \lambda \underbrace{\sum_{(i,j)} b_{ij}^2}_{\text{Bias}}$$

Proposition (informal). Under certain conditions:

- When there is no noise, our estimator with $\lambda = 0$ is consistent.
- When there is no bias, our estimator with $\lambda = \infty$ is equivalent to taking the sample mean.

Sample mean is not consistent

Minimax optimal

How to choose hyperparameter λ?

Natural idea: Cross-validation

Challenge…

Cross-validation to choose λ: Naïve approach

- Partition all evaluations (i, j)'s into training and validation sets
- For each λ:
  - On training set estimate $\hat{x}$ and $\hat{b}$ as minimizers of

$$\sum_{(i,j)\in\text{Train}}\left(y_{ij} - x_i \;\; - b_{ij}\right)^2 + \lambda \sum_{(i,j)\in\text{Train}} b_{ij}^2$$

  - On validation set, evaluate $\sum_{(i,j)\in\text{Validation}}\left(y_{ij} - \hat{x}_i \;\; - \hat{b}_{ij}\right)^2$
- Choose the λ with the smallest (residual) validation error

What goes wrong?

Problem with naïve crossvalidation

Model: $y_{ij} = x_i^* + \mathrm{b}_{ij} + \mathrm{noise}_{ij}$

- On training set, estimate $\hat{x}_i$ and $\{\hat{b}_{ij}\}_{(i,j)\in \text{Train}}$
- Want to compute residual in validation set: $\sum_{(i,j)\in \text{Validation}} (y_{ij} - \hat{x}_i - \hat{b}_{ij})^2$
- But the training set gives $\{\hat{b}_{ij}\}_{(i,j)\in \text{Train}}$ and **not** $\{\hat{b}_{ij}\}_{(i,j)\in \text{Validation}}$

Cross-validation to choose λ

Idea 2.0: Use knowledge of partial ordering of biases to

(i) appropriately choose a train-test split and

(ii) carefully interpolate $\{\hat{b}_{ij}\}_{(i,j)\in\text{Train}}$ to get $\{\hat{b}_{ij}\}_{(i,j)\in\text{Validation}}$

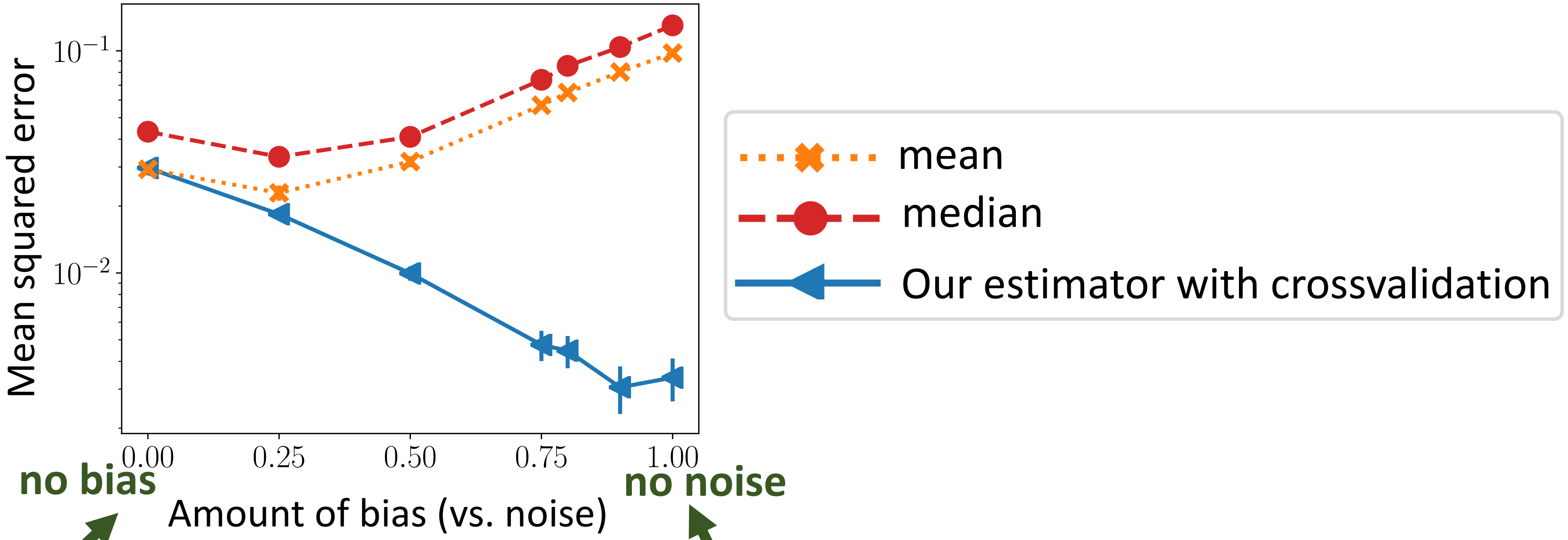
Theorem (informal). Under certain conditions:

- When there is no noise, $\hat{\mathrm{x}}_{\mathrm{CV}} \to \hat{\mathrm{x}}^{(\lambda=0)}$
- When there is no bias, $\hat{\mathrm{x}}_{\mathrm{CV}} \to \hat{\mathrm{x}}^{(\lambda=\infty)}$

Our cross-validation successfully recovers the two extremal cases.

