On Testing for Biases in Peer Review

Ivan Stelmakh

joint work with Nihar B. Shah and Aarti Singh

Machine Learning Department
Carnegie Mellon University

Double-Blind vs Single-Blind

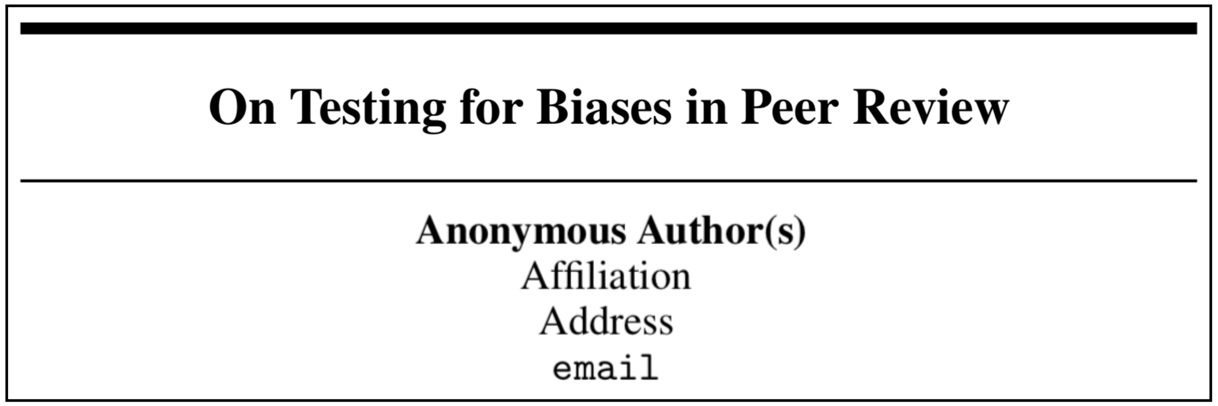
On Testing for Biases in Peer Review

Anonymous Author(s)
Affiliation
Address
`email`

On Testing for Biases in Peer Review

Ivan Stelmakh, Nihar B. Shah and Aarti Singh
School of Computer Science
Carnegie Mellon University
`{stiv,nihars,aarti}@cs.cmu.edu`

Double-Blind vs Single-Blind



On Testing for Biases in Peer Review

Anonymous Author(s)
Affiliation
Address
email

On Testing for Biases in Peer Review

Ivan Stelmakh, Nihar B. Shah and Aarti Singh
School of Computer Science
Carnegie Mellon University
{stiv,nihars,aarti}@cs.cmu.edu

- Lot of debate on gender/race/fame/… biases in single-blind peer review

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; Tomkins et al., 2017

Double-Blind vs Single-Blind

On Testing for Biases in Peer Review

Anonymous Author(s)
Affiliation
Address
`email`

On Testing for Biases in Peer Review

Ivan Stelmakh, Nihar B. Shah and Aarti Singh
School of Computer Science
Carnegie Mellon University
`{stiv,nihars,aarti}@cs.cmu.edu`

- Lot of debate on gender/race/fame/… biases in single-blind peer review
- Many conferences use single-blind review

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; Tomkins et al., 2017

Double-Blind vs Single-Blind

On Testing for Biases in Peer Review

Anonymous Author(s)
Affiliation
Address
`email`

On Testing for Biases in Peer Review

Ivan Stelmakh, Nihar B. Shah and Aarti Singh
School of Computer Science
Carnegie Mellon University
`{stiv,nihars,aarti}@cs.cmu.edu`

- Lot of debate on gender/race/fame/… biases in single-blind peer review
- Many conferences use single-blind review
- «Where is the evidence of bias in my academic community?»

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; Tomkins et al., 2017

Double-Blind vs Single-Blind

On Testing for Biases in Peer Review

Anonymous Author(s)
Affiliation
Address
email

On Testing for Biases in Peer Review

Ivan Stelmakh, Nihar B. Shah and Aarti Singh
School of Computer Science
Carnegie Mellon University
{stiv,nihars,aarti}@cs.cmu.edu

- Lot of debate on gender/race/fame/… biases in single-blind peer review
- Many conferences use single-blind review
- «Where is the evidence of bias in my academic community?»

