

On Testing for Biases in Peer Review

Ivan Stelmakh, Nihar Shah and Aarti Singh

**Carnegie
Mellon
University**

Double-Blind vs Single-Blind



Lady Justice. Court of Final Appeal, Hong Kong



Lady Justice. Supreme Court, Moscow

Blank, 1991; Seeber & Bacchelli, 2017; Snodgrass, 2006; Largent & Snodgrass, 2016; Okike et al., 2016; Budden et al., 2008; Webb et al., 2008; Hill & Provost, 2003

Background

- In 2017 Tomkins, Zhang and Heavlin conducted a remarkable semi-randomized controlled trial during the peer review for the WSDM conference to test for biases in single-blind peer review
- They found biases in favour of papers authored by
 - ◆ Researchers from top universities
 - ◆ Researchers from top companies
 - ◆ Famous authors
- WSDM switched to the double-blind setup in 2018, but many CS theory conferences still use single-blind peer review

The focus of this work is on principled design of the methods to test for biases within conference peer review

Problem setup and notation

Papers



$$\longrightarrow w_j \in \{-1, 1\}$$

In this talk we assume that there is only one protected attribute

w - **protected attribute** Equals 1 if paper's authors belong to a category of interest.

If we are interested in gender biases, then $w = 1$ if a paper has a female lead author and $w = -1$ otherwise

Reviewers



Accept

Reject

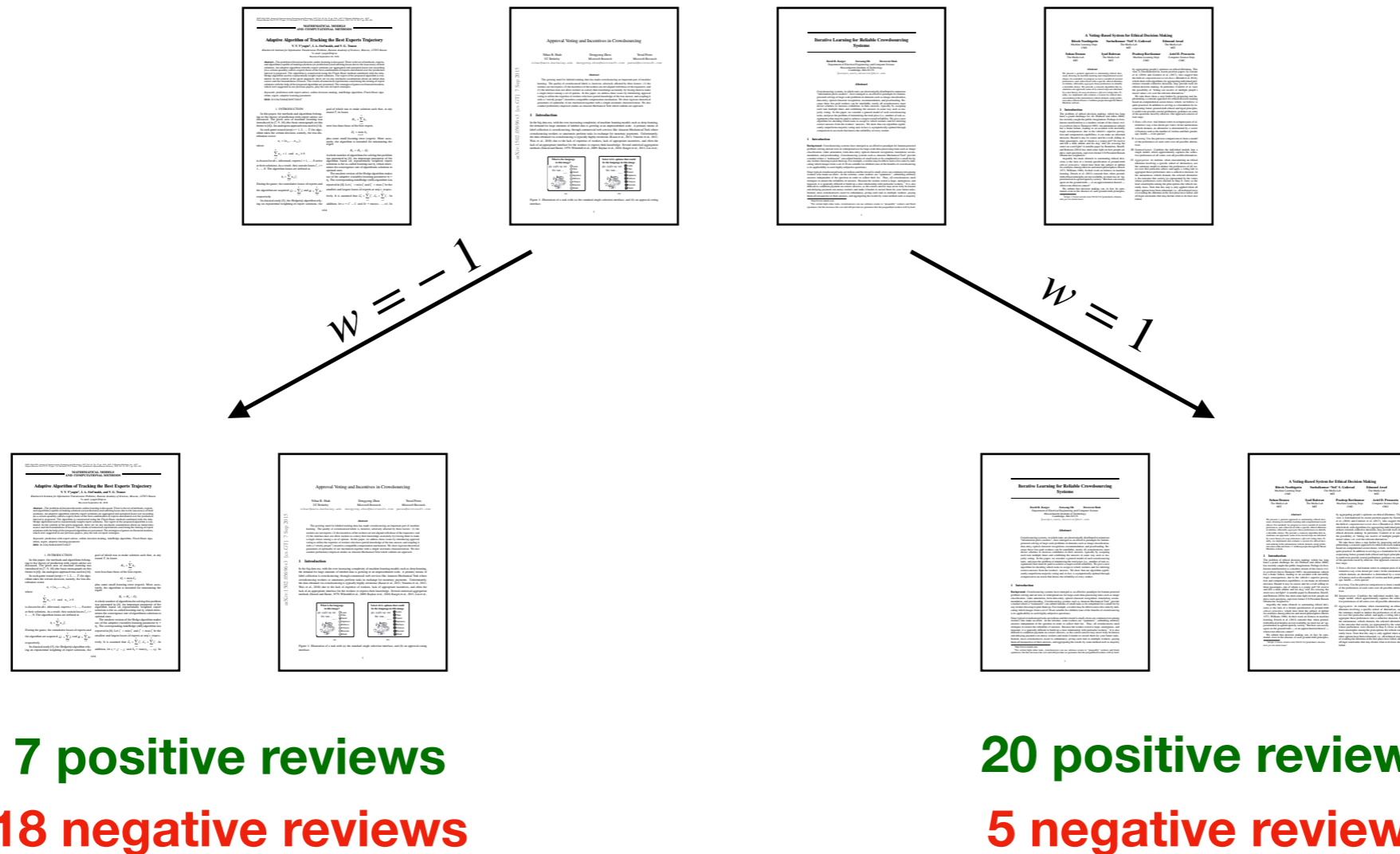
Reviewers in SB setup **observe** the protected attribute

Reviewers in DB setup **do not observe** the protected attribute

Question of interest

Are reviewers in SB setup biased against or in favour of papers from the category of interest?

Naive approach



Can we conclude that there is a bias in favour of papers with $w = 1$?

NO!

Think about category of interest being **Nobel laureates**.