Semi-synthetic experiments

- Indiana University Bloomington
- 10 sessions of a course
- Simulate bias and noise using real grading statistics



no bias

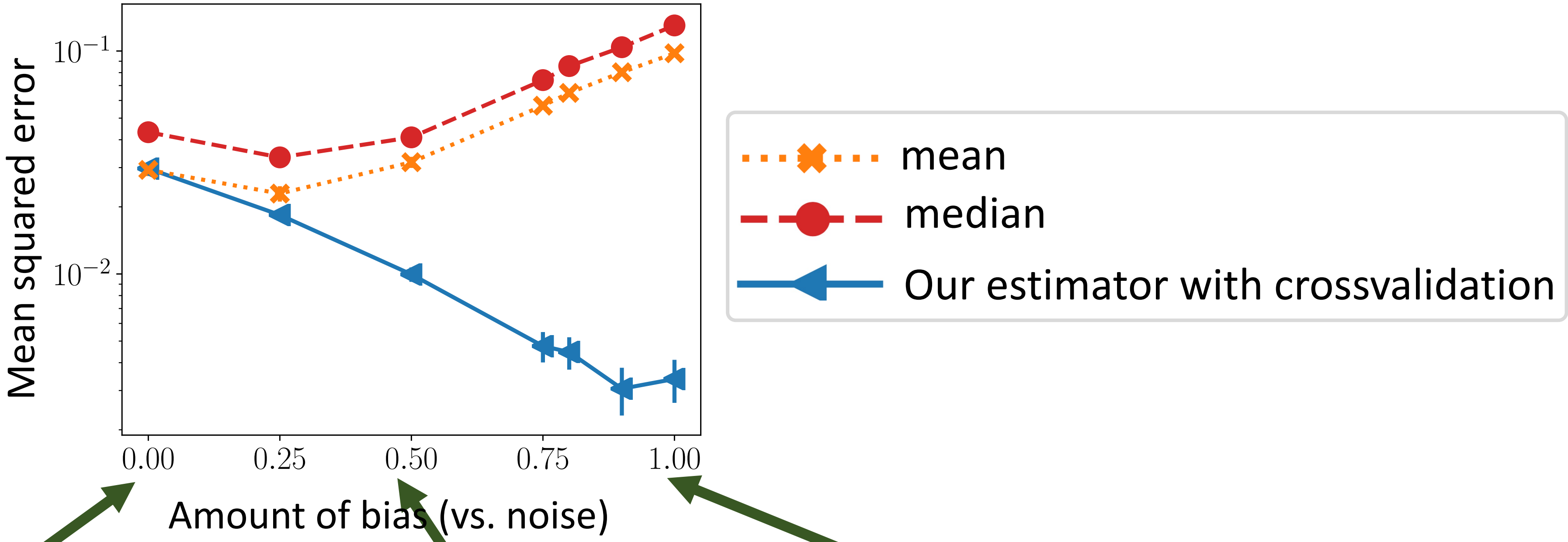
no noise

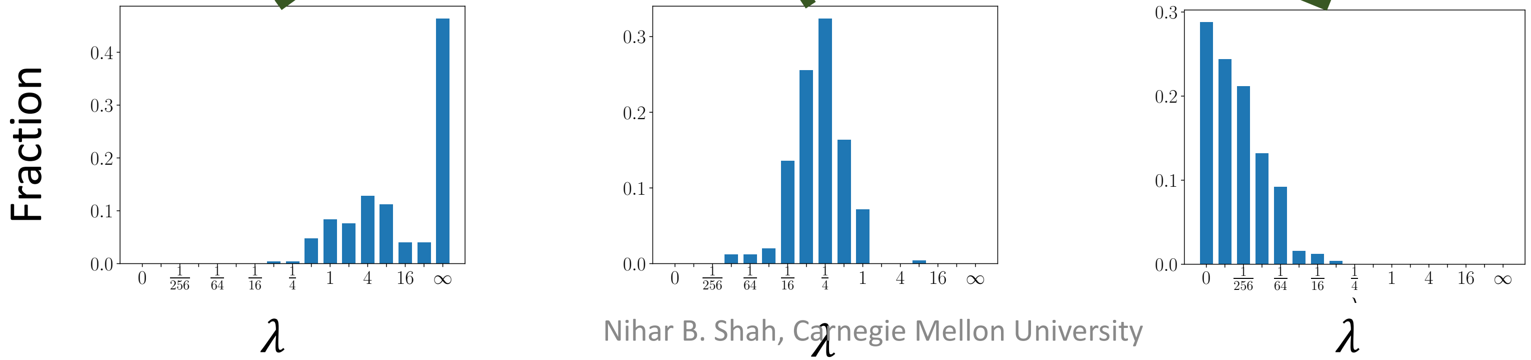
all estimators work well

our estimator significantly outperforms mean & median

Semi-synthetic experiments

- Indiana University Bloomington
- 10 sessions of a course
- Simulate bias and noise using real grading statistics