Our focus is on tools to test for biases in single-blind conference peer review

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; Tomkins et al., 2017

Remarkable WSDM'17 Experiment

Tomkins, Zhang and Heavlin, 2017

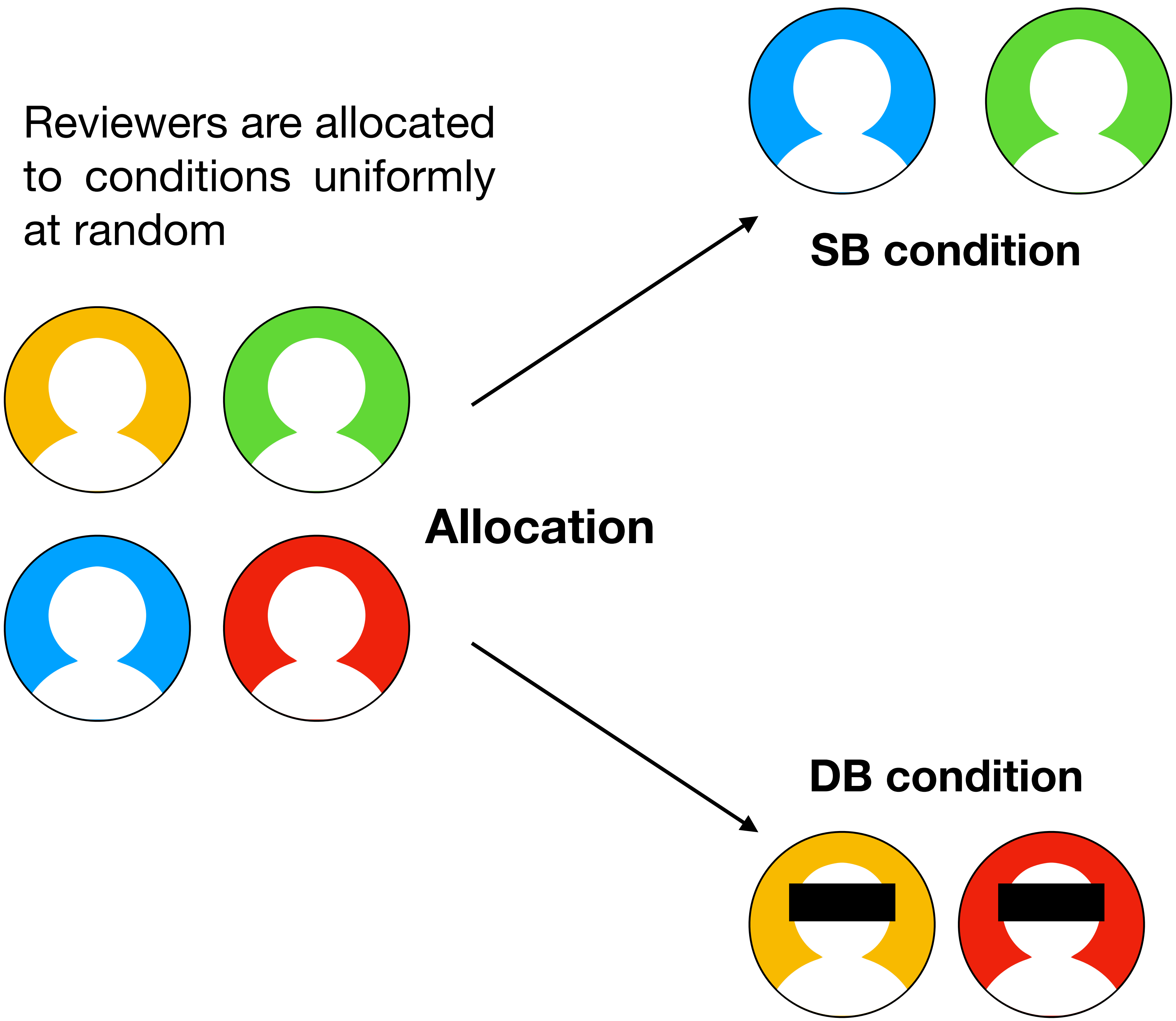
Remarkable WSDM'17 Experiment

Tomkins, Zhang and Heavlin, 2017

Remarkable WSDM’17 Experiment

Tomkins, Zhang and Heavlin, 2017

Reviewers are allocated to conditions uniformly at random

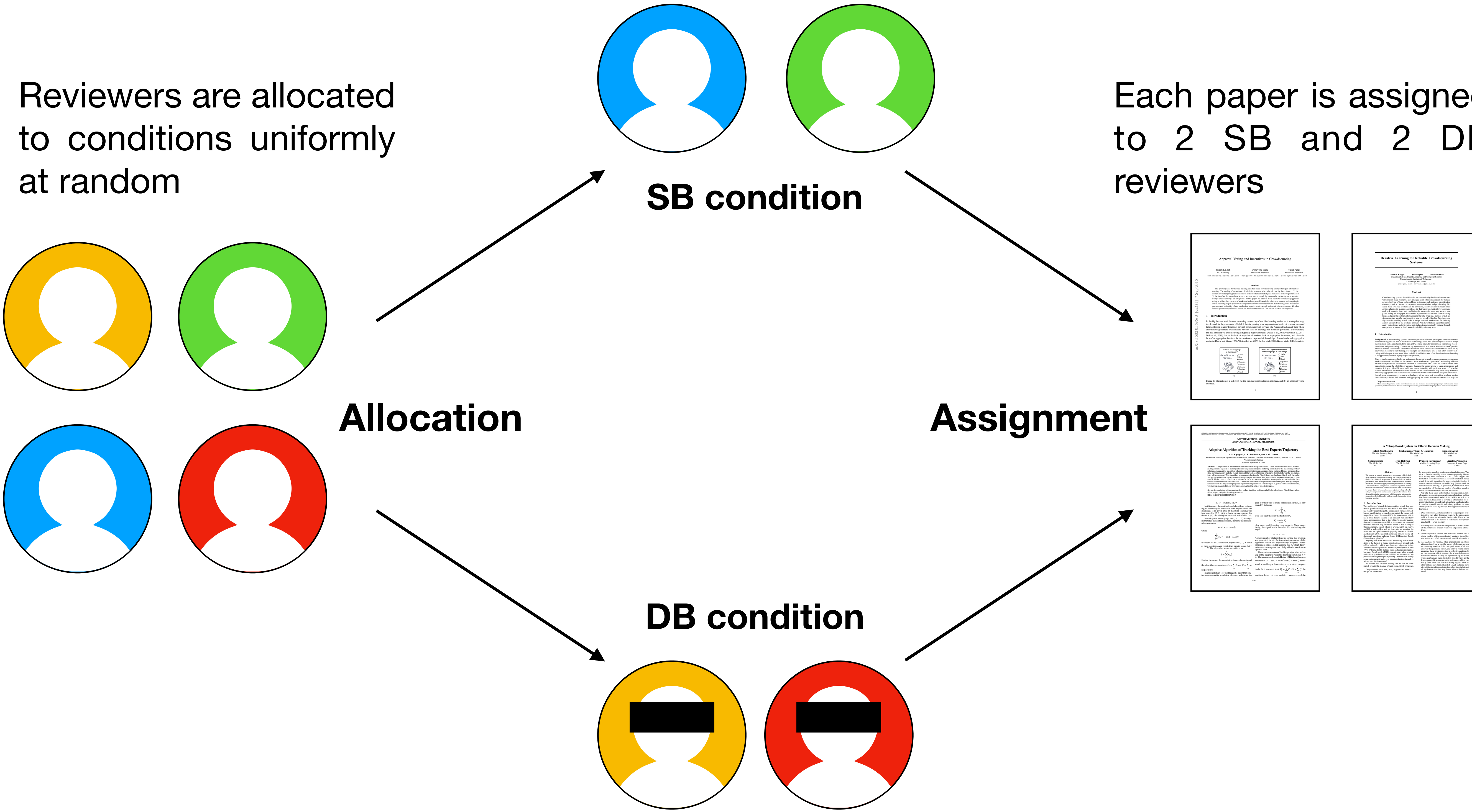


Remarkable WSDM'17 Experiment

Tomkins, Zhang and Heavlin, 2017

Reviewers are allocated to conditions uniformly at random

Each paper is assigned to 2 SB and 2 DB reviewers



Remarkable WSDM'17 Experiment

Tomkins, Zhang and Heavlin, 2017

Results of the experiment

Remarkable WSDM'17 Experiment

Tomkins, Zhang and Heavlin, 2017

Results of the experiment

- SB review induces biases in favour of papers authored by

Remarkable WSDM'17 Experiment

Tomkins, Zhang and Heavlin, 2017

Results of the experiment

- SB review induces biases in favour of papers authored by
  - ✦ Researchers from top universities

Remarkable WSDM'17 Experiment

Tomkins, Zhang and Heavlin, 2017

Results of the experiment

- SB review induces biases in favour of papers authored by
  - ✦ Researchers from top universities
  - ✦ Researchers from top companies

