Making Scientific Peer Review Scientific

Ivan Stelmakh

August 2022
CMU-ML-22-105

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee
Nihar Shah, Co-chair
Aarti Singh, Co-chair
Larry Wasserman
Eric Horvitz (Microsoft)

Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy

This research was supported by the United States Department of the Interior award D17AP00001, the National Science Foundation awards CCF1563918 and CCF1763734, and the Office of Naval Research award N000142212181.

Keywords: peer review, fairness, accuracy, incentives, policy.

To my Pittsburgh friends without whom I would not complete this thesis in 5 years...
And to my Moscow friends without whom I would complete this thesis much earlier.

Abstract

Nowadays many important applications such as hiring, university admissions, and scientific peer review rely on the *collective* efforts of a large number of individuals. These applications often operate at an extremely large scale which creates both *opportunities* and *challenges*. On the *opportunity* side, the large amount of data generated in these applications enables a novel data science perspective on the classical problem of decision-making. On the *challenge* side, in many of these applications, human decision-makers need to interact with various interfaces and algorithms, and follow various policies. When not carefully designed, such interfaces, algorithms, and policies may lead to unintended consequences. Identifying and overcoming such unintended consequences is an important research problem. In this thesis, we explore these opportunities and tackle these challenges with a general goal of *understanding and improving distributed human decision-making in a principled manner*.

One application where the need for improvement is especially strong is scientific peer review. On the one hand, peer review is the backbone of academia, and scientific community agrees on the importance of improvement of the system. On the other hand, peer review is a microcosm of distributed decision-making that features a complex interplay between *noise*, *bias*, and *incentives*. Thus, insights learned from this specific domain apply to many other areas where similar problems arise. All in all, in this thesis, we aim at developing a principled approach towards scientific peer review—an important prerequisite for fair, equitable, and efficient progression of science.

The three broad challenges that arise in peer review are noise, bias, and incentives. In this thesis, we work on each of these challenges:

- **Noise and reviewer assignment.** A suitable choice of reviewers is a cornerstone of peer review: poor assignment of reviewers to submissions may result in a large amount of noise in decisions. Nowadays, the scale of many publication venues makes it infeasible to manually assign reviewers to submissions. Thus, stakeholders rely on algorithmic support to automate this task. Our work demonstrates that when such algorithmic support is not designed with application-specific constraints in mind, it can result in unintended consequences, compromising fairness and accuracy of the process. More importantly, we make progress in developing better algorithms by (i) designing an assignment algorithm with strong theoretical guarantees and reliable practical performance, and (ii) collecting a dataset that enables other researchers to develop better algorithms for estimating expertise of reviewers in reviewing submissions.

- **Bias and policies.** Human decision-making is susceptible to various biases, including identity-related biases (e.g., race and gender) and policy-related biases (e.g., primacy effect). To counteract these biases in peer review, it is crucial to design peer-review policies in an evidence-based manner. With this motivation, we conduct a series of real-world experiments to collect evidence that informs stakeholders in their policy decisions. Our work reveals that while some of the commonplace biases (e.g., herding) are not present in peer review, there are other application-specific biases (e.g., resubmission bias) that significantly impact decisions. Additionally, we demonstrate that reliable testing for biases in peer review often requires novel statistical tools as off-the-shelf techniques may result in false conclusions.

- **Incentives and reviewing.** Honesty is a core value of science and peer review is built on the assumption of honesty of everyone involved in the process. However, fierce competition in the academic job market and the large power a single reviewer has over an outcome of a submission create incentives for reviewers to consciously or subconsciously deviate from honest behavior. Our work offers (i) tools to test for such deviations, (ii) empirical evidence of the presence of wrong incentives, and (iii) potential solutions on how to incentivize reviewers to put more effort in writing high-quality reviews.

Contents

1 Introduction **1**

I Noise and Reviewer Assignment **8**

2 A Gold Standard Dataset for the Reviewer Assignment Problem **9**
1 Introduction . . . 9
2 Related literature . . . 10
3 Data collection pipeline . . . 11
4 Data exploration . . . 13
5 Experimental setup . . . 14
6 Results . . . 16
7 Discussion . . . 19

3 Fair and Accurate Reviewer Assignment **21**
1 Introduction . . . 21
2 Related literature . . . 22
3 Problem setting . . . 25
4 Reviewer assignment algorithm . . . 28
5 Approximation guarantees . . . 32
6 Objective-score model . . . 35
7 Subjective-score model . . . 40
8 Experiments . . . 42
9 Proofs . . . 51
10 Discussion . . . 62
Appendix . . . 63

II Bias and Policies **69**

4 Identity-Related Biases in Single-Blind Peer Review **70**
1 Introduction . . . 70
2 Preliminaries . . . 72
3 Problems with the past approach . . . 73
4 Novel framework to test for biases . . . 76
5 Proposed solution . . . 78
6 Analysis . . . 82
7 Discussion . . . 87
Appendix . . . 88

5 Identity-Related Biases in Double-Blind Peer Review **114**
1 Introduction . . . 114
2 Related work . . . 115
3 Methods . . . 116
4 Main results . . . 120
5 Discussion . . . 124
Appendix . . . 125

6 The Novice Reviewers' Bias against Resubmissions in Conference Peer Review **128**
1 Introduction . . . 128
2 Related literature . . . 130
3 Experimental setup . . . 131
4 Analysis of the experiment . . . 133
5 Discussion . . . 136

7 A Large Scale Randomized Controlled Trial on Herding in Peer-Review Discussions **140**
1 Introduction . . . 140
2 Methods . . . 141
3 Results of the experiment . . . 144
4 Discussion . . . 147
Appendix . . . 147

III Incentives and Reviewing **154**

8 Detecting Strategic Behaviour in Peer Assessment **155**
1 Introduction . . . 155
2 Related literature . . . 156
3 Problem formulation . . . 157
4 Testing procedure . . . 160
5 Experiment to elicit strategic behaviour . . . 162
6 Evaluation of the test . . . 165
7 Discussion . . . 167
Appendix . . . 167

9 A Study on Citation Bias in Peer Review **174**
1 Introduction . . . 174
2 Related literature . . . 175
3 Methods . . . 176
4 Results . . . 181
5 Discussion . . . 182
Appendix . . . 185

10 A Novice-Reviewer Experiment to Address Scarcity of Qualified Reviewers in Large Conferences **191**
1 Introduction . . . 191
2 Methodology . . . 193
3 Evaluation . . . 196
4 Discussion . . . 202
Appendix . . . 207

IV Conclusions **211**

11 Conclusions **212**

Chapter 1

Introduction

The life of an individual is determined by decisions and judgements made by the individual themselves and by other people. A long line of work in psychology (Asch, 1951; Tversky and Kahneman, 1974; Shafir et al., 1993; Gilovich et al., 2002), economics (Friedman and Savage, 1948; Keeney and Raiffa, 1976; Kahneman and Tversky, 1979; Bertrand and Mullainathan, 2004), and philosophy (where the studies of decision-making date back to Aristotle) scrutinizes the principles of individual human judgement and decision-making.

More recently, intensification of various processes—increased number of job-seeking applicants, growth in popularity of higher education, huge demand for carefully-annotated data—has resulted in an increased burden on decision-makers. Indeed, it is no longer possible for a single recruiter to evaluate all candidates, a single university admission officer cannot handle thousands of applications, and a single human cannot annotate even a moderately sized dataset. Therefore, nowadays in many real-life applications the decision-making is distributed across a large group of individuals who collectively work towards the common goal. Examples of such distributed decision-making include crowdsourcing, admission and hiring committees, and scholarly peer-review systems.

The amount of data generated in these distributed systems is often much larger than that generated when a single individual is responsible for all the decisions. Hence, the prevalence of such systems allows to tackle the problem of human decision-making from the perspective of data science, and, more generally, computer science. On the other hand, the process of collective decision-making depends crucially on the design of the system used to implement the process. In other words, decision-makers need to interact with various interfaces and algorithms, and the design of these interfaces and algorithms may have a direct impact on the quality of the final decisions. Thus, it is instrumental to use the arsenal of computer science to not only understand the properties of distributed decision-making systems but also to design these systems in a principled evidence-based manner. Thus, in this thesis we aim at pursuing the following broad research direction:

Understanding and principled design of distributed human decision-making systems

Scientific Peer Review

One of the most important applications that rely on distributed decision-making is scientific peer review—the backbone of academia (Smith, 2006; Price and Flach, 2017). Peer review is used to assess research work for competence, significance and originality, and employs experts working in the same field to conduct these evaluations (Brown, 2004; Bornmann, 2011). Across many fields of science, peer review is regarded as a tool to ensure high standards of published research (Mulligan et al., 2013) and improve the quality of research articles (Taylor and Francis group, 2015). Overall, most scientists agree that peer review is a crucial mechanism for scientific communications (Ware, 2016).

Beyond the scientific community, publication in a peer-reviewed venue is also interpreted as a quality-assurance sign by the media and general public (Smith, 2006). For instance, a single research article

published in a prestigious journal that claimed that vaccines may predispose to behavioral regression in children (Wakefield et al., 1998) received strong media coverage and resulted in a drop in acceptance of vaccination (DeStefano and Chen, 1999). Although the findings of that article were quickly refuted by the scientific community and the study was eventually retracted (Rao and Andrade, 2011), the claims made therein were causing vaccination fears for a long time (Larson et al., 2011). Thus, an important role of peer review is to ensure that research findings are not misinterpreted by society.

Finally, in addition to being the cornerstone for the dissemination of completed research, peer review nowadays plays a crucial role in shaping the directions of future research: it is used by funding bodies around the world (including US agencies NSF and NIH, and European Research Council) to distribute multi-billion dollar budgets through grants and awards. Thus, the review process should be able to identify the most promising research directions in order to spend taxpayers' money in the most effective way.

With all the roles peer review plays in the progression of science, it is extremely important to ensure that the peer-review process constitutes a "*mechanism for rational, fair, and objective decision-making*" (Jefferson et al., 2002). However, an observation made by Rennie (2016) in his Nature commentary indicates that there is a large gap between the current and ideal states of peer review:

> "*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.***"

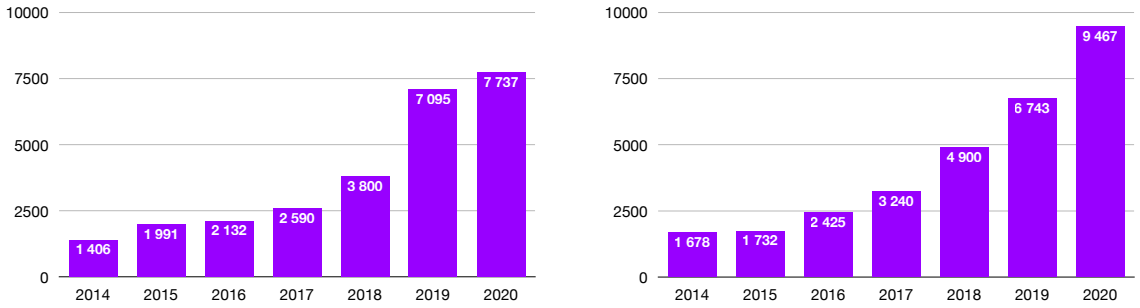
The opinion of Rennie is supported by anecdotal and empirical evidence that identifies various shortcomings of the review system: gender bias (Bernstein, 2015; Tomkins et al., 2017), strategic behavior (Anderson et al., 2007; Langford, 2008; Akst, 2010), wrong incentives (COPE, 2018; Van Noorden, 2020), and many others (see overview by Shah 2022). Importantly, in addition to the short-term impact on the outcome of a particular paper or a grant proposal (Thurner and Hanel, 2011), these shortcomings may have far-reaching consequences on the career trajectories of researchers due to the widespread prevalence of the rich-get-richer effect in academia (Merton, 1968; Triggle and Triggle, 2007; Squazzoni and Gandelli, 2012; Thorngate and Chowdhury, 2014).

Overall, peer review is an application that is very important to improve. Simultaneously, it features various challenging problems related to distributed decision-making, and insights learned from this specific domain apply to many other areas where similar problems arise. With this motivation, *we aim at developing a principled approach towards scientific peer review—an important prerequisite for fair, equitable, and efficient progression of science.* Specifically, we focus on three broad challenges that arise in peer review: noise, bias, and incentives.

Conference peer review The aforementioned problems with peer review are universal across various fields of science. However, peer review in computer science has been put under additional strain due to the nearly-exponential growth in the number of submissions received by leading venues (Figure 1). As a result, computer science conferences faced the urgent need for algorithmic support. Thus, for concreteness, in this thesis, we focus the discussion on computer science conferences which are considered to be at least as prestigious as top journals in the area and are frequently the terminal venue of publication. That said, we underscore that most of the tools, techniques, and insights we develop also apply to journal peer review.

Noise and Reviewer Assignment

Handling noise is one of the key challenges in learning from people (Shah, 2017; Kahneman et al., 2021). In peer review, noise manifests in erroneous evaluations of submissions under review. To minimize this noise, it is of utmost importance to assign papers to the right reviewers (Black et al., 1998; Thurner and Hanel, 2011; Bianchi and Squazzoni, 2015). Even a small fraction of incorrect reviews can have significant adverse effects on the quality of the published scientific standard (Thurner and Hanel, 2011) and dominate the



(a) AAAI (b) NeurIPS

Figure 1: Number of submissions to two flagship computer science conferences. The figure is adapted from the survey by Shah, 2022.

benefits yielded by the peer-review process that may have high standards otherwise (Squazzoni and Gandelli, 2012). Thus, the importance of the assignment stage of the peer-review process cannot be overestimated; quoting Rodriguez et al. (2007):

> *"One of the first and potentially most important stage is the one that attempts to distribute submitted manuscripts to competent referees."*

Historically, the allocation of papers to reviewers was manually conducted by journal editors or conference program chairs. However, given the massive scale of many conferences such as NeurIPS and AAAI, these assignments are largely performed in an automated manner. For instance, NeurIPS 2016 assigned 5 out of 6 reviewers per paper using an automated process (Shah et al., 2018). This problem of automated reviewer assignment in conferences forms the focus of the first part of this thesis.

Conceptually, the assignment problem consists of two stages:

- *Similarity computation* The key component of the assignment are *similarities*: for each (submission, reviewer) pair, a quantity that captures the competence of the reviewer in reviewing the submission needs to be defined in order to formalize the notion of the assignment quality.
- *Constrained optimization* Having similarities computed, the goal is to assign papers to referees, optimizing some notion of the assignment quality subject to certain load constraints.

Our work aims at developing a holistic approach towards the assignment stage. In that, Part I presents two contributions towards this broad goal that we now discuss in more detail.

Similarity computation In Chapter 2, we focus on the problem of similarity computation. While there are several algorithms developed for this problem (Charlin and Zemel, 2013; OpenReview, 2022), these algorithms have not been juxtaposed in a principled comparison. Consequently, three flagship computer science conferences—ICML, NeurIPS, and ACL—rely on three different algorithms to compute similarities. The key challenge towards comparing the existing algorithms is the lack of ground truth data and *we address this challenge by collecting a novel dataset of reviewers' expertise.* Our dataset comprises 58 computer science researchers, ranging from graduate students to senior professors, each of whom self-reported expertise in reviewing 5–10 papers they have read previously. We release the collected dataset and encourage researchers in natural language processing (NLP) and other areas to use this data and design more accurate algorithms for similarity computation. We also use this data to compare several popular algorithms currently used in practice and come up with evidence-based recommendations for stakeholders. This chapter is based on joint work with John Wieting, Graham Neubig, and Nihar Shah (forthcoming).

Constrained optimization In Chapter 3, we consider the problem of automated assignment of papers to reviewers when similarities are given. Specifically, we focus on a dual objective of fairness and statistical

accuracy. Our fairness objective is to maximize the review quality of the most disadvantaged paper, in contrast to the commonly used objective of maximizing the total quality over all papers. We design an assignment algorithm based on an incremental max-flow procedure that we prove is near-optimally fair. Our statistical accuracy objective is to ensure correct recovery of the papers that should be accepted. We provide a sharp minimax analysis of the accuracy of the peer-review process for a popular objective-score model as well as for a novel subjective-score model that we propose. Our analysis proves that our proposed assignment algorithm also leads to a near-optimal statistical accuracy. Finally, we design a novel experiment that allows for an objective comparison of various assignment algorithms, and overcomes the inherent difficulty posed by the absence of a ground truth in experiments on peer-review. The results of this experiment as well as of other experiments on synthetic and real data corroborate the theoretical guarantees of our algorithm. The algorithm we developed in this chapter was used in the assignment stage of the ICML 2020 conference and significantly outperformed the conventional algorithm in terms of our fairness objective while being competitive in terms of the objective of the conventional algorithm. This chapter is based on joint work with Nihar Shah, and Aarti Singh (Stelmakh et al., 2021a).

Bias and Policies

Bias is another characteristic of human thinking that can compromise the fairness and accuracy of decisions (Kahneman, 2011). To alleviate the impact of biases, it is extremely important to design policies and procedures of the peer-review process in a principled manner. As is the case for any scientific system, an essential mechanism for principled development of peer review is the feedback loop[1]. However, in peer review, policies established by the organizers are rarely evaluated in experiments due to logistical costs and other difficulties. Hence, the feedback loop in peer review is broken, resulting in the process being designed in an unscientific manner.

In Part II of this thesis, we describe a series of theoretical and experimental works that quantitatively investigate the presence of several biases in peer review and inform stakeholders in making policy decisions. In that, we focus on two types of biases:

- *Identity-related biases (Chapters 4 and 5)* Gender, race, age and other identity-related biases are prevalent in human decisions (Bendick et al., 1996; Bertrand and Mullainathan, 2004; Bendick and Nunes, 2011; Quillian et al., 2017). Several policies have been proposed to alleviate these biases in peer review, including hiding author identities from reviewers (double-blind peer review) and banning authors from posting their preprints online. Our work aims at helping organizers to estimate the efficacy of these policies in order to make informed decisions.
- *Policy-related biases (Chapters 6 and 7)* Throughout the review process, reviewers need to complete various tasks, interacting with different interfaces, framings, and policies. Research in psychology has demonstrated the importance of careful design of such interfaces, framings, and policies as otherwise they can bias human decisions (Tversky and Kahneman, 1974, 1981; Gilovich et al., 2002). Thus, we conduct real-world experiments to guide the principled design of the review process with policy-related biases in mind.

Let us now discuss the content of Part II in more detail.

Identity-related biases In Chapter 4, we contribute to a long-standing debate on whether exposing author identities to reviewers induces biases against certain groups, and our focus is on designing tests to detect the presence of such biases. Our starting point is a remarkable work by Tomkins, Zhang and Heavlin which conducted a controlled, large-scale experiment to investigate existence of biases in the peer reviewing of the WSDM conference. We present two sets of results: the first set of results is negative, and pertains to the statistical tests and the experimental setup used in the work of Tomkins et al. We show that the test employed therein does not guarantee control over false alarm probability and under correlations between relevant variables coupled with any of the following conditions, with high probability, can declare a presence of

[1]Feedback loop is the process of testing new ideas and adjusting them based on the outcome of the test.

bias when it is in fact absent: (a) measurement error, (b) model mismatch, (c) reviewer calibration. Moreover, we show that the setup of their experiment may itself inflate false alarm probability if (d) bidding is performed in non-blind manner or (e) popular reviewer assignment procedure is employed. Our second set of results is positive and is built around a novel approach to testing for biases that we propose. We present a general framework for testing for biases in (single vs. double blind) peer review. We then design hypothesis tests that under minimal assumptions guarantee control over false alarm probability and non-trivial power even under conditions (a)–(c) as well as propose an alternative experimental setup which mitigates issues (d) and (e). Finally, we show that no statistical test can improve over the non-parametric tests we consider in terms of the assumptions required to control for the false alarm probability. This chapter is based on joint work with Nihar Shah and Aarti Singh (Stelmakh et al., 2019).

While single-blind conferences debate whether they should switch to the double-blind mode or not, double-blind conferences have engaged in debates over whether to allow authors to post their papers online on arXiv or elsewhere during the review process. Independently, some authors of research papers face the dilemma of whether to put their papers on arXiv due to its pros and cons. In Chapter 5, we report the results of the study that substantiates this debate and dilemma via quantitative measurements. Specifically, we conducted surveys of reviewers in two top-tier double-blind computer science conferences—ICML 2021 (5361 submissions and 4699 reviewers) and EC 2021 (498 submissions and 190 reviewers). Our two main findings are as follows. First, more than a third of the reviewers self-report searching online for a paper they are assigned to review. Second, outside the review process, we find that preprints from better-ranked affiliations see a weakly higher visibility, with a correlation of 0.06 in ICML and 0.05 in EC. In particular, papers associated with the top-10-ranked affiliations had a visibility of approximately 11% in ICML and 22% in EC, whereas the remaining papers had a visibility of 7% and 18% respectively. This chapter is based on joint work with Charvi Rastogi, Xinwei Shen, Marina Meila, Federico Echenique, Shuchi Chawla, and Nihar Shah (Rastogi et al., 2022b).

Policy-related biases Moving to the policy-related biases, Chapter 6 is motivated by an observation that modern machine learning and computer science conferences are experiencing a surge in the number of submissions that challenges the quality of peer review as the number of competent reviewers is growing at a much slower rate. To curb this trend and reduce the burden on reviewers, several conferences have started encouraging or even requiring authors to declare the previous submission history of their papers. Such initiatives have been met with skepticism among authors, who raise the concern about a potential bias in reviewers' recommendations induced by this information. In this chapter, we investigate whether reviewers exhibit a bias caused by the knowledge that the submission under review was previously rejected at a similar venue, focusing on a population of novice reviewers who constitute a large fraction of the reviewer pool in leading machine learning and computer science conferences. We design and conduct a randomized controlled trial closely replicating the relevant components of the peer-review pipeline with 133 reviewers (master's, junior PhD students, and recent graduates of top US universities) writing reviews for 19 papers. The analysis reveals that reviewers indeed become negatively biased when they receive a signal about paper being a resubmission, giving almost 1 point lower overall score on a 10-point Likert item ($\Delta = -0.78,\ 95\%\ \text{CI} = [-1.30, -0.24]$) than reviewers who do not receive such a signal. Looking at specific criteria scores (originality, quality, clarity and significance), we observe that novice reviewers tend to underrate quality the most. This chapter is based on joint work with Nihar Shah, Aarti Singh, and Hal Daumé III (Stelmakh et al., 2021d).

Finally, in Chapter 7 we focus on the dynamics of discussions between reviewers and investigate the presence of herding behaviour therein. Specifically, we aim to understand whether reviewers and discussion chairs get disproportionately influenced by the first argument presented in the discussion when (in case of reviewers) they form an independent opinion about the paper before discussing it with others. In conjunction with the review process of ICML 2020, we design and execute a randomized controlled trial that involves 1,544 papers and 2,797 reviewers with the goal of testing for the conditional causal effect of the discussion initiator's opinion on the outcome of a paper. Our experiment reveals no evidence of herding in peer-review discussions. This observation is in contrast with past work that has documented an undue influence of the first piece of information on the final decision (e.g., anchoring effect) and analyzed herding behaviour in other applications (e.g., financial markets). Regarding policy implications, the absence of the herding effect

suggests that the current status quo of the absence of a unified policy towards discussion initiation does not result in an increased arbitrariness of the resulting decisions. This chapter is based on joint work with Charvi Rastogi, Nihar Shah, Aarti Singh, and Hal Daumé III (Stelmakh et al., 2020).

Incentives and Reviewing

Even when the process is optimally designed to handle noise and bias, the overall quality of decisions is contingent upon reviewers being honest and motivated to write high-quality reviews. While there is no doubt that most reviewers honestly invest their time and effort to advance science, large workloads (McCook, 2006) and strong competition in the academic job market (Alberts et al., 2014) may create wrong incentives for reviewers, leading to superficial reviewing (Teixeira da Silva and Al-Khatib, 2017) or even strategic behavior (Resnik et al., 2008; Langford, 2012a).

In Part III of the thesis, we focus on the problem of incentives and motivation of reviewers, considering two aspects:

- *Deviations from honest behavior (Chapters 8 and 9)* Empirical evidence documents various cases in which reviewers deviate from honest behavior to achieve personal benefits (Resnik et al., 2008; Fong and Wilhite, 2017; COPE, 2018). To help venue organizers in making informed decisions on what interventions need to be taken to minimize the impact of strategic behavior, we design tools and conduct real-world studies to quantify the presence of strategic behavior in peer review.
- *Quality of reviews (Chapter 10)* To keep up with the growth in the number of submissions (Figure 1), it is necessary to enlarge the pool of reviewers. The key challenge on this way is to not sacrifice the quality of reviews and our work investigates an interplay between motivation and review quality.

Let us now discuss the content of Part III in more detail.

Deviations from honest behavior In Chapter 8, we consider the issue of strategic behaviour in competitive peer review (e.g., when the number of accepted papers is predetermined and the sets of authors and reviewers coincide). In this setting, reviewers may be incentivized to misreport evaluations in order to improve their own final standing. Our focus is on designing methods for detection of such manipulations. Specifically, we consider a setting in which reviewers evaluate a subset of papers submitted to a conference and output rankings that are later aggregated to form a final ordering. In that, we investigate a statistical framework for this problem and design principled tests for detecting strategic behaviour. We prove that our tests have strong false alarm guarantees and evaluate their detection ability in practical settings. For this, we design and conduct an experiment that elicits strategic behaviour from subjects and release a dataset of patterns of strategic behaviour that may be of independent interest. We use the collected data to perform a series of real and semi-synthetic evaluations that reveal a strong detection power of our tests. This chapter is based on joint work with Nihar Shah and Aarti Singh (Stelmakh et al., 2021b).

In Chapter 9, we investigate incentives that are not related to the outcome of submissions authored by the reviewer. For this, we observe that citations play an important role in researchers' careers as a key factor in evaluation of scientific impact. Many anecdotes advice authors to exploit this fact and cite prospective reviewers to try obtaining a more positive evaluation for their submission. In this chapter, we investigate if such a *citation bias* actually exists: Does the citation of a reviewer's own work in a submission cause them to be positively biased towards the submission? In conjunction with the review process of two flagship conferences in machine learning and algorithmic economics, we execute an observational study to test for citation bias in peer review. In our analysis, we carefully account for various confounding factors such as paper quality and reviewer expertise, and apply different modeling techniques to alleviate concerns regarding the model mismatch. Overall, our analysis involves 1,314 papers and 1,717 reviewers and detects citation bias in both venues we consider. In terms of the effect size, by citing a reviewer's work, a submission has a non-trivial chance of getting a higher score from the reviewer: an expected increase in the score is approximately 0.23 on a 5-point Likert item. For reference, a one-point increase of a score by a single reviewer improves the position of a submission by 11% on average. This chapter is based on joint work with Charvi Rastogi, Ryan Liu, Shuchi Chawla, Federico Echenique, and Nihar Shah (Stelmakh et al., 2022).

Quality of reviews Chapter 10 is dedicated to reviewer recruiting with a focus on the scarcity of qualified reviewers in large conferences. Specifically, we design a procedure for (i) recruiting highly-motivated reviewers from the population not typically covered by major conferences and (ii) guiding them through the reviewing pipeline. In conjunction with ICML 2020 we recruit a small set of reviewers through our procedure and compare their performance with the general population of ICML reviewers. Our experiment reveals that a combination of the recruiting and guiding mechanisms allows for a principled enhancement of the reviewer pool and results in reviews of superior quality compared to the conventional pool of reviews as evaluated by senior members of the program committee (meta-reviewers). This chapter is based on joint work with Nihar Shah, Aarti Singh, and Hal Daumé III (Stelmakh et al., 2021c).

The discussion in this thesis is focused on peer review, but we note that the high-level insights and intuitions are also applicable to many other areas, including hiring, university admissions, and crowdsourcing. Three parts that follow below and chapters within these parts are written in a mostly independent fashion, allowing the reader to choose any order of reading them or to skip any of them without a significant loss in context.