Feedback: Open problems

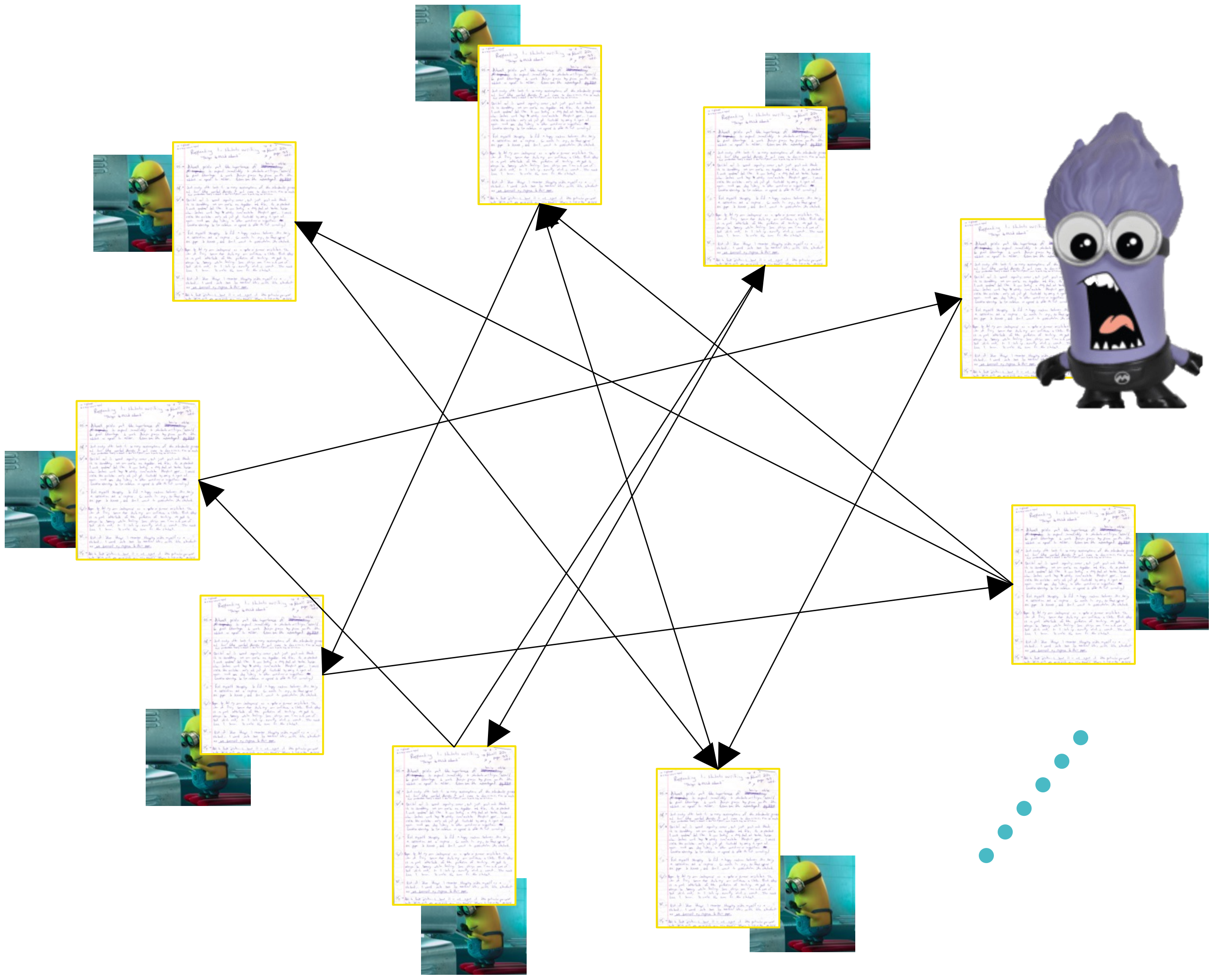
- Trialed for >1000 submissions
  - Clever experiments and publicly-released data with "ground truth" for this problem?

- Guarantees (and possibly new estimators):
  - Sample complexity guarantees
  - Guarantees for non-extremal points

- More nuances in the model

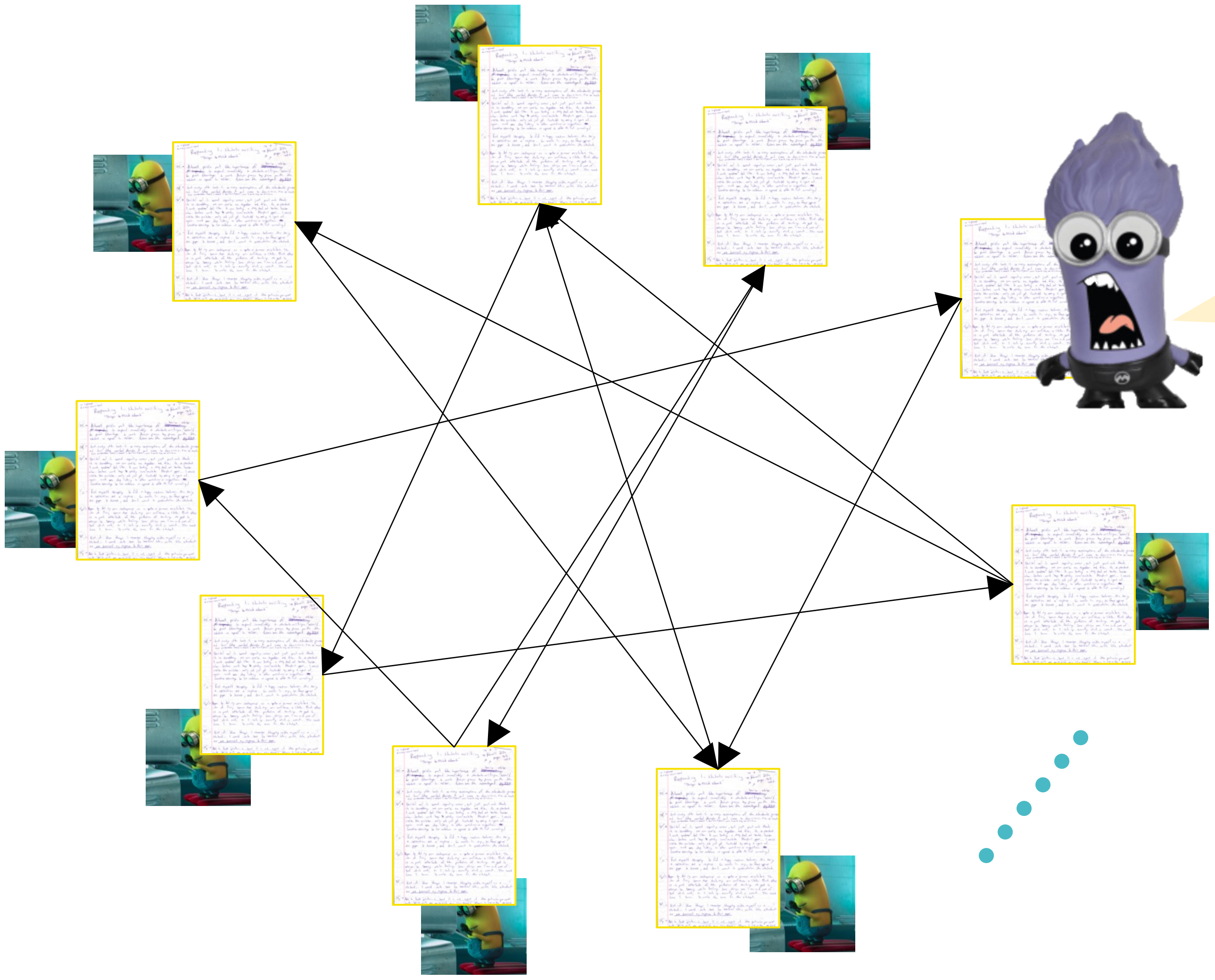
- What incentive structure does this lead to?