Remarkable WSDM'17 Experiment

Tomkins, Zhang and Heavlin, 2017

Results of the experiment

- SB review induces biases in favour of papers authored by
  - ✦ Researchers from top universities
  - ✦ Researchers from top companies
  - ✦ Famous researchers

Remarkable WSDM’17 Experiment

Tomkins, Zhang and Heavlin, 2017

Results of the experiment

- SB review induces biases in favour of papers authored by
  - ✦ Researchers from top universities
  - ✦ Researchers from top companies
  - ✦ Famous researchers
- No bias against female authors observed, but meta-analysis detects it

Remarkable WSDM'17 Experiment

Tomkins, Zhang and Heavlin, 2017

Results of the experiment

- SB review induces biases in favour of papers authored by
  - Researchers from top universities
  - Researchers from top companies
  - Famous researchers
- No bias against female authors observed, but meta-analysis detects it
- WSDM switched to double-blind peer review in 2018

Our Work

Our Work

Peer review setup has many idiosyncrasies and requires utmost care when making policy-changing conclusions.

Our Work

Peer review setup has many idiosyncrasies and requires utmost care when making policy-changing conclusions.

Negative results

We uncover a number of **issues in the methodology** of the past work and show that one **cannot use off-the-shelf procedures** to test for biases

Our Work

Peer review setup has many idiosyncrasies and requires utmost care when making policy-changing conclusions.

Negative results

We uncover a number of **issues in the methodology** of the past work and show that one **cannot use off-the-shelf procedures** to test for biases

Positive results

We design a **principled approach towards testing** for biases in peer review

Testing Paradigm

Testing Paradigm

- **False alarm.** Claiming the bias when the bias is **absent**

Testing Paradigm

- **False alarm.** Claiming the bias when the bias is **absent**
- **Correct detection.** Claiming the bias when the bias is **present**

Testing Paradigm

- **False alarm.** Claiming the bias when the bias is **absent**
- **Correct detection.** Claiming the bias when the bias is **present**

Reliable testing

maximize **probability of correct detection**

s.t. **probability of false alarm < 0.05**

Testing Paradigm

- **False alarm.** Claiming the bias when the bias is **absent**
- **Correct detection.** Claiming the bias when the bias is **present**

Reliable testing

maximize **probability of correct detection**

s.t. **probability of false alarm < 0.05**

Control over false alarm probability is of utmost importance

Negative Results

Negative Results

Ingredient 1: **Correlation** between paper quality and author category

Negative Results

Ingredient 1: **Correlation** between paper quality and author category

For example, *famous authors may write stronger than average papers*

Negative Results

Ingredient 1: **Correlation** between paper quality and author category

For example, *famous authors may write stronger than average papers*

Ingredient 2: **Any of the following factors**

Negative Results

Ingredient 1: **Correlation** between paper quality and author category

For example, *famous authors may write stronger than average papers*

Ingredient 2: **Any of the following factors**

- Noise in reviews

Negative Results

Ingredient 1: **Correlation** between paper quality and author category

For example, *famous authors may write stronger than average papers*

Ingredient 2: **Any of the following factors**

- Noise in reviews
- Subjectivity of reviewers

Negative Results

Ingredient 1: **Correlation** between paper quality and author category

For example, *famous authors may write stronger than average papers*

Ingredient 2: **Any of the following factors**

- Noise in reviews
- Subjectivity of reviewers
- Reviewer miscalibration

Negative Results

Ingredient 1: **Correlation** between paper quality and author category

For example, *famous authors may write stronger than average papers*

Ingredient 2: **Any of the following factors**

- Noise in reviews
- Subjectivity of reviewers
- Reviewer miscalibration
- Standard bidding procedure

Negative Results

Ingredient 1: **Correlation** between paper quality and author category

For example, *famous authors may write stronger than average papers*

Ingredient 2: **Any of the following factors**

- Noise in reviews
- Subjectivity of reviewers
- Reviewer miscalibration
- Standard bidding procedure
- Popular paper-reviewer matching algo