Hypothesis testing framework

Data: reviewers decisions and papers authorship information

Null hypothesis (absence of bias)

H_0 : knowledge of the protected attribute does not change reviewers attitude to the paper

Alternative hypothesis (presence of bias)

H_1 : knowledge of the protected attribute makes reviewers more harsh to papers with $w_j = -1$ and more lenient to papers with $w_j = 1$

Type-I error

Claiming **presence** of bias when the bias is **absent**

Type-II error

Claiming **absence** of bias when the bias is **present**

For a given α our goal is to design a test that **provably keeps Type-I error rate below α and subject to this has low Type-II error**

Past approach: parametrization of objectivity

Tomkins, Zhang and Heavlin, 2017

Setup of the experiment

Reviewers are allocated to conditions uniformly at random



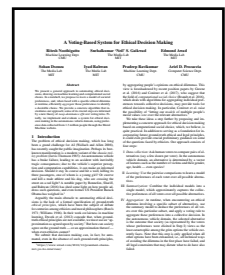
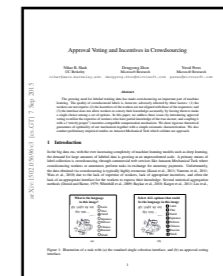
Allocation



SB condition

For simplicity, assume that each paper is assigned to 1 SB and 1 DB reviewer

Assignment
(any algorithm)



DB condition



Testing procedure

Objective score model



q

Each paper has **unknown** parameter that represents a quality of the paper.

Double blind estimate

q



\hat{q}

DB reviewers estimate paper quality

- DB reviewers do not have access to protected attribute
- Estimated quality \hat{q} is not biased

Testing procedure

$w_j \in \{-1, 1\}$ - protected attribute

$\pi_{ij}^{(sb)} \in [0, 1]$ - probability that reviewer i recommends acceptance of paper j in SB setup

Logistic model for bias

$$\forall i, j$$
$$\log \frac{\pi_{ij}^{(sb)}}{1 - \pi_{ij}^{(sb)}} = \beta_0 + \beta_1 q_j + \beta_2 w_j \quad (1)$$

$\beta_0, \beta_1, \beta_2$ - unknown coefficients to be estimated from data

Null hypothesis (absence of bias)

$$H_0 : \beta_2 = 0$$

Alternative hypothesis (presence of bias)

$$H_1 : \beta_2 \neq 0$$

Outline of the Tomkins et al. test

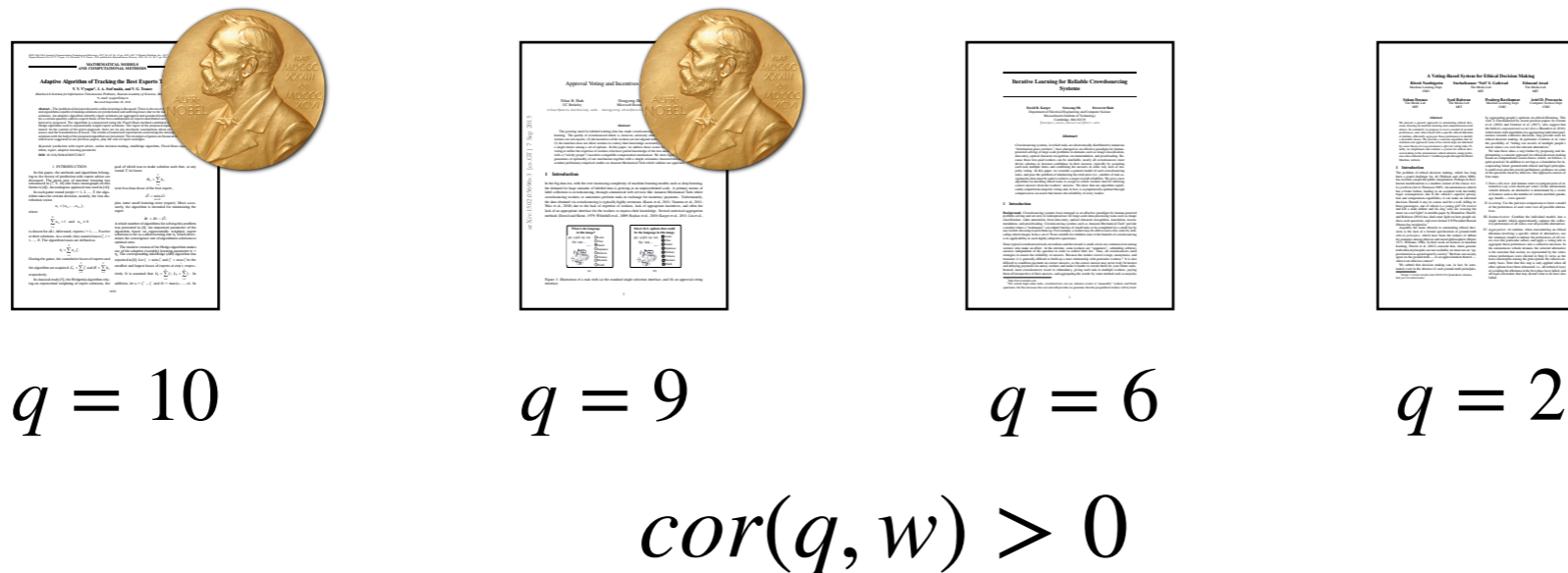
- Use DB reviewers to estimate q_j for each paper j and use plugin estimates \hat{q}_j in (1)
- Fit observed **accept/reject** decisions of SB reviewers into logistic model
- Use Wald test at the level α to test if $\beta_2 = 0$

Negative results

Key ingredient

Recall that for some categories of authors it is natural to expect that their papers have above average qualities (the Nobel laureates example)

Formally, this intuition can be expressed as a non-zero correlation between w and q .