Part I

Noise and Reviewer Assignment

Chapter 2

A Gold Standard Dataset for the Reviewer Assignment Problem

1 Introduction

Assignment stage is the most automated stage of the peer-review process. The key component of existing approaches to the submission-reviewer assignment is the notion of the *assignment score*:[1] a quantity that captures the expertise of the reviewer in reviewing the submission for each (submission, reviewer) pair. Several algorithms for computing assignment scores have been already proposed and used in practice. These algorithms rely on (i) textual content of submissions and reviewers' past papers, (ii) subject areas of submissions and reviewers, and (iii) other sources of information to estimate expertise of reviewers in reviewing submissions (we review these algorithms in Sections 2 and 5). However, the development of expertise-estimation algorithms has not been following the standard scientific path of iteratively improving the algorithms based on their practical performance. Instead, different existing algorithms are developed independently and are used in parallel without a clear notion of their relative performance: for example, three flagship machine learning conferences—ICML, NeurIPS, and ACL—rely on three different expertise-estimation algorithms.

The main hurdle towards the principled comparison of the expertise-estimation algorithms, and, more generally, towards the development of better algorithms is the absence of *gold standard data*. Indeed, the practical performance of any machine learning model depends heavily on the quality of data it is trained and evaluated on (Garbage in, garbage out, Babbage, 1864). However, there is no high-quality dataset of reviewers' expertise in reviewing submissions that is openly available to researchers. Moreover, peculiarities of the review process make it challenging to collect such a dataset: while data from actual review processes often contains self-reported evaluations (both ex-ante and ex-post) of expertise in reviewing submissions (Stelmakh et al., 2021c, 2022), this data (i) usually cannot be released without compromising the privacy of reviewers, and (ii) may be biased or noisy (see additional discussion in Section 2).

In this chapter, *we address this challenge and collect a dataset of reviewers' expertise that can facilitate the progress in the reviewer assignment problem.* Specifically, we conduct a survey of 58 computer science researchers whose experience level ranges from graduate students to senior professors. In the survey, we ask participants to report their expertise in reviewing 5-10 computer science papers they read over the last year.

Contributions Overall, our contributions are threefold:

- Fist, we collect and release a high-quality dataset of reviewers' expertise that can be used for training and/or evaluation of expertise estimation algorithms.
- Second, we use the collected dataset to compare existing expertise-estimation algorithms and inform organizers in making a principled choice for their venue.

[1] In this section, we use terms *assignment score* and *similarity* interchangeably.

- Third and finally, we conduct an exploratory analysis that highlights areas of improvement for existing algorithms and our insights can be used to develop better algorithms to improve peer review.

Let us now make two important remarks. First, while our dataset focuses on computer science, other communities may also use it to evaluate existing or develop new expertise-estimation algorithms. Indeed, to evaluate an algorithm from another domain on our data, it is sufficient to fine-tune the algorithm on profiles of computer science researchers crawled from semantic scholar and then evaluate it on our dataset.

Second, we underscore that the dataset we release is not set in stone. Instead, we release an initial version and encourage readers of this thesis to participate in our survey and contribute their data to the dataset. By collecting more samples, we enable more fine-grained comparisons and also improve the diversity of the dataset in terms of both population of participants and subject areas of papers. The survey is available at:

`https://forms.gle/rUV8hikwDRXZ3BTNA`

and we will be updating the released version regularly.

2 Related literature

In this section, we discuss relevant past studies. We begin with an overview of works that report comparisons of different expertise-estimation algorithms. We then provide a brief discussion of the procedure used in modern computer science conferences to compute assignment scores that are eventually used to allocate reviewers to submissions. Finally, we conclude with a list of works that design algorithms to automate other aspects of the reviewer assignment.

Evaluation of expertise-estimation algorithms OpenReview team (OpenReview, 2022) uses a heuristic approach to evaluate algorithms. In that, they consider papers authored by a number of researchers, remove one of these papers from the corpus, and predict expertise of each researcher in reviewing the selected paper. The performance of an algorithm then is measured as a fraction of times the author of the selected paper is predicted to be among the top reviewers for this paper. This heuristic, however, may lead to noisy results as algorithms that accurately predict the authorship relationship (and hence do well according to this approach) are not guaranteed to accurately estimate expertise in reviewing submissions authored by other researchers.

Rodriguez and Bollen (2008) and Anjum et al. (2019) rely on a different approach of querying expertise evaluations from reviewers and comparing predictions of the algorithms against these evaluations. In that, Rodriguez and Bollen (2008) rely on *ex-ante* bids—preferences of reviewers in reviewing submissions made in advance of reviewing. In contrast, Anjum et al. (2019) rely on *ex-post* evaluations of expertise made by reviewers after reviewing the submissions. Both of these works conduct small-scale evaluations to compare algorithms (Rodriguez and Bollen (2008) employ 102 papers and 69 reviewers, Anjum et al. (2019) employ 20 papers and 33 reviewers). However, both works use sensitive data that cannot be released without compromising the privacy of reviewers. Additionally, *ex-ante* evaluations of Rodriguez and Bollen (2008) may be not very accurate as (i) bids may contaminate expertise judgments with *willingness* to review submissions and (ii) bids are based on a very brief acquaintance with papers. On the other hand, *ex-post* data by Anjum et al. (2019) is collected for papers assigned to reviewers using a specific expertise-estimation algorithm. Thus, while collected evaluations have high precision, they may also have low recall if the employed expertise-estimation algorithm erroneously assigned low expertise scores to some (paper, reviewer) pairs as evaluations of expertise for such papers were not observed.

Mimno and McCallum (2007) use a clever idea to collect a dataset that can be released publicly. For this, they use 148 papers accepted to the NeurIPS 2006 conference and 364 reviewers from the NeurIPS 2005 conference and ask human annotators (independent established researchers) to evaluate expertise for a selected subset of 650 (paper, reviewer) pairs. While this approach results in a publicly available dataset, we note that external expertise judgments may also be noisy as judges may have incomplete information about the expertise of reviewers.

In this work, we aim at collecting a novel dataset of reviewer expertise that (i) can be released publicly and (ii) contains accurate self-evaluations of expertise that are based on a deep understanding of the paper and are not biased towards any existing expertise-estimation algorithm.

Assignment scores in conferences In modern conferences, assignment scores are typically computed by combining two types of input:

- *Initial automated estimates* First, an expertise-estimation algorithm is used to compute initial estimates. Many algorithms have been proposed for this task (Mimno and McCallum, 2007; Rodriguez and Bollen, 2008; Charlin and Zemel, 2013; Liu et al., 2014; Tran et al., 2017; Anjum et al., 2019; OpenReview, 2022) and we provide more details on several algorithms used in flagship computer science conferences in Section 5.
- *Human corrections* Second, automated estimates are corrected by reviewers who can read abstracts of submissions and report bids—preferences in reviewing the submissions. A number of works focus on the bidding stage and (i) explore the optimal strategy to assist reviewers in navigating the pool of thousands submissions (Fiez et al., 2020) or (ii) protect the system from strategic bids made by colluding reviewers willing to get assigned to each other's paper (Jecmen et al., 2020; Wu et al., 2021; Jecmen et al., 2022).

Combining these two types of input in a principled manner is a non-trivial task. As a result, different conferences use different strategies (Shah, 2022) and there is a lack of empirical or theoretical evidence that would guide venue organizers in their decisions.

Automation of the assignment stage At high level, automated assignment stage consists of two steps: first, assignment scores are computed; second, reviewers are allocated to submissions such that some notion of assignment quality (formulated in terms of the assignment scores) is maximized. In this chapter, we focus on the first step of the process. However, for completeness, we now mention several works that design algorithms for the second step.

A popular notion of assignment quality is the cumulative assignment score, that is, the sum of the assignment scores across all assigned reviewers and papers. An algorithm pursuing such an objective is implemented in the widely employed TPMS assignment algorithm (Charlin and Zemel, 2013) and similar ideas are explored in many papers (Goldsmith and Sloan, 2007; Tang et al., 2010; Long et al., 2013). While the cumulative objective is a natural choice, it has been observed that it can discriminate some submissions by assigning all irrelevant reviewers to a subset of submissions when a more balanced assignment exists (Garg et al., 2010). Thus, a number of works has explored the idea of assignment fairness, aiming at producing more balanced assignments (Kobren et al., 2019; Stelmakh et al., 2021a). Finally, other works explore the ideas of envy-freeness (Tan et al., 2021; Payan, 2022) and diversity (Li et al., 2015).

3 Data collection pipeline

In this section, we describe the process of data collection. Before we delve into details, we note that in this work we target computer science as the primary application area and tailor our data-collection process accordingly. We reiterate, however, that other communities may also use our dataset by fine-tuning their existing algorithms on profiles of computer science researchers and applying them to our dataset. In addition, we note that the data collection and release was performed under the approval of an institutional review board, with appropriate consent by the participants.

Gold standard We begin with a discussion of the gold standard data for computation of assignment scores. The gold standard dataset should satisfy two desiderata. First, it should comprise accurate evaluations of expertise of researchers in reviewing papers. In this work, we rely of self-evaluations of expertise. Thus, to collect high quality of data, it is important to ensure that researchers are familiar with papers for which they evaluate their expertise. Additionally, the dataset should be constructed such that it can be released publicly without disclosing any sensitive information.

Let us now discuss our approach to recruiting participants and obtaining accurate estimates of their expertise in reviewing papers included in the dataset.

Participant recruiting We recruited participants using a combination of several channels that are typically employed to recruit reviewers for computer science conferences:

- *Mailing lists* First, we sent recruiting emails to relevant mailing lists of several universities and research departments of companies.
- *Social media* Second, two authors of this work posted a call for participation on their Twitter accounts.
- *Personal communication* Third, we sent personal invites to researchers from the network of authors of this work.

To ensure that the pool of participants is limited to computer science researchers, we introduced a screening criterion requiring that prospective participants have at least one paper published in the broad area of computer science.

Overall, for the initial version of the dataset we release in this work, we managed to recruit 58 participants, all of whom passed the screening. Among the aforementioned three channels, personal communication ended up being the most successful and most of the participants joined the data-collection process after receiving the personalized request.

Expertise evaluations The key idea of our approach to expertise evaluation is to ask participants to *evaluate their expertise in reviewing papers they read in the past.* Indeed, after reading a paper, a researcher is at the best possible position to evaluate whether they have the right background—both in terms of the techniques used in the paper and in terms of the broader research area of the paper—to judge the quality of the paper. With this motivation, participants of the survey were asked to:

> *Recall 5-10 papers in their broad research area that they read to a reasonable extent in the last year and tell us their expertise in reviewing these papers.*

In more detail, the choice of papers was constrained by two minor conditions:

- The papers reported by a participant should not be authored by them
- The papers reported by a participant should be freely available online

In addition to these constraints, we gave several recommendations to the participants in order to make the dataset more diverse and useful for the research purposes:

- First, we asked participants to choose papers that cover the whole spectrum of their expertise with some papers being well-separated (e.g., very high expertise and very low expertise) and some papers being nearly-tied (e.g., two medium-expertise papers).
- Second, we recommended participants to avoid ties in their evaluations. To help participants comply with this recommendation, we implemented evaluation on a 1 to 5 scale with a 0.25 step size. Thus, participants were able to report papers with small differences in expertise.
- Third, we asked participants to come up with papers that they think may be tricky for existing expertise estimation algorithms. In that, we relied on the commonsense understandings and did not instruct participants on the inner-workings of these algorithms.

Overall, the time needed for participants to contribute to the dataset is estimated to be 5–10 minutes (one minute per paper). The interface of the survey is available at `https://forms.gle/rUV8hikwDRXZ3BTNA`.

Data release Following the procedure outlined above, we collected responses from 58 researchers. These responses constitute an initial version of the dataset that we release in this work. Each entry in the dataset corresponds to a participant and comprises evaluations of their expertise in reviewing papers of their choice. For each paper and each participant, we provide representations that are sufficient to start working on our dataset:

Total number of participants: 58

Characteristic	Quantity	Value
Gender	% Males	78
Affiliation	% Carnegie Mellon	40
	% Google	7
Country	% USA	74
Position	% PhD student	45
	% Faculty	28
	% Post-PhD (non-faculty)	12
Experience	Min # publications	2
	Max # publications	492
	Mean # publications	54

Table 1: Demography of participants. For the first four characteristics, quantities represent percentages of the one or more most popular classes in the dataset. Note that classification is done manually based on publicly available information and may not be errorless. For the last characteristic, quantities are computed based on Semantic Scholar profiles.

- *Participant* Each participant is represented by their Semantic Scholar ID, name, and complete bibliography crawled from Semantic Scholar on May 1, 2022.
- *Paper* Each paper, including papers from participants' bibliographies, is represented by its Semantic Scholar ID, title, abstract, list of authors, publication year, and arXiv identifier. Additionally, papers from participants' responses are supplied with links to publicly available PDFs.

4 Data exploration

In this section we explore the collected data and present various characteristics of the dataset. The next sections will then detail the results of using this data to benchmark various popular algorithms.

4.1 Participants

We begin with a discussion of the pool of the survey participants and Table 1 displays its key characteristics. First, we note that all participants work in the broad area of computer science and have a rich publication profile (at least two papers published, with the mean of 54 papers). In many subareas of computer science, including machine learning and artificial intelligence, having two papers is usually sufficient to join the reviewer pool of flagship conferences. Given that approximately 85% of participants either have PhD or are in the process of getting the degree, we conclude that most of the researchers who contributed to our dataset are eligible to review for computer science conferences.

Second, we caveat that most of the participants are male researchers affiliated with US-based organizations, with about 40% of all participants being affiliated with Carnegie Mellon University. Hence, the population of participants is not necessarily representative of the general population of the machine learning and computer science communities. We encourage researchers to be aware of this limitation when using our dataset. That said, we note that the data collection process does not finish with the publication of this work and we will be updating the dataset as new responses come. We also encourage readers to contribute 5–10 minutes of their time to fill out the survey `https://forms.gle/rUV8hikwDRXZ3BTNA` and make the dataset more representative.

Total number of papers: 463

Characteristic	Quantity	Value
Open Access	# Listed on semantic scholar	462
	# Listed on arXiv	411
	# PDF available publicly	457
Research Areas	# Computer science	459
Publication Year	% Before 2020	25
	% On or after 2020	75

Table 2: Characteristics of the papers in the released dataset. Most of the papers are available online and belong to the broad area of computer science.

4.2 Papers

Next, we describe the set of papers that constitute our dataset. Overall, participants evaluated their expertise in reviewing 463 unique papers. A total of 12 papers appeared in reports of two participants, 1 paper was mentioned by three participants, and the remaining papers were mentioned by one participant each.

Table 2 presents several characteristics of the pool of papers included to our dataset. First, we note that all but one of the papers are listed on Semantic Scholar, enabling expertise-estimation algorithms developed on the dataset to query additional information about the papers from the Semantic Scholar database. Additionally, most of the papers (99%) have their PDFs publicly available, thereby allowing algorithms to use full texts of papers to compute similarities.

Semantic Scholar has a built-in tool to identify research areas of the papers (Wade, 2022). According to this tool, 99% of the papers included to our dataset belong to the broad area of Computer Science—the target area for our data-collection procedure. The remaining four papers belong to the neighboring fields of mathematics, philosophy, and the computational branch of biology. To conclude, we note that approximately 75% of all papers in our dataset are published on or after 2020, ensuring that our dataset contains recent papers that expertise-estimation algorithms encounter in practice.

4.3 Evaluations of expertise

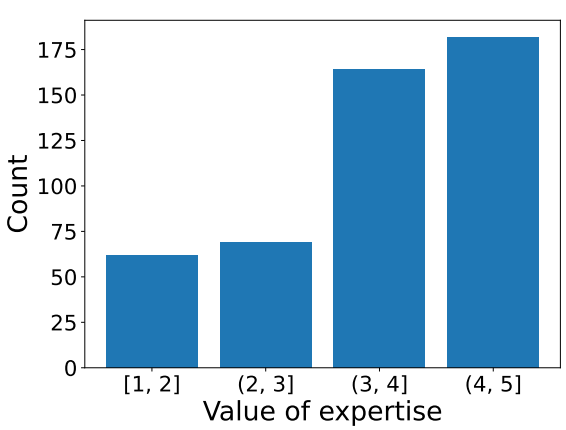
Finally, we proceed to the key aspect of our dataset—evaluations of expertise in reviewing the papers reported by participants. All but one participant reported expertise in reviewing at least 5 papers with the mean number of papers per participant being 8.2 and the total number of expertise evaluations being 477.

Figure 1 provides visualization of expertise evaluations made by participants. First, Figure 1a displays the histogram of expertise values. Observe that while a large fraction of reported papers belong to the expertise area of participants (expertise score larger than 3), about a third of evaluations are made for papers that reviewers are not competent in (expertise score 3 or lower).

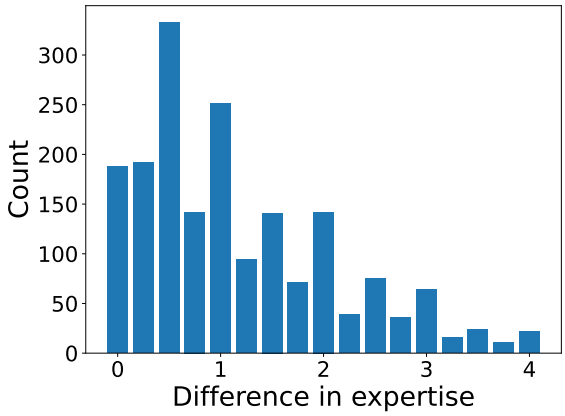
Second, Figure 1b shows the distributions of pairwise differences in expertise evaluations made by the same reviewer. To build this figure, for each participant we considered all pairs of papers in their report. Next, we pooled the absolute values of the pairwise differences in expertise across participants. We then plotted the histogram of these differences in the figure. Observe that the distribution in Figure 1b is heavy tailed, suggesting that our dataset is suitable for evaluating the accuracy of expertise-estimation methods both at a coarse level (large differences between the values of expertise) and fine level (small differences between the values of expertise).

5 Experimental setup

In this section we describe the setup of experiments on our dataset.



(a) Histogram of expertise values.



(b) Histogram of differences in expertise evaluations.

Figure 1: Distribution of expertise scores reported by participants.

5.1 Metric

We begin with defining a metric that we use in this work to evaluate performance of the algorithms. For this, we rely on the Kendall's Tau distance that is closely related to the widely used Kendall's Tau rank correlation coefficient (Kendall, 1938).

Specifically, consider any algorithm $\mathcal{A}$ that produces predictions s of reviewers' expertise in reviewing a given set of papers.[2] Each participant of our study r reported their expertise $\varepsilon \in \{1, 1.25, 1.5, \dots, 5\}$ in reviewing m_r papers: $\mathcal{P}_r = \{p_1, p_2, \dots, p_{m_r}\}$. The values of expertise $\mathcal{E}_r = \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_{m_r}\}$ induce a partial ordering over the set of papers $\mathcal{P}_r$. Similarly, predictions of expertise of participant r in reviewing papers $\mathcal{P}_r$ produced by the algorithm $\mathcal{S} = \{s_1, s_2, \dots, s_{m_r}\}$ induce another partial ordering over $\mathcal{P}_r$. Using these two partial orderings, we define the unnormalized loss of algorithm $\mathcal{A}$ with respect to participant r as follows:

$$L_r(\mathcal{A}) = \sum_{\substack{i,j=1 \\ i<j}}^{m_r} \left(\underbrace{\mathbb{I}\{(s_i - s_j) \times (\varepsilon_i - \varepsilon_j) < 0\}}_{\text{disagreement}} \times \big|\varepsilon_i - \varepsilon_j\big| \;+\; \underbrace{\mathbb{I}\{(s_i - s_j) \times (\varepsilon_i - \varepsilon_j) = 0\}}_{\text{tie}} \times \frac{1}{2}\big|\varepsilon_i - \varepsilon_j\big| \right).$$

In words, for each pair of papers (i, j) reported by participant r, the algorithm is not penalized when the ordering of papers induced by the predicted expertise $\{s_i, s_j\}$ agrees with the ground truth expertise-based ordering $\{\varepsilon_i, \varepsilon_j\}$. When two orderings disagree, the algorithm is penalized by the difference of expertise reported by the participant ($|\varepsilon_i - \varepsilon_j|$). Finally, when scores computed by the algorithm indicate a tie while expertise scores are different, the algorithm is penalized by half the difference in expertise ($1/2|\varepsilon_i - \varepsilon_j|$).

Having the unnormalized loss with respect to a participant defined, we now define the overall loss $L \in [0, 1]$ of the algorithm as follows:

$$L(\mathcal{A}) = \frac{\sum\limits_{r} L_r}{\sum\limits_{r} \sum\limits_{\substack{i,j=1 \\ i<j}}^{m_r} \big|\varepsilon_i - \varepsilon_j\big|}. \tag{2.1}$$

In words, we take the sum of unnormalized losses and normalize this sum by the loss achieved by the adversarial algorithm that reverses the ground-truth ordering of expertise (that is, sets $s = -\varepsilon$). Overall, our loss L takes values from 0 to 1 with lower values indicating better performance.

[2] Our metric is agnostic to the range of predicted values of expertise s as long as larger values indicate larger predicted expertise.

5.2 Algorithms

In this work, we evaluate several algorithms that we now discuss. All of these algorithms operate with (i) the list of submissions for which assignment scores need to be computed and (ii) reviewers' profiles that include past publications of reviewers. Conceptually, all algorithms predict reviewers' expertise by evaluating textual overlap between each submission and papers in each reviewer's publication profile. Let us now provide more detail on how this idea is implemented in each of the algorithms under consideration.

Trivial baseline First, we consider a trivial baseline that ignores the content of submissions and reviewers' profiles when computing the assignment scores: for each (participant, paper) pair, the TRIVIAL algorithm predicts the score s to be 1.

Toronto Paper Matching System (TPMS) The TPMS algorithm (Charlin and Zemel, 2013) is widely used by flagship conferences such as ICML and AAAI and is based on the TF-IDF similarities. While exact implementation is not publicly available, in our experiments we use an open-source implementation by Xu et al. (2019) which implements the basic TF-IDF logic of TPMS.

OpenReview algorithms OpenReview (`https://www.openreview.net`) is a conference-management system used by NeurIPS, ICLR, and many other machine learning conferences. It offers a family of algorithms for measuring expertise of reviewers in reviewing submissions. In this work, we evaluate the following algorithms:

- ELMO. This algorithm relies on general-purpose **E**mbeddings from **L**anguage **Mo**dels (Peters et al., 2018) to compute textual similarities between submissions and reviewers' past papers.
- SPECTER. This algorithm employs more specialized document-level embeddings of scientific documents (Cohan et al., 2020). Specifically, SPECTER explores the citation graph to construct embeddings that are useful for a variety downstream tasks, including similarity computation that we focus on in this work.
- SPECTER+MFR. Finally, SPECTER+MFR further enhances SPECTER. Instead of constructing a single embedding of each paper, it construct multiple embeddings that correspond to different facets of the paper. These embeddings are then used to compute the assignment scores.

We use implementations of these methods that are available on the OpenReview GitHub page[3] and execute them with default parameters (configuration files are available in supplementary materials).

6 Results

In this section, we report the results of evaluation of algorithms described in Section 5.2. First, we juxtapose all algorithms on our data (Section 6.1). Second, we use the TPMS algorithm to explore various aspects of the expertise-estimation problem (Section 6.2).

6.1 Comparison of the algorithms

Our first set of results compares the performance of the existing expertise-estimation algorithms. To run these algorithms on our data, we need to make some modeling choices faced by conference organizers in practice:

- *Paper representation.* First, in their inner-workings, expertise-estimation algorithms operate with some representation of the paper content. Possible choices of representations include: (i) title of the paper, (ii) title and abstract, and (iii) full text of the paper. We choose option (ii) as this option is often used in real conferences and is supported by all algorithms we consider in this work. Thus, to predict expertise,

[3] `https://github.com/openreview/openreview-expertise`

Algorithm	Loss	95% CI for Loss	Δ with TPMS	95% CI for Δ
Trivial	0.50	—	—	—
TPMS	0.28	[0.23, 0.33]	—	—
ELMo	0.34	[0.29, 0.40]	+0.06	[0.00, 0.13]
Specter	0.27	[0.21, 0.34]	−0.01	[−0.06, 0.04]
Specter+MFR	**0.24**	[0.18, 0.30]	−0.04	[−0.09, 0.01]

Table 3: Comparison of expertise-estimation algorithms on the collected data. All algorithms operate with reviewer profiles consisting of the 20 most recent papers and use titles and abstracts of papers. Lower values of loss are better. A positive (respectively, negative) value of Δ indicates that the algorithm performs worse (respectively, better) than TPMS.

algorithms are provided with the title and abstract of each paper (papers reported by participants and papers in their publication profiles).

- *Reviewer profiles.* The second important parameter is the choice of papers to include in reviewers' profiles. In real conferences, this choice is often left to reviewers who can manually select the papers they find representative of their expertise. In our experiments, we construct reviewer profiles automatically by using the 20 most recent papers from their Semantic Scholar profiles. If a reviewer has less than 20 papers published, we include all of them in their profile. Our choice of the reviewer profile size is governed by the observation that the mean length of the reviewer profile in NeurIPS 2022 is 16.5 papers. By setting the maximum number of papers to 20, we achieve the mean profile length of 14.8, thereby operating with the amount of information close to that available to algorithms in real conferences.

Statistical aspects To build reviewer profiles, we use publication years to order papers by recency, where we break ties uniformly at random. Thus, the content of reviewer profiles depends on randomness. To average this randomness out, we repeat the procedure of profile construction and similarity prediction 10 times, and report the mean loss over these iterations. That said, we note that the observed variability due to the randomness in the construction of reviewer profiles is negligible (standard deviation over all iterations is less than 0.005).

The pointwise performance estimates obtained by the procedure above depend on the selection of participants who contributed to our dataset. To quantify the associated level of uncertainty, we now compute 95% confidence intervals as follows. For 1,000 iterations, we create a new *reviewer pool* by sampling participants with replacement and recomputing the loss of each algorithm on the bootstrapped set of reviewers. To save computation time, we do not reconstruct reviewer profiles for each of these iterations as the uncertainty associated with the construction of reviewer profiles is small. Instead, we reuse profiles constructed to obtain pointwise estimates.

Finally, we build additional confidence intervals for the *difference* in the performance of the algorithms. Indeed, even when the losses of the algorithms fluctuate with the choice of the reviewer pool, the *relative difference* in performance of a pair of algorithms may be stable. To verify this intuition, we use the procedure above to build confidence intervals for the difference in performance between the TPMS algorithm and each of the OpenReview algorithms. TPMS is chosen as a baseline for this comparison due to its simplicity.

Results of the comparison Table 3 displays results of the comparison. The first pair of columns presents the loss of each algorithm on our dataset and the associated confidence intervals. The third and the forth columns investigate the relative difference in performance between the non-trivial algorithms. In that, the table displays the differences in performance between TPMS and each of the OpenReview algorithms together with the associated confidence intervals.

First, we note that all algorithms we consider in this work considerably outperform the TRIVIAL baseline, confirming that texts of papers are indeed useful in evaluating the expertise of researchers.

Second, comparing three algorithms from the OpenReview toolkit, we note that SPECTER+MFR and SPECTER outperform ELMO. The former two algorithms rely on domain-specific embeddings while ELMO uses general-purpose embeddings. Thus, the nature of the text similarity computation task in the academic context may be sufficiently different from that in other domains.

The third observation is the most surprising. TPMS algorithm is much simpler than other non-trivial algorithms: in contrast to ELMO, SPECTER, and SPECTER+MFR, it does not rely on carefully learned embeddings and can be efficiently executed on CPU. However, TPMS is competitive against complex SPECTER and SPECTER+MFR and even outperforms ELMO. While the performance of SPECTER and SPECTER+MFR algorithms could, in principle, be improved if these algorithms were additionally fine-tuned in a dataset-specific manner, the off-the-shelf performance of these algorithms is only marginally better than that of a much simpler TPMS algorithm.

To conclude the discussion of the results, we note that the confidence intervals for the performance of the algorithms, as well as for the relative differences, are wide. Thus, the observations we make in this section regarding the differences between TPMS, SPECTER, and SPECTER+MFR may not be statistically significant. It is, therefore, crucial to increase the size of our dataset to enable more principled comparison and identify more fine-grained differences between the algorithms.

6.2 The role of modeling choices

In the beginning of Section 6.1 we made two modeling choices pertaining to (i) representations of the papers provided to expertise-estimation algorithms and (ii) the size of reviewers' profiles used by these algorithms. In this section, we investigate these two questions in more detail.