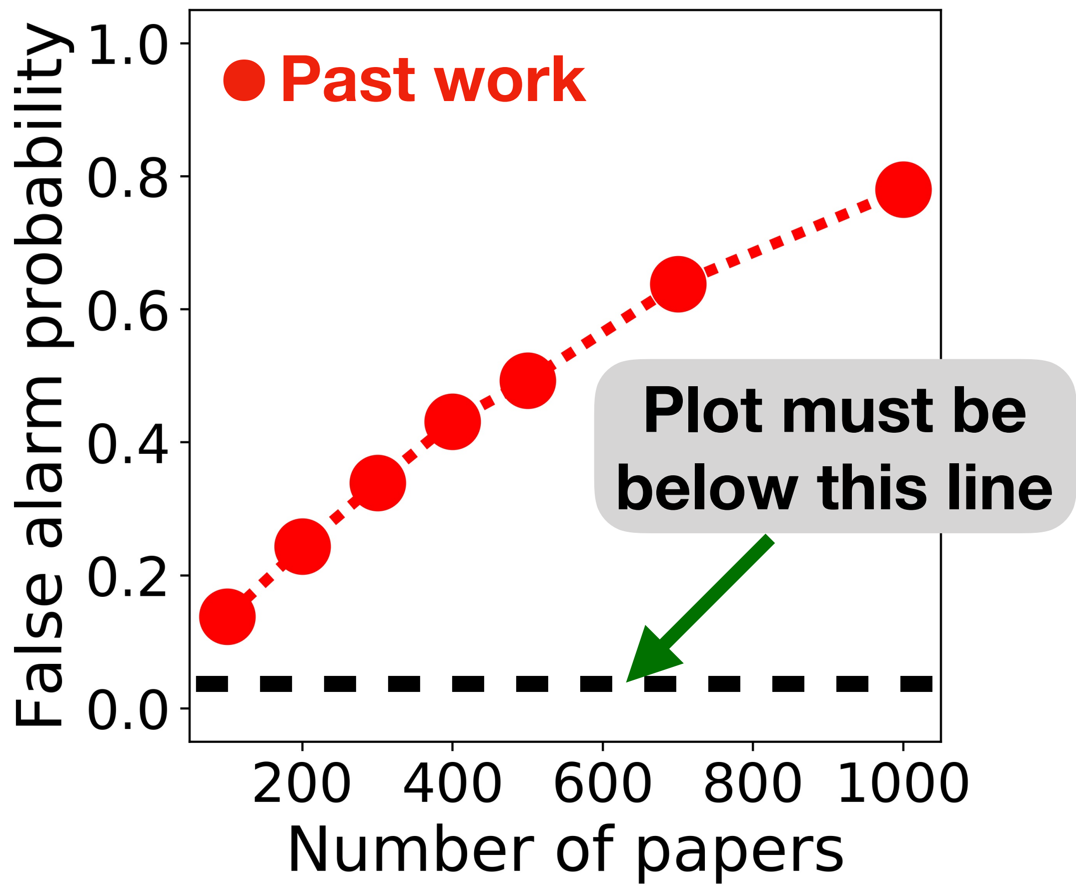
Negative Results

Ingredient 1: **Correlation** between paper quality and author category

For example, *famous authors may write stronger than average papers*

Ingredient 2: **Any of the following factors**

- Noise in reviews
- Subjectivity of reviewers
- Reviewer miscalibration
- Standard bidding procedure
- Popular paper-reviewer matching algo



Idiosyncrasies of peer review make testing difficult and break false alarm guarantees of the past work

Positive Results

Positive Results

Novel experimental setup

Minimal changes to the standard peer-review process. Accommodates **bidding** and **any paper-reviewer matching algo**

Positive Results

Novel experimental setup

Minimal changes to the standard peer-review process. Accommodates **bidding** and **any paper-reviewer matching algo**

Novel test for biases

Minimal assumptions on behaviour of reviewers. We **do not assume** absence of **noise**, **subjectivity** or **miscalibration**