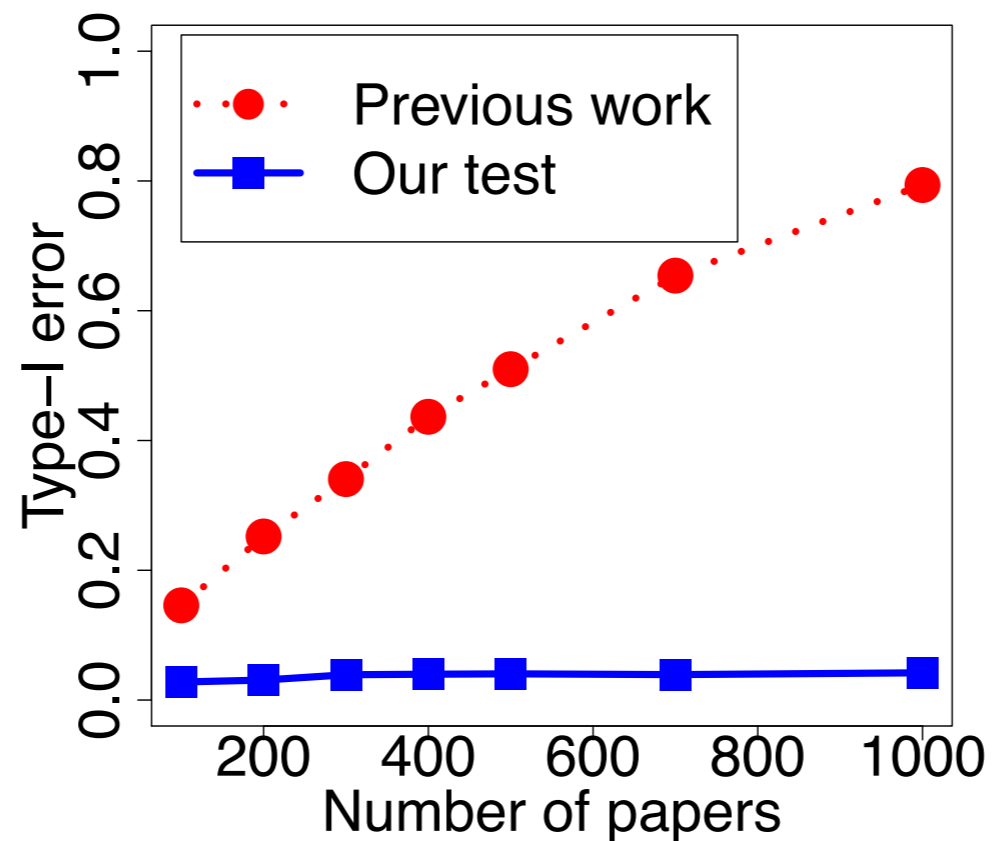


Correlation itself doesn't cause issues, but we identify several conditions where it can be **significantly harmful**

Reviewers' noise

Reviewers are noisy and hence \hat{q}_j should be seen as a **noisy estimate** of q_j

In presence of correlations, the noise in covariate measurement **may undermine Type-I error guarantees** of the Tomkins et al. test

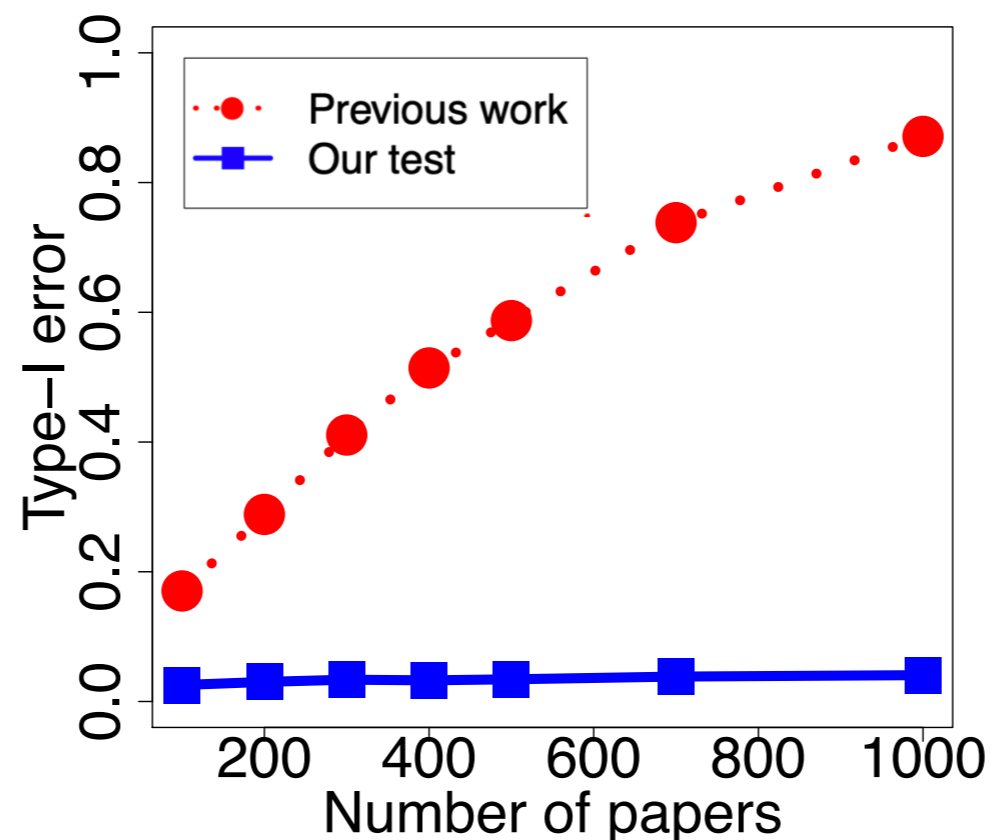


Comparison of Type-I error rates when DB reviewers are noisy. Valid test must have Type-I error below $\alpha = 0.05$

Observe that the issue exacerbates as sample size grows!