- *Question 1 (paper representation).* Some expertise-estimation algorithms are designed to work with titles and/or abstracts of papers (e.g., SPECTER) while others can also incorporate the full texts of the manuscripts (e.g., TPMS). Consequently, there is a potential trade-off between accuracy and computation time. Indeed, richer representations are envisaged to result in higher accuracy, but are also associated with an increased demand for computational power. As with the choice of the algorithm itself, there is no guidance on what amount of information should be supplied to the algorithm as the gains from using more information are not quantified. With this motivation, our first question is:

  > *What are the benefits of providing richer representations of papers to expertise-estimation algorithms?*

- *Question 2 (reviewer profile).* The second important choice is the size of reviewers' profiles. On the one hand, by including only very recent papers in reviewers' profiles, conference organizers are at risk of not using enough data to obtain accurate values of expertise. On the other hand, old papers may not accurately represent the current expertise of researchers and hence may result in noise when used to compute expertise. Thus, our second question is:

  > *What is the optimal number of the most recent papers to include in the profiles of reviewers?*

To investigate these questions, we choose the TPMS algorithm as the workhorse to perform additional evaluations. We make this choice for two reasons. First, TPMS can work with all possible choices of the paper representation: title only, title and abstract, and full text of the paper. In contrast, other methods do not support the full-text mode. Second, TPMS is fast to execute, enabling us to compute expertise for hundreds of parameter configurations in a reasonable time.

Having the algorithm chosen, we vary the number of papers included in the reviewers' profiles from 1 to 20. For each value of the profile length, we consider three representations of the paper content: (i) title, (ii) title+abstract, and (iii) full text of the paper. Overall, for each combination of parameters, we construct reviewer profiles and predict similarities using the approach introduced in Section 6.1. The only exception is

Paper Representation	Loss	95% CI for Loss	Δ with Title+Abstract	95% CI for Δ
Title	0.34	[0.29, 0.41]	+0.07	[0.03, 0.12]
Title+Abstract	0.27	[0.22, 0.33]	—	—
Full Text	0.24	[0.18, 0.30]	−0.03	[−0.09, 0.03]

Table 4: Performance of the TPMS algorithm with 10 most recent papers included in reviewers' profiles and with different choices of the paper representation. Lower values of loss are better. A positive (respectively, negative) value of Δ indicates that the corresponding choice of the paper representation leads to a weaker (respectively, stronger) performance.

that we repeat the procedure for averaging out the randomness in the profile creation 5 times (instead of 10) to save the computation time.

Papers and reviewer profiles Before presenting the results, we note that in this section, we conduct all evaluations focusing on papers that have PDFs publicly available. In that, we remove 6 papers from the dataset as they are not available online (see Table 2). Similarly, we limit reviewer profiles to papers whose semantic scholar profiles contain links to arXiv. One of the participants did not have any such papers and we also exclude them from the dataset.

Results of the additional evaluations Figure 2 and Table 4 display the results of the additional evaluations. In that, Figure 2 displays the pointwise loss of the TPMS algorithm for each choice of parameters. To save computation time, we do not build confidence intervals for each combination of parameters. Instead, Table 4 sets the number of papers in reviewers' profiles to 10 and presents confidence intervals for losses incurred by the algorithm under different choices of paper representations. Let us now make two observations.

First, paper abstracts are very useful in improving the quality of expertise prediction as compared to titles alone. That said, adding full texts of the papers does not result in a strong increase in performance. Overall, the choice of title and abstract to represent the content of a paper may be sufficient in practice, balancing accuracy with computational efficiency as handling the full texts of submissions may require significant additional resources with only a marginal gain in accuracy.

Second, the loss curves plateau once reviewer profiles include 8 or more of their most recent papers. Additional increase of the profile length does not impact the quality of predictions. Thus, in practice, reviewers may be instructed to include 10 representative papers to their profile which for most of the active researchers amounts to the number of papers published in 1-3 years.

7 Discussion

In this work, we collect a novel dataset of reviewers' expertise in reviewing papers. In contrast to datasets collected in previous works, our dataset (i) can be released publicly, and (ii) contains evaluations of expertise made by scientists who have actually read the papers for their own research purposes. We use this dataset to juxtapose several existing expertise-estimation algorithms and help venue organizers in choosing an algorithm in an evidence-based manner.

We emphasize again that the dataset we release in this work is just an initial version and we keep the data-collection process open to increase the sample size of the dataset and make it more representative. Thus, we encourage readers of this thesis to contribute 5–10 minutes of their time and report their expertise in reviewing papers:

`https://forms.gle/rUV8hikwDRXZ3BTNA`

Second, we note that the difference in performance between the simple TPMS algorithm and more advanced Specter and Specter+MFR algorithms is quite small. An important continuation of the present

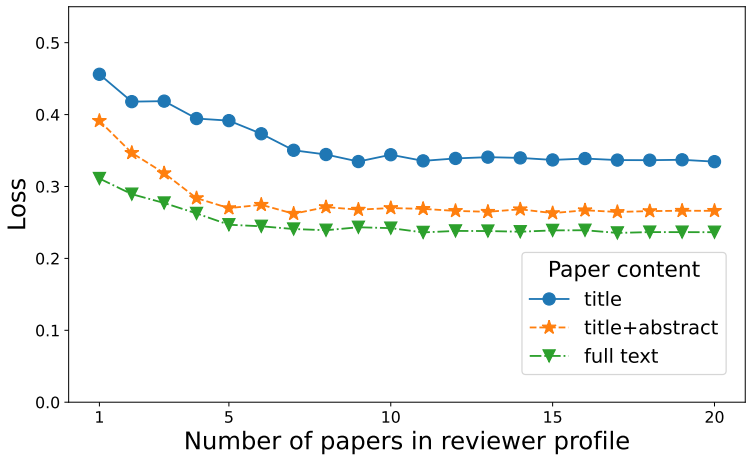


Figure 2: Impact of different choices of parameters on the quality of predicted similarities. Confidence intervals are not shown (see Table 4).

work would be to investigate this phenomenon and evaluate whether the performance of SPECTER and SPECTER+MFR significantly improves if these algorithms are fine-tuned on our dataset.

Finally, our dataset can be used to develop new expertise-estimation algorithms. Thus, we encourage researchers from the natural language processing and other communities to use our data in order to improve peer review.

Chapter 3

Fair and Accurate Reviewer Assignment

1 Introduction

In this chapter, we consider a problem of assigning papers to reviewers when similarities between reviewers and papers are given. Specifically, we study this problem with a dual goal of fairness and accuracy in mind. By fairness, we specifically consider the notion of max-min fairness which is studied in various branches of science and engineering (Rawls, 1971; Lenstra et al., 1990; Hahne, 1991; Lavi et al., 2003; Bonald et al., 2006; Asadpour and Saberi, 2010). In our context of reviewer assignments, max-min fairness posits maximizing the review-quality of the paper with the least qualified reviewers. The max-min fair assignment guarantees that no paper is discriminated against in favor of more lucky counterparts. That is, even the most ambivalent paper with a small number of reviewers being competent enough to evaluate its merits will receive as good treatment as possible. The max-min fair assignment also ensures that in *any other assignment* there exists at least one paper with the fate at least as bad as the fate of the most disadvantaged paper in the aforementioned fair assignment.

Alongside, we also consider the requirement of statistical accuracy. One of the main goals of the conference peer-review process is to select the set of "top" papers for acceptance. Two key challenges towards this goal are to handle the noise in the reviews and subjective opinions of the reviewers; we accommodate these aspects in terms of existing (Ge et al., 2013; McGlohon et al., 2010; Dai et al., 2012) and novel statistical models of reviewer behavior. Prior works on the reviewer assignment problem (Long et al., 2013; Garg et al., 2010; Karimzadehgan et al., 2008; Tang et al., 2010) offer a variety of algorithms that optimize the assignment for certain deterministic objectives, but do not study their assignments from the lens of statistical accuracy. In contrast, our goal is to design an assignment algorithm that can simultaneously achieve both the desired objectives of fairness and statistical accuracy.

We make several contributions towards this problem. We first present a novel algorithm, which we call PeerReview4All, for assigning reviewers to papers. Our algorithm is based on a construction of multiple candidate assignments, each of which is obtained via an incremental execution of max-flow algorithm on a carefully designed flow network. These assignments cater to different structural properties of the similarities and a judicious choice between them provides the algorithm appealing properties.

Our second contribution is an analysis of the fairness objective that our PeerReview4All algorithm can achieve. We show that our algorithm is optimal, up to a constant factor, in terms of the max-min fairness objective. Furthermore, our algorithm can adapt to the underlying structure of the given similarity data between reviewers and papers and in various cases yield better guarantees including the exact optimal solution in certain scenarios. Finally, after optimizing the outcome for the most worst-off paper and fixing the assignment for that paper, our algorithm aims at finding the most fair assignment for the next worst-off paper and proceeds in this manner until the assignment for each paper is fixed.

As a third contribution, we show that our PeerReview4All algorithm results in strong statistical guarantees in terms of correctly identifying the top papers that should be accepted. We consider a popular statistical model (Ge et al., 2013; McGlohon et al., 2010; Dai et al., 2012) which assumes existence of some true objective score for every paper. We provide a sharp analysis of the minimax risk in terms of "incorrect" accept/reject decisions, and show that our PeerReview4All algorithm leads to a near-optimal solution.

Fourth, noting that paper evaluations are typically subjective (Kerr et al., 1977; Mahoney, 1977; Ernst and Resch, 1994; Bakanic et al., 1987; Lamont, 2009), we propose a novel statistical model capturing subjective opinions of reviewers, which may be of independent interest. We provide a sharp minimax analysis under this subjective setting and prove that our assignment algorithm PeerReview4All is also near-optimal for this subjective-score setting.

Our fifth and final contribution comprises empirical evaluations. We designed and conducted an experiment on the Amazon Mechanical Turk crowdsourcing platform to objectively compare the performance of different reviewer-assignment algorithms. The design of the experiment is done carefully to circumvent the challenge posed by the absence of a ground truth in peer review settings, so that we can evaluate accuracy objectively. In addition to the MTurk experiment, we provide an extensive evaluation of our algorithm on synthetic data, provide an evaluation on a reconstructed similarity matrix from the ICLR 2018 conference, and report the results of the experiment on real conference data conducted by Kobren et al. (2019). The results of these experiments highlight the promise of PeerReview4All in practice, in addition to the theoretical benefits discussed elsewhere in the paper. The dataset pertaining to the MTurk experiment, as well as the code for our PeerReview4All algorithm, are available on the website of the author of this thesis.

The remainder of this chapter is organized as follows. We discuss related literature in Section 2. In Section 3, we present the problem setting formally with a focus on the objective of fairness. In Section 4 we present our PeerReview4All algorithm. We establish deterministic approximation guarantees on the fairness of our PeerReview4All algorithm in Section 5. We analyze the accuracy of our PeerReview4All algorithm under an objective-score model in Section 6, and introduce and analyze a subjective score model in Section 7. We empirically evaluate the algorithm in Section 8 using synthetic and real-world experiments. We then provide the proofs of all the results in Section 9. We conclude the chapter with a discussion in Section 10.

2 Related literature

The reviewer assignment process consists of two steps. First, a "similarity" between every (paper, reviewer) pair that captures the competence of the reviewer for that paper is computed. These similarities are computed based on various factors such as the text of the submitted paper, previous papers authored by reviewers, reviewers' bids and other features. Second, given the notion of good assignment, specified by the program chairs, papers are allocated to reviewers, subject to constraints on paper/reviewer loads. This work focuses on the second step (assignment), assuming the first step of computing similarities as a black box. In this section, we give a brief overview of the past literature on both of the steps of the reviewer-assignment process.

Computing similarities. The problem of identifying similarities between papers and reviewers is well-studied in data mining community. For example, Mimno and McCallum (2007) introduce a novel topic model to predict reviewers' expertise. Liu et al. (2014) use the random walk with restarts model to incorporate both expertise of reviewers and their authority in the final similarities. Co-authorship graphs (Rodriguez and Bollen, 2008) and more general bibliographic graph-based data models (Tran et al., 2017) give appealing methods which do not require a set of reviewers to be pre-determined by conference chair. Instead, these methods recommend reviewers to be recruited, which might be particularly useful for journal editors.

One of the most widely used automated assignment algorithms today is the Toronto Paper Matching System or TPMS (Charlin and Zemel, 2013) which also computes estimations of similarities between submitted papers and available reviewers using techniques in natural language processing. These scores might be enhanced with reviewers' self-accessed expertise adaptively queried from them in an automatic manner.

Our work uses these similarities as an input for our assignment algorithm, and considers the computation of these similarity values as a given black box.

Cumulative goal functions. With the given similarities, much of past work on reviewer assignments develop algorithms to maximize the cumulative similarity, that is, the sum of the similarities across all assigned reviewers and all papers. Such an objective is pursued by the organizers of SIGKDD conference (Flach et al., 2010) and by the widely employed TPMS assignment algorithm (Charlin and Zemel, 2013). Various other popular conference management systems such as EasyChair (`easychair.org`) and HotCRP (`hotcrp.com`) and several other papers (see Long et al. 2013; Charlin et al. 2012; Goldsmith and Sloan 2007; Tang et al. 2010 and references therein) also aim to maximize various cumulative functionals in their automated reviewer assignment procedures. In the sequel, we argue however that optimizing such cumulative objectives is not fair — in order to maximize them, these algorithms may discriminate against some subset of papers. Moreover, it is the non-mainstream submissions that are most likely to be discriminated against. With this motivation, we consider a notion of fairness instead.

Fairness. In order to ensure that no papers are discriminated against, we aim at finding a *fair assignment* — an assignment that ensures that the most disadvantaged paper gets as competent reviewers as possible. The issue of fairness is partially tackled by Hartvigsen et al. (1999), where they necessitate every paper to have at least one reviewer with expertise higher than certain threshold, and then maximize the value of that threshold. However, this improvement only partially solves the issue of discrimination of some papers: having assigned one strong reviewer to each paper, the algorithm may still discriminate against some papers while assigning remaining reviewers. Given that nowadays large conferences such as NeurIPS and ICML assign 4-6 reviewers to each paper, a careful assessment of the paper by one strong reviewer might be lost in the noise induced by the remaining weak reviews. In the present study, we measure the quality of assignment with respect to any particular paper as sum similarity over reviewers assigned to that paper. Thus, the fairness of assignment is the minimum sum similarity across all papers; we call an assignment fair if it maximizes the fairness. We note that assignment computed by our PeerReview4All algorithm is guaranteed to have *at least as large* max-min fairness as that proposed by Hartvigsen et al. (1999).

Benferhat and Lang (2001) discuss different approaches to selection of the "optimal" reviewer assignment. Together with considering a cumulative objective, they also note that one may define the optimal assignment as an assignment that minimizes a disutility of the most disadvantaged reviewer (paper). This approach resembles the notion of max-min fairness we study in this chapter, but Benferhat and Lang (2001) do not propose any algorithm for computing the fair assignment.

The notion of max-min fairness was formally studied in context of peer-review by Garg et al. (2010). While studying a similar objective, our work develops both conceptual and theoretical novelties which we highlight here. First, Garg et al. (2010) measure the fairness in terms of reviewers' bids — for every reviewer they compute a value of papers assigned to that reviewer based on her/his bids and maximize the minimum value across all reviewers. While satisfying reviewers is a useful practice, we consider fairness towards the papers in their review to be of utmost importance. During a bidding process reviewers have limited time resources and/or limited access to papers' content to evaluate their relevance, and hence reviewers' bids alone are not a good proxy towards the measure of fairness. In contrast, in this work we consider similarities — scores that are designed to represent a competence of reviewer in assessing a paper. Besides reviewers' bids, similarities are computed based on the full text of the submissions and papers authored by reviewer and can additionally incorporate various factors such as quality of previous reviews, experience of reviewer and other features that cannot be self-assessed by reviewers.

The assignment algorithm proposed in Garg et al. (2010) works in two steps. In the first step, the problem is set up as an integer programming problem and a linear programming relaxation is solved. The second step involves a carefully designed rounding procedure that returns a valid assignment. The algorithm is guaranteed to recover an assignment whose fairness is within a certain additive factor from the best possible assignment. However, the fairness guarantees provided in Garg et al. (2010) turn out to be vacuous for various similarity matrices. As we discuss later in the paper, this is a drawback of the algorithm itself and not an artifact of their guarantees. In contrast, we design an algorithm with multiplicative approximation factor that is guaranteed to always provide a non-trivial approximation which is at most constant factor away from the optimal.

Next, Garg et al. (2010) consider fairness of the assignment as an eventual metric of the assignment quality. However, we note that the main goal of the conference paper reviewing process is an accurate acceptance of

the best papers. Thus, in the present work we both theoretically and empirically study the impact of the fairness of the assignment on the quality of the acceptance procedure.

Finally, although Garg et al. (2010) present their algorithm for the case of discrete reviewer's bids, we note that this assumption can be relaxed to allow real-valued similarities with a continuous range as in our setting. In this work we refer to the corresponding extension of their algorithm as the Integer Linear Programming Relaxation (ILPR) algorithm.

Fair division. A direction of research that is relevant to our work studies the problem of fair division where max-min fairness is extensively developed. The seminal work of Lenstra et al. (1990) provides a constant factor approximation to the minimum makespan scheduling problem where the goal is to assign a number of jobs to the unrelated parallel machines such that the maximal running time is minimized. Recently Asadpour and Saberi (2010); Bansal and Sviridenko (2006) proposed approximation algorithms for the problem of assigning a number of indivisible goods to several people such that the least happy person is as happy as possible. However, we note that techniques developed in these papers cannot be directly applied for reviewer assignments problem in peer review due to the various idiosyncratic constraints of this problem. In contrast to the classical formulation studied in these works, our problem setting requires each paper to be reviewed by a fixed number of reviewers and additionally has constraints on reviewers' loads. Such constraints allow us to achieve an approximation guarantee that is independent of the total number of papers and reviewers, and depends only on λ, the number of reviewers required per paper, as $\frac{1}{\lambda}$. In contrast, the approximation factor of Asadpour and Saberi (2010) gets worse at a rate of $\frac{1}{\sqrt{m}\log^3 m}$, where m is a number of persons (papers in our setting).

Statistical aspects. Different statistical aspects related to conference peer-review have been studied in the literature. McGlohon et al. (2010) and Dai et al. (2012) studied aggregation of consumers ratings to generate a ranking of restaurants or merchants. They come up with objective score model of reviewer which we also use in this work. Ge et al. (2013) also use similar model of reviewer and propose a Bayesian approach to calibrating reviewer' scores, which allows to incorporate different biases in context of conference peer-review. Sajjadi et al. (2016) empirically compare different methods of score aggregation for peer grading of homeworks. Peer grading is a related problem to conference peer review, with the key difference that the questions and answers ("papers") are more closed-ended and objective. They conclude that although more sophisticated methods are praised in the literature, the simple averaging algorithm demonstrates better performance in their experiment. Another interesting observation they make is an edge of cardinal grades over ordinal in their setup. In this work we also consider the conferences with cardinal grading scheme of submissions.

To the best of our knowledge, no prior works on conference peer-review has studied the entire pipeline — from assignment to acceptance — from a statistical point of view. In this work we take the first steps to close this gap and provide a strong minimax analysis of naïve yet interesting procedure of determining top k papers. Our findings suggest that higher fairness of the assignment leads to better quality of acceptance procedure. We consider both the objective score model (Ge et al., 2013; McGlohon et al., 2010; Dai et al., 2012) and a novel subjective-score model that we propose in the present work.

Coverage and Diversity. For completeness, we also discuss several related works that study reviewer assignment problem.

Li et al. (2015) present a greedy algorithm that tries to avoid assigning a group of stringent reviewers or a group of lenient reviewers to a submission, thus maintaining diversity of the assignment in terms of having different combinations of reviewers assigned to different papers.

Another way to ensure diversity of the assignment is proposed by Liu et al. (2014). Instead of designing the special assignment algorithm, they try to incentivize the diversity by special construction of similarities. Besides incorporating expertise and authority of reviewers in similarities, they add an additional term to the optimization problem which balances similarities by increasing scores for reviewers from different research areas.

Karimzadehgan et al. (2008) consider topic coverage as an objective and propose several approaches to maintain broad coverage, requiring reviewers assigned to paper being expert in different subtopics covered by the paper. They empirically verify that given a paper and a set of reviewers, their algorithms lead to better coverage of paper's topics as compared to baseline technique that assigns reviewers based on some measure of

similarity between text of submission and papers authored by reviewers, but does not do topic matching.

A similar goal is formally studied by Long et al. (2013). They measure the coverage of the assignment in terms of the total number of distinct topics of papers covered by the assigned reviewers. They propose a constant factor approximation algorithm that benefits from a sub-modular nature of the objective. As we show in Appendix A3, the techniques of Long et al. (2013) can be combined with our proposed algorithm to obtain an assignment which maintains not only fairness, but also a broad topic coverage.

Research on peer review. The explosion in the number of submissions in many conferences has spurred research in computer science on improving peer review. In addition to problems of fairness and accuracy of the reviewer-paper assignment process, there are a number of challenges in peer review which are addressed in the literature to various extents. These include problems of bias (Tomkins et al., 2017; Stelmakh et al., 2019), miscalibration (Ge et al., 2013; Roos et al., 2011; Flach et al., 2010; Wang and Shah, 2019), subjectivity (Noothigattu et al., 2020), strategic behavior (Balietti et al., 2016; Xu et al., 2019), and others (Lawrence and Cortes, 2014; Gao et al., 2019). Of particular interest is the work by Fiez et al. (2020) which optimizes the process by which reviewers can bid on which papers they prefer to review. In most automated reviewer-paper assignment systems, the bids and the text-matching similarities are then combined (Shah et al., 2018) to form the similarities used to compute the assignment. The bidding and the reviewer-paper assignments are executed separately in current systems, and given the intrinsic relations between the two, it is of interest to jointly design the two systems in the future.

3 Problem setting

In this section we present the problem setting formally with a focus on the objective of fairness. (We introduce the statistical models we consider in Sections 6 and 7.)

3.1 Preliminaries and notation

Given a collection of $m \geq 2$ papers, suppose that there exists a true, unknown total ranking of the papers. The goal of the program chair (PC) of the conference is to recover top k papers, for some pre-specified value $k < m$. In order to achieve this goal, the PC recruits $n \geq 2$ reviewers and asks each of them to read and evaluate some subset of the papers. Each reviewer can review a limited number of papers. We let μ denote the maximum number of papers that any reviewer is willing to review. Each paper must be reviewed by λ distinct reviewers. In order to ensure this setting is feasible, we assume that $n\mu \geq m\lambda$. In practice, λ is typically small (2 to 6) and hence should conceptually be thought of as a constant.

The PC has access to a similarity matrix $S = \{s_{ij}\} \in [0,1]^{n \times m}$, where s_{ij} denotes the similarity between any reviewer $i \in [n]$ and any paper $j \in [m]$.[1] These similarities are representative of the envisaged quality of the respective reviews: a higher similarity between any reviewer and paper is assumed to indicate a higher competence of that reviewer in reviewing that paper (this assumption is formalized later). We do not discuss the design of such similarities, but often they are provided by existing systems (Charlin and Zemel, 2013; Mimno and McCallum, 2007; Liu et al., 2014; Rodriguez and Bollen, 2008; Tran et al., 2017).

Our focus is on the assignment of papers to reviewers. We represent any assignment by a matrix $A \in \{0,1\}^{n \times m}$, whose $(i,j)^{\text{th}}$ entry is 1 if reviewer i is assigned paper j and 0 otherwise. We denote the set of reviewers who review paper j under an assignment A as $\mathcal{R}_A(j)$. We call an assignment *feasible* if it respects the (μ, λ) conditions on the reviewer and paper loads. We denote the set of all feasible assignments as $\mathcal{A}$:

$$\mathcal{A} := \Big\{ A \in \{0,1\}^{n \times m} \mid \sum_{i \in [n]} A_{ij} = \lambda \ \forall j \in [m], \sum_{j \in [m]} A_{ij} \leq \mu \ \forall i \in [n] \Big\}.$$

Our goal is to design a reviewer-assignment algorithm with a two-fold objective: (i) fairness to all papers, (ii) strong statistical guarantees in terms of recovering the top papers.

[1] Here, we adopt the standard notation $[\nu] = \{1, 2, \ldots, \nu\}$ for any positive integer ν.

From a statistical perspective, we assume that when any reviewer i is asked to evaluate any paper j, then she/he returns score $y_{ij} \in \mathbb{R}$. The end goal of the PC is to accept or reject each paper. In this work we consider a simplified yet indicative setup. We assume that the PC wishes to accept the k "top" papers from the set of m submitted papers. We denote the "true" set of top k papers as $\mathcal{T}_k^*$. While the PC's decisions in practice would rely on several additional factors including the text comments by reviewers and the discussions between them, in order to quantify the quality of any assignment we assume that the top k papers are chosen through some estimator $\widehat{\theta}$ that operates on the scores provided by the reviewers. Such an estimator can be used in practice to serve as a guide to the program committee in order to help reduce their load. These acceptance decisions can be described by the chosen assignment and estimator $\left(A, \widehat{\theta}\right)$. We denote the set of accepted papers under an assignment A and estimator $\widehat{\theta}$ as $\mathcal{T}_k = \mathcal{T}_k\left(A, \widehat{\theta}\right)$. The PC then wishes to maximize the probability of recovering the set $\mathcal{T}_k^*$ of top k papers.

Although the goal of exact recovering of top k papers is appealing, given the large number of papers submitted to a conference such as ICML and NeurIPS, this goal might be too optimistic. Another alternative is to recover top k papers allowing for a certain Hamming error tolerance $t \in \{0, \ldots, k-1\}$. For any two subsets $\mathcal{M}_1, \mathcal{M}_2$ of $[m]$, we define their Hamming distance to be the number of items that belong to exactly one of the two sets — that is

$$\mathcal{D}_H\left(\mathcal{M}_1, \mathcal{M}_2\right) = \operatorname{card}\left(\{\mathcal{M}_1 \cup \mathcal{M}_2\} \setminus \{\mathcal{M}_1 \cap \mathcal{M}_2\}\right). \tag{3.1}$$

The goal of PC under this scenario is to choose a pair $\left(A, \widehat{\theta}\right)$ such that for the given error tolerance parameter t, the probability $\mathbb{P}\left\{\mathcal{D}_H\left(\mathcal{T}_k, \mathcal{T}_k^*\right) > 2t\right\}$ is minimized. We return to more details on the statistical aspects later in the paper.

3.2 Fairness objective

An assignment objective that is popular in past papers (Charlin and Zemel, 2013; Charlin et al., 2012; Taylor, 2008) is to maximize the cumulative similarity over all papers. Formally, these works choose an assignment $A \in \mathcal{A}$ which maximizes the quantity

$$G^S\left(A\right) := \sum_{j=1}^{m} \sum_{i \in \mathcal{R}_A(j)} s_{ij}. \tag{3.2}$$

An assignment algorithm that optimizes this objective (3.2) is implemented in the widely used Toronto Paper Matching System (Charlin and Zemel, 2013). We will refer to the feasible assignment that maximizes the objective (3.2) as A^{TPMS} and denote the algorithm which computes A^{TPMS} as TPMS.

We argue that the objective (3.2) does not necessarily lead to a *fair* assignment. The optimal assignment can discriminate some papers in order to maximize the cumulative objective. To see this issue, consider the following example.

Consider a toy problem with $n = m = 3$ and $\mu = \lambda = 1$, with a similarity matrix shown in Table 1. In this example, paper c is easy to evaluate, having non-zero similarities with all the reviewers, while papers a and b are more specific and weak reviewer 2 has no expertise in reviewing them. Reviewer 1 is an expert and is able to assess all three papers. Maximizing total sum of similarities (3.2), the TPMS algorithm will assign reviewers 1, 2, and 3 to papers a, b, and c respectively. Observe that under this assignment, paper b is assigned a reviewer who has insufficient expertise to evaluate the paper. On the other hand, the alternative assignment which assigns reviewers 1, 2, and 3 to papers a, c, and b respectively ensures that every paper has a reviewer with similarity at least $1/5$. This "fair" assignment does not discriminate against papers a and b for improving the review quality of the already benefitting paper c.