Author-identity Bias



Joint work with:
Ivan Stelmakh
Aarti Singh

Author-identity 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…

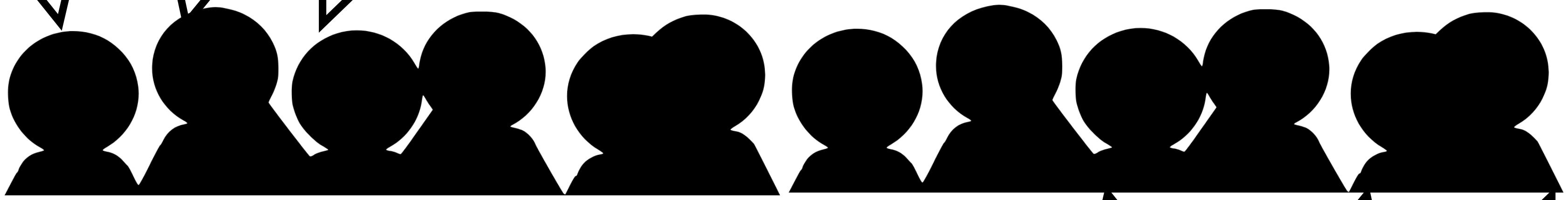
A Principled Interpretation of Minion Speak

Anonymous Authors
Anonymous Affiliation

In this paper we present a new understanding of…

Lot of debate!

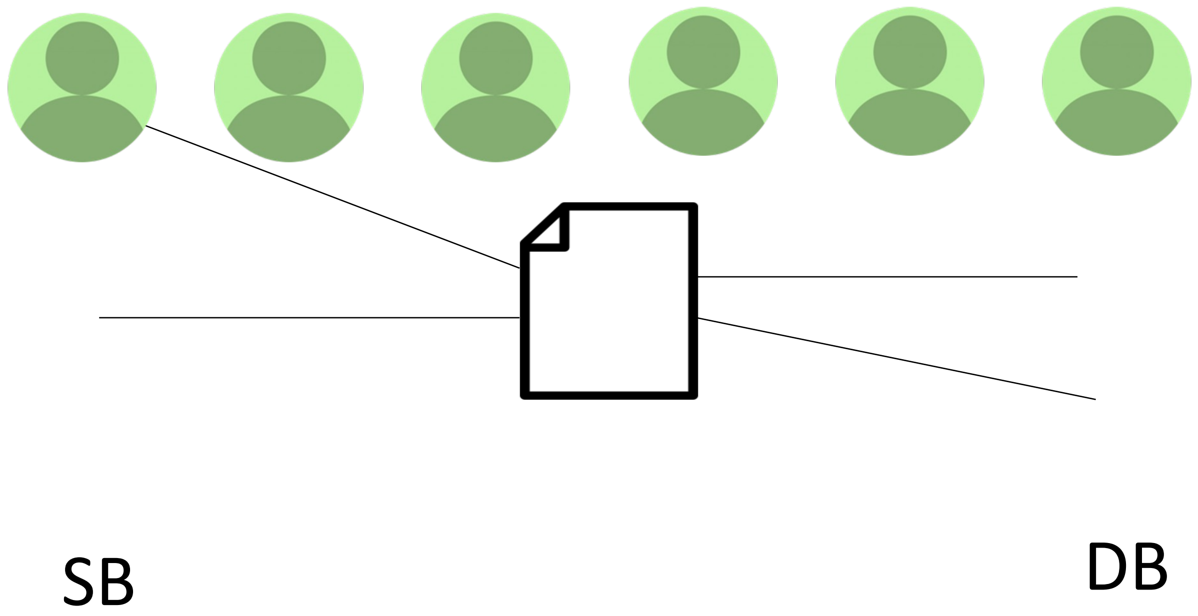
Single blind can lead to gender/fame/race/... biases



Where is the evidence of bias in my research community?

How to rigorously test for biases in peer review ?

WSDM'17 experiment: Setup



- 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: Attributes

Test for biases pertaining to *author attributes*:

- Famous author
- Top university
- Top company
- At least one woman author
- 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]

This was our starting point…

In the simulations in the next few slides, their test designed to operate at P(type I error) ≤ 0.05