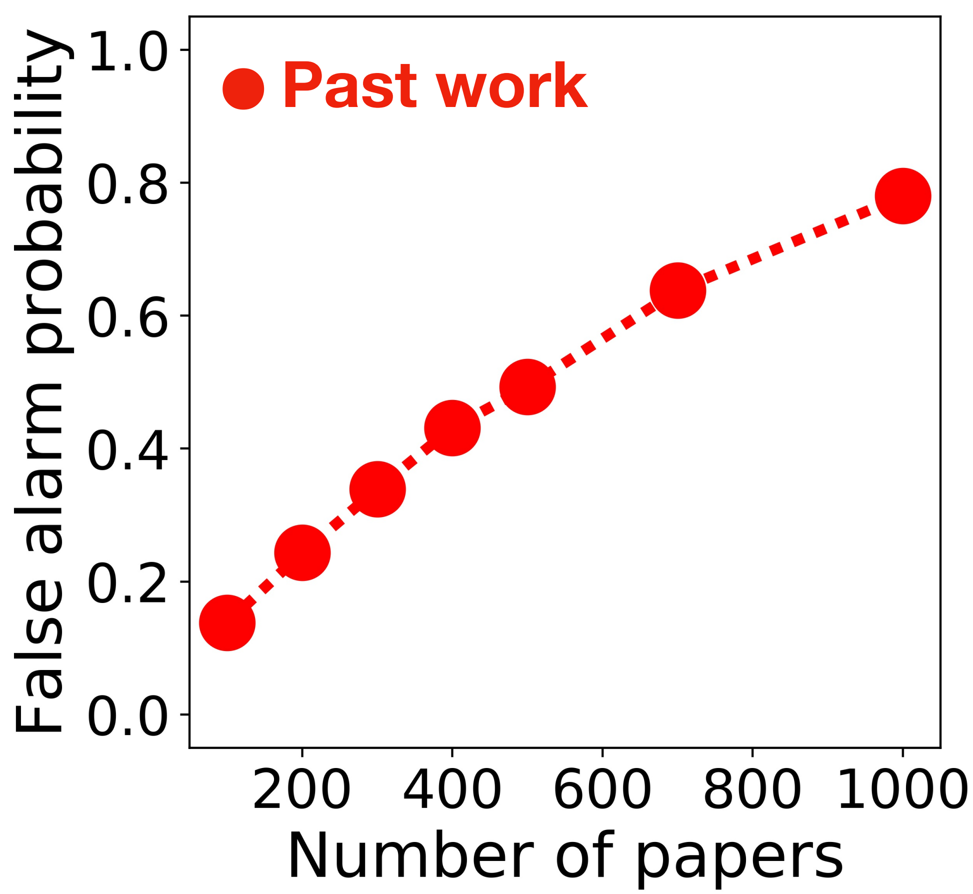
Positive Results

Novel experimental setup

Minimal changes to the standard peer-review process. Accommodates **bidding** and **any paper-reviewer matching algo**

Novel test for biases

Minimal assumptions on behaviour of reviewers. We **do not assume** absence of **noise**, **subjectivity** or **miscalibration**