With this motivation, we now formally describe the notion of fairness that we aim to optimize in this work. Inspired by the notion of max-min fairness in a variety of other fields (Rawls, 1971; Lenstra et al., 1990; Hahne, 1991; Lavi et al., 2003; Bonald et al., 2006; Asadpour and Saberi, 2010), we aim to find a feasible

	Paper a	Paper b	Paper c
Reviewer 1	1	1	1
Reviewer 2	0	0	1/5
Reviewer 3	1/4	1/4	1/2

Table 1: Example similarity.

assignment $A \in \mathcal{A}$ to maximize the following objective Γ^S for given similarity matrix S:

$$\Gamma^S(A) = \min_{j \in [m]} \sum_{i \in \mathcal{R}_A(j)} s_{ij}. \tag{3.3}$$

The assignment optimal for (3.3) maximizes the minimum sum similarity across all the papers. In other words, for *every other assignment* there exists some paper which has the same or lower sum similarity. Returning to our example, the objective (3.3) is maximized when reviewers 1, 2, and 3 are assigned to papers a, c, and b respectively.

Our reviewer assignment algorithm presented subsequently guarantees the aforementioned fair assignment. Importantly, while aiming at optimizing (3.3), our algorithm does even more — having the assignment for the worst-off paper fixed, it finds an assignment that satisfies the second worst-off paper, then the next one and so on until all papers are assigned.

It is important to note that similarities s_{ij} obtained by different techniques (Charlin and Zemel, 2013; Mimno and McCallum, 2007; Rodriguez and Bollen, 2008; Tran et al., 2017) all have different meanings. Therefore, the PC might be interested to consider a slightly more general formulation and aim to maximize

$$\Gamma^S_f(A) = \min_{j \in [m]} \sum_{i \in \mathcal{R}_A(j)} f(s_{ij}), \tag{3.4}$$

for some reasonable choice of monotonically increasing function $f : [0, 1] \to [0, \infty]$.[2] While the same effect might be achieved by redefining $s'_{ij} = f(s_{ij})$ for all $i \in [n]$, $j \in [m]$, this formulation underscores the fact that assignment procedure is not tied to any particular method of obtaining similarities. Different choices of f represent the different views on the meaning of similarities. As a short example, let us consider $f(s_{ij}) = \mathbb{I}\{s_{ij} > \zeta\}$ for some $\zeta > 0$.[3] This choice stratifies reviewers for each paper into strong (similarity higher than ζ) and weak. The fair assignment would be such that the most disadvantaged paper is assigned to as many strong reviewers as possible. We discuss other variants of f later when we come to the statistical properties of our algorithm. In what follows we refer to the problem of finding reviewer assignment that maximizes the term (3.4) as the *fair assignment problem*.

Unfortunately, the assignment optimal for (3.4) is hard to compute for any reasonable choices of function f. Garg et al. (2010) showed that finding a fair assignment is an NP-hard problem even if $f(s) \in \{1, 2, 3\}$ and $\lambda = 2$.

With this motivation, in the next section we design a reviewer assignment algorithm that seeks to optimize the objective (3.4) and provide associated approximation guarantees. We will refer to a feasible assignment that exactly maximizes $\Gamma^S_f(A)$ as A^{HARD}_f and denote the algorithm that computes A^{HARD}_f as Hard. When the function f is clear from context, we drop the subscript f and denote the Hard assignment as A^{HARD} for brevity.

Finally we note that for our running example (Table 1 above), the ILPR algorithm (Garg et al., 2010), despite trying to optimize fairness of the assignment, also returns an unfair assignment A^{ILPR} which coincides with A^{TPMS}. The reason for this behavior lies in the inner-working of the ILPR algorithm: a linear

[2] We allow $f(s_{ij}) = \infty$. When reviewer with similarity ∞ is assigned to paper, she/he is able to perfectly access the quality of the paper.

[3] We use $\mathbb{I}$ to denote the indicator function, that is, $\mathbb{I}\{x\} = 1$ if x is true and $\mathbb{I}\{x\} = 0$ otherwise.

programming relaxation splits reviewers 1 and 2 in two and makes them review both paper a and paper b. During the rounding stage, reviewer 1 is assigned to either paper a or paper b, ensuring that the remaining paper will be reviewed by reviewer 2. Given that reviewer 2 has zero similarity with both papers a and b, the fairness of the resulting assignment will be 0. Such an issue arises more generally in the ILPR algorithm and is discussed in more detail subsequently in Section 5.3 and in Appendix A1.1.

4 Reviewer assignment algorithm

In this section we first describe our PeerReview4All algorithm followed by an illustrative example.

4.1 Algorithm

A high level idea of the algorithm is the following. For every integer $\kappa \in [\lambda]$, we try to assign each paper to κ reviewers with maximum possible similarities while respecting constraints on reviewer loads. We do so via a carefully designed "subroutine" that is explained below. Continuing for that value of κ, we complement this assignment with $(\lambda - \kappa)$ additional reviewers for each paper. Repeating the procedure for each value of $\kappa \in [\lambda]$, we obtain λ candidate assignments each with λ reviewers assigned to each paper, and then choose the one with the highest fairness. The assignment at this point ensures guarantees of worst-case fairness (3.4). We then also optimize for the second worst-off paper, then the third worst-off paper and so on in the following manner. In the assignment at this point, we find the most disadvantaged papers and permanently fix corresponding reviewers to these papers. Next, we repeat the procedure described above to find the most fair assignment among the remaining papers, and so on. By doing so, we ensure that our final assignment is not susceptible to bottlenecks which may be caused by irrelevant papers with small average similarities.

The higher-level idea behind the aforementioned subroutine to obtain the candidate assignment for any value of $\kappa \in [\lambda]$ is as follows. The subroutine constructs a layered flow network graph with one layer for reviewers and one layer for papers, that captures the similarities and the constraints on the paper/reviewer loads. Then the subroutine incrementally adds edges between (reviewer, paper) pairs in decreasing order of similarity and stops when the paper load constraints are met (each paper can be assigned to κ reviewers using only edges added at this point). This iterative procedure ensures that the papers are assigned reviewers with approximately the highest possible similarities.

We formally present our main algorithm as Algorithm 1 and the subroutine as Subroutine 1. In what follows, we walk the reader through the steps in the subroutine and the algorithm in more detail.

Subroutine. A key component of our algorithm is a construction of a flow network in a sequential manner in Subroutine 1. The subroutine takes as input, among other arguments, the set $\mathcal{M}$ of papers that are not yet assigned and the required number of reviewers per paper $\kappa \leq \lambda$. The goal of the subroutine is to assign each paper in $\mathcal{M}$ with κ reviewers, respecting the reviewer load constraints, in a way that minimum similarity across all paper-reviewer pairs in resulting assignment is maximized.

The output of the subroutine is an assignment (represented by variable A) which is initially set as empty (Step 1). The subroutine begins (Step 2) with a construction of a directed acyclic graph (a "flow network") comprising 4 layers in the following order: a source, all reviewers, all papers in $\mathcal{M}$, and a sink. An edge may exist only between consecutive layers. The edges between the first two layers control the reviewers' workloads and edges between the last two layers represent the number of reviews required by the papers. Finally, costs of the all edges in this initial construction are set to 0. Note that in subsequent steps, the edges are added only between the second and third layers. Thus, the maximum flow in the network is at most $|\mathcal{M}|\kappa$.

The crux of the subroutine is to incrementally add edges one at a time between the layers, representing the reviewers and papers, in a carefully designed manner (Steps 3 and 4). The edges are added in order of decreasing similarities. These edges control a reviewer-paper relationship: they have a unit capacity to ensure that any reviewer can review any paper at most once and their costs are equal to the similarity between the corresponding (reviewer, paper) pair.

After adding each edge, the subroutine (Step 5) tests whether a max-flow of size $|\mathcal{M}|\kappa$ is feasible. Note that a feasible flow of size $|\mathcal{M}|\kappa$ corresponds to a feasible assignment: by construction of the flow network

Subroutine 1 PeerReview4All Subroutine

Input: $\kappa \in [\lambda]$: number of reviewers required per paper
$\mathcal{M}$: set of papers to be assigned
$S \in (\{-\infty\} \cup [0,1])^{n \times |\mathcal{M}|}$: similarity matrix
$(\mu^{(1)}, \ldots, \mu^{(n)}) \in [\mu]^n$: reviewers' maximum loads

Output: Reviewer assignment A

Algorithm:

1. Initialize A to an empty assignment
2. Initialize the flow network:
   - **Layer 1:** one vertex (source)
   - **Layer 2:** one vertex for every reviewer $i \in [n]$, and directed edges of capacity $\mu^{(i)}$ and cost 0 from the source to every reviewer
   - **Layer 3:** one vertex for every paper $j \in \mathcal{M}$
   - **Layer 4:** one vertex (sink), and directed edges of capacity κ and cost 0 from each paper to the sink
3. Find (reviewer, paper) pair (i, j) such that the following two conditions are satisfied:
   - the corresponding vertices i and j are not connected in the flow network
   - the similarity s_{ij} is maximal among the pairs which are not connected (ties are broken arbitrarily)

   and call this pair (i', j')
4. Add a directed edge of capacity 1 and cost $s_{i'j'}$ between nodes i' and j'
5. Compute the max-flow from source to sink, if the size of the flow is strictly smaller than $|\mathcal{M}|\kappa$, then go to Step 3
6. If there are multiple possible max-flows, choose any one arbitrarily (or use any heuristic such as max-flow with max cost)
7. For every edge (i, j) between layers 2 (reviewers) and 3 (papers) which carries a unit of flow in the selected max-flow, assign reviewer i to paper j in the assignment A

---

described earlier, we know that the reviewer and paper load constraints are satisfied. The capacity of each edge in our flow network is a non-negative integer, thereby guaranteeing that the max-flow is an integer, that it can be found in polynomial time, and that the flow in every edge is a non-negative integer under the max-flow. Once the max-flow of size $|\mathcal{M}|\kappa$ is reached, the subroutine stops adding edges. At this point, it is ensured that the value of the lowest similarity in the resulting assignment is maximized.

Finally, the subroutine assigns each paper to κ reviewers, using only the "high similarity" edges added to the network so far (Steps 6 and 7). The existence of the corresponding assignment is guaranteed by max-flow in the network being equal to $|\mathcal{M}|\kappa$. There may be more than one feasible assignments that attain the max-flow. While any of these assignments would suffice from the standpoint of optimizing the worst-case fairness objective (3.4), the PC may wish to make a specific choice for additional benefits and specify the heuristic to pick the max-flow in Step 6 of the subroutine. For example, if the max-flow with the maximum cost is selected, then the resulting assignment nicely combines fairness with the high average quality of the assignment. Another choice, discussed in Appendix A3, helps with broad topic coverage of the assignment. Importantly, the approximation guarantees established in Theorem 1 and Corollary 1, as well as statistical guarantees from Sections 6 and 7 hold for any max-flow assignment chosen in Steps 6 and 7.

For comparison, we note that the TPMS algorithm can equivalently be interpreted in this framework as follows. The TPMS algorithm would first *connect all reviewers to all papers* in layers 2 and 3 of the flow graph. It will then compute a max-flow with max cost in this fully connected flow network and make reviewer-paper assignments corresponding to the edges with unit flow between layers 2 and 3. In contrast, our sequential construction of the flow graph prevents papers from being assigned to weak reviewers and is crucial towards ensuring the fairness objective.

Algorithm 1 PeerReview4All Algorithm

Input: $\lambda \in [n]$: number of reviewers required per paper
$S \in [0,1]^{n \times m}$: similarity matrix
$\mu \in [m]$: reviewers' maximum load
f: transformation of similarities
Output: Reviewer assignment A_f^{PR4A}
Algorithm:

1. Initialize $\overline{\mu} = (\mu, \ldots, \mu) \in [\mu]^n$
   A_f^{PR4A}, A_0 : empty assignments
   $\mathcal{M} = [m]$: set of papers to be assigned
2. For $\kappa = 1$ to λ
   (a) Set $\overline{\mu}^{\mathrm{tmp}} = \overline{\mu}, S^{\mathrm{tmp}} = S$
   (b) Assign κ reviewers to every paper using subroutine: $A_\kappa^1 = \mathrm{Subroutine}(\kappa, \mathcal{M}, S^{\mathrm{tmp}}, \overline{\mu}^{\mathrm{tmp}})$
   (c) Decrease $\overline{\mu}^{\mathrm{tmp}}$ for every reviewer by the number of papers she/he is assigned in A_κ^1, set corresponding similarities in S^{tmp} to $-\infty$
   (d) Run subroutine with adjusted $\overline{\mu}^{\mathrm{tmp}}$ and S^{tmp} to assign remaining $\lambda - \kappa$ reviewers to every paper: $A_\kappa^2 = \mathrm{Subroutine}(\lambda - \kappa, \mathcal{M}, S^{\mathrm{tmp}}, \overline{\mu}^{\mathrm{tmp}})$
   (e) Create assignment A_κ such that for every pair (i,j) of reviewer $i \in [n]$ and paper $j \in \mathcal{M}$, reviewer i is assigned to paper j if she/he is assigned to this paper in either A_κ^1 or A_κ^2
3. Choose $\widetilde{A} \in \arg\max_{\kappa \in [\lambda] \cup \{0\}} \Gamma_f^S(A_\kappa)$ with ties broken arbitrarily
4. For every paper $j \in \mathcal{J}^* := \arg\min_{\ell \in \mathcal{M}} \sum_{i \in \mathcal{R}_{\widetilde{A}}(\ell)} f(s_{i\ell})$, assign all reviewers $\mathcal{R}_{\widetilde{A}}(j)$ to paper j in A_f^{PR4A}
5. For every reviewer $i \in [n]$, decrease $\mu^{(i)}$ by the number of papers in $\mathcal{J}^*$ assigned to i
6. Delete columns corresponding to the papers $\mathcal{J}^*$ from S and $\widetilde{A}$, update $\mathcal{M} = \mathcal{M} \backslash \mathcal{J}^*$
7. Set $A_0 = \widetilde{A}$
8. If $\mathcal{M}$ is not empty, go to Step 2

Algorithm. The algorithm calls the subroutine iteratively and uses the outputs of these iterates in a carefully designed manner. Initially, all papers belong to a set $\mathcal{M}$ which represents papers that are not yet assigned. The algorithm repeats Steps 2 to 7 until all papers are assigned. In every iteration, for every value of $\kappa \in [\lambda]$, the algorithm first calls the subroutine to assign κ reviewers to each paper from $\mathcal{M}$ (Step 2b), and then adjusts reviewers' capacities and the similarity matrix (Step 2c) to prevent any reviewer being assigned to the same paper twice. Next, the subroutine is called again (Step 2d) to assign another $(\lambda - \kappa)$ reviewers to each paper. As a result, after completion of Step 2, λ feasible candidate assignments $A_1, \ldots, A_\lambda$ are constructed. Each assignment $A_\kappa, \kappa \in [\lambda]$, is guaranteed (through the Step 2b) to maximize the minimum similarity across pairs (i,j) where $j \in \mathcal{M}$ and reviewer i is among κ strongest reviewers assigned to paper j in A_κ; and (through the Steps 2d and 2e) to have each paper assigned with exactly λ reviewers.

In Step 3, the algorithm chooses the assignment with the highest fairness (3.4) among the λ candidate assignments and the assignment A_0 from the previous iteration (empty in the first iteration). Note that since A_0 is also included in the maximizer, the fairness cannot decrease in subsequent iterations.

In the chosen assignment, the algorithm identifies the papers that are most disadvantaged, and fixes the assignment for these papers (Step 4). The assignment for these papers will not be changed in any subsequent step. The next steps (Steps 5 and 6) update the auxiliary variables to account for this assignment that is fixed — decreasing the corresponding reviewer capacities and removing these assigned papers from the set $\mathcal{M}$. Step 7 then keeps a track of the present assignment $\widetilde{A}$ for use in subsequent iterations, ensuring that fairness cannot decrease as the algorithm proceeds.

Remarks. We make a few additional remarks regarding the PeerReview4All algorithm.

1. Computational cost: A naïve implementation of the PeerReview4All algorithm has a computational complexity $\widetilde{\mathcal{O}}\left(\lambda(m+n)m^2n\right)$. We give more details on implementation and computational aspects in Appendix A2.

2. Variable reviewer or paper loads: More generally, the PeerReview4All algorithm allows for specifying different loads for different reviewers and/or papers. For general paper loads, we consider $\kappa \leq \max_{j\in[m]} \lambda^{(j)}$ and define the capacity of edge between node corresponding to any paper j and sink as $\min\{\kappa, \lambda^{(j)}\}$.

3. Incorporating conflicts of interest: One can easily incorporate any conflict of interest between any reviewer and paper by setting the corresponding similarity to $-\infty$.

4. Topic coverage: The techniques developed in Long et al. (2013) can be employed to modify our algorithm in a way that it first ensures fairness and then, among all approximately fair assignments, picks one that approximately maximizes the number of distinct topics of papers covered. We discuss this modification in Appendix A3.

4.2 Example

To provide additional intuition behind the design of the algorithm, we now present an example that we also use in the next section to explain our approximation guarantees.

Let for a moment assume that $f(s) = s$ and let ζ be a constant close to 1. Consider the following two scenarios:

(S1) The optimal assignment A^{HARD} is such that all the papers are assigned to reviewers with high similarity:

$$\min_{i\in\mathcal{R}_{A^{\mathrm{HARD}}}(j)} s_{ij} > \zeta \qquad \forall j \in [m]. \tag{3.5}$$

(S2) The optimal assignment A^{HARD} is such that there are some "critical" papers which have $\eta < \lambda$ assigned reviewers with similarities higher than ζ and the remaining assigned reviewers with small similarities. All other papers are assigned to λ reviewers with similarity higher than ζ.

Intuitively, the first scenario corresponds to an ideal situation since there exists an assignment such that each paper has λ competent reviewers (with similarity $\zeta \approx 1$). In contrast, in the second scenario, even in the fair assignment, some papers lack expert reviewers. Such a scenario may occur, for example, if some non-mainstream papers were submitted to a conference. This case entails identifying and treating these disadvantaged papers as well as possible. To be able to find the fair assignment in both scenarios, the assignment algorithm should distinguish between them and adapt its behavior to the structure of similarity matrix. Let us track the inner-workings of PeerReview4All algorithm to demonstrate this behaviour.

We note that by construction, the fairness of the resulting assignment A^{PR4A} is determined in the first iteration of Steps 2 to 7 of Algorithm 1, so we restrict our attention to $\mathcal{M} = [m]$. First, consider scenario (S1). The subroutine called with parameter $\kappa = \lambda$ will add edges to the flow network until the maximal flow of size $m\lambda$ is reached. Since the optimal assignment A^{HARD} is such that the lowest similarity is higher than ζ, the last edge added to the flow network will have similarity at least ζ, implying that the fairness of the candidate assignment A_λ, which is a lower bound for the fairness of resulting assignment, will be at least $\lambda\zeta$. Given that ζ is close to one, we conclude that in this case algorithm is able to recover an assignment which is at least very close to optimal.

Now, let us consider scenario (S2). In this scenario, the subroutine called with $\kappa = \lambda$ may return a poor assignment. Indeed, since there is a lack of competent reviewers for critical papers, there is no way to assign each paper with λ reviewers having a high minimum similarity in the assignment. However, the subroutine called with parameter $\kappa = \eta$ will find η strong reviewers for each paper (including the critical papers), thereby leading to a fairness $\Gamma^S\left(A^{\mathrm{PR4A}}\right) \geq \eta\zeta$. The obtained lower bound guarantees that the assignment recovered by the PeerReview4All algorithm is also close to the optimal, because in the fair assignment A^{HARD} some papers have only η strong reviewers.

This example thus illustrates how the PeerReview4All algorithm can adapt to the structure of the similarity matrix in order to guarantee fairness, as well as other guarantees that are discussed subsequently in the paper.

5 Approximation guarantees

In this section we provide guarantees on the fairness of the reviewer-assignment by our algorithm. We first establish guarantees on the max-min fairness objective introduced earlier (Section 5.1). We subsequently show that our algorithm optimizes not only the worst-off paper but recursively optimizes all papers (Section 5.2). We then conclude this section on deterministic approximation guarantees with a comparison to past literature (Section 5.3).

5.1 Max-min fairness

We begin with some notation that will help state our main approximation guarantees. For each value of $\kappa \in [\lambda]$, consider the reviewer-assignment problem but where each paper requires κ (instead of λ) reviews (each reviewer still can review up to μ papers). Let us denote the family of all feasible assignments for this problem as $\mathcal{A}_\kappa$. Now define the quantities

$$\begin{aligned} s^*_\kappa &:= \max_{A \in \mathcal{A}_\kappa} \min_{j \in [m]} \min_{i \in \mathcal{R}_A(j)} s_{ij}, \\ s^*_0 &:= \max_{i \in [n]} \max_{j \in [m]} s_{ij}, \quad \text{and} \\ s^*_\infty &:= \min_{i \in [n]} \min_{j \in [m]} s_{ij}. \end{aligned} \tag{3.6}$$

Intuitively, for *every* assignment from the family $\mathcal{A}_\kappa$, the quantity s^*_κ upper bounds the minimum similarity for any assigned (reviewer, paper) pair. It also means that the value s^*_κ is achievable by some assignment in $\mathcal{A}_\kappa$. The value s^*_0 captures the value of the largest entry in the similarity matrix S and gives a trivial upper bound $\Gamma^S_f(A) \le \lambda f(s^*_0)$ for every feasible assignment $A \in \mathcal{A}$. Likewise, the value s^*_∞ captures the smallest entry in the similarity matrix S and yields a lower bound $\Gamma^S_f(A) \ge \lambda f(s^*_\infty)$ for every feasible assignment $A \in \mathcal{A}$.

We are now ready to present the main result on the approximation guarantees for the PeerReview4All algorithm as compared to the optimal assignment A^{HARD}.

Theorem 1. *Consider any feasible values of (n, m, λ, μ), any monotonically increasing function $f : [0, 1] \to [0, \infty]$, and any similarity matrix S. The assignment A^{PR4A}_f given by the* PeerReview4All *algorithm guarantees the following lower bound on the fairness objective* (3.4)*:*

$$\frac{\Gamma^S_f\left(A^{PR4A}_f\right)}{\Gamma^S_f\left(A^{HARD}_f\right)} \ge \frac{\max_{\kappa \in [\lambda]} \left(\kappa f(s^*_\kappa) + (\lambda - \kappa) f(s^*_\infty)\right)}{\min_{\kappa \in [\lambda]} \left((\kappa - 1) f(s^*_0) + (\lambda - \kappa + 1)\, f(s^*_\kappa)\right)} \tag{3.7a}$$

$$\ge 1/\lambda. \tag{3.7b}$$

Remarks. The numerator of (3.7a) is a lower bound on the fairness of the assignment returned by our algorithm. It is important to note that if $\lambda = 1$, that is, if we only need to assign one reviewer for each paper, then our PeerReview4All Algorithm finds exact solution for the problem, recovering the classical results of Garfinkel (1971) as a special case.

In practice, the number of reviewers λ required per paper is a small constant (typically set as 3), and in that case, our algorithm guarantees a constant factor approximation. Note that the fraction in the right hand side of (3.7a) can become $0/0$ or ∞/∞, and in both cases it should be read as 1.

The bound (3.7a) can be significantly tighter than $1/\lambda$, as we illustrate in the following example.

Example. Consider two scenarios (S1) and (S2) from Section 4.2, and consider $f(s) = s$. One can see that under scenario (S1), we have $s^*_\lambda \ge \zeta$. Setting $\kappa = \lambda$ in the numerator and $\kappa = 1$ in the denominator of the bound (3.7a), and recalling that $\zeta \approx 1$, we obtain:

$$\frac{\Gamma^S\left(A^{\text{PR4A}}\right)}{\Gamma^S\left(A^{\text{HARD}}\right)} \ge \frac{\zeta}{s^*_1} \approx 1,$$

where we have also used the fact that $s_1^* \leq 1$. Let us now consider the second scenario (S2) in the example of Section 4.2. In this scenario, since each paper can be assigned to η strong reviewers with similarity higher than ζ, we have $s_\eta^* = \zeta \approx 1$. We then also have $s_0^* \leq 1$. Moreover, there are some papers which have only η strong reviewers in optimal assignment A^{HARD}, and hence we have $s_{\eta+1}^* \ll s_0^*$. Setting $\kappa = \eta$ in the numerator and $\kappa = \eta + 1$ in the denominator of the bound (3.7a), some algebraic simplifications yield the bound

$$\frac{\Gamma^S\left(A^{\mathrm{PR4A}}\right)}{\Gamma^S\left(A^{\mathrm{HARD}}\right)} \geq \frac{\eta s_\eta^* + (\lambda - \eta)s_\infty^*}{\eta s_0^* + (\lambda - \eta)s_{\eta+1}^*} \geq \frac{s_\eta^*}{s_0^*} - \frac{(\lambda - \eta)}{\eta}\frac{s_{\eta+1}^*}{s_0^*} \approx 1.$$

We now briefly provide more intuition on the bound (3.7a) by interpreting it in terms of specific steps in the algorithm. Setting $f(s) = s$, let us consider the first iteration of the algorithm. Recalling the definition (3.6) of s_κ^*, the PeerReview4All subroutine called with parameter κ on Step 2b finds an assignment such that all the similarities are at least s_κ^*. This guarantee in turn implies that the fairness of the corresponding assignment A_κ is at least $\kappa s_\kappa^* + (\lambda - \kappa)s_\infty^*$, thereby giving rise to the numerator of (3.7a). The denominator is an upper bound of the fairness of the optimal assignment A^{HARD}. The expression for any value of κ is obtained by simply appealing to the definition of s_κ^* which is defined in terms of the optimal assignment. By definition (3.6) of s_κ^*, for every feasible assignment A exists at least one paper such that at most $\kappa - 1$ of the assigned reviewers are of similarity larger than s_κ^*. Thus, the fairness of the optimal assignment is upper-bounded by the sum similarity of the paper that has $\kappa - 1$ reviewers with similarity s_0^* (the highest possible similarity), and $\lambda - \kappa + 1$ reviewers with similarity s_κ^*.

Finally, one may wonder whether optimizing the objective (3.2) as done by prior works (Charlin and Zemel, 2013; Charlin et al., 2012) can also guarantee fairness. It turns out that this is not the case (see the example in Table 1 for intuition), and optimizing the objective (3.2) is not a suitable proxy towards the fairness objective (3.4). In Appendix A1.2 we show that in general the fairness objective value of the TPMS algorithm which optimizes (3.2) may be *arbitrarily bad* as compared to that attained by our PeerReview4All algorithm.

In Appendix A1.3 we show that the analysis of the approximation factor of our algorithm is tight in a sense that there exists a similarity matrix for which the bound (3.7b) is met with equality. That said, the approximation factor of our PeerReview4All algorithm can be much better than $\frac{1}{\lambda}$ for various other similarity matrices, as demonstrated in examples (S1) and (S2).

5.2 Beyond worst case

The previous section established guarantees for the PeerReview4All algorithm on the fairness of the assignment in terms of the worst-off paper. In this section we formally show that the algorithm does more: having the assignment for the worst-off paper fixed, the algorithm then satisfies the second worst-off paper, and so on.

Recall that Algorithm 1 iteratively repeats Steps 2 to 7. In fact, the first time that Step 3 is executed, the resulting intermediate assignment $\widetilde{A}$ achieves the max-min guarantees of Theorem 1. However, the algorithm does not terminate at this point. Instead, it finds the most disadvantaged papers in the selected assignment and fixes them in the final output A_f^{PR4A} (Step 4), attributing these papers to reviewers according to $\widetilde{A}$. Then it repeats the entire procedure (Steps 2 to 7) again to identify and fix the assignment for the most disadvantaged papers among the remaining papers and so on until the all papers are assigned in A_f^{PR4A}. We denote the total number of iterations of Steps 2 to 7 in Algorithm 1 as p ($\leq m$). For any iteration $r \in [p]$, we let $\mathcal{J}_r$ be the set of papers which the algorithm, in this iteration, fixes in the resulting assignment. We also let $\widetilde{A}_r, r \in [p]$, denote the assignment selected in Step 3 of the r^{th} iteration. Note that eventually all the papers are fixed in the final assignment A_f^{PR4A}, and hence we must have $\bigcup_{r\in[p]} \mathcal{J}_r = [m]$.