Characteristic 0: Correlations between quality of papers and certain attributes

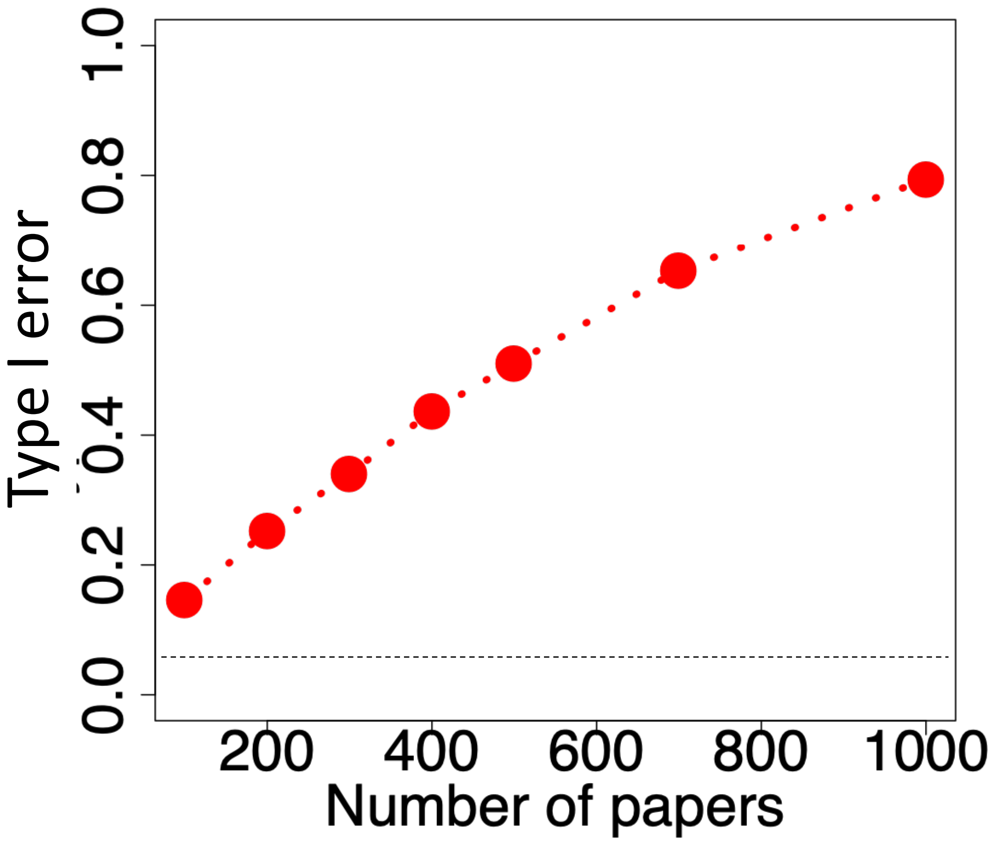
- Famous author
- Top university
- Top company

Combined with other characteristics…

Characteristic 1: Reviews are noisy

Reviewers are noisy (and hence DB reviews are a noisy estimate of "intrinsic" value q_p of any paper p)