Model mismatch

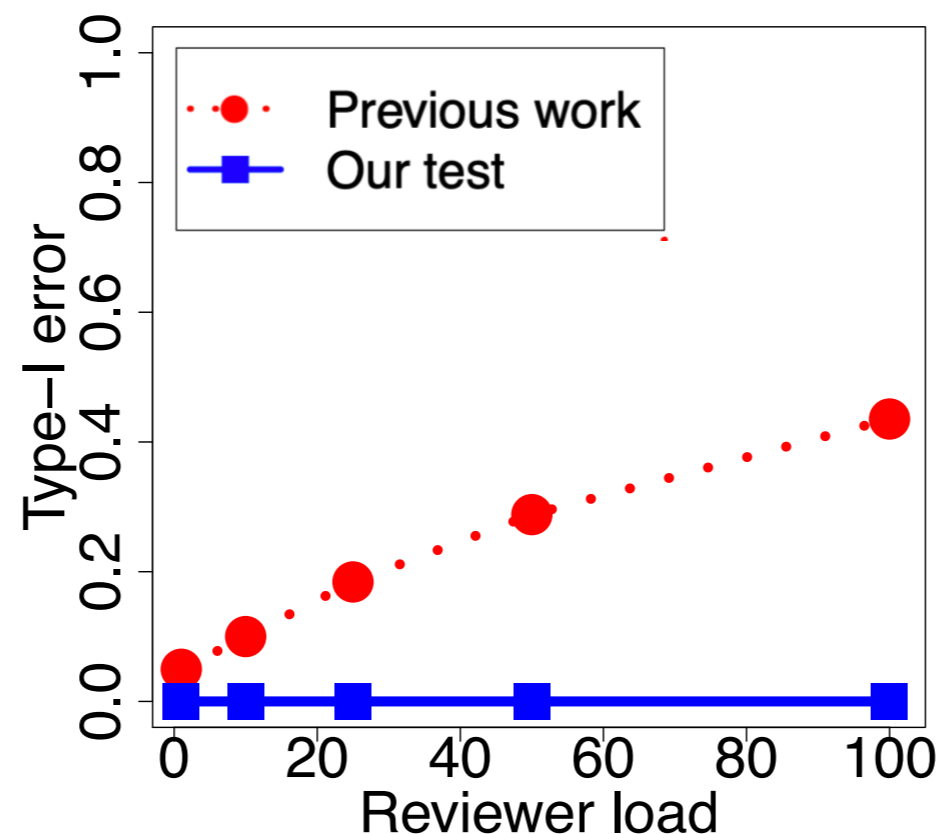
Reasonable violation of the parametric model may **break Type-I error guarantees** of the test from the past work



Comparison of Type-I error rates under violation of parametric model. Valid test must have Type-I error below $\alpha = 0.05$

Miscalibration

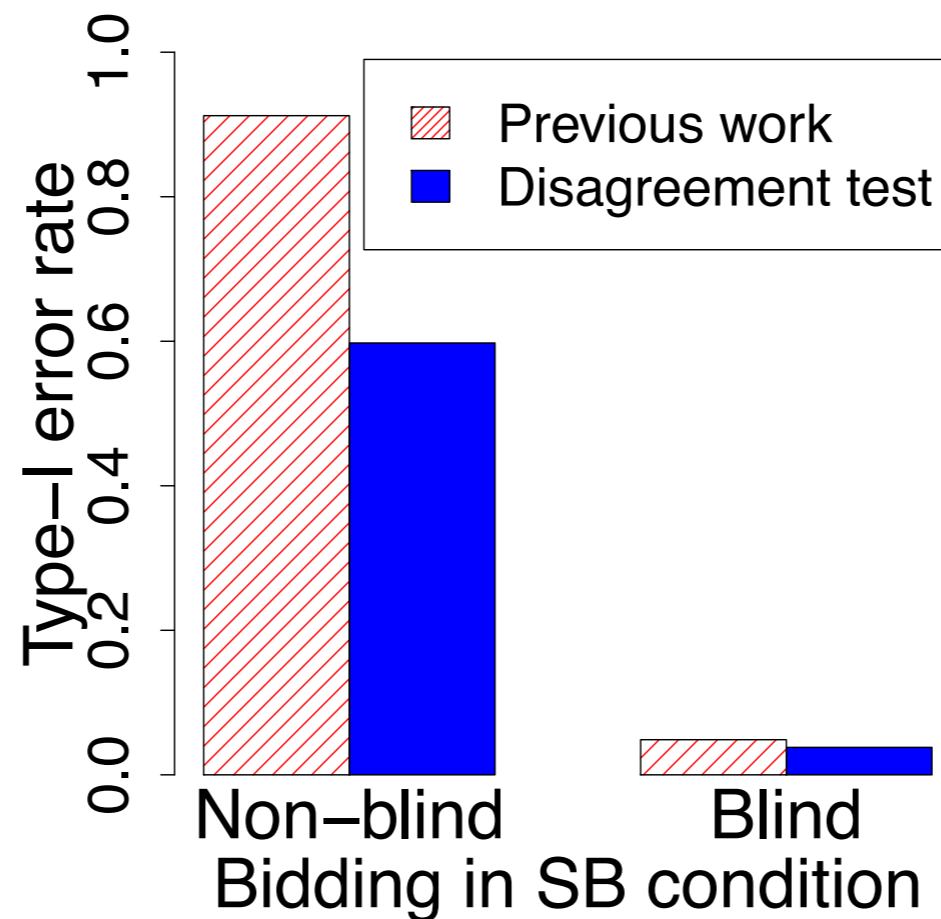
Parametric model assumes that reviews given by the same reviewer to different papers are independent. In practice, **this assumption may be violated due to spurious correlations introduced by reviewers' miscalibration**



Comparison of Type-I error rates for specific pattern of reviewers' miscalibration. Valid test must have Type-I error below $\alpha = 0.05$