Once papers are fixed in the final output A_f^{PR4A}, the assignment for these papers are not changed any more. Thus, at the end of each iteration $r \in [p]$ of Steps 2 to 7, the algorithm deletes (Step 6) the columns of similarity matrix that correspond to the papers fixed in this iteration. For example, at the end of the first

iteration, columns which correspond to $\mathcal{J}_1$ are deleted from S. For each iteration $r \in [p]$, we let S_r denote the similarity matrix at the beginning of the iteration. Thus, we have $S_1 = S$, because at the beginning of the first iteration, no papers are fixed in the final assignment A_f^{PR4A}.

Moving forward, we are going to show that for every iteration $r \in [p]$, the sum similarity of the worst-off papers $\mathcal{J}_r$ (which coincides with the fairness of $\widetilde{A}_r$) is close to the best possible, given the assignment for the all papers fixed in the previous iterations. As in Theorem 1, we will compare the fairness $\Gamma_f^S\left(\widetilde{A}_r\right)$ with the fairness of the optimal assignment that HARD algorithm would return if called at the beginning of the r^{th} iteration. We stress that for every $r \in [p]$, the HARD algorithm assigns papers $\bigcup_{l=r}^{p} \mathcal{J}_l$ and respects the constraints on reviewers' loads, adjusted for the assignment of papers $\bigcup_{l=1}^{r-1} \mathcal{J}_l$ in A_f^{PR4A}. We denote the corresponding assignment as $A_f^{\text{HARD}}(\mathcal{J}_{\{r:p\}})$. Note that $A_f^{\text{HARD}}(\mathcal{J}_{\{1:p\}}) = A_f^{\text{HARD}}$. The following corollary summarizes the main result of this section:

Corollary 1. *For any integer $r \in [p]$, the assignment $\widetilde{A}_r$, selected by the* PEERREVIEW4ALL *algorithm in Step 3 of the r^{th} iteration, guarantees the following lower bound on the fairness objective* (3.4)*:*

$$\frac{\Gamma_f^S\left(\widetilde{A}_r\right)}{\Gamma_f^S\left(A_f^{HARD}(\mathcal{J}_{\{r:p\}})\right)} \geq \frac{\max\limits_{\kappa\in[\lambda]}\left(\kappa f(s_\kappa^*) + (\lambda-\kappa) f(s_\infty^*)\right)}{\min\limits_{\kappa\in[\lambda]}\left((\kappa-1) f(s_0^*) + (\lambda-\kappa+1)\, f(s_\kappa^*)\right)} \geq 1/\lambda, \tag{3.8}$$

where values $s_\kappa^*, \kappa \in \{0,\ldots,\lambda\} \cup \{\infty\}$, are defined with respect to the similarity matrix S_r and constraints on reviewers' loads adjusted for the assignment of papers $\bigcup_{l=1}^{r-1} \mathcal{J}_l$ in A_f^{PR4A}.

The corollary guarantees that each time the algorithm fixes the assignment for some papers $j \in \mathcal{M}$ in A_f^{PR4A}, the sum similarity for these papers (which is smallest among papers from $\mathcal{M}$) is close to the optimal fairness, where optimal fairness is conditioned on the previously assigned papers. In case $r = 1$, the bound (3.8) coincides with the bound (3.7) from Theorem 1. Hence, once the assignment for the most worst-off papers is fixed, the PEERREVIEW4ALL algorithm adjusts maximum reviewers' loads and looks for the most fair assignnment of the remaining papers.

5.3 Comparison to past literature

In this section we discuss how the approximation results established in previous sections relate to the past literature.

First, we note that the assignment A_1, computed in Step 2 in the first iteration of Steps 2 to 7 of Algorithm 1, recovers the assignment of Hartvigsen et al. (1999), thus ensuring that our algorithm is *at least as fair* as theirs. Second, if the goal is to assign only one reviewer ($\lambda = 1$) to each of the papers, then our PEERREVIEW4ALL algorithm finds the optimally fair assignment and recovers the classical result of Garfinkel (1971).

In the remainder of this section, we provide a comparison between the guarantees of the PEERREVIEW4ALL algorithm established in Theorem 1 and the guarantees of the ILPR algorithm (Garg et al., 2010). Rewriting the results of Garg et al. (2010) in our notation, we have the bound:

$$\frac{\Gamma_f^S\left(A_f^{\text{ILPR}}\right)}{\Gamma_f^S\left(A_f^{\text{HARD}}\right)} \geq \frac{\Gamma_f^S\left(A_f^{\text{HARD}}\right) - \left(f(s_0^*) - f(s_\infty^*)\right)}{\Gamma_f^S\left(A_f^{\text{HARD}}\right)} = 1 - \frac{f(s_0^*) - f(s_\infty^*)}{\Gamma_f^S\left(A_f^{\text{HARD}}\right)}. \tag{3.9}$$

Note that our bound (3.7) for our PEERREVIEW4ALL algorithm is multiplicative and bound for the ILPR algorithm is additive which makes them incomparable in a sense that neither one dominates another. However,

we stress the following differences. First, if we assume f to be upper-bounded by one, then assignment A^{ILPR} satisfies the bound

$$\Gamma_f^S\left(A_f^{\mathrm{ILPR}}\right) \geq \Gamma_f^S\left(A_f^{\mathrm{HARD}}\right) - 1. \tag{3.10}$$

This bound gives a nice additive approximation factor — for a large value of the optimal fairness $\Gamma_f^S\left(A_f^{\mathrm{HARD}}\right)$, the constant additive factor is negligible. However, if the optimal fairness is small, which can happen if some papers do not have a sufficient number of high-expertise reviewers, then the lower bound on the fairness of the ILPR assignment (3.10) becomes negative, making the guarantees vacuous as any arbitrary assignment will achieve a non-negative fairness. Note that this issue is not an artifact of the analysis but is inherent in the ILPR algorithm itself, as we demonstrate in the example presented in Table 1 and in Appendix A1.1. In contrast, our algorithm in the worst case has a multiplicative approximation factor $1/\lambda$ ensuring that it always returns a non-trivial assignment.

This discrepancy becomes more pronounced if the function f is allowed to be unbounded, and the similarities are significantly heterogeneous. Suppose there is some reviewer $i \in [n]$ and paper $j \in [m]$ such that $f(s_{ij}) \gg \Gamma_f^S\left(A^{\mathrm{HARD}}\right)$. Then the bound (3.9) for the ILPR algorithm again becomes vacuous, while the bound (3.7) for the PeerReview4All algorithm continues to provide a non-trivial approximation guarantee.

Finally, we note that the bound (3.9) is also extended by Garg et al. (2010) to obtain guarantees on the fairness for the second worst-off paper and so on.

6 Objective-score model

We now turn to establishing statistical guarantees for our PeerReview4All algorithm from Section 4. We begin by considering an "objective" score model which we borrow from past works.

6.1 Model setup

The objective-score model assumes that each paper $j \in [m]$ has a true, unknown quality $\theta_j^* \in \mathbb{R}$ and each reviewer $i \in [n]$ assigned to paper j gives her/his estimate y_{ij} of θ_j^*. The eventual goal is to estimate top k papers according to true qualities $\theta_j^*, j \in [m]$. Following the line of works by Ge et al. (2013); McGlohon et al. (2010); Dai et al. (2012); Sajjadi et al. (2016), we assume the score y_{ij} given by any reviewer $i \in [n]$ to any paper $j \in [m]$ to be independently and normally distributed around the true paper qualities:

$$y_{ij} \sim \mathcal{N}\left(\theta_j^*, \sigma_{ij}^2\right). \tag{3.11}$$

Note that McGlohon et al. (2010); Dai et al. (2012) and Sajjadi et al. (2016) consider the restricted setting with $\sigma_{ij} = \sigma_i$ for all $(i, j) \in [n] \times [m]$, which implies that the variance of the reviewers' scores depends only on the reviewer, but not on the paper reviewed. We claim that this assumption is not appropriate for our peer-review problem: conferences today (such as ICML and NeurIPS) cover a wide spectrum of research areas and it is not reasonable to expect the reviewer to be equally competent in all of the areas.

In our analysis, we assume that the noise variances are some function of the underlying computed similarities.[4] We assume that for any $i \in [n]$ and $j \in [m]$, the noise variance

$$\sigma_{ij}^2 = h(s_{ij}),$$

for some monotonically decreasing function $h : [0, 1] \to [0, \infty)$. We assume that this function h is known; this assumption is reasonable as the function can, in principle, be learned from the data from the past conferences.

We note that the model (3.11) does not consider reviewers' biases. However, some reviewers might be more stringent while others are more lenient. This difference results in score of any reviewer i for any paper j being centered not at θ_j^*, but at $(\theta_j^* + b_i)$. A common approach to reduce biases in reviewers' scores is a

[4]Recall that the similarities can capture not only affinity in research areas but may also incorporate the bids or preferences of reviewers, past history of review quality, etc.

post-processing. For example, Ge et al. (2013) compared different statistical models of reviewers in attempt to calibrate the biases; the techniques developed in that work may be extended to the reviewer model (3.11). Thus, we leave that bias term out for simplicity.

6.2 Estimator

Given a valid assignment $A \in \mathcal{A}$, the goal of an estimator is to recover the top k papers. A natural way to do so is to compute the estimates of true paper scores θ^*_j and return top k papers with respect to these estimated scores. The described estimation procedure is a significantly simplified version of what is happening in the real-world conferences. Nevertheless, this fully-automated procedure may serve as a guideline for area chairs, providing a first-order estimate of the total ranking of submitted papers. In what follows, we refer to any estimator as $\widehat{\theta}$ and to the estimated score of any paper j as $\widehat{\theta}_j$. Specifically, we consider the following two estimators:

- Maximum likelihood estimator (MLE) $\widehat{\theta}^{\mathrm{MLE}}$

$$\widehat{\theta}^{\mathrm{MLE}}_j = \frac{1}{\sum\limits_{i \in \mathcal{R}_A(j)} \frac{1}{\sigma^2_{ij}}} \sum_{i \in \mathcal{R}_A(j)} \frac{y_{ij}}{\sigma^2_{ij}} \sim \mathcal{N}\left(\theta^*_j, \frac{1}{\sum\limits_{i \in \mathcal{R}_A(j)} \frac{1}{\sigma^2_{ij}}}\right). \tag{3.12}$$

  Under the model (3.11), $\widehat{\theta}^{\mathrm{MLE}}_j$ is known to have minimal variance across all linear unbiased estimations. The choice of $\widehat{\theta}^{\mathrm{MLE}}$ follows a paradigm that more experienced reviewers should have higher weight in decision making.

- Mean score estimator (MEAN) $\widehat{\theta}^{\mathrm{MEAN}}$

$$\widehat{\theta}^{\mathrm{MEAN}}_j = \frac{1}{\lambda} \sum_{i \in \mathcal{R}_A(j)} y_{ij} \sim \mathcal{N}\left(\theta^*_j, \frac{1}{\lambda^2} \sum_{i \in \mathcal{R}_A(j)} \sigma^2_{ij}\right). \tag{3.13}$$

  The mean score estimator is convenient in practice because it is not tied to the assumed statistical model, and in the past has been found to be predictive of final acceptance decisions in peer-review settings such as National Science Foundation grant proposals (Cole et al., 1981) and homework grading (Sajjadi et al., 2016). This observation is supported by the program chair of ICML 2012 John Langford, who notices in his blog (Langford, 2012b) that in ICML 2012 the decisions on the acceptance were "surprisingly uniform as a function of average score in reviews".

6.3 Analysis

Here we present statistical guarantees for both $\widehat{\theta}^{\mathrm{MLE}}$ and $\widehat{\theta}^{\mathrm{MEAN}}$ estimators and for both exact top k recovery and recovery under a Hamming error tolerance.

6.3.1 Exact top k recovery

Let us use (k) and $(k+1)$ to denote the indices of the papers that are respectively ranked k^{th} and $(k+1)^{\mathrm{th}}$ according to their true qualities. Similar to the past work by Shah and Wainwright (2015) on top k item recovery, a central quantity in our analysis is a *k-separation threshold* Δ_k defined as:

$$\Delta_k := \theta^*_{(k)} - \theta^*_{(k+1)} > 0. \tag{3.14}$$

Intuitively, if the difference between k^{th} and $(k+1)^{\mathrm{th}}$ papers is large enough, it should be easy to recover top k papers. To formalize this intuition, for any value of a parameter $\delta \geq 0$, consider a family $\mathcal{F}_k$ of papers' scores

$$\mathcal{F}_k(\delta) := \left\{(\theta_1, \ldots, \theta_m) \in \mathbb{R}^m \middle| \theta_{(k)} - \theta_{(k+1)} \geq \delta\right\}. \tag{3.15}$$

For the first half of this section, we assume that function h is bounded, that is, $h : [0,1] \to [0,1]$.[5] This assumption implicitly assumes that every reviewer $i \in [n]$ can provide a minimum level of expertise while reviewing any paper $j \in [m]$ even if she/he has zero similarity $s_{ij} = 0$ with that paper.

In addition to the gap Δ_k, the hardness of the problem also depends on the similarities between reviewers and papers. For instance, if all reviewers have near-zero similarity with all the papers, then recovery is impossible unless the gap is extremely large. In order to quantify the tractability of the problem in terms of the similarities we introduce the following set $\mathcal{S}$ of families of similarity matrices parameterized by a non-negative value q:

$$\mathcal{S}(q) := \left\{ S \in [0,1]^{n \times m} \middle| \Gamma^S_{1-h}\left(A^{\mathrm{HARD}}_{1-h}\right) \geq q \right\}. \tag{3.16}$$

In words, if similarity matrix S belongs to $\mathcal{S}(q)$, then the fairness of the optimally fair (with respect to $f = 1 - h$) assignment is at least q.

Finally, we define a quantity τ_q that captures the quality of approximation provided by PeerReview4All:

$$\tau_q := \inf_{S \in \mathcal{S}(q)} \frac{\Gamma^S_{1-h}\left(A^{\mathrm{PR4A}}_{1-h}\right)}{\Gamma^S_{1-h}\left(A^{\mathrm{HARD}}_{1-h}\right)}. \tag{3.17}$$

Note that Theorem 1 gives lower bounds on the value of τ_q.

Having defined all the necessary notation, we are ready to present the first result of this section on recovering the set of top k papers $\mathcal{T}^*_k$.

Theorem 2. *(a) For any $\epsilon \in (0, 1/4)$, $q \in [\lambda(1 - h(0)), \lambda]$ and any monotonically decreasing $h : [0,1] \to [0,1]$, if $\delta > \frac{2\sqrt{2}}{\lambda}\sqrt{(\lambda - q\tau_q)\ln\frac{m}{\sqrt{\epsilon}}}$, then for $\left(A, \widehat{\theta}\right) \in \left\{ \left(A^{PR4A}_{1-h}, \widehat{\theta}^{MEAN}\right), \left(A^{PR4A}_{h^{-1}}, \widehat{\theta}^{MLE}\right) \right\}$*

$$\sup_{\substack{(\theta^*_1,\ldots,\theta^*_m) \in \mathcal{F}_k(\delta) \\ S \in \mathcal{S}(q)}} \mathbb{P}\left\{ \mathcal{T}_k\left(A, \widehat{\theta}\right) \neq \mathcal{T}^*_k \right\} \leq \epsilon. \tag{3.18}$$

(b) Conversely, for any continuous strictly monotonically decreasing $h : [0,1] \to [0,1]$ and any $q \in [\lambda(1 - h(0)), \lambda]$, there exists a universal constant $c > 0$ such that if $m > 6$ and $\delta < \frac{c}{\lambda}\sqrt{(\lambda - q)\ln m}$, then

$$\sup_{S \in \mathcal{S}(q)} \inf_{\left(\widehat{\theta}, A \in \mathcal{A}\right)} \sup_{(\theta^*_1,\ldots,\theta^*_m) \in \mathcal{F}_k(\delta)} \mathbb{P}\left\{ \mathcal{T}_k\left(A, \widehat{\theta}\right) \neq \mathcal{T}^*_k \right\} \geq \frac{1}{2}.$$

Remarks. 1. The PeerReview4All assignment algorithm thus leads to a strong minimax guarantee on the recovery of the top k papers: the upper and lower bounds differ by at most a $\tau_q \geq \frac{1}{\lambda}$ term in the requirement on δ and constant pre-factor. Also note that as discussed in Section 5.1, approximation factor τ_q of the PeerReview4All algorithm can be much better than $1/\lambda$ for various similarity matrices.

2. In addition to quantifying the performance of PeerReview4All, an important contribution of Theorem 2 is a sharp minimax analysis of the performance of *every* assignment algorithm. Indeed, the approximation ratio τ_q (3.17) can be defined for any assignment algorithm, by substituting corresponding assignment instead of A^{PR4A}_{1-h}. For example, if one has access to the optimal assignment A^{HARD} (e.g., by using PeerReview4All if $\lambda = 1$) then we will have corresponding approximation ratio $\tau_q = 1$ thereby yielding bounds that are sharp up to constant pre-factors.

3. While on one hand the estimator $\widehat{\theta}^{\mathrm{MLE}}$ is preferred over $\widehat{\theta}^{\mathrm{MEAN}}$ when model (3.11) is correct, on the other hand, if $h(s) \in [0,1]$, then the estimator $\widehat{\theta}^{\mathrm{MEAN}}$ is more robust to model mismatches.

4. The technical assumption $q \in [\lambda(1 - h(0)), \lambda]$ is made without loss of any generality, because values of q outside this range are vacuous. In more detail, for any similarity matrix $S \in [0,1]^{n \times m}$, it must be

[5] More generally, we could consider bounded function h with range $[0, c]$ for some $c > 0$. Without loss of generality, we set $c = 1$ which can always be achieved by appropriate scaling.

that $\Gamma^S_{1-h}\left(A^{\mathrm{HARD}}_{1-h}\right) \geq \lambda\,(1-h(0))$. Moreover, the co-domain of function h comprises only non-negative real values, implying that $\Gamma^S_{1-h}\left(A^{\mathrm{HARD}}_{1-h}\right) \leq \lambda$ for any similarity matrix $S \in [0,1]^{n\times m}$.

5. The upper bound of the theorem holds for a slightly more general model of reviewers — reviewers with sub-Gaussian noise. Formally, in addition to the Gaussian noise model (3.11), the proof of Theorem 2(a) also holds for the following class of distributions of the score y_{ij}:

$$y_{ij} = \theta^*_{ij} + sG\left(h(s_{ij})\right), \tag{3.19}$$

where $sG\left(\sigma^2\right)$ is an arbitrary mean zero sub-Gaussian random variable with scale parameter σ^2.

The conditions of Theorem 2 require function h to be bounded. We now relax our earlier boundedness assumption on h and consider $h : [0,1] \to [0,\infty)$.

In what follows we restrict our attention to MLE estimator $\widehat{\theta}^{\mathrm{MLE}}$ which represents the paradigm that reviewers with higher similarity should have more weight in the final decision. In order to demonstrate that our PeerReview4All algorithm is able to adapt to different structures of similarity matrices — from hard cases when optimal assignment provides only one strong reviewer for some of the papers, to ideal cases when there are λ strong reviewers for every paper — let us consider the following set $\mathcal{S}_\kappa$ of families of similarity matrices parametrized by a non-negative value v and integer parameter $\kappa \in [\lambda]$:

$$\mathcal{S}_\kappa(v) := \left\{ S \in [0,1]^{n\times m} \middle| s^*_\kappa \geq v \right\}. \tag{3.20}$$

Here s^*_κ is as defined in (3.6).

In words, the parameter v defines the notion of strong reviewer while parameter κ denotes the maximum number of strong (with similarity higher than v) reviewers that can be assigned to each paper without violating the (μ,λ) conditions.

Then the following adaptive analogue of Theorem 2 holds:

Corollary 2. *(a) For any $\epsilon \in (0,1/4)$, $v \in [0,1]$, $\kappa \in [\lambda]$ and any monotonically decreasing $h : [0,1] \to [0,\infty)$, if $\delta > 2\sqrt{2}\sqrt{\frac{h(v)h(0)}{\kappa h(0)+(\lambda-\kappa)h(v)}}\ln\frac{m}{\sqrt{\epsilon}}$, then*

$$\sup_{\substack{(\theta^*_1,\ldots,\theta^*_m)\in\mathcal{F}_k(\delta)\\ S\in\mathcal{S}_\kappa(v)}} \mathbb{P}\left\{\mathcal{T}_k(A^{PR4A}_{h^{-1}},\widehat{\theta}^{MLE}) \neq \mathcal{T}^*_k\right\} \leq \epsilon.$$

(b) Conversely, for any continuous strictly monotonically decreasing $h : [0,1] \to [0,\infty)$, any $v \in [0,1]$, and any $\kappa \in [\lambda]$, there exists a universal constant $c > 0$ such that if $m > 6$ and $\delta \leq c\sqrt{\frac{h(v)h(0)}{\kappa h(0)+(\lambda-\kappa)h(v)}}\ln m$, then

$$\sup_{S\in\mathcal{S}_\kappa(v)} \inf_{\left(\widehat{\theta},A\in\mathcal{A}\right)} \sup_{(\theta^*_1,\ldots,\theta^*_m)\in\mathcal{F}_k(\delta)} \mathbb{P}\left\{\mathcal{T}_k(A,\widehat{\theta}) \neq \mathcal{T}^*_k\right\} \geq \frac{1}{2}.$$

Remarks. 1. Observe that there is no approximation factor in the upper bound. Thus, the PeerReview4All algorithm together with $\widehat{\theta}^{\mathrm{MLE}}$ are simultaneously minimax optimal up to a constant pre-factor in classes of similarity matrices $\mathcal{S}_\kappa(v)$ for all $\kappa \in [\lambda]$, $v \in [0,1]$.

2. Corollary 2(a) remains valid for generalized sub-Gaussian model of reviewer (3.19).

3. Corollary 2 together with Theorem 2 show that our PeerReview4All algorithm produces the assignment $A^{\mathrm{PR4A}}_{h^{-1}}$ which is simultaneously minimax (near-)optimal for various classes of similarity matrices. We thus see that our PeerReview4All algorithm is able to adapt to the underlying structure of similarity matrix S in order to construct an assignment in which even the most disadvantaged paper gets reviewers with sufficient expertise to estimate the true quality of the paper.

6.3.2 Approximate recovery under Hamming error

Although our ultimate goal is to recover set $\mathcal{T}_k^*$ of top k papers exactly, we note that often scores of boundary papers are close to each other so it may be impossible to distinguish between the k^{th} and $(k+1)^{\text{th}}$ papers in the total ranking. Thus, a more realistic goal would be to try to accept papers such that the set of accepted papers is in some sense "close" to the set $\mathcal{T}_k^*$. In this work we consider the standard notion of Hamming distance (3.1) as a measure of closeness. We are interested in minimizing the quantity:

$$\mathbb{P}\left\{\mathcal{D}_H\left(\mathcal{T}_k\left(A,\widehat{\theta}\right),\mathcal{T}_k^*\right)>2t\right\}$$

for some user-defined value of $t \in [k-1]$.

Similar to the exact recovery setup, the key role in the analysis is played by generalized separation threshold (compare with equation 3.14):

$$\Delta_{k,t} := \theta^*_{(k-t)} - \theta^*_{(k+t+1)},$$

where $(k-t)$ and $(k+t+1)$ are indices of papers that take $(k-t)^{\text{th}}$ and $(k+t+1)^{\text{th}}$ positions respectively in the underlying total ranking. For any value of $\delta > 0$ we consider the following generalization of the set $\mathcal{F}_k(\delta)$ defined in (3.15):

$$\mathcal{F}_{k,t}(\delta) := \left\{(\theta_1,\dots,\theta_m)\in\mathbb{R}^m \middle| \theta_{(k-t)} - \theta_{(k+t+1)} \geq \delta\right\}.$$

Also recall the family of matrices $\mathcal{S}(q)$ from (3.16) and the approximation factor τ_q from (3.17) for any parameter q. With this notation in place, we now present the analogue of Theorem 2 in case of approximate recovery under the Hamming error.

Theorem 3. *(a) For any $\epsilon \in (0,1/4)$, $q \in [\lambda(1-h(0)),\lambda]$, $t \in [k-1]$, and any monotonically decreasing $h:[0,1]\to[0,1]$, if $\delta > \frac{2\sqrt{2}}{\lambda}\sqrt{(\lambda - q\tau_q)\ln\frac{m}{\sqrt{\epsilon}}}$, then for $\left(A,\widehat{\theta}\right) \in \left\{\left(A_{1-h}^{PR4A},\widehat{\theta}^{MEAN}\right),\left(A_{h^{-1}}^{PR4A},\widehat{\theta}^{MLE}\right)\right\}$*

$$\sup_{\substack{(\theta_1^*,\dots,\theta_m^*)\in\mathcal{F}_{k,t}(\delta)\\ S\in\mathcal{S}(q)}} \mathbb{P}\left\{\mathcal{D}_H\left(\mathcal{T}_k\left(A,\widehat{\theta}\right),\mathcal{T}_k^*\right)>2t\right\} \leq \epsilon.$$

(b) Conversely, for any continuous strictly monotonically decreasing $h:[0,1]\to[0,1]$, any $q\in[\lambda(1-h(0)),\lambda]$, and any $0<t<k$, there exists a universal constant $c>0$ such that for given constants $\nu_1\in(0;1)$ and $\nu_2\in(0,1)$ if $2t \leq \frac{1}{1+\nu_2}\min\left\{m^{1-\nu_1},k,m-k\right\}$ and $\delta \leq \frac{c}{\lambda}\sqrt{(\lambda-q)\nu_1\nu_2\ln m}$, then for m larger than some (ν_1,ν_2)-dependent constant,

$$\sup_{S\in\mathcal{S}(q)}\inf_{\left(\widehat{\theta},A\in\mathcal{A}\right)}\sup_{\left(\theta_1^*,\dots,\theta_m^*\right)\in\mathcal{F}_{k,t}(\delta)} \mathbb{P}\left\{\mathcal{D}_H\left(\mathcal{T}_k\left(A,\widehat{\theta}\right),\mathcal{T}_k^*\right)>2t\right\} \geq \frac{1}{2}.$$

Remarks. This theorem provides a strong minimax characterization of the PeerReview4All algorithm for approximate recovery. Note that upper and lower bounds differ by the approximation factor τ_q, which is at most $\frac{1}{\lambda}$, and a pre-factor which depends only on the constants ν_1 and ν_2.

To conclude the section, we state the result for the family $\mathcal{S}_\kappa(v)$ of similarity matrices defined in (3.20) for any parameter v, showing that adaptive behavior of PeerReview4All algorithm (Corollary 2) also carries over to the Hamming error metric.

Corollary 3. *(a) For any $\epsilon\in(0,1/4)$, $v\in[0,1]$, $\kappa\in[\lambda]$, $t\in[k-1]$, and any monotonically decreasing $h:[0,1]\to[0,\infty)$, if $\delta > 2\sqrt{2}\sqrt{\frac{h(v)h(0)}{\kappa h(0)+(\lambda-\kappa)h(v)}\ln\frac{m}{\sqrt{\epsilon}}}$, then*

$$\sup_{\substack{(\theta_1^*,\dots,\theta_m^*)\in\mathcal{F}_{k,t}(\delta)\\ S\in\mathcal{S}_\kappa(v)}} \mathbb{P}\left\{\mathcal{D}_H\left(\mathcal{T}_k\left(A_{h^{-1}}^{PR4A},\widehat{\theta}^{MLE}\right),\mathcal{T}_k^*\right)>2t\right\} \leq \epsilon.$$

(b) Conversely, for any continuous strictly monotonically decreasing $h : [0,1] \to [0,\infty)$, any $v \in [0,1]$, $\kappa \in [\lambda]$ and any $t \in [k-1]$, there exists a universal constant $c > 0$ such that for given constants $\nu_1 \in (0;1)$ and $\nu_2 \in (0,1)$ if $2t \le \frac{1}{1+\nu_2} \min\left\{m^{1-\nu_1}, k, m-k\right\}$ and $\delta \le c\sqrt{\frac{h(v)h(0)}{\kappa h(0)+(\lambda-\kappa)h(v)}\nu_1\nu_2 \ln m}$, then for m larger than some (ν_1,ν_2)-dependent constant,

$$\sup_{S\in\mathcal{S}_\kappa(v)} \inf_{\left(\widehat{\theta}, A\in\mathcal{A}\right)} \sup_{\left(\theta_1^*,\ldots,\theta_m^*\right)\in\mathcal{F}_{k,t}(\delta)} \mathbb{P}\left\{\mathcal{D}_H\left(\mathcal{T}_k\left(A,\widehat{\theta}\right), \mathcal{T}_k^*\right) > 2t\right\} \ge \frac{1}{2}.$$

The results established in this section thus show that our PeerReview4All algorithm produces an assignment which is minimax (near-)optimal for both exact and approximate recovery of the top k papers.