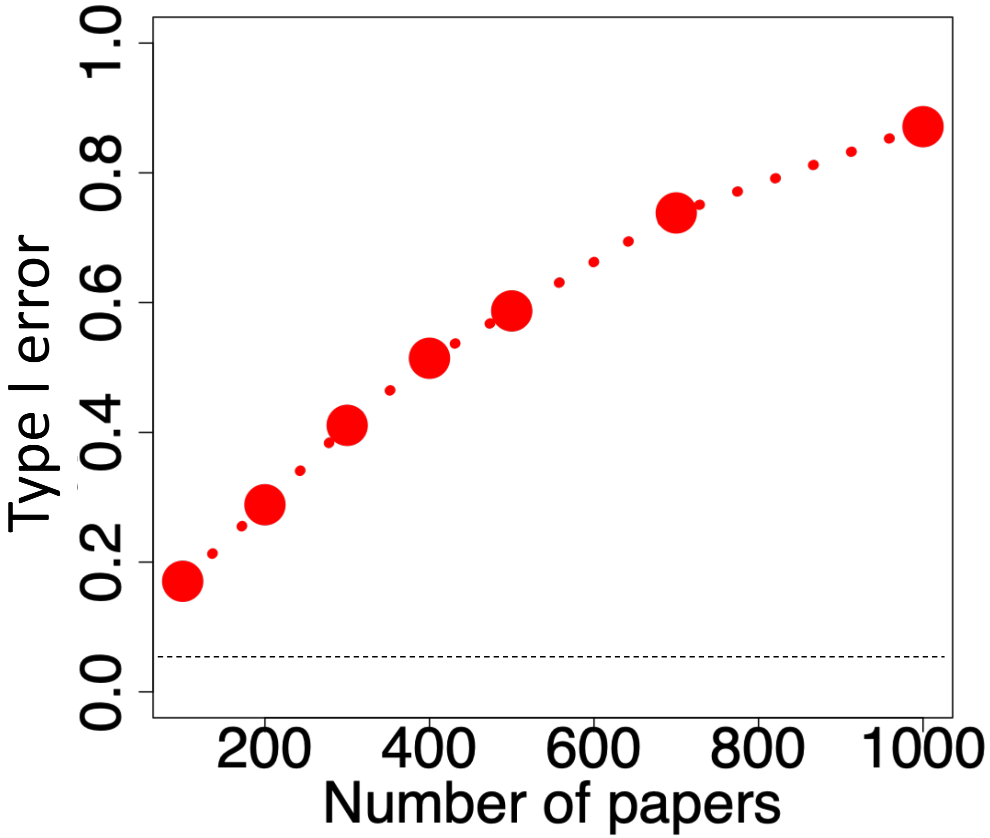
Must ensure: P(type I error) ≤ 0.05



Characteristic 2: Model complexity

Human evaluations may be more complex than the simple parametric/logistic model

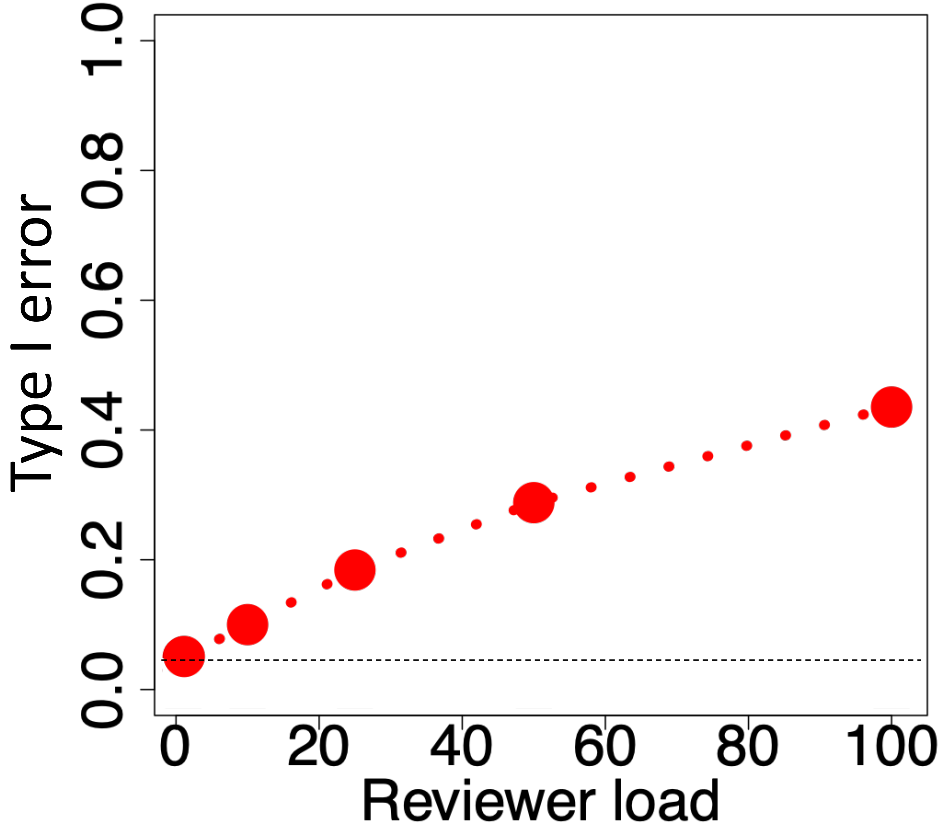
Must ensure: P(type I error) $\leq$ 0.05



Characteristic 3: 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(type I error) $\leq$ 0.05



[Mitliagkas et al. 2011, Ammar et al. 2012, Freund et al. 2003, Brenner et al. 2005, Flach et al. 2010, Roos et al. 2011, Mackay et al. 2017]

Characteristic 4: Bidding

	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			

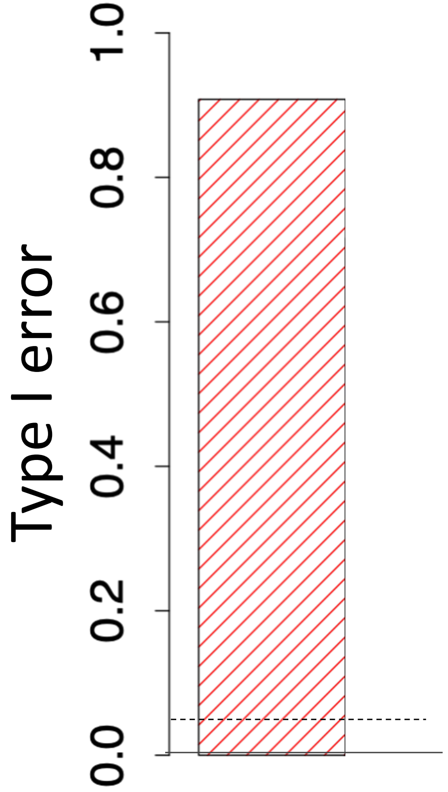
Reviewers indicate which papers they would like or not like to review

[Section 3.1.3 of bit.ly/PeerReviewOverview]

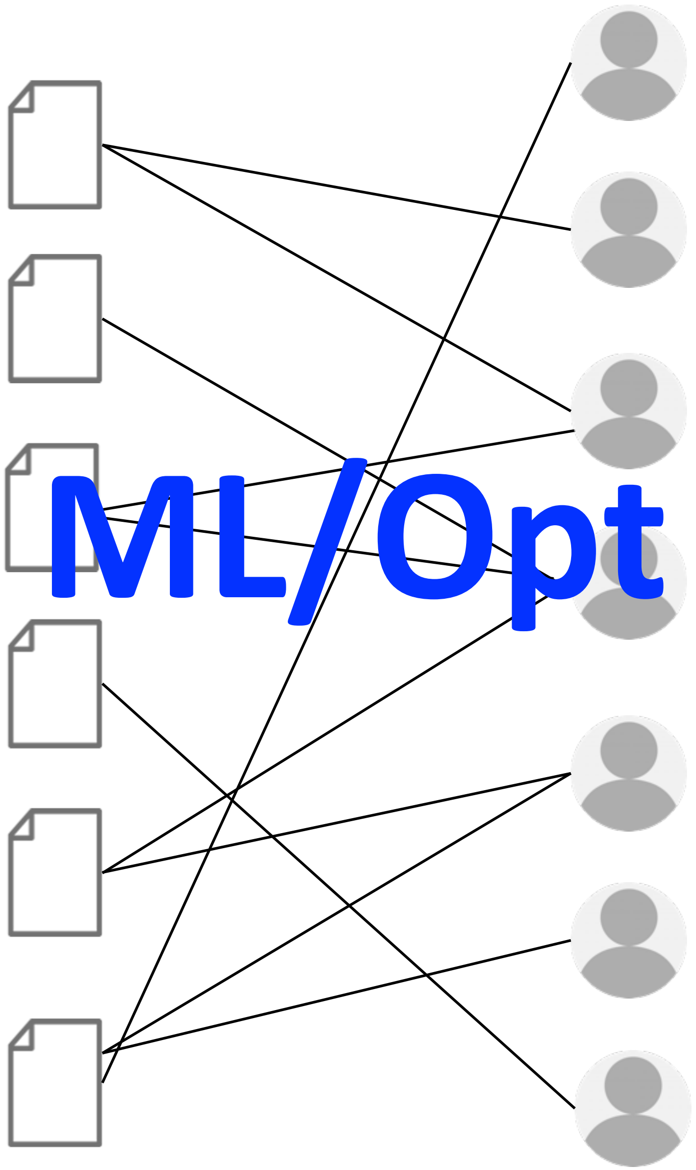
Characteristic 4: Bidding

Asymmetric bidding: SB reviewers observe author identities and DB reviewers do not