Bidding

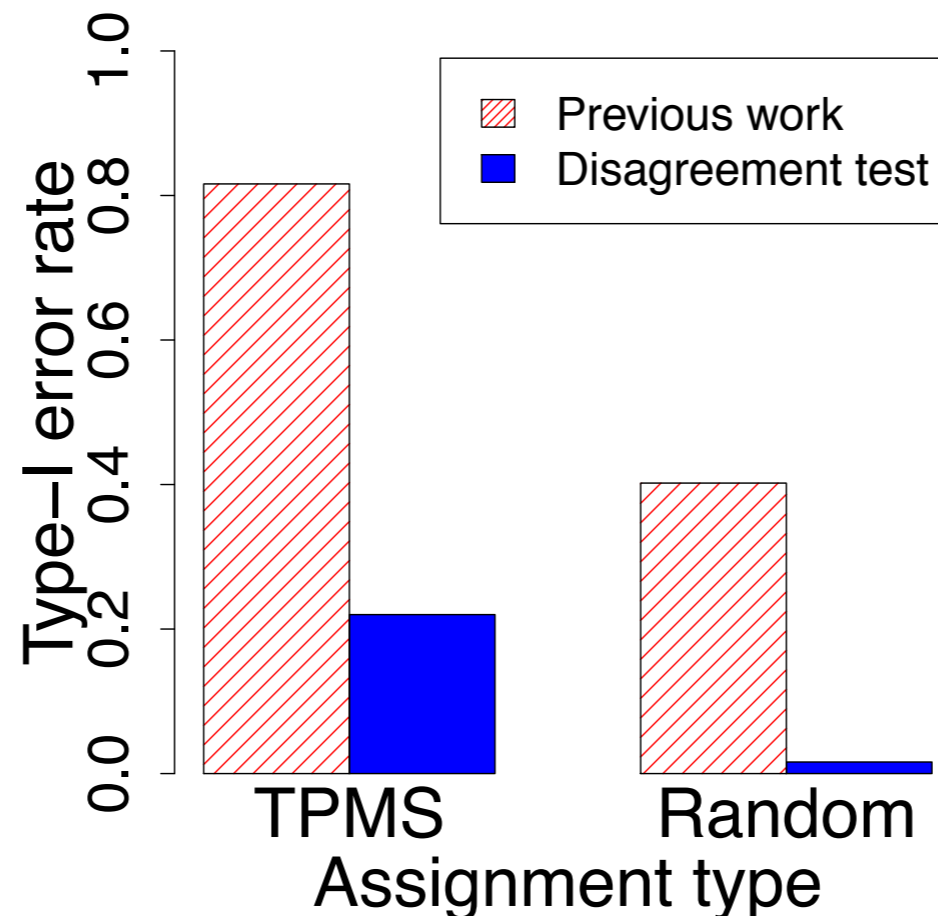
If SB reviewers observe protected attributes during the bidding state and DB reviewers do not, the **testing for biases in decisions is hard**



Comparison of Type-I error rates for specific pattern of reviewers' bidding. Valid test must have Type-I error below $\alpha = 0.05$

Non-random assignment

Under the experimental setup of Tomkins et al., if reviewers are assigned to papers using popular TPMS assignment algorithm, then **controlling Type-I error rate is hard**



Comparison of Type-I error rates for specific pattern of similarity matrix. Valid test must have Type-I error below $\alpha = 0.05$

Observe that even our robust test is unable to control for the Type-I error when assignment of papers to reviewers is not random.

**Our approach:
deparametrization of subjectivity**

Setup of the experiment

Reviewers are allocated to conditions uniformly at random

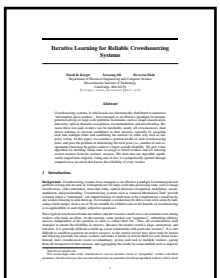
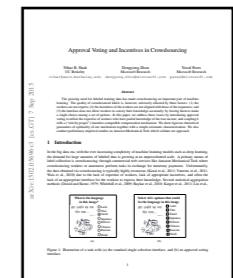


Allocation

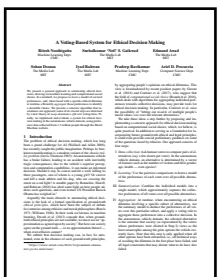


SB condition

For simplicity, assume that each paper is assigned to 1 SB and 1 DB reviewer



Random Assignment



DB condition

For today's talk only.



Our formulation

$\pi_{ij}^{(sb)} \in [0,1]$ - probability that reviewer i recommends acceptance of paper j in **SB setup**

$\pi_{ij}^{(db)} \in [0,1]$ - probability that reviewer i recommends acceptance of paper j in **DB setup**

Absence of bias. There is no difference in behaviour of SB and DB reviewers

$$H_0 : \pi_{ij}^{(sb)} = \pi_{ij}^{(db)} \quad \forall i, j$$

Presence of bias. Each reviewer is more harsh (resp. lenient) to papers with $w = 1$ (resp. $w = -1$) in SB condition than in DB condition

$$H_1 : \begin{cases} \pi_{ij}^{(sb)} \leq \pi_{ij}^{(db)} & \text{if } w_j = 1 \\ \pi_{ij}^{(sb)} \geq \pi_{ij}^{(db)} & \text{if } w_j = -1 \end{cases}$$