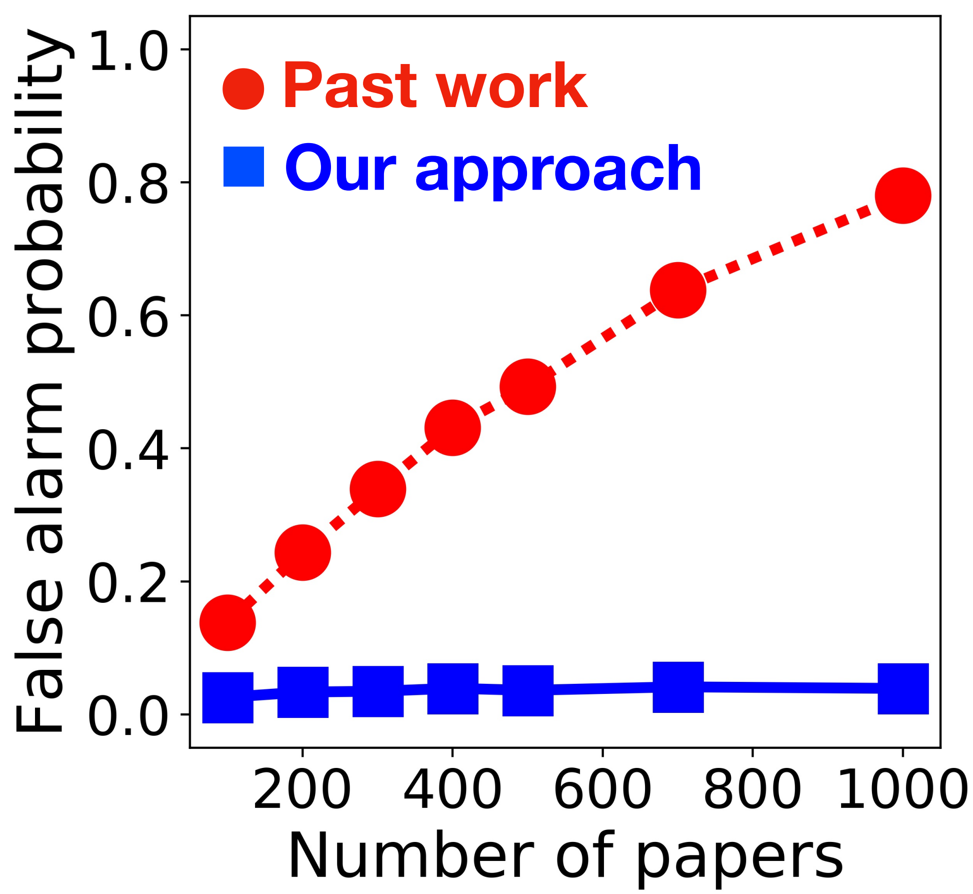
Positive Results

Novel experimental setup

Minimal changes to the standard peer-review process. Accommodates **bidding** and **any paper-reviewer matching algo**

Novel test for biases

Minimal assumptions on behaviour of reviewers. We **do not assume** absence of **noise**, **subjectivity** or **miscalibration**



We design a principled approach towards testing for biases with strong rigorous guarantees on false alarm control