7 Subjective-score model

In the previous section, we analyzed the performance of our PeerReview4All assignment algorithm under a model with objective scores. Indeed, various past works on peer-review (as well as various other domains of machine learning) assume existence of some "true" objective scores or ranking of the underlying items (papers). However, in practice, reviewers' opinions on the quality of any paper are typically highly subjective (Kerr et al., 1977; Mahoney, 1977; Ernst and Resch, 1994; Bakanic et al., 1987; Lamont, 2009). Even two highly experienced researchers with vast experience and expertise may have considerably differing opinions about the contributions of a paper. Following this intuition, we wish to move away from the assumption of some true objective scores $\{\theta_j^*\}_{j\in[m]}$ of the paper.

With this motivation, in this section we develop a novel model to capture such subjective opinions and present a statistical analysis of our assignment algorithm under this subjective-score model.

7.1 Model

The key idea behind our subjective score model is to separate out the subjective part in any reviewer's opinion from the noise inherent in it. Our model is best described by first considering a hypothetical situation where every reviewer *spends an infinite time and effort on reviewing every paper, gaining a perfect expertise in the field of that paper and a perfect understanding of the paper's content.* We let $\widetilde{\theta}_{ij} \in \mathbb{R}$ denote the score that this fully competent version of reviewer $i \in [n]$ would provide to paper $j \in [m]$, and denote the matrix of reviewers subjective scores as $\widetilde{\Theta} = \left\{\widetilde{\theta}_{ij}\right\}_{i\in[n], j\in[m]}$. Continuing momentarily in this hypothetical world, when all the reviewers are fully competent in evaluating all the papers, every feasible reviewer-assignment is of the same quality since there is no noise in the reviewers' scores. Since all reviewers have an equal, full competence, a natural choice of scoring any paper $j \in [m]$ is to take the mean score provided by the fully competent reviewers who review that paper:

$$\widetilde{\theta}_j^\star(A) := \frac{1}{\lambda} \sum_{i\in\mathcal{R}_A(j)} \widetilde{\theta}_{ij}. \tag{3.21}$$

Let us now exit our hypothetical world and return to reality. In a real conference peer-review setting the reviews will be noisy. Following the previous noise assumptions, we assume that score of any reviewer $i \in [n]$ for any paper $j \in [m]$ that she/he reviews is distributed as

$$y_{ij} \sim \mathcal{N}(\widetilde{\theta}_{ij}, h(s_{ij})),$$

for some known continuous strictly monotonically decreasing function $h : [0,1] \to [0,1]$. Under this model, the higher the similarity s_{ij}, the better the score y_{ij} represents the subjective score $\widetilde{\theta}_{ij}$ which reviewer $i \in [n]$ would give to paper $j \in [m]$ if she/he had infinite expertise.

The goal under this model is to assign reviewers to papers such that reviewers are of enough ability to convey their opinions $\widetilde{\theta}_{ij}$ from the hypothetical full-competence world to the real world with scores y_{ij}. In

other words, the goal of the assignment is to ensure the recovery of the top k papers in terms of the mean full-competence subjective scores $\{\widetilde{\theta}_j^\star\}_{j\in[m]}$.

7.2 Analysis

In this section we present statistical guarantees for $\widehat{\theta}^{\mathrm{MEAN}}$ in context of subjective-score model.

7.2.1 Exact top k recovery

Since the true scores for any reviewer-paper pair are subjective, and since we are interested in mean full-competence subjective scores, a natural choice for estimating $\{\widetilde{\theta}_j^\star\}$ from the actual provided scores $\{y_{ij}\}$ is the averaging estimator $\widehat{\theta}^{\mathrm{MEAN}}$ which for every paper $j \in [m]$ estimates $\widetilde{\theta}_j^\star$ as $\widehat{\theta}_j^{\mathrm{MEAN}} = \frac{1}{\lambda} \sum\limits_{i \in \mathcal{R}_A(j)} y_{ij}$. Having defined the model and estimator, we now provide a sharp minimax analysis for the subjective-score model. In order to state our main result, we recall the family of similarity matrices $\mathcal{S}(q)$ defined earlier in (3.16) and the approximation ratio τ_q defined in (3.17), both parameterized by some non-negative value q.

Note that the notion of the k-separation threshold (3.14) does not carry over directly from the objective score model to the subjective score model. The reason is that the ranking now is induced by the assignment and changes as we change the assignment. Consequently, we introduce the following family of papers' scores that are governed by the assignment A and parametrized by a positive real value δ:

$$\mathcal{F}_k(A,\delta) = \left\{ \widetilde{\Theta} \in \mathbb{R}^{n\times m} \middle| \widetilde{\theta}^\star_{(k)}(A) - \widetilde{\theta}^\star_{(k+1)}(A) \geq \delta \right\}. \tag{3.22}$$

Since in this section we consider only mean score estimator $\widehat{\theta}^{\mathrm{MEAN}}$, we omit index $1-h$ from A^{PR4A}_{1-h}, but always imply that assignment A^{PR4A} is built with respect to the function $1-h$. For every feasible assignment A, we augment the notation $\mathcal{T}_k^*$ with $\mathcal{T}_k^\star\left(A, \widetilde{\theta}^\star(A)\right)$ to highlight that the set of the top k papers is induced by the assignment A. Let us now present the main result of this section.

Theorem 4. *(a) For any $\epsilon \in (0,1/4)$, $q \in [\lambda(1-h(0)), \lambda]$ and any monotonically decreasing $h:[0,1]\to[0,1]$, if $\delta > \frac{2\sqrt{2}}{\lambda}\sqrt{(\lambda - q\tau_q)\ln\frac{m}{\sqrt{\epsilon}}}$, then*

$$\sup_{\substack{\widetilde{\Theta}\in\mathcal{F}_k(A^{PR4A},\delta)\\ S\in\mathcal{S}(q)}} \mathbb{P}\left\{\mathcal{T}_k(A^{PR4A},\widehat{\theta}^{MEAN}) \neq \mathcal{T}_k^\star\left(A^{PR4A},\widetilde{\theta}^\star(A^{PR4A})\right)\right\} \leq \epsilon.$$

(b) Conversely, for any continuous strictly monotonically decreasing $h:[0,1]\to[0,1]$ and any $q \in [\lambda(1-h(0)),\lambda]$, there exists a universal constant $c>0$ such that if $m>6$ and $\delta < \frac{c}{\lambda}\sqrt{(\lambda-q)\ln m}$, then

$$\sup_{S\in\mathcal{S}(q)} \inf_{\left(\widehat{\theta},A\in\mathcal{A}\right)} \sup_{\widetilde{\Theta}\in\mathcal{F}_k(A,\delta)} \mathbb{P}\left\{\mathcal{T}_k(A,\widehat{\theta}) \neq \mathcal{T}_k^\star\left(A,\widetilde{\theta}^\star(A)\right)\right\} \geq \frac{1}{2}.$$

We thus see that our assignment algorithm PeerReview4All not only leads to the strong guarantees under the objective-score model but simultaneously also under the setting where the opinions of reviewers may be subjective.

7.2.2 Approximate recovery under Hamming error

We now present guarantees for approximate recovering under the Hamming error for the PeerReview4All algorithm. We generalize the family of score matrices (3.22), for which we consider any integer error tolerance parameter $t \in \{0,\ldots,k-1\}$ and any any feasible assignment A. Then we define the following family of subjective papers' scores, parameterized by non-negative value δ:

$$\mathcal{F}_{k,t}(A,\delta) = \left\{ \widetilde{\Theta} \in \mathbb{R}^{n\times m} \middle| \widetilde{\theta}^\star_{(k-t)}(A) - \widetilde{\theta}^\star_{(k+t+1)}(A) \geq \delta \right\}.$$

Observe that the class $\mathcal{F}_{k,t}(A,\delta)$ coincides with the class $\mathcal{F}_k(\delta)$ from (3.22) when $t=0$.

Theorem 5. *(a) For any $\epsilon \in (0, 1/4)$, $q \in [0, \lambda]$, $t \in [k-1]$, and any monotonically decreasing $h : [0, 1] \to [0, 1]$, if $\delta > \frac{2\sqrt{2}}{\lambda}\sqrt{(\lambda - q\tau_q)\ln\frac{m}{\sqrt{\epsilon}}}$, then*

$$\sup_{\substack{\widetilde{\Theta}\in\mathcal{F}_{k,t}(A^{PR4A},\delta)\\ S\in\mathcal{S}(q)}} \mathbb{P}\left\{\mathcal{D}_H\left(\mathcal{T}_k\left(A^{PR4A},\widehat{\theta}\right),\mathcal{T}_k^*\left(A^{PR4A},\widetilde{\theta}^\star(A^{PR4A})\right)\right) > 2t\right\} \le \epsilon.$$

Conversely, for any continuous strictly monotonically decreasing $h : [0, 1] \to [0, 1]$, any $q \in [\lambda(1 - h(0)), \lambda]$, and any $0 < t < k$, there exists a universal constant $c > 0$ such that for given constants $\nu_1 \in (0, 1)$ and $\nu_2 \in (0, 1)$ if $2t \le \frac{1}{1+\nu_2}\min\left\{m^{1-\nu_1}, k, m-k\right\}$ and $\delta \le \frac{c}{\lambda}\sqrt{(\lambda - q)\nu_1\nu_2 \ln m}$, then for m larger than some (ν_1, ν_2)-dependent constant,

$$\sup_{S\in\mathcal{S}(q)}\inf_{(\widehat{\theta},A\in\mathcal{A})}\sup_{\widetilde{\Theta}\in\mathcal{F}_{k,t}(A,\delta)} \mathbb{P}\left\{\mathcal{D}_H\left(\mathcal{T}_k\left(A,\widehat{\theta}\right),\mathcal{T}_k^*\left(A,\widetilde{\theta}^\star(A)\right)\right) > 2t\right\} \ge \frac{1}{2}.$$

Similar to Theorem 4, Theorem 5 shows that PeerReview4All algorithm is minimax optimal up to a constant pre-factor and approximation factor *given* that reviewers' subjective scores $\widetilde{\Theta}$ belong to the class $\mathcal{F}_{k,t}(A, \delta)$.

8 Experiments

In this section we conduct empirical evaluations of the PeerReview4All algorithm and compare it with the TPMS (Charlin and Zemel, 2013), ILPR (Garg et al., 2010) and Hard algorithms. Our implementation of the PeerReview4All algorithm picks max-flow with maximum cost in Step 6 of Subroutine 1.

Previous work on the conference paper assignment problem (Garg et al., 2010; Long et al., 2013; Karimzadehgan et al., 2008; Tang et al., 2010) conducted evaluations of the proposed algorithms in terms of various objective functions that measure the quality of the assignment. For example, Garg et al. (2010) compared fairness from reviewers' perspective using the number of satisfied bids as a criteria. While these evaluations allow to compare algorithms in terms of particular objective, we note that the main goal of the peer-review system is to accept the best papers. It is not straightforward whether an improvement of some other objective will lead to the improvement of the quality of the paper acceptance process.

In contrast to the prior works, in this section we not only consider the fairness objective (Subsections 8.2 and 8.3), but also design experiments (Subsections 8.1 and 8.4) to directly evaluate the accuracy resulting from the assignment procedures.

8.1 Synthetic simulations

We begin with synthetic simulations. We consider the instance of the reviewer assignment problem with $m = n = 100$ and $\lambda = \mu = 4$. We select the moderate values of m and n to keep track of the optimal assignment A^{HARD} which we find as a solution of the corresponding integer linear programming problem. For every real-valued constant c, we denote the matrix with all entries being equal to c as $\mathbf{c}$. Similarly, we denote the matrix with entries independently sampled from a Beta distribution with parameters (α, β) as $\mathcal{B}(\alpha, \beta)$.

We consider the objective-score model of reviewers (3.11) with $h(s) = 1 - s$ together with estimator $\widehat{\theta}^{\mathrm{MLE}}$. Thus, assignments A^{PR4A}, A^{ILPR} and A^{HARD} aim to optimize $\Gamma^S_{(1-s)^{-1}}(A)$ while assignment A^{TPMS} aims to maximize the cumulative sum of similarities $G^S(A)$ as defined in (3.2).

In what follows we simulate the following problem instances:

(C1) **Non-mainstream papers.** There are $m_1 = 80$ conventional papers for which there exist $n_1 = 80$ expert reviewers with high similarity, and $m_2 = 20$ non-mainstream papers for which all the reviewers have similarity smaller than or equal to 0.5. There are also $n_2 = 20$ weak reviewers who have moderate

	Fairness $\Gamma^S_{(1-s)^{-1}}(A)$					Sum of Similiarities $G^S(A)$				
	Case 1	Case 2	Case 3	Case 4	Case 5	Case 1	Case 2	Case 3	Case 4	Case 5
A^{TPMS}	4.7	5.1	13.3	4.0	10.9	300	168	295	296	311
A^{HARD}	8.0	13.1	26.6	14.0	10.9	296	162	232	234	175
A^{ILPR}	8.0	5.0	4.0	14.0	10.9	296	165	188	293	296
A^{PR4A}	8.0	13.1	22.0	6.5	10.9	296	166	239	290	309

Table 2: Comparison of assignment produced by PeerReview4All, Hard, ILPR and TPMS algorithms in terms of the fairness and the sum of similarities (higher values are better).

similarities with papers from the first group and low similarities with papers from the second group. The similarities are given by the block matrix:

$$S_1 = \left[\begin{array}{c|c} \mathbf{0.9} & \mathbf{0.5} \\ \hline \underbrace{\mathbf{0.5}}_{80} & \underbrace{\mathbf{0.15}}_{20} \end{array}\right] \begin{matrix} \} 80 \\ \} 20 \end{matrix}$$

(C2) **Many weak reviewers.** In this scenario there are $n_1 = 25$ strong reviewers with high similarity with every paper and $n_2 = 75$ weak reviewers with small similarity with every paper:

$$S_2 = \left[\begin{array}{c} \mathbf{0.8} + 0.2 \times \mathcal{B}(1,3) \\ \hline \underbrace{\mathbf{0.1} + 0.2 \times \mathcal{B}(1,3)}_{100} \end{array}\right] \begin{matrix} \} 25 \\ \} 75 \end{matrix}$$

(C3) **Few super-strong reviewers.** The following example tests the algorithms in scenario when some small number of the reviewers are much stronger than the others. Similarities for this scenario are given by the block matrix:

$$S_3 = \left[\begin{array}{c|c} \mathbf{0.98} & \mathbf{0.9} \\ \hline \mathbf{0} & \mathbf{0.7} \\ \hline \underbrace{\mathbf{0.9}}_{60} & \underbrace{\mathbf{0.9}}_{40} \end{array}\right] \begin{matrix} \} 10 \\ \} 50 \\ \} 40 \end{matrix}$$

(C4) **Adverse case.** Having analyzed the inner working of our PeerReview4All algorithm, we construct a similarity matrix which is hard for the algorithm to compute the fair assignment.[6]

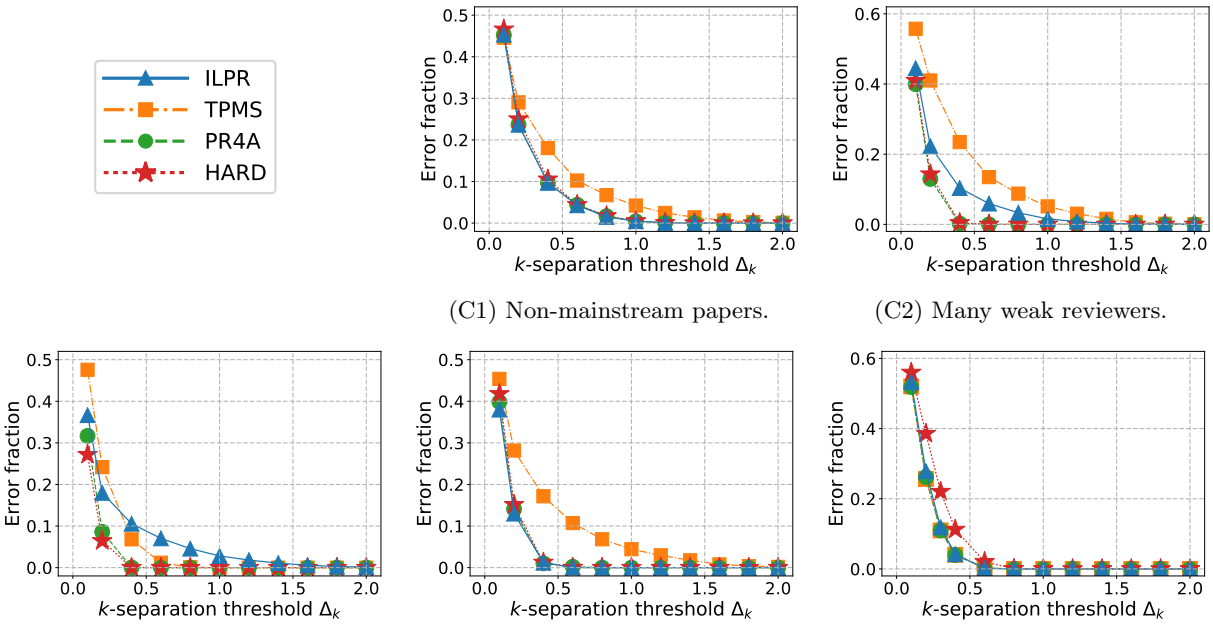
(C5) **Sparse similarities.** Each entry of similarity matrix S_5 is zero with probability 0.8 or otherwise is drawn independently and uniformly at random from $[0.1, 0.9]$.

8.1.1 Fairness

In this section we analyze the quality of assignments produced by PeerReview4All, Hard, ILPR and TPMS algorithms and for all the five cases described above. The results are summarized in Table 2 where we compute the measures of fairness $\Gamma^S_{(1-s)^{-1}}(A)$ and the conventional sum of similarities $G^S(A)$ for each of the assignments.

The results in Table 2 show that in all five cases PeerReview4All algorithm finds an assignment A^{PR4A} with at least as much fairness as A^{TPMS}. At the same time, the max cost heuristic that we use in Step 6 of Subroutine 1 helps the average quality (total sum similarity) of the assignment A^{PR4A} to be either close to or larger than average quality of both A^{ILPR} and A^{HARD}.

[6] We do not give an explicit expression of the matrix S_4 for this case, due to its complicated structure.



(C1) Non-mainstream papers.

(C2) Many weak reviewers.

(C3) Few super-strong reviewers.

(C4) Adverse case.

(C5) Sparse similarities.

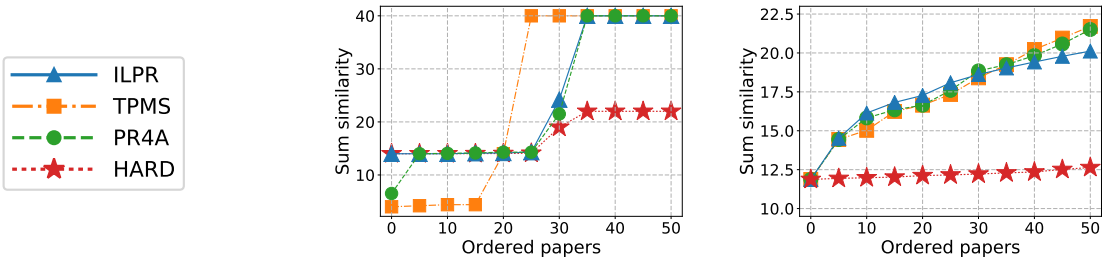
Figure 1: Fraction of papers incorrectly accepted by $\widehat{\theta}^{\mathrm{MLE}}$ based on assignments produced by PeerReview4All, Hard, ILPR and TPMS for different values of the separation threshold.

In Case (C1), the TPMS algorithm sacrifices the quality of reviewers for non-mainstream papers, assigning them to weak reviewers. In contrast, all other algorithms assign four best possible reviewers to these unconventional papers in order to maintain fairness. In Case (C2), the PeerReview4All and Hard algorithms assign one strong reviewer for each paper while TPMS, in attempt to maximize the value of its goal function, assigns strong reviewers according to their highest similarities which leads to an unfair assignment. The ILPR algorithm fails to find a fair assignment in Cases (C2) and (C3): the poor performance of ILPR algorithm is caused by the fact that some of the reviewers in our examples have similarities close to maximal, making the value of $f(s) = \frac{1}{1-s}$ large, which, in turn, makes the approximation guarantee (3.9) of ILPR algorithm weak. In Case (C4), the PeerReview4All algorithm was unable to recover the fair assignment. Instead, the assignment within approximation ratio $1/3$, which is a bit better than the worst case $1/\lambda = 1/4$ approximation, was discovered. Finally, in Case (C5), the all algorithms managed to recover fair assignment. However, we note that the total sum similarity of the A^{HARD} assignment is low as compared to other algorithms. The reason is that the corresponding solution of the integer linear programming problem in the Hard algorithm is optimized for the fairness towards the worst-off paper and does not try to continue optimization, once the assignment for that paper is fixed. In contrast, both PeerReview4All and ILPR algorithms try to maximize the fate of the second worst-off paper, when the assignment for the most worst-off paper is fixed.

8.1.2 Statistical accuracy

As we have pointed out, the main goal of the assignment procedure is to ensure the acceptance of the k best papers $\mathcal{T}_k^*$. While in real conferences the acceptance process is complicated and involves discussions between reviewers and/or authors, here we consider a simplified scenario. Namely, we assume an objective-score model defined in Section 6 and reviewer model (3.11) with $h(s) = 1 - s$.

The experiment executes 1,000 iterations of the following procedure. We randomly choose $k = 20$ indices



(C4) Adverse case. (C5) Sparse similarities.

Figure 2: Sum similarity for the 50 most worst-off papers in assignments produced by PeerReview4All, Hard, ILPR and TPMS.

of the "true best" papers $\mathcal{T}_k^* = \{j_1, \dots, j_k\} \subset [m]$. Each of these papers $j \in \mathcal{T}_k^*$ is assigned score $\theta_j^* = 1$, while for each of the remaining papers $j \in [m] \backslash \mathcal{T}_k^*$ we set $\theta_j^* = 1 - \Delta_k$, where $\Delta_k \in (0, 2]$. Next, given the similarity matrix S, we compute assignments A^{PR4A}, A^{HARD}, A^{ILPR} and A^{TPMS}. For each of these assignments we compute the estimations of the set of top k papers using the $\widehat{\theta}^{\text{MLE}}$ estimator and calculate the fraction of wrongly accepted papers.

For every similarity matrix $S_r, r \in [5]$, and for every value of $\Delta_k \in \{0.1k \mid k \in [20]\}$, we compute the mean of the obtained values over the 1,000 iterations. Figure 1 summarizes the dependence of the fraction of incorrectly accepted papers on the value of separation threshold Δ_k for all five cases (C1)-(C5).

The obtained results suggest that the increase in fairness of the assignment leads to an increase in the accuracy of the acceptance procedure, provided that the average sum similarity of the assignment does not decrease dramatically. The PeerReview4All algorithm significantly outperforms TPMS both in terms of fairness and in terms of fraction of incorrectly accepted papers for the first four cases. The low fairness of assignments computed by ILPR in Cases (C2) and (C3) lead to the large fraction of errors in the acceptance procedure. As we noted earlier, the ILPR algorithm has weak approximation guarantees when the function f is allowed to be unbounded. In section 8.4 we will consider the mean score estimator ($f(s) = s$) which is more suitable scenario for ILPR algorithm.

Interestingly, in Case (C4), the PeerReview4All algorithm recovers sub-optimal assignment in terms of fairness, but still performs well in terms of the accuracy of the acceptance procedure. To understand this effect, for each of the assignments $A^{\text{TPMS}}, A^{\text{HARD}}, A^{\text{ILPR}}$ and A^{PR4A} we compute the sum similarity for all papers in the assignments and plot these values for 50 the most worst-off papers in each of the assignment in Figure 2. Despite the inability of PeerReview4All to find the fair assignment for the most worst-off paper, Corollary 1 guarantees that sum similarities for the remaining papers will not be too far from the optimal, and we see this aspect in Figure 2(C4). As one can see, the sum similarity for all but tiny fraction of papers in A^{PR4A} is large enough, thus ensuring the low fraction of incorrectly accepted papers.

Finally, note that in Case (C5), the Hard algorithm, while having optimal fairness, has a lower accuracy as compared to other algorithms. As Figure 2(C5) demonstrates, the Hard algorithm does not optimize for the second worst off paper and recovers sub-optimal assignment for all but the most disadvantaged paper. In contrast, as Figure 2 suggests, the ILPR and PeerReview4All algorithms do not stop their work after the most disadvantaged paper is satisfied, but instead continue to optimize the assignment for the remaining papers and eventually ensure not only fairness, but also high average quality of the assignment.

8.2 Experiment on the approximation of ICLR similarity matrix

In absence of publicly available similarity matrices from conferences, we are unable to compare the assignment computed by the PeerReview4All algorithm to the actual conference assignment. To circumvent this issue, we use an approximate version of the similarity matrix from the International Conference on Learning Representations (ICLR'18) that was constructed by Xu et al. (2019) and compare the performance of the PeerReview4All and TPMS assignment algorithms on this matrix.

Algorithm	Fairness $\Gamma^S(A)$	Mean sum of sim. $G^S(A)$
A^{TPMS}	0.12	0.413
A_1^{PR4A} (one iteration)	0.15 (+25%)	0.408 (−1%)
A^{PR4A} (full)	0.15 (+25%)	0.406 (−2%)

Table 3: Results of the experiment on the approximation of ICLR'18 similarity matrix. Values in brackets represent relative changes as compared to the TPMS assignment.

8.2.1 Matrix construction

The similarity matrix we use for comparison was constructed by Xu et al. (2019) as follows. OpenReview (`openreview.net`) — increasingly popular conference management system — maintains a public database of all papers (with author identities being visible) submitted to the ICLR'18 conference, thereby giving access to the pool of submissions. Next, it was assumed that all authors of submissions are simultaneously reviewers and that there are no additional reviewers. The publication profiles of reviewers were constructed by scraping the data from databases of scientific publications. Finally, the open-source code (`bitbucket.org/lcharlin/tpms/`) and the material of the original paper (Charlin and Zemel, 2013) were used to compute the similarity matrix according to the TPMS procedure.

The process outlined above resulted in the similarity matrix S that has $n = 2435$ reviewers and $m = 911$ papers. Additionally, it was assumed that any reviewer has a conflict of interests with the submitted papers that she/he has authored; these conflicts are represented by a binary matrix C whose $(i, j)^{\text{th}}$ entry equals 1 if and only if reviewer i has a conflict with paper j. Similarity matrix S possesses a considerable heterogeneity as demonstrated by some papers having mean similarity with non-conflicting reviewers almost four times larger than others.

The large size of the similarity matrix makes computation of the optimally fair assignment infeasible, and hence in this section we do not compute the Hard assignment. Additionally, our implementation of the ILPR assignment algorithms was computationally inefficient and in absence of the publicly available source code we also exclude this algorithm from comparison.

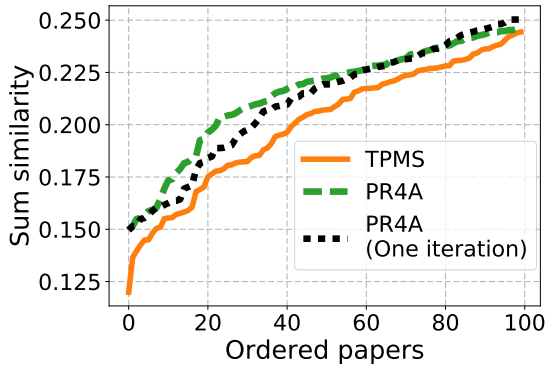
8.2.2 Evaluation

Having defined the similarity matrix and matrix of conflicts, we compute assignments of papers to reviewers with $\lambda = 4$ (each paper is assigned to 4 reviewers) and $\mu = 2$ (each reviewer is allocated at most 2 papers) using the TPMS and PeerReview4All assignment algorithms with the identity transformation function $f(s) = s$. In addition to the standard load constraints, we require that no paper is assigned to a conflicting reviewer. Finally, as pointed out in Section 5.2, the fairness guarantees of Theorem 1 are achieved after the first iteration of Steps 2 to 7 of Algorithm 1. Hence, we include the corresponding assignment for comparison and denote it as A_1^{PR4A}.