At least one inequality is strict

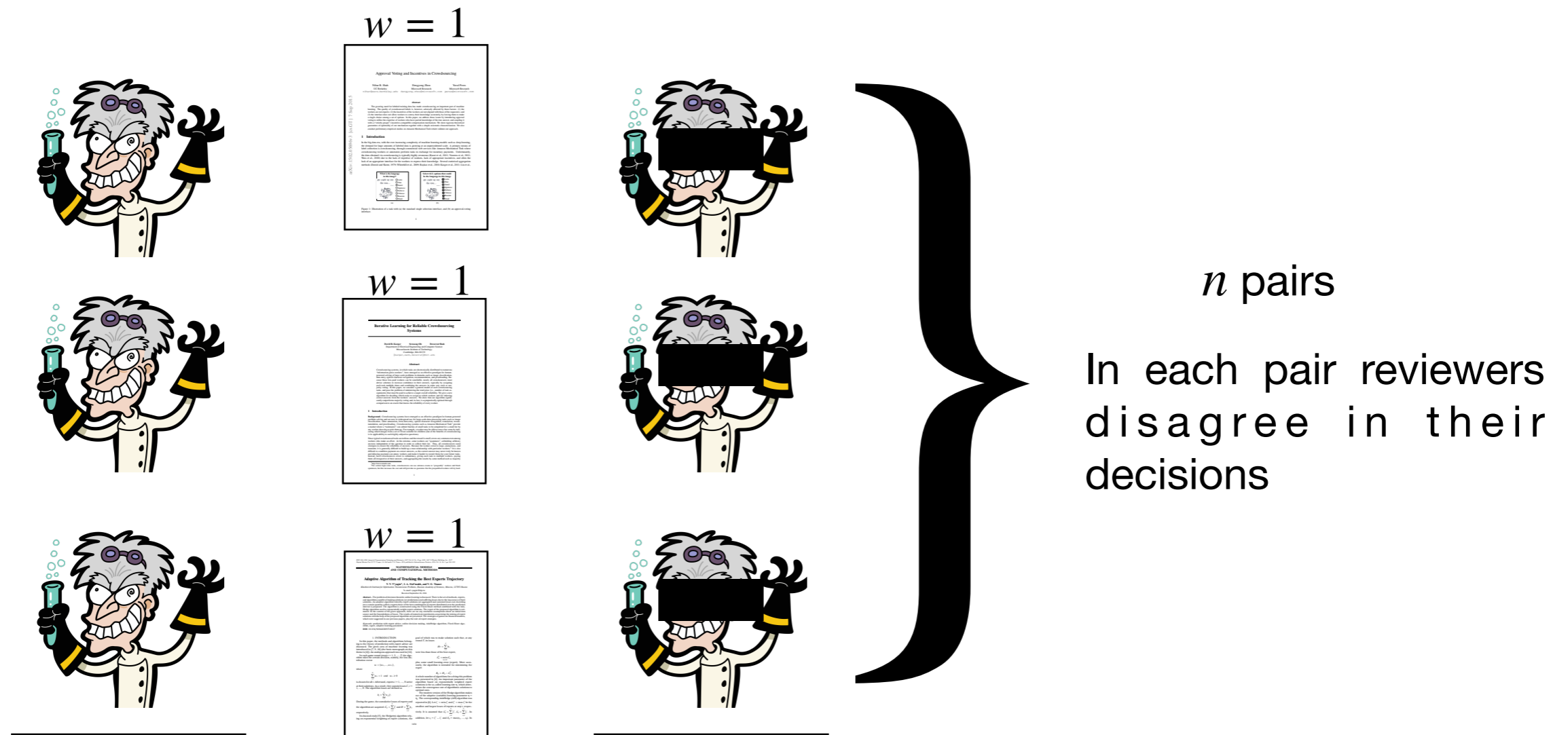
- **Subjective score model.** We do not assume existence of true underlying scores and hence allow for subjectivity
- **Non-parametric model.** We do not assume any parametric relationship that describes reviewers' behaviour

Our formulation generalizes the formulation of the past work

Testing procedure. Intuition

Assume that the bias is absent

$$H_0 : \pi_{ij}^{(sb)} = \pi_{ij}^{(db)} \quad \forall i, j$$



How many **accepts** do you expect in SB condition?

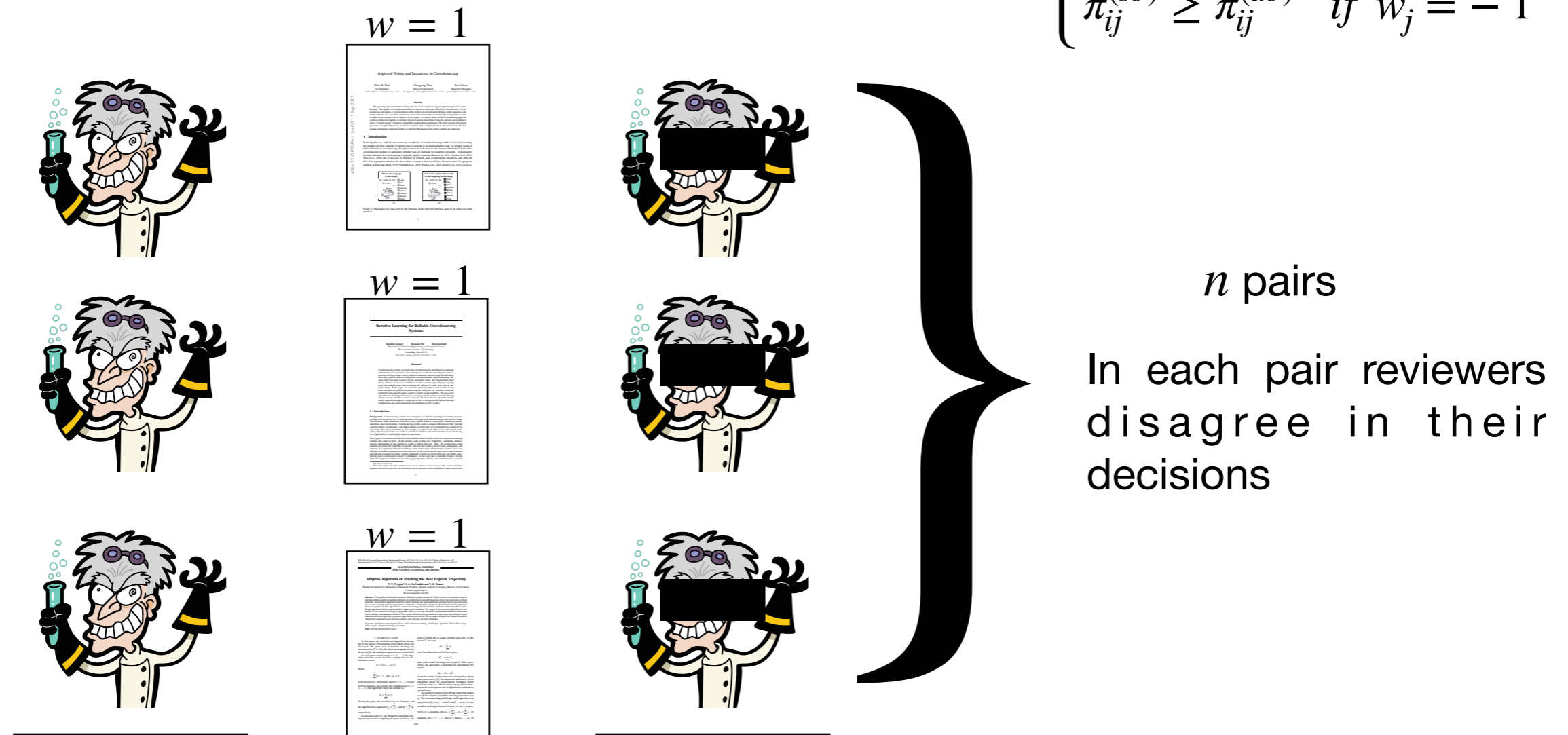
How many **accepts** do you expect in DB condition?