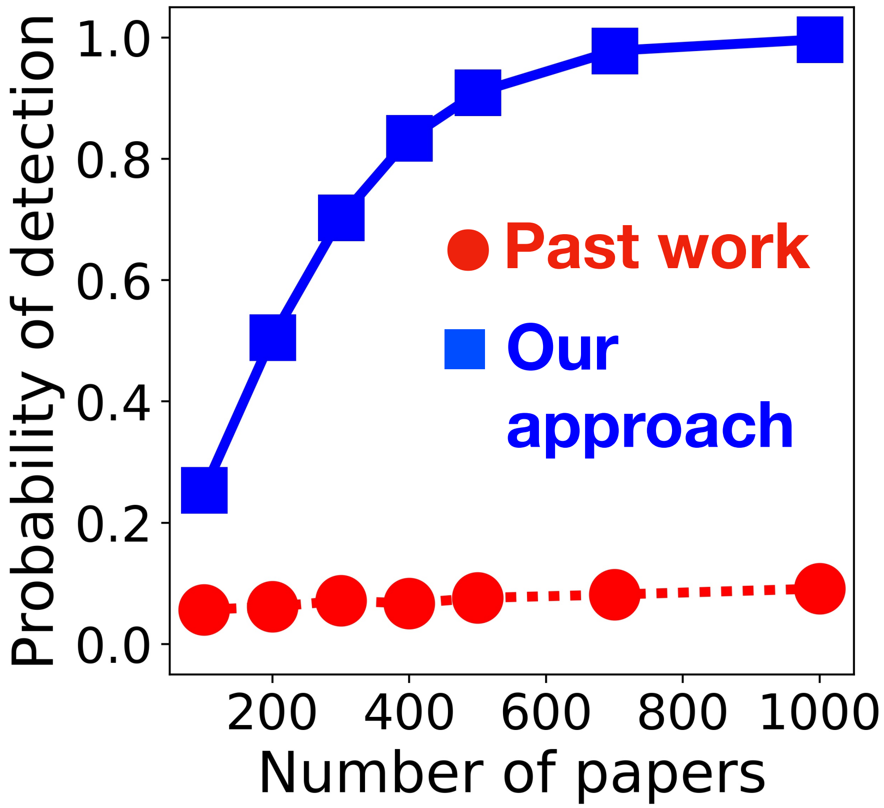
Correct Detection

Correct Detection

Our test also performs well in detecting the bias

Correct Detection

Our test also performs well in detecting the bias

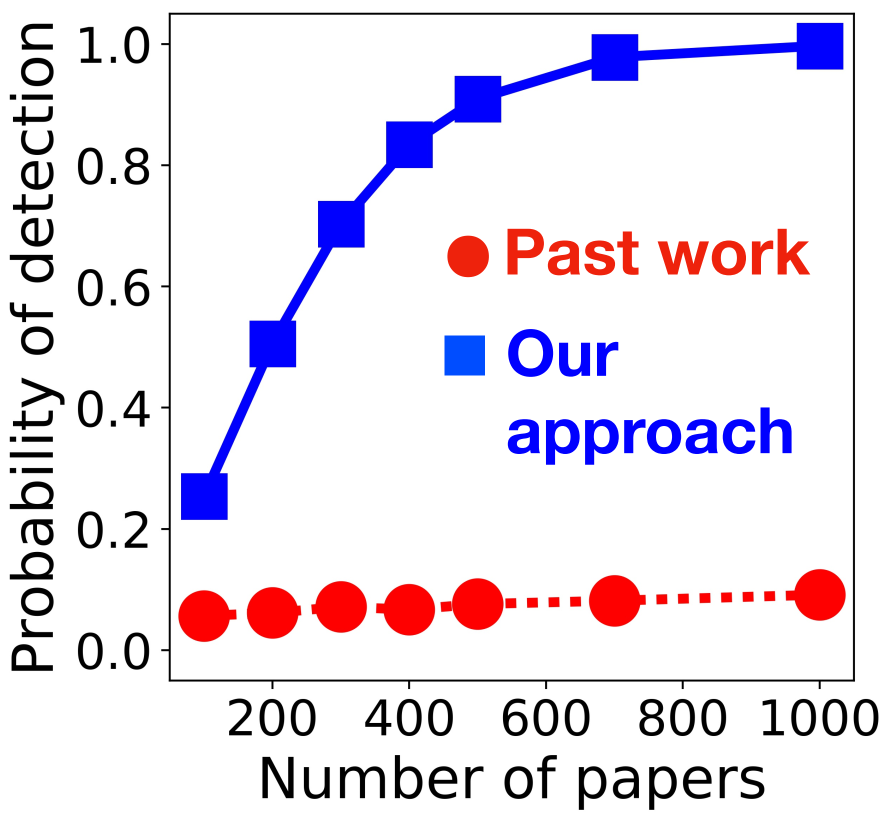


Correlation + noise

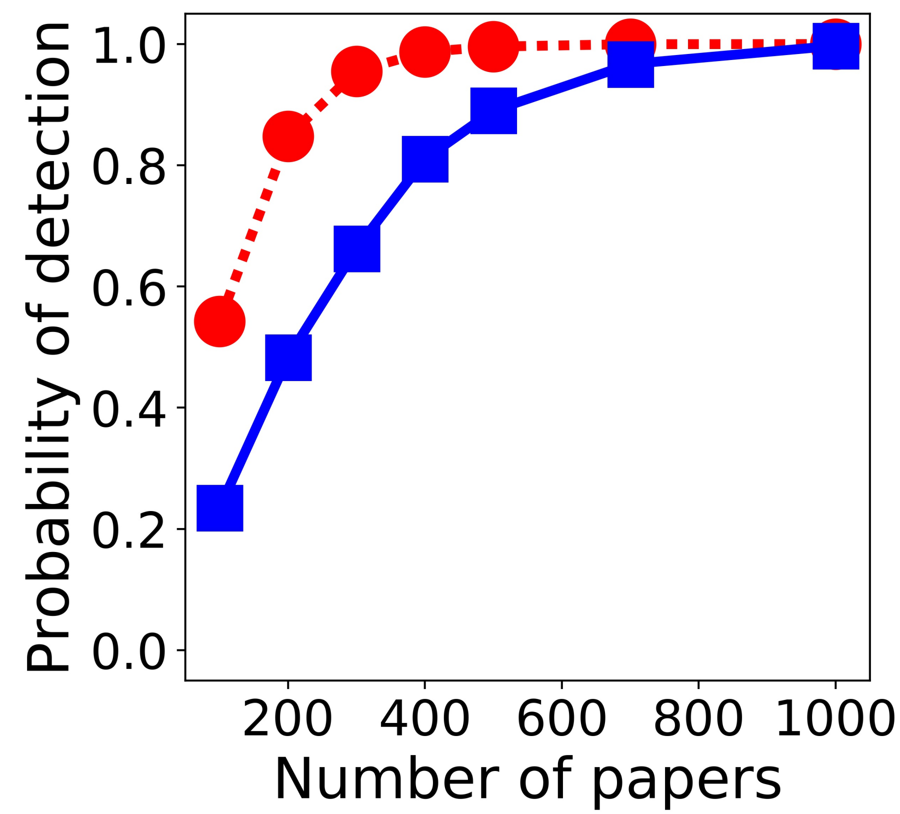
- Much higher probability of detection in «hard» cases where the past work fails

Correct Detection

Our test also performs well in detecting the bias



Correlation + noise



All assumptions of the past work are satisfied

- Much higher probability of detection in «hard» cases where the past work fails
- Not too much loss in power when the assumptions made in the past work are exactly met

Want to Know More?

Please come to the poster session!

5PM @ East Exhibition Hall B + C, #115