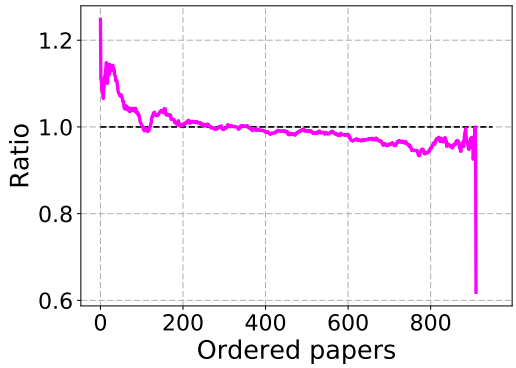
Table 3 summarizes the results of the experiment, comparing the resulting assignments in terms of fairness (3.3) and cumulative similarity (3.2). We see that the fairness of the assignment computed by the PeerReview4All algorithm is significantly higher than the fairness of the TPMS algorithm. Similar to the case of synthetic simulations, the max cost heuristic used in Step 6 of Subroutine 1 helps our algorithm to maintain a high value of cumulative similarity, which is only marginally below the optimal value.

The large size of the similarity matrix at hands makes evaluation of the optimal fairness achieved by A^{HARD} computationally prohibitive. However, we can still find an upper bound on $\Gamma^S\left(A^{\text{HARD}}\right)$ by dropping reviewer load constraints and allowing all reviewers to review unlimited number of papers. The resulting bound allows us to compute a lower bound on the approximation ratio of the PeerReview4All algorithm:

$$\frac{\Gamma^S\left(A^{\text{PR4A}}\right)}{\Gamma^S\left(A^{\text{HARD}}\right)} \geq 0.98,$$



(a) Sum similarity for 100 most disadvantaged papers in each assignment.



(b) Ratio of ordered sum similarities in A^{PR4A} to ordered sum similarities in A^{TPMS}.

Figure 3: Comparison on the approximation of ICLR'18 similarity matrix.

which shows that in practice the approximation factor of the PeerReview4All algorithm can be much better than the worst-case approximation factor $\frac{1}{\lambda}$ guaranteed by Theorem 1.

Continuing the analysis, for each of the assignments A^{TPMS}, A_1^{PR4A} and A^{PR4A} we compute the sum similarity for all papers in the assignments and plot these values for 100 the most worst-off papers in each of the assignment in Figure 3a. This figure demonstrates that while the fairness guarantees of Theorem 1 can be achieved by a single iteration of Steps 2 to 7, subsequent iterations help to improve the assignment for the second worst-off paper and so on. Finally, for each of the assignments A^{TPMS} and A^{PR4A} we sort papers in order of increasing sum similarity of assigned reviewers and plot the ratios (PeerReview4All to TPMS) of these sums in Figure 3b. Figure 3b shows that the PeerReview4All algorithm indeed balances the assignment by improving the quality for the worst-off papers at the expense of decreasing the quality for the most benefiting papers.

8.3 Experiment on MIDL and CVPR similarity matrices

Subsequent to the publication of the first version of this work, a follow-up paper by Kobren et al. (2019) has been published. There authors propose two novel assignment algorithms that also aim at ensuring the fairness of the assignment. In that work, the PeerReview4All algorithm with the identity transformation function ($f(s) = s$) was compared with other assignment algorithms on similarity matrices from three real conferences: Medical Imaging and Deep Learning Conference (MIDL), and two editions of the Conference on Computer Vision and Pattern Recognition (CVPR'17 and CVPR'18). With the kind permission of Ari Kobren, we describe the results of their experiments in which our algorithm was evaluated.

8.3.1 Brief discussion of the algorithms by Kobren et al.

We begin with a brief theoretical comparison of the PeerReview4All algorithm with the algorithms proposed by Kobren et al. (2019). Recall that the PeerReview4All algorithm aims at optimizing fairness of the assignment (3.3) and does not directly optimize for the total sum similarity. However, when in its inner workings the algorithm faces a choice between different suitable similarity matrices (Step 6 of the Subroutine 1), it can heuristically optimize for the total sum similarity by using the max cost heuristic. In contrast, Kobren et al. (2019) consider a problem of optimizing for the total sum similarity of the assignment *with an additional constraint of each paper having the sum similarity larger than some threshold* T, which can be specified by user or found by the binary search. They design two novel algorithms which we refer to as FairIr and FairFlow.

Given a feasible instance of the reviewer assignment problem, the FairIr algorithm is able to compute the assignment with the optimal value of the total sum similarity, violating the fairness constraints by an

Conference	Parameters	Algorithm	Time (s)	Fairness $\Gamma^S(A)$	Mean sum of sim. $G^S(A)$
MIDL	$n = 177,\ m = 118$ $\lambda = 3,\ \mu = 4$	A^{TPMS}	0.1	0.90	1.71
		A^{PR4A}	293.8	0.92	1.67
		A^{FairIr}	1.6	0.93	1.71
		A^{FairFlow}	1.2	0.94	1.68
CVPR'17	$n = 1373,\ m = 2623$ $\lambda = 3,\ \mu = 6$	A^{TPMS}	47	0	2.08
		A_1^{PR4A}	3251	0.77	1.96
		A^{FairIr}	595	0.27	2.05
		A^{FairFlow}	225	0.77	1.69
CVPR'18	$n = 2840,\ m = 5062$ $\lambda = 3,\ \mu = 9$	A^{TPMS}	257	1.37	22.23
		A_1^{PR4A}	8684	12.68	21.48
		A^{FairIr}	3786	7.19	22.18
		A^{FairFlow}	910	11.12	17.98

Table 4: Results of the experiment conducted by Kobren et al. on similarity matrices from real conferences. On large instances only a single iteration of the PeerReview4All algorithm was computed and the corresponding assignment is denoted A_1^{PR4A}.

additive factor which is upper bounded by the maximum entry of the similarity matrix. In that, fairness guarantees of FairIr are equivalent to those of ILPR (and hence may become vacuous when similarity matrix is significantly heterogeneous), but additionally the FairIr algorithm achieves the highest possible value of sum similarity.[7] The FairFlow algorithm is a heuristic which does not have theoretical guarantees, but in return has much lower computational complexity.

Another difference between PeerReview4All and the algorithms proposed by Kobren et al. (2019) is that both FairIr and FairFlow allow to specify a *lower bound on reviewer load*, thereby ensuring that each reviewer reviews at least some number of papers. In our work, we do not study such constraints and PeerReview4All does not support such constraints as is. Hence, below we report only those comparisons in which our algorithm was evaluated by Kobren et al. (2019), that is, the comparisons in which the lower bound on reviewer load was not enforced.

Overall, the FairIr and FairFlow algorithms aim at balancing the fairness and the total sum similarity of the assignment. By choosing an appropriate heuristic in Step 6 of the Subroutine 1, one can ensure that PeerReview4All also heuristically optimizes for the total sum similarity. Let us now report the experimental results of Kobren et al. (2019) that allows to compare the algorithms on both objectives of fairness and total sum similarity.

8.3.2 Summary of the experiments

The key summary statics of the Kobren et al. (2019) experiments are represented in Table 4.[8] For each similarity matrix, the assignments respecting the corresponding paper and reviewer load constraints were computed by the TPMS, PeerReview4All, FairIr and FairFlow algorithms. These assignments were then compared based on (a) running time of the algorithm, (b) fairness of the assignment and (c) mean sum similarity of the assignment. First, we notice that our naive implementation of the PeerReview4All algorithm is significantly slower than all other algorithms, and for large instances only a single iteration of the algorithm can be computed in a reasonable time (recall that even one iteration is sufficient to satisfy the fairness guarantees of Theorem 1). Nonetheless, even on the largest instance with more than 5,000 papers the running time of the first iteration of our algorithm took less than three hours which is still feasible given

[7] Observe that this value is lower than those achieved by TPMS as FairIr has additional constraint on the fairness of the assignment.

[8] We omit some statistics which are not of direct interest (for example, max sum similarity in the assignment).

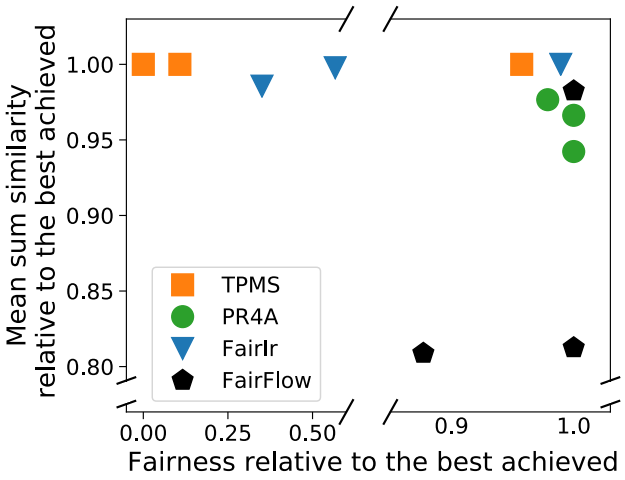


Figure 4: Visualization of comparison of the algorithms based on fairness and total sum similarity.

that the full assignment procedure needs to be run only once in the conference timeline.

The remaining two dimensions of comparison represent two notions of quality of the assignment: fairness and total sum similarity. Ideally, we would like to have an algorithm which simultaneously optimizes both of these notions. Figure 4 visualizes the comparison of the algorithms and is constructed as follows. For each of the three experiments, we compute the maximum value of fairness achieved by any of the algorithms. Using this value, for each algorithm we compute its "competetiveness" as the fairness achieved by that algorithm divided by the maximum fairness. We then repeat the same for the total sum similarity. As a result, in each experiment the performance of each algorithm can be represented as a data point in two-dimensional space where x-axis represents the competitiveness in terms of fairness and y-axis represents the competitiveness in terms of the total sum similarity.

Figure 4 demonstrates that in each of the three experiments the PeerReview4All algorithm (even with one iteration) was able to achieve maximum or close-to-maximum values of both fairness and total sum similarity. In contrast, each of the other algorithms under consideration achieved considerably lower value of either fairness or total sum similarity in two out of three experiments.

Overall, we conclude that while being considerably (but not prohibitively) slower than other algorithms, PeerReview4All managed to achieve the best balance of fairness and total sum similarity, despite optimizing the latter objective only heuristically.

8.4 Experiment on Amazon Mechanical Turk

Even if peer-review data from conferences was available to us, it would not allow for an objective evaluation of any assignment algorithm with respect to accuracy of the acceptance procedure. There are two reasons for this hinderance: (a) No ground truth ranking is available; and (b) The data contains only reviews that correspond to one particular assignment and has missing reviews for other assignments.

In this section we present an experiment which we carefully design to overcome the fundamental issues with objective empirical evaluations of reviewer assignments. Our experiment allows us to directly measure the accuracy of final decisions to evaluate any assignment. We execute our experiment on the Amazon Mechanical Turk (`mturk.com`) crowdsourcing platform.

8.4.1 Design of experiment

We designed the experiment in a manner that allows us to objectively evaluate the performance of any assignment algorithm. Specifically, the experiment should provide us access to some similarities between reviewers and papers, execute any assignment algorithm, and eventually objectively evaluate the final outcome.

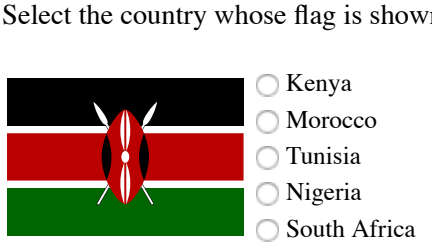


Figure 5: Question interface

The experiment considers crowdsourcing workers as reviewers and a number of general knowledge questions as papers. Specifically, 80 workers were recruited and presented with a list of 60 flags of different countries. The workers were asked to determine the country of each flag, choosing one of five options for each question. The interface of the task is represented in Figure 5. Unknown to the worker, the 60 countries comprised 10 countries each from 6 different geographic regions. Three participants did not attempt some of the questions and their responses were discarded from the dataset. The dataset is available on the website of the author of this thesis.

8.4.2 Evaluation

After obtaining the data from Amazon Mechanical Turk, we executed the following procedure for 1,000 iterations. In each of the 6 regions, we first split the 10 questions into two sets: a "gold standard" set of 8 questions chosen uniformly at random and an "unresolved" set comprising the 2 remaining questions. The set of all 12 unresolved questions are analogous to papers in the peer-review setting ($m = 12$). We computed the similarity of any worker to any paper (question) as the fraction of questions that the worker answered correctly among the 8 gold standard questions for the region corresponding to that paper (question). Having computed the similarities, we selected $n = 40$ of the workers uniformly at random and created five assignments $A^{\text{TPMS}}, A^{\text{PR4A}}$, A^{ILPR}, A^{HARD} and A^{RAND}, with identity transformation function $f(s) = s$, where A^{RAND} is a random feasible assignment. In each of these assignments, every question was answered by $\lambda = 3$ workers and every worker answered at most $\mu = 2$ questions. Finally, for each assignment, we computed the answers for the remaining $m = 12$ questions by taking a majority vote of the responses from workers assigned to each question. Ties are also considered as mistakes.

At the end of all iterations, we computed the fraction of questions whose final answers are estimated incorrectly under the five assignments as well as the mean fairness $\Gamma^S(A)$ and conventional sum of similarities $G^S(A)$. We summarize the results in Table 5. We see that all non-trivial algorithms significantly outperform random assignment. However, A^{TPMS} incurs about 8% increased error as compared to A^{PR4A}.

Similar to Case (C5) of synthetic experiments, the optimally fair assignment A^{HARD} turns out to incur larger fraction of errors as compared to approximations A^{PR4A} and A^{ILPR}. The reason is that the assignment A^{HARD} maximizes the quality of the assignment with respect to the most "disadvantaged" question, but in contrast to A^{PR4A} and A^{ILPR}, does not care about the fate of remaining questions.

We also see that A^{PR4A} slightly outperforms A^{ILPR} in terms of the fraction of errors while having slightly smaller average fairness. One reason for this is that in parallel with $\Gamma^S\left(A^{\text{PR4A}}\right)$ being close to optimal, PeerReview4All algorithm managed to achieve the high value of conventional sum of similarities, thus maintaining a balance between the fairness $\Gamma^S(A)$ and the global objective $G^S(A)$.

We find these observations to be of notable interest for the actual conference peer-review scenarios. The task of identifying flags in the experiment involved a rather homogeneous set of similarities (in the sense that each worker either knew many or only few flags) where optimizing (3.2) or (3.3) would yield similar results. In contrast, the significantly higher heterogeneity in peer-review, the presence of many non-mainstream papers as well as both very strong and very weak reviewers, is expected to further amplify the observed improvements offered by the PeerReview4All algorithm as compared to TPMS and ILPR.

Algorithm	Error fraction	Error increase	Mean fairness $\Gamma^S(A)$	Mean sum of sim. $G^S(A)$
A^{RAND}	0.394	+275%	6.4	171.1
A^{TPMS}	0.113	+8%	20.8	274.6
A^{HARD}	0.110	+5%	21.9	269.8
A^{ILPR}	0.108	+3%	21.7	270.4
A^{PR4A}	0.105	—	21.6	272.9

Table 5: Results of the experiment on Amazon Mechanical Turk.

9 Proofs

We now present the proofs of our main results.

9.1 Proof of Theorem 1

We prove the result in three steps. First, we establish a lower bound on the fairness of the PeerReview4All algorithm. Then we establish an upper bound on the fairness of the optimal assignment. Finally, we combine these bounds to obtain the result (3.7).

Lower bound for the PeerReview4All algorithm.
We show a lower bound for the intermediate assignment $\widetilde{A}$ at Step 3 during the first iteration of Steps 2 to 7. We denote this particular assignment as $\widetilde{A}_1$. Note that in Step 4 we fix the assignment for $\widetilde{A}_1$'s worst-off papers into the final output, and hence we have $\Gamma_f^S\left(\widetilde{A}_1\right) \geq \Gamma_f^S\left(A_f^{\text{PR4A}}\right)$. On the other hand, by keeping track of A_0 (Step 7), we ensure that in all of the subsequent iterations of Steps 2 to 7, the temporary assignment $\widetilde{A}$ will be at least as fair as $\widetilde{A}_1$, which implies $\Gamma_f^S\left(\widetilde{A}_1\right) = \Gamma_f^S\left(A_f^{\text{PR4A}}\right)$.

Getting back to the first iteration of Steps 2 to 7, we note that when Step 2 is completed, we have λ assignments $A_1, \ldots, A_\lambda$ as candidates. Notice that for every $\kappa \in [\lambda]$, assignment A_κ is constructed with a two-step procedure by joining the outputs A_κ^1 and A_κ^2 of Subroutine 1. Recalling the definition (3.6) of s_κ^*, we now show that for every value of $\kappa \in [\lambda]$, the assignment A_κ^1 satisfies:

$$\min_{j \in [m]} \min_{i \in \mathcal{R}_{A_\kappa^1}(j)} s_{ij} = s_\kappa^*.$$

Consider any value of $\kappa \in [\lambda]$. The definition of s_κ^* ensures that there exist an assignment, say A^*, which assigns κ reviewers to each paper in a way that minimum similarity in this assignment equals s_κ^*. Now note that Subroutine 1, called in Step 2b of the algorithm, adds edges to the flow network in order of decreasing similarities. Thus, at the time all edges with similarity higher or equal to s_κ^* are added, we have that no edges with similarity smaller that s_κ^* are added, and that all edges which correspond to the assignment A^* are also added to the network. Thus, a maximum flow of size $m\kappa$ is achieved and hence each assigned (reviewer, paper) pair has similarity at least s_κ^*.

Recalling that s_∞^* is the lowest similarity in similarity matrix S, one can deduce that $\Gamma_f^S(A_\kappa) \geq \kappa f(s_\kappa^*) + (\lambda - \kappa) f(s_\infty^*)$ due to the monotonicity of f. Consequently, we have

$$\Gamma_f^S\left(A_f^{\text{PR4A}}\right) \geq \Gamma_f^S(A_\kappa) \geq \kappa f(s_\kappa^*) + (\lambda - \kappa) f(s_\infty^*), \tag{3.23}$$

for all $\kappa \in [\lambda]$. Taking a maximum over all values of $\kappa \in [\lambda]$ concludes the proof.

Upper bound for the optimal assignment A_f^{HARD}.
Consider any value of $\kappa \in [\lambda]$. By definition (3.6) of s_κ^*, for any feasible assignment $A \in \mathcal{A}$, there exists some paper $j_\kappa^* \in [m]$ for which at most $(\kappa - 1)$ reviewers have similarity strictly greater than s_κ^*. Let us now

consider assignment A_f^{HARD} and corresponding paper j_κ^*. This paper is assigned to at most $(\kappa-1)$ reviewers with similarity greater than s_κ^* and to at least $(\lambda-\kappa+1)$ reviewers with similarity smaller or equal to s_κ^*. Recalling that s_0^* is the largest possible similarity, we conclude that due to monotonicity of f, the following upper bound holds:

$$\Gamma_f^S\left(A_f^{\mathrm{HARD}}\right) = \min_{j\in[m]} \sum_{i\in\mathcal{R}_{A_f^{\mathrm{HARD}}}(j)} f(s_{ij}) \le \sum_{i\in\mathcal{R}_{A_f^{\mathrm{HARD}}}(j_\kappa^*)} f(s_{ij_\kappa^*}) \le (\kappa-1)\, f(s_0^*) + (\lambda-\kappa+1)\, f(s_\kappa^*). \quad (3.24)$$

Taking a minimum over all values of $\kappa\in[\lambda]$, then yields an upper bound on the fairness of A_f^{HARD}.

Putting it together.
To conclude the argument, it remains to plug in the obtained bounds (3.23) and (3.24) into ratio $\frac{\Gamma_f^S\left(A_f^{\mathrm{PR4A}}\right)}{\Gamma_f^S\left(A_f^{\mathrm{HARD}}\right)}$:

$$\frac{\Gamma_f^S\left(A_f^{\mathrm{PR4A}}\right)}{\Gamma_f^S\left(A_f^{\mathrm{HARD}}\right)} \ge \frac{\max\limits_{\kappa\in[\lambda]}\left(\kappa f(s_\kappa^*) + (\lambda-\kappa)\, f(s_\infty^*)\right)}{\min\limits_{\kappa\in[\lambda]}\left((\kappa-1) f(s_0^*) + (\lambda-\kappa+1)\, f(s_\kappa^*)\right)}.$$

Setting $\kappa=1$ in both numerator and denominator and recalling that $f(s)\ge 0\ \forall s\in[0,1]$, we obtain a worst-case approximation in terms of required paper load: $\frac{\Gamma^S\left(A^{\mathrm{PR4A}}\right)}{\Gamma^S\left(A^{\mathrm{HARD}}\right)} \ge \frac{1}{\lambda}$.

9.2 Proof of Corollary 1

Let us pause the PeerReview4All algorithm at the beginning of the r^{th} iteration of Steps 2 to 7 and inspect its state.

- The set $\mathcal{M}$ consists of papers that are not yet assigned:

$$\mathcal{M} = [m]\backslash\left(\bigcup_{l=1}^{r-1}\mathcal{J}_l\right).$$

- The vector of reviewers' loads $\overline{\mu}$ is adjusted with respect to assigned papers. For every reviewer $i\in[n]$, we have:

$$\overline{\mu}_i = \mu - \mathrm{card}\left(\left\{j\in\bigcup_{l=1}^{r-1}\mathcal{J}_l \,\middle|\, i\in\mathcal{R}_{A_f^{\mathrm{PR4A}}}(j)\right\}\right).$$

- The similarity matrix S_r consists of columns of the initial similarity matrix S which correspond to papers in $\mathcal{M}$.

The only thing that connects the algorithm with the previous iterations is the assignment A_0, computed in Step 7 of the previous iteration. However, we note that the sum similarity for the worst-off papers, determined in Step 4 of the current iteration (in other words, fairness of $\widetilde{A}_r$), is lower-bounded by the largest fairness of the candidate assignments $A_1,\dots,A_\lambda$, which are computed in Step 2.

We now repeat the proof of Theorem 1 with the following changes. Instead of the similarity matrix S, we use the updated matrix S_r; instead of considering all papers m we consider only papers from $\mathcal{M}$; instead of assuming that each reviewer $i\in[n]$ can review at most μ papers, we allow reviewer $i\in[n]$ to review at most $\overline{\mu}_i$ papers. Hence, we arrive to the bound (3.7) on the fairness of $\widetilde{A}_r$, where A^{HARD} should be read as $A^{\mathrm{HARD}}(\mathcal{M}) = A^{\mathrm{HARD}}\left(\mathcal{J}_{\{r:p\}}\right)$ and values $s_\kappa^*, \kappa\in\{0,\dots,\lambda\}\cup\{\infty\}$ are computed for similarity matrix S_r and constraints on reviewers' loads $\overline{\mu}$. Thus, we obtain (3.8) and conclude the proof of the corollary.

9.3 Proof of Theorem 2

Before we prove the theorem, let us formulate an auxiliary lemma which will help us show the claimed upper bound. We give the proof of this lemma subsequently in Section 9.3.3.

Lemma 1. *Consider any valid assignment $A \in \mathcal{A}$ and any estimator $\widehat{\theta} \in \left\{\widehat{\theta}^{MLE}, \widehat{\theta}^{MEAN}\right\}$. Then for every $\delta > 0$, the error incurred by $\widehat{\theta}$ is upper bounded as*

$$\sup_{\left(\theta_1^*,\ldots,\theta_m^*\right)\in\mathcal{F}_k(\delta)} \mathbb{P}\left\{\mathcal{T}_k\left(A,\widehat{\theta}\right) \neq \mathcal{T}_k^*\right\} \leq k(m-k)\exp\left\{-\left(\frac{\delta}{2\widetilde{\sigma}(A,\widehat{\theta})}\right)^2\right\},$$

where

$$\widetilde{\sigma}^2(A,\widehat{\theta}) = \begin{cases} \max\limits_{j\in[m]}\left(\sum\limits_{i\in\mathcal{R}_A(j)} \frac{1}{\sigma_{ij}^2}\right)^{-1} & \text{if } \widehat{\theta} = \widehat{\theta}^{MLE} \\ \max\limits_{j\in[m]}\left(\frac{1}{\lambda^2}\sum\limits_{i\in\mathcal{R}_A(j)} \sigma_{ij}^2\right) & \text{if } \widehat{\theta} = \widehat{\theta}^{MEAN}. \end{cases}$$

9.3.1 Proof of upper bound

First, recall from (3.13) the distribution of $\widehat{\theta}_j^{\mathrm{MEAN}}, j \in [m]$. Then the PeerReview4All algorithm called with $f = 1-h$ simultaneously tries to maximize the fairness of the assignment with respect to f and minimize the maximum variance of the estimated scores $\widehat{\theta}_j^{\mathrm{MEAN}}, j \in [m]$. Similarly, the choice of $f = h^{-1}$ ensures that together with optimizing the corresponding fairness, the algorithm also minimizes the maximum variance of $\widehat{\theta}_j^{\mathrm{MLE}}, j \in [m]$, defined in (3.12). Thus, the choice of the estimator defines the choice of the transformation function f which minimizes the maximum variance of the estimated scores. To maintain brevity, we denote $A_{\mathrm{MEAN}} = A_{1-h}^{\mathrm{PR4A}}$, $A_{\mathrm{MLE}} = A_{h^{-1}}^{\mathrm{PR4A}}$, $A_{\mathrm{MEAN}}(j) = \mathcal{R}_{A_{\mathrm{MEAN}}}(j)$ and $A_{\mathrm{MLE}}(j) = \mathcal{R}_{A_{\mathrm{MLE}}}(j)$.