We expect approximately equal number of accepts

Testing procedure. Intuition

Assume that the bias is present

$$H_1 : \begin{cases} \pi_{ij}^{(sb)} \leq \pi_{ij}^{(db)} & \text{if } w_j = 1 \\ \pi_{ij}^{(sb)} \geq \pi_{ij}^{(db)} & \text{if } w_j = -1 \end{cases}$$

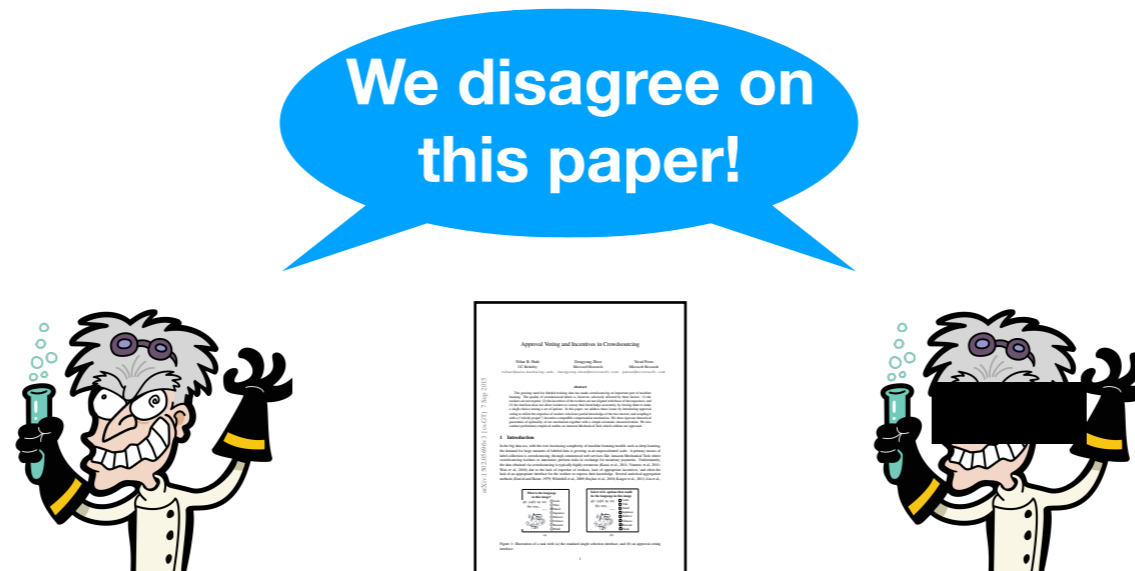


How many **accepts** do you expect in SB condition?

How many **accepts** do you expect in DB condition?

We expect more accepts from DB reviewers

Testing procedure



Disagreement test

- Find a set of triples (SB reviewer, DB reviewer, Paper) such that
 - ♦ Each reviewer appears in at most one triple
 - ♦ There are 'enough' papers with $w = -1$ and $w = 1$
- Condition on triples where reviewers disagree in their decisions
- Run permutation test at the level α to look for 'trends' in the remaining triples