Must ensure: P(type I error) ≤ 0.05



Characteristic 5: Non-random assignment

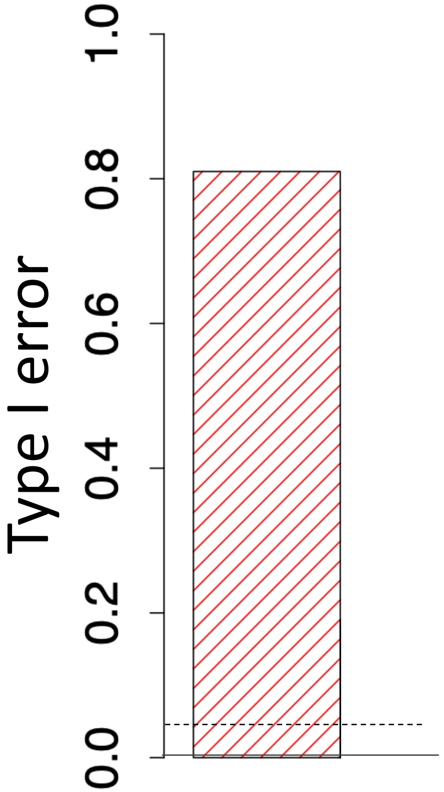


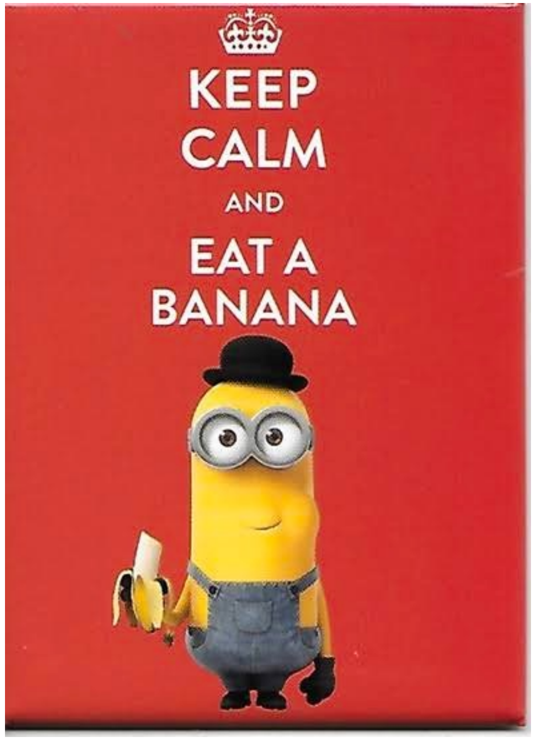
[Section 3 of bit.ly/PeerReviewOverview]

Characteristic 5: Non-random assignment

Assignment of reviewers to papers is **not** random

Must ensure: P(type I error) $\leq$ 0.05





Let’s address this.

Formulation

$\pi_{rp}^{(\mathrm{sb})}$ = P(reviewer *r* accepts of paper *p* in **SB setup)**

$\pi_{rp}^{(\mathrm{db})}$ = P(reviewer *r* accepts of paper *p* in **DB setup)**

Absence of bias. No difference in behavior of SB and DB reviewers

$$H_0 \colon \pi_{rp}^{(\mathrm{sb})} = \pi_{rp}^{(\mathrm{sb})} \quad \forall r, p$$

Presence of bias. Reviewers in SB are more harsh (or lenient) than those in DB for papers in certain group.

$$H_1 \colon \begin{array}{ll} \pi_{rp}^{(\mathrm{sb})} \leq \pi_{rp}^{(\mathrm{db})} & \text{if paper p is in group} \\ \pi_{rp}^{(\mathrm{sb})} \geq \pi_{rp}^{(\mathrm{db})} & \text{if paper p not in group} \end{array}$$

and at least one inequality is strict.

- No assumption of existence of any “true scores”
- Non-parametric model

Experiment design and test

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: Assigning 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

Our guarantees

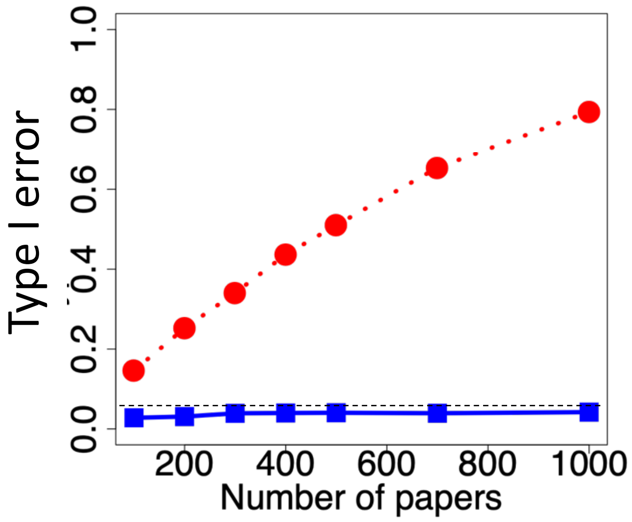
Theorem (informal)

Our experimental setup and test controls the false alarm probability at any given level $\alpha \in (0,1)$ and has asymptotic probability of detection of 1.