Main Result

Power — ability to detect bias when it is present

Informal theorem

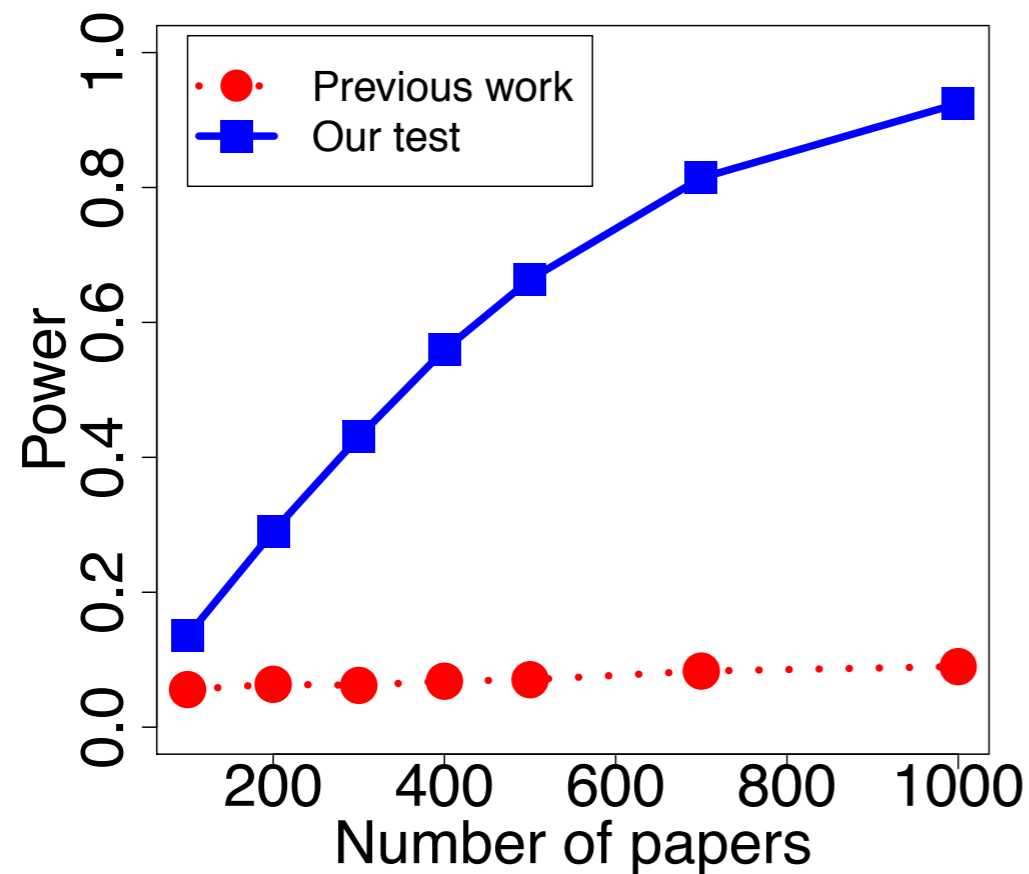
Under our formulation of the bias testing problem, if the assignment of reviewers to papers is performed at random, the disagreement test controls for the Type-I error rate at any given level $\alpha \in (0,1)$ and has non-trivial power.

Remark

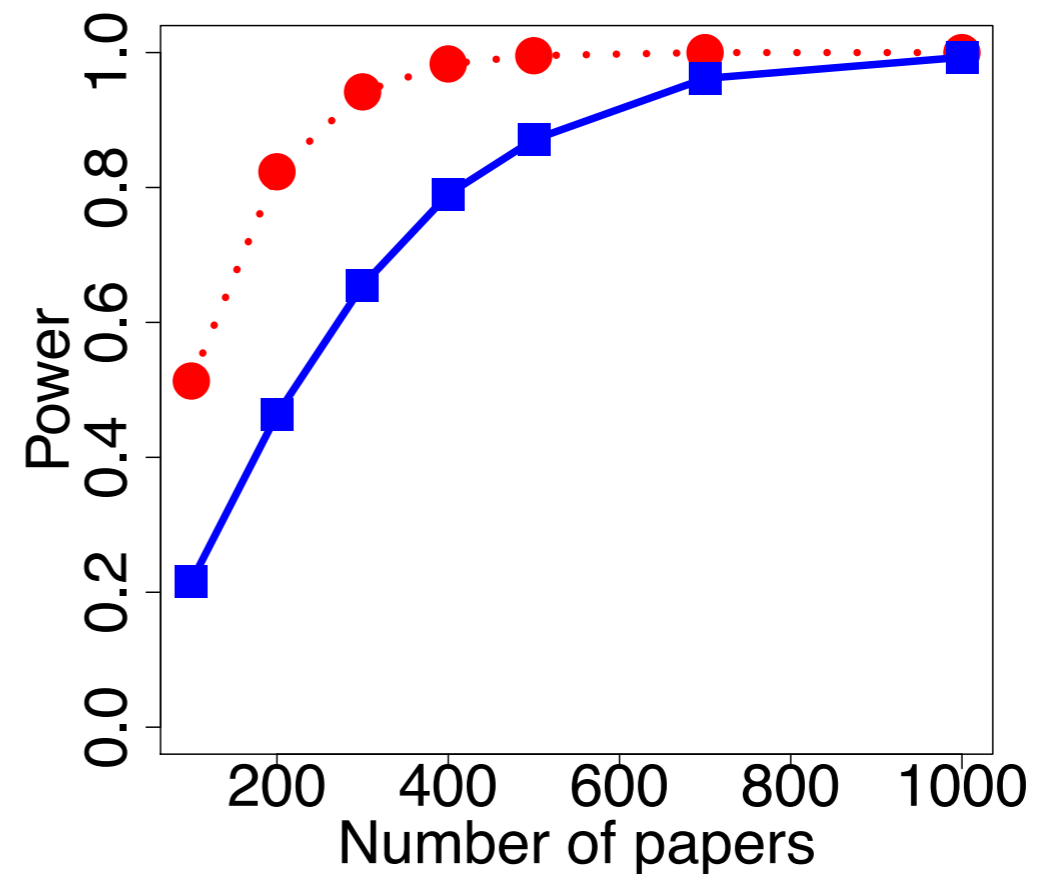
In the full version of the paper we omit the requirement of the random assignment, thereby ensuring that the test can be run irrespective of what the assignment algorithm is used

Power of the test

Power — ability to detect bias when it is present



DB reviewers are noisy



All assumptions of the previous work are satisfied

Synthetic simulations. Higher values are better. Bias is present in both cases. Error bars are too small to be visible.

Outline

- Past approaches to test for biases in peer review are **at risk of being unreliable** under plausible violations of strong assumptions
- We design the Disagreement test that **provably controls for the Type-I error rate** under significantly weaker assumptions

More in the full paper

- New experimental procedure that **allows for any assignment algorithm** to be used
- General case of **more than one protected attribute**
- Generalization of the bias testing problem in case **SB condition itself may change the behaviour** of reviewers even under absence of bias