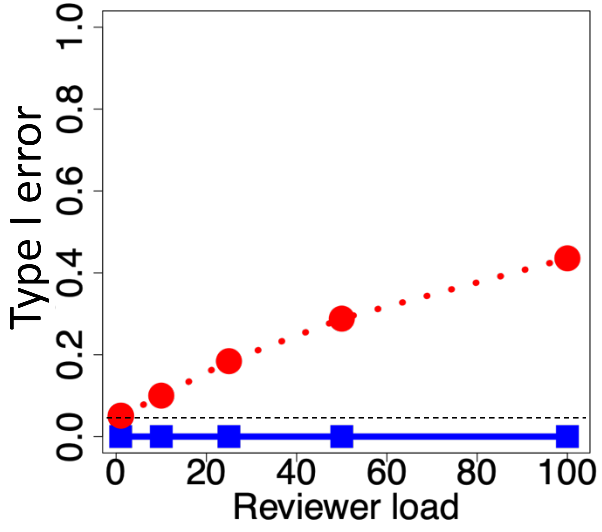
Type I error control

● **Tomkins et al.**
■ **Our work**

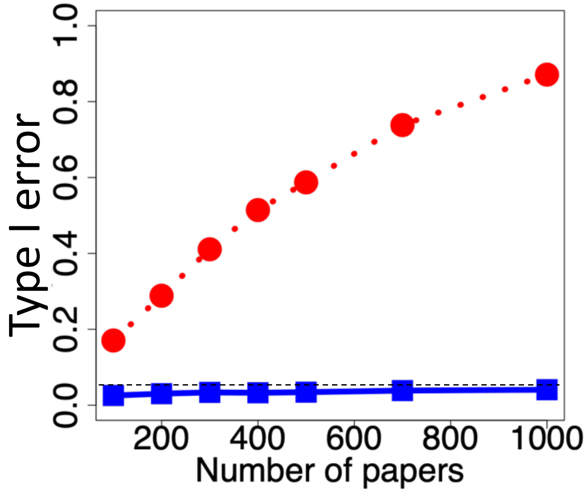
Reviews are noisy



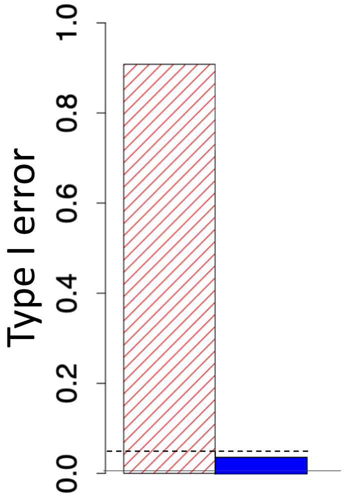
Intra-reviewer dependency



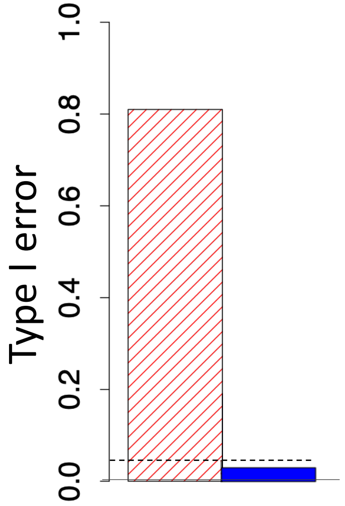
Model complexity



Bidding



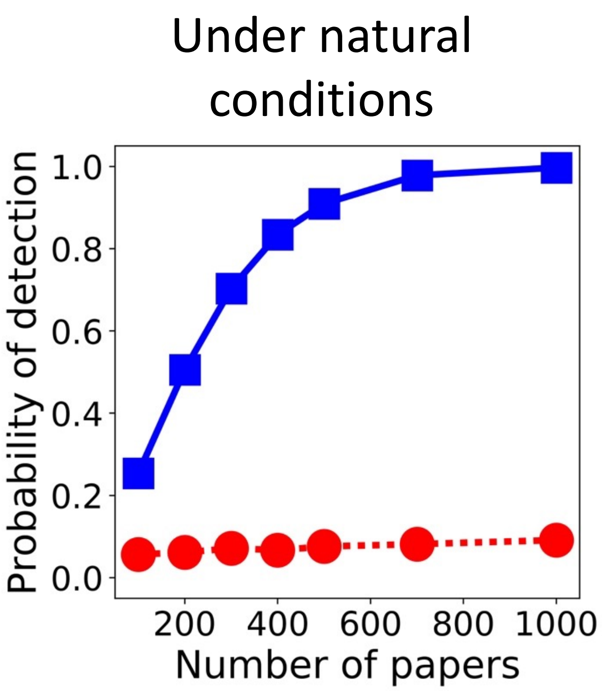
Non-random assignment

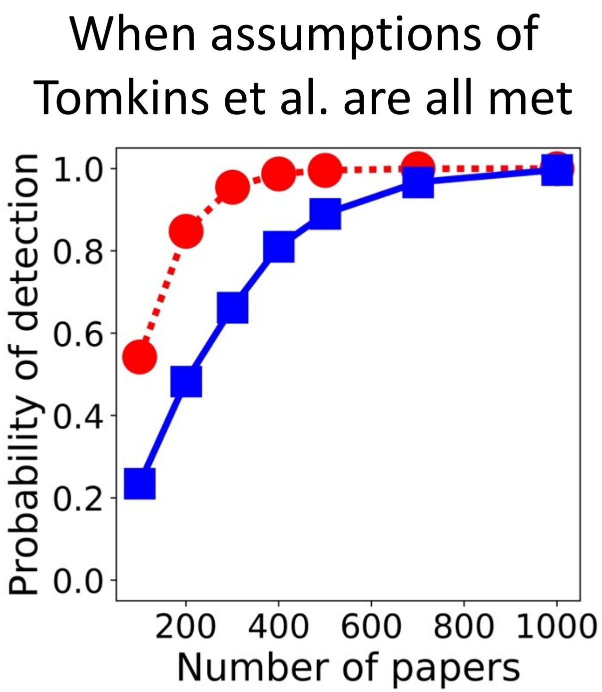


Non-trivial detection power

● **Tomkins et al.**

■ **Our work**





Open problems

- Better theoretical guarantees on power for given type I error
- arXiv playing spoilsport? [Rastogi et al. 2022]
- Biases in other review components such as program committee meetings and discussions [Teplitskiy et al. 2019]
- Biases in text [Manzoor et al. 2021]

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

Conclusions

- **Many sources of biases and unfairness in peer review**

- **Urgent need to revamp peer review, at scale**
  - Lot at stake: Careers, Scientific progress

- **Lots of open problems!**
  - Exciting
  - Theoretical / Applied / Conceptual
  - Challenging
  - **Impactful**

Overview article: bit.ly/PeerReviewOverview

Merci! Questions?

Feel free to reach out: nihars@cs.cmu.edu