Let now $S \in \mathcal{S}(q)$. We begin with the pair of assignment and estimator $\left(A_{\mathrm{MEAN}}, \widehat{\theta}^{\mathrm{MEAN}}\right)$. Notice that for arbitrary feasible assignment $A \in \mathcal{A}$ and estimator $\widehat{\theta}^{\mathrm{MEAN}}$,

$$\begin{aligned}\widetilde{\sigma}^2(A,\widehat{\theta}^{\mathrm{MEAN}}) &= \max_{j\in[m]}\left(\frac{1}{\lambda^2}\sum_{i\in\mathcal{R}_A(j)}\sigma_{ij}^2\right) = \frac{1}{\lambda^2}\max_{j\in[m]}\left(\sum_{i\in\mathcal{R}_A(j)} 1-(1-h(s_{ij}))\right) \\ &= \frac{1}{\lambda^2}\left(\lambda - \min_{j\in[m]}\sum_{i\in\mathcal{R}_A(j)}(1-h(s_{ij}))\right) = \frac{1}{\lambda^2}\left(\lambda - \Gamma_{1-h}^S(A)\right).\end{aligned}$$

Now we can write

$$\begin{aligned}\sup_{S\in\mathcal{S}(q)} \widetilde{\sigma}^2(A_{\mathrm{MEAN}},\widehat{\theta}^{\mathrm{MEAN}}) &= \frac{1}{\lambda^2}\left(\lambda - q\inf_{S\in\mathcal{S}(q)}\frac{\Gamma_{1-h}^S(A_{\mathrm{MEAN}})}{q}\right) \\ &\leq \frac{1}{\lambda^2}\left(\lambda - q\inf_{S\in\mathcal{S}(q)}\frac{\Gamma_{1-h}^S(A_{\mathrm{MEAN}})}{\Gamma_{1-h}^S(A_{1-h}^{\mathrm{HARD}})}\right) \\ &= \frac{\lambda - q\tau_q}{\lambda^2}.\end{aligned}$$

Using Lemma 1, we conclude the proof for the mean score estimator:

$$\sup_{\substack{(\theta_1^*,\dots,\theta_m^*)\in\mathcal{F}_k(\delta)\\ S\in\mathcal{S}(q)}} \mathbb{P}\left\{\mathcal{T}_k\left(A_{\mathrm{MEAN}},\widehat{\theta}^{\mathrm{MEAN}}\right)\neq\mathcal{T}_k^*\right\} \leq k(m-k)\exp\left\{-\left(\frac{\delta}{2\sup\limits_{S\in\mathcal{S}(q)}\widetilde{\sigma}(A_{\mathrm{MEAN}},\widehat{\theta}^{\mathrm{MEAN}})}\right)^2\right\} \tag{3.25}$$

$$\leq m^2\exp\left\{-\frac{\lambda^2\delta^2}{4\left(\lambda-q\tau_q\right)}\right\} \leq m^2\exp\left\{-\ln\frac{m^2}{\epsilon}\right\} \leq \epsilon. \tag{3.26}$$

Let us now consider the pair $\left(A_{\mathrm{MLE}},\widehat{\theta}^{\mathrm{MLE}}\right)$. It suffices to show that

$$\sup_{S\in\mathcal{S}(q)}\widetilde{\sigma}^2(A_{\mathrm{MLE}},\widehat{\theta}^{\mathrm{MLE}}) \leq \sup_{S\in\mathcal{S}(q)}\widetilde{\sigma}^2(A_{\mathrm{MEAN}},\widehat{\theta}^{\mathrm{MEAN}}). \tag{3.27}$$

Let us consider $S\in\mathcal{S}(q)$. Recall from the proof of Theorem 1 that the fairness of the resulting assignment is determined in the first iteration of Steps 2 to 7. After completion of Step 2, we have λ candidate assignments $A_1,\dots,A_\lambda$. Observe that Subroutine 1 in Step 6 uses the same heuristic for both A_{MEAN} and A_{MLE}. Hence, the λ candidate assignments yielded when PeerReview4All constructs A_{MEAN} coincide with the candidate assignments yielded when PeerReview4All constructs A_{MLE}. Depending on the choice of f, in Step 3 the algorithm picks one assignment that maximizes fairness (3.4) with respect to f. Thus,

$$\Gamma^S_{1-h}\left(A_{\mathrm{MEAN}}\right)=\max_{\kappa\in[\lambda]}\Gamma^S_{1-h}\left(A_\kappa\right) \quad\text{and}\quad \Gamma^S_{h^{-1}}\left(A_{\mathrm{MLE}}\right)=\max_{\kappa\in[\lambda]}\Gamma^S_{h^{-1}}\left(A_\kappa\right). \tag{3.28}$$

Hence, we have

$$\begin{aligned}\widetilde{\sigma}^2(A_{\mathrm{MLE}},\widehat{\theta}^{\mathrm{MLE}}) &= \max_{j\in[m]}\left(\sum_{i\in A_{\mathrm{MLE}}(j)}\frac{1}{\sigma_{ij}^2}\right)^{-1} = \max_{j\in[m]}\left(\frac{1}{\sum\limits_{i\in A_{\mathrm{MLE}}(j)}\frac{1}{h(s_{ij})}}\right)\\ &= \frac{1}{\Gamma^S_{h^{-1}}\left(A_{\mathrm{MLE}}\right)} \leq \frac{1}{\Gamma^S_{h^{-1}}\left(A_{\mathrm{MEAN}}\right)}.\end{aligned}$$

where the last ineqaulity is due to (3.28). Recalling the definition of the fairness (3.4) and using Jensen's inequality, we continue:

$$\begin{aligned}\widetilde{\sigma}^2(A_{\mathrm{MLE}},\widehat{\theta}^{\mathrm{MLE}}) &\leq \max_{j\in[m]}\left(\frac{1}{\lambda^2}\sum_{i\in A_{\mathrm{MEAN}}(j)}h(s_{ij})\right) = \max_{j\in[m]}\left(\frac{\lambda-\sum\limits_{i\in A_{\mathrm{MEAN}}(j)}(1-h(s_{ij}))}{\lambda^2}\right)\\ &= \frac{\lambda-\Gamma^S_{1-h}\left(A_{\mathrm{MEAN}}\right)}{\lambda^2} = \widetilde{\sigma}^2(A_{\mathrm{MEAN}},\widehat{\theta}^{\mathrm{MEAN}}).\end{aligned}$$

Taking a supremum over all $S\in\mathcal{S}(q)$, we obtain (3.27) which together with Lemma 1 and the first part of the statement concludes the proof.

9.3.2 Proof of lower bound

Proof of our lower bound is based on Fano's inequality (Cover and Thomas, 2005) which provides a lower bound for probability of error in L-ary hypothesis testing problems.

Without loss of generality we assume that $k\leq\frac{1}{2}m$. Otherwise, the result will hold by symmetry of the problems.

We first claim that there exists a value $s \in [0,1]$ such that $h(s) = 1 - \frac{q}{\lambda}$. Indeed, by assumptions of the theorem, h is continuous strictly monotonically decreasing function and $\frac{q}{\lambda} \geq 1 - h(0)$. Thus, $h(0) \geq 1 - \frac{q}{\lambda}$. On the other hand, if $h(1) > 1 - \frac{q}{\lambda}$, then for every similarity matrix S we have

$$\Gamma^S_{1-h}(A) \leq \lambda(1 - h(1)) < q.$$

The last inequality contradicts with the definition (3.16) of $\mathcal{S}(q)$, verifying that

$$h(0) \geq 1 - \frac{q}{\lambda} \geq h(1).$$

Given that h is continuous strictly monotonically decreasing function, we conclude that these exists $s = h^{-1}\left(1 - \frac{q}{\lambda}\right) \in [0,1]$.

Consider the similarity matrix $\widetilde{S} = \left\{h^{-1}\left(1 - \frac{q}{\lambda}\right)\right\}^{n \times m}$. Observe that $\widetilde{S} \in \mathcal{S}(q)$, since every feasible assignment $A \in \mathcal{A}$ has fairness

$$\Gamma^{\widetilde{S}}_{1-h}(A) = \min_{j \in [m]} \sum_{i \in \mathcal{R}_A(j)} (1 - h(s_{ij})) = \min_{j \in [m]} \sum_{i \in \mathcal{R}_A(j)} \left\{1 - h\left(h^{-1}\left(1 - \frac{q}{\lambda}\right)\right)\right\} = q.$$

Thus, in any feasible assignment each paper $j \in [m]$ receives λ reviewers with similarity exactly $h^{-1}\left(1 - \frac{q}{\lambda}\right)$.

To apply Fano's inequality, we need to reduce our problem to a hypothesis testing problem. To do so, let us introduce the set $\mathcal{P}$ of $(m - k + 1)$ instances of the paper accepting/rejecting problem: every problem instance in this set has the same similarity matrix $\widetilde{S}$, but differs in the set of top k papers $\mathcal{T}^*_k$. We now consider the problem of distinguishing between these problem instances, which is equivalent to the problem of correctly recovering the top k papers. More concretely, we denote the $(m - k + 1)$ problem instances as, $\mathcal{P} = \{1, 2, \ldots, m - k + 1\}$, where for any problem $\ell \in \mathcal{P}$ the set of top k papers is denoted as $\mathcal{T}^*_k(\ell)$ and set as $\{1, 2, \ldots, k-1\} \cup \{k - 1 + \ell\}$. The true quality of any paper $j \in [m]$ in any problem instance $\ell \in \mathcal{P}$ is

$$\theta^*_j(\ell) = \begin{cases} \delta & \text{if} \quad j \in \mathcal{T}^*_k(\ell) \\ 0 & \text{otherwise,} \end{cases}$$

thereby ensuring that $(\theta^*_1(\ell), \ldots, \theta^*_m(\ell)) \in \mathcal{F}_k(\delta)$, for every instance $\ell \in \mathcal{P}$.

Let P denote a random variable which is uniformly distributed over elements of $\mathcal{P}$. Then given $P = \ell$, we denote a random matrix of reviewers' scores as $Y^{(\ell)} \in \mathbb{R}^{\lambda \times m}$ whose $(r,j)^{\text{th}}$ entry is a score given by reviewer $i_r, r \in [\lambda]$, assigned to paper j and

$$Y^{(\ell)}_{rj} \sim \begin{cases} \mathcal{N}\left(\delta, 1 - \frac{q}{\lambda}\right) & \text{if} \quad j \in \mathcal{T}^*_k(\ell) \\ \mathcal{N}\left(0, 1 - \frac{q}{\lambda}\right) & \text{otherwise.} \end{cases} \tag{3.29}$$

We denote the distribution of random matrix $Y^{(\ell)}$ as $\mathbb{P}^{(\ell)}$. Note that $Y^{(\ell)}$ does not depend on the selected assignment $A \in \mathcal{A}$. Indeed, recall from (3.11), that assignment A affects only variances of observed scores. On the other hand, for any reviewer $i \in [n]$ and for any paper $j \in [m]$, the score y_{ij} has variance $1 - \frac{q}{\lambda}$. Thus, for any feasible assignment A and any $\ell \in \mathcal{P}$, the distribution of random matrix Y^{ℓ} has the form (3.29).

Now let us consider the problem of determining the index $P = \ell \in \mathcal{P}$, given the observation $Y^{(\ell)}$ following the distribution $\mathbb{P}^{(\ell)}$. Fano's inequality provides a lower bound for probability of error of every estimator $\varphi : \mathbb{R}^{\lambda \times m} \to \mathcal{P}$ in terms of Kullback-Leibler divergence between distributions $\mathbb{P}^{(\ell_1)}$ and $\mathbb{P}^{(\ell_2)}$ $(\ell_1 \neq \ell_2,\ \ell_1, \ell_2 \in [m - k + 1])$:

$$\mathbb{P}\{\varphi(Y) \neq P\} \geq 1 - \frac{\max\limits_{\ell_1 \neq \ell_2 \in \mathcal{P}} \mathrm{KL}\left[\mathbb{P}^{(\ell_1)} || \mathbb{P}^{(\ell_2)}\right] + \log 2}{\log\left(\mathrm{card}(\mathcal{P})\right)}, \tag{3.30}$$

where $\mathrm{card}(\mathcal{P})$ denotes the cardinality of $\mathcal{P}$ and equals $(m - k + 1)$ for our construction.

Let us now derive an upper bound on the quantity

$$\max_{\ell_1 \neq \ell_2 \in \mathcal{P}} \mathrm{KL}\left[\mathbb{P}^{(\ell_1)} || \mathbb{P}^{(\ell_2)}\right]. \tag{3.31}$$

First, note that for each $\ell \in [m-\kappa+1]$, entries of $Y^{(\ell)}$ are independent. Second, for arbitrary $\ell_1 \neq \ell_2$, the distributions of $Y^{(\ell_1)}$ and $Y^{(\ell_2)}$ differ only in two columns. Thus,

$$\mathrm{KL}\left[\mathbb{P}^{(\ell_1)} || \mathbb{P}^{(\ell_2)}\right] = \lambda\left\{\mathrm{KL}\left[\mathcal{N}\left(\delta, 1-\frac{q}{\lambda}\right) || \mathcal{N}\left(0, 1-\frac{q}{\lambda}\right)\right] + \mathrm{KL}\left[\mathcal{N}\left(0, 1-\frac{q}{\lambda}\right) || \mathcal{N}\left(\delta, 1-\frac{q}{\lambda}\right)\right]\right\}.$$

Some simple algebraic manipulations yield:

$$\mathrm{KL}\left[\mathcal{N}\left(\delta, 1-\frac{q}{\lambda}\right) || \mathcal{N}\left(0, 1-\frac{q}{\lambda}\right)\right] = \mathrm{KL}\left[\mathcal{N}\left(0, 1-\frac{q}{\lambda}\right) || \mathcal{N}\left(\delta, 1-\frac{q}{\lambda}\right)\right] = \frac{\delta^2}{2\left(1-\frac{q}{\lambda}\right)}. \tag{3.32}$$

Finally, substituting (3.32) in (3.30), for $m > 6$ and for a sufficiently small constant c, we have

$$\mathbb{P}\left\{\varphi(Y) \neq P\right\} \geq 1 - \frac{\frac{\lambda^2\delta^2}{\lambda - q} + \log 2}{\log\left(m-k+1\right)} \geq 1 - \frac{c^2 \ln m + 1}{\log\left(\frac{m}{2}+1\right)} \geq \frac{1}{2}.$$

This lower bound implies

$$\sup_{S \in \mathcal{S}(q)} \inf_{\left(\widehat{\theta}, A \in \mathcal{A}\right)} \sup_{\left(\theta_1^*, \ldots, \theta_m^*\right) \in \mathcal{F}_k(\delta)} \mathbb{P}\left\{\mathcal{T}_k\left(A, \widehat{\theta}\right) \neq \mathcal{T}_k^*\right\} \geq \frac{1}{2}.$$

9.3.3 Proof of Lemma 1

First, let $\widehat{\theta} = \widehat{\theta}^{\mathrm{MEAN}}$. Then given a valid assignment A, the estimates $\widehat{\theta}_j^{\mathrm{MEAN}}, j \in [m]$, are distributed as

$$\widehat{\theta}_j^{\mathrm{MEAN}} \sim \mathcal{N}\left(\theta_j^*, \frac{1}{\lambda^2}\sum_{i \in \mathcal{R}_A(j)} \sigma_{ij}^2\right) = \mathcal{N}\left(\theta_j^*, \bar{\sigma}_j^2\right),$$

where we have defined $\bar{\sigma}_j^2 = \frac{1}{\lambda^2}\sum\limits_{i \in \mathcal{R}_A(j)} \sigma_{ij}^2$. Now let us consider two papers j_1, j_2 such that j_1 belongs to the top k papers $\mathcal{T}_k^*$ and $j_2 \notin \mathcal{T}_k^*$. The probability that paper j_2 receives higher score than paper j_1 is upper bounded as

$$\begin{aligned}\mathbb{P}\left\{\widehat{\theta}_{j_1}^{\mathrm{MEAN}} \leq \widehat{\theta}_{j_2}^{\mathrm{MEAN}}\right\} &= \mathbb{P}\left\{\left(\widehat{\theta}_{j_1}^{\mathrm{MEAN}} - \widehat{\theta}_{j_2}^{\mathrm{MEAN}}\right) - \mathbb{E}\left\{\widehat{\theta}_{j_1}^{\mathrm{MEAN}} - \widehat{\theta}_{j_2}^{\mathrm{MEAN}}\right\} \leq -\mathbb{E}\left\{\widehat{\theta}_{j_1}^{\mathrm{MEAN}} - \widehat{\theta}_{j_2}^{\mathrm{MEAN}}\right\}\right\} \\ &\overset{(i)}{\leq} \exp\left\{-\frac{\left(\mathbb{E}\left\{\widehat{\theta}_{j_1}^{\mathrm{MEAN}} - \widehat{\theta}_{j_2}^{\mathrm{MEAN}}\right\}\right)^2}{2\left(\bar{\sigma}_{j_1}^2 + \bar{\sigma}_{j_2}^2\right)}\right\} \overset{(ii)}{\leq} \exp\left\{-\left(\frac{\delta}{2\widetilde{\sigma}(A, \widehat{\theta}^{\mathrm{MEAN}})}\right)^2\right\},\end{aligned}$$

where inequality (i) is due to Hoeffding's inequality, and inequality (ii) holds because $\mathbb{E}\left\{\widehat{\theta}_{j_1}^{\mathrm{MEAN}} - \widehat{\theta}_{j_2}^{\mathrm{MEAN}}\right\} = \theta_{j_1}^* - \theta_{j_2}^* \geq \delta$ and $\widetilde{\sigma}^2(A, \widehat{\theta}^{\mathrm{MEAN}}) = \max\limits_{j \in [m]} \bar{\sigma}_j^2$. The estimator makes a mistake if and only if at least one paper from $\mathcal{T}_k^*$ receives lower score than at least one paper from $[m] \backslash \mathcal{T}_k^*$. A union bound across every paper from $\mathcal{T}_k^*$, paired with $(m-k)$ papers from $[m] \backslash \mathcal{T}_k^*$, yields our claimed result.

Let us now consider $\widehat{\theta} = \widehat{\theta}^{\mathrm{MLE}}$. Then it is not hard to see that

$$\widehat{\theta}_j^{\mathrm{MEAN}} \sim \mathcal{N}\left(\theta_j^*, \left(\sum_{i \in \mathcal{R}_A(j)} \frac{1}{\sigma_{ij}^2}\right)^{-1}\right) = \mathcal{N}\left(\theta_j^*, \bar{\sigma}_j^2\right),$$

where we denoted $\bar{\sigma}_j^2 = \left(\sum\limits_{i \in \mathcal{R}_A(j)} \frac{1}{\sigma_{ij}^2}\right)^{-1}$. Proceeding in a manner similar to the proof for the averaging estimator yields the claimed result.

9.4 Proof of Corollary 2

The proof of Corollary 2 follows along similar lines as the proof of Theorem 2.

9.4.1 Proof of upper bound

Let us consider some $\kappa \in [\lambda]$ and $S \in \mathcal{S}_\kappa(v)$. We apply Lemma 1 to proof the upper bound and in order to do so, we need to derive an upper bound on $\widetilde{\sigma}(A^{\mathrm{PR4A}}_{h^{-1}}, \widehat{\theta}^{\mathrm{MLE}})$.

$$\begin{aligned}\widetilde{\sigma}^2(A^{\mathrm{PR4A}}_{h^{-1}}, \widehat{\theta}^{\mathrm{MLE}}) &= \max_{j\in[m]} \left(\sum_{i \in \mathcal{R}_{A^{\mathrm{PR4A}}_{h^{-1}}}(j)} \frac{1}{\sigma^2_{ij}} \right)^{-1} = \left(\min_{j\in[m]} \sum_{i \in \mathcal{R}_{A^{\mathrm{PR4A}}_{h^{-1}}}(j)} h^{-1}(s_{ij}) \right)^{-1} \\ &\leq \frac{1}{\frac{\kappa}{h(v)} + \frac{\lambda-\kappa}{h(0)}} = \frac{h(v)h(0)}{\kappa h(0) + (\lambda-\kappa)h(v)}.\end{aligned}$$

Thus,

$$\sup_{S\in\mathcal{S}_\kappa(v)} \widetilde{\sigma}^2(A^{\mathrm{PR4A}}_{h^{-1}}, \widehat{\theta}^{\mathrm{MLE}}) \leq \frac{h(v)h(0)}{\kappa h(0) + (\lambda-\kappa)h(v)}. \tag{3.33}$$

It remains to apply Lemma 1 to complete our proof, and we do so by applying the chain of arguments (3.25) and (3.26) to the bound (3.33), where the pair $(A^{\mathrm{PR4A}}_{1-h}, \widehat{\theta}^{\mathrm{MEAN}})$ in (3.25) and (3.26) is substituted with the pair $(A^{\mathrm{PR4A}}_{h^{-1}}, \widehat{\theta}^{\mathrm{MLE}})$.

9.4.2 Proof of lower bound

To prove the lower bound, we use the Fano's ineqaulity in the same way as we did when proved Theorem 2(b). However, we now need to be more careful with construction of working similarity matrix $\widetilde{S} \in \mathcal{S}_\kappa(v)$.

As in the proof of Theorem 2(b), we assume $k \leq \frac{m}{2}$. If the converse holds, than the result holds by symmetry of the problem. Next, consider arbitrary feasible assignment $\widetilde{A} \in \mathcal{A}_\kappa$. Recall, that $\mathcal{A}_\kappa$ consists of assignments which assign each paper $j \in [m]$ to κ instead of λ reviewers such that each reviewer reviews at most μ papers.

Now we define a similarity matrix $\widetilde{S}$ as follows:

$$s_{ij} = \begin{cases} v & \text{if} \quad i \in \mathcal{R}_{\widetilde{A}}(j) \\ 0 & \text{otherwise.} \end{cases} \tag{3.34}$$

Thus, for each paper $j \in [m]$ there exist exactly κ reviewers with non-zero similarity v and in every feasible assignment $A \in \mathcal{A}$ each paper $j \in [m]$ is assigned to at most κ reviewers with non-zero similarity. Note that $\widetilde{S} \in \mathcal{S}_\kappa(v)$.

Now let us consider the set of $(m-k+1)$ problem instances $\mathcal{P}$ defined in Section 9.3.2. For every feasible assignment $A \in \mathcal{A}$, if $Y^{(A,\ell)}$ is a matrix of observed reviewers' scores for instance $\ell \in \mathcal{P}$, then $(r,j)^{\text{th}}$ entry of $Y^{(A,\ell)}$ follows the distribution

$$Y^{(A,\ell)}_{rj} = \begin{cases} \mathcal{N}\big(\delta \times \mathbb{I}\left\{j \in \mathcal{T}^*_k(\ell)\right\}, h(v)\big) & \text{if} \quad \widetilde{A}_{i_r j} = 1 \\ \mathcal{N}\big(\delta \times \mathbb{I}\left\{j \in \mathcal{T}^*_k(\ell)\right\}, h(0)\big) & \text{if} \quad \widetilde{A}_{i_r j} = 0, \end{cases} \tag{3.35}$$

where $i_r, r \in [\lambda]$ is reviewer assigned to paper j in assignment A.

We denote the distribution of random matrix $Y^{(A,\ell)}$ as $\mathbb{P}^{(A,\ell)}$. Note that in contrast to the proof of Theorem 2, here $Y^{(A,\ell)}$ does depend on the selected assignment $A \in \mathcal{A}$. Thus, instead of (3.31), we need to derive an upper bound on the quantity

$$\sup_{A\in\mathcal{A}} \max_{\ell_1 \neq \ell_2 \in \mathcal{P}} \mathrm{KL}\left[\mathbb{P}^{(A,\ell_1)} || \mathbb{P}^{(A,\ell_2)}\right].$$

First, note that for each $\ell \in [m-k+1]$ and for each feasible assignment $A \in \mathcal{A}$, the entries of $Y^{(A,\ell)}$ are independent. Second, for arbitrary $\ell_1 \neq \ell_2$, the distributions of $Y^{(A,\ell_1)}$ and $Y^{(A,\ell_2)}$ differ only in two columns. Thus, for any feasible assignment $A \in \mathcal{A}$, we have

$$\begin{aligned}\mathrm{KL}\left[\mathbb{P}^{(A,\ell_1)}||\mathbb{P}^{(A,\ell_2)}\right] &\leq \gamma_{\ell_1}\mathrm{KL}\left[\mathcal{N}\big(\delta,h(v)\big)||\mathcal{N}\big(0,h(v)\big)\right] + (\lambda-\gamma_{\ell_1})\,\mathrm{KL}\left[\mathcal{N}\big(\delta,h(0)\big)||\mathcal{N}\big(0,h(0)\big)\right] \\ &\quad + \gamma_{\ell_2}\mathrm{KL}\left[\mathcal{N}\big(0,h(v)\big)||\mathcal{N}\big(\delta,h(v)\big)\right] + (\lambda-\gamma_{\ell_2})\,\mathrm{KL}\left[\mathcal{N}\big(0,h(0)\big)||\mathcal{N}\big(\delta,h(0)\big)\right] \qquad (3.36)\\ &= (\gamma_{\ell_1}+\gamma_{\ell_2})\,\frac{\delta^2}{2h(v)} + (2\lambda-\gamma_{\ell_1}-\gamma_{\ell_2})\,\frac{\delta^2}{2h(0)}, \qquad (3.37)\end{aligned}$$

where γ_{ℓ_1} is the number of reviewers with similarity v assigned to paper $(k-1+\ell_1)$ in A and γ_{ℓ_2} is the number of reviewers with similarity v assigned to paper $(k-1+\ell_2)$. By construction of similarity matrix $\widetilde{S}$, for each $\ell \in [m-k+1]$ and for each $A \in \mathcal{A}$, we have $\gamma_\ell \leq \kappa$. Note that two summands in (3.37) are proportional to a convex combination of $\frac{\delta^2}{2h(v)}$ and $\frac{\delta^2}{2h(0)}$. Moreover, by monotonicity of h, we have $\frac{\delta^2}{2h(v)} \geq \frac{\delta^2}{2h(0)}$, and hence

$$\sup_{A\in\mathcal{A}}\max_{\ell_1\neq\ell_2\in\mathcal{P}}\mathrm{KL}\left[\mathbb{P}^{(A,\ell_1)}||\mathbb{P}^{(A,\ell_2)}\right] \leq \frac{\kappa\delta^2}{h(v)} + \frac{(\lambda-\kappa)\,\delta^2}{h(0)} = \delta^2\left(\frac{\kappa h(0)+(\lambda-\kappa)\,h(v)}{h(v)h(0)}\right).$$

Applying Fano's ineqaulity (3.30), we conclude that for all feasible assignments $A \in \mathcal{A}$, if $m > 6$ and universal constant c is sufficiently small, then

$$\mathbb{P}\left\{\varphi(Y) \neq P\right\} \geq 1 - \frac{\delta^2\left(\frac{\kappa h(0)+(\lambda-\kappa)h(v)}{h(v)h(0)}\right)+\log 2}{\log\left(m-k+1\right)} \geq 1 - \frac{c^2\ln m+1}{\log\left(\frac{m}{2}+1\right)} \geq \frac{1}{2}.$$

This bound thus implies

$$\sup_{S\in\mathcal{S}_\kappa(v)}\inf_{\left(\widehat{\theta},A\in\mathcal{A}\right)}\sup_{\left(\theta_1^*,\ldots,\theta_m^*\right)\in\mathcal{F}_k(\delta)}\mathbb{P}\left\{\mathcal{T}_k\left(A,\widehat{\theta}\right)\neq\mathcal{T}_k^*\right\} \geq \frac{1}{2}.$$

9.5 Proof of Theorem 3

Before we prove the theorem, we state an auxiliary proposition which will help us to prove a lower bound.

Lemma 2 (Shah and Wainwright, 2015)**.** *Let $t > 0$ be an integer such that $2t \leq \frac{1}{1+\nu_2}\min\left\{m^{1-\nu_1}, k, m-k\right\}$ for some constants $\nu_1, \nu_2 \in (0;1)$ and m is larger than some (ν_1,ν_2)-dependent constant. Then there exist a set of binary strings $\left\{b^1, b^2, \ldots, b^L\right\} \subseteq \{0,1\}^{m/2}$ with cardinality $L > \exp\left\{\frac{9}{10}\nu_1\nu_2 t\log m\right\}$ such that*

$$\mathcal{D}_H\left(b^{\ell_1}, \mathbf{0}_{m/2}\right) = 2(1+\nu_2)t \quad \textit{and} \quad \mathcal{D}_H\left(b^{\ell_1}, b^{\ell_2}\right) > 4t \quad \forall \ell_1 \neq \ell_2 \in [L]$$

The proof of Lemma 2 relies on a coding-theoretic result due to Levenshtein (1971) which gives a lower bound on the number of codewords of fixed length m and Hamming weights c_1 with Hamming distance between each pair of codewords higher than c_2.

9.5.1 Proof of upper bound

Without loss of generality we assume that the true underlying ranking of the papers is $1, 2, \ldots, k, \ldots, m$. We prove the claim for pair $\left(A_{1-h}^{\mathrm{PR4A}}, \widehat{\theta}^{\mathrm{MEAN}}\right)$ below, and proof for $\left(A_{h^{-1}}^{\mathrm{PR4A}}, \widehat{\theta}^{\mathrm{MLE}}\right)$ follows from the proof of the corresponding part of Theorem 2(a).

From the proof of Lemma 1 and Section 9.3.1, we know that under conditions of the theorem, for every paper $j_1 \leq k-t$ and for every paper $j_2 \geq k+t+1$,

$$\sup_{S\in\mathcal{S}(q)}\mathbb{P}\left\{\widehat{\theta}_{j_1}^{\mathrm{MEAN}} - \widehat{\theta}_{j_2}^{\mathrm{MEAN}} \leq 0\right\} \leq \exp\left\{-\left(\frac{\delta}{2\sup\limits_{S\in\mathcal{S}(q)}\widetilde{\sigma}(A_{1-h}^{\mathrm{PR4A}},\widehat{\theta}^{\mathrm{MEAN}})}\right)^2\right\} \qquad (3.38)$$

where

$$\sup_{S\in\mathcal{S}(q)} \widetilde{\sigma}^2(A_{1-h}^{\mathrm{PR4A}}, \widehat{\theta}^{\mathrm{MEAN}}) \leq \frac{\lambda - \tau_q q}{\lambda^2}. \tag{3.39}$$

Taking a union bound across every paper from the top $(k-t)$ papers, paired with the bottom $(m-k-t)$ papers, we obtain

$$\sup_{S\in\mathcal{S}(q)} \mathbb{P}\left\{\exists j_1 \leq k-t, j_2 \geq k+t+1 \quad \text{such that } \widehat{\theta}_{j_1}^{\mathrm{MEAN}} \leq \widehat{\theta}_{j_2}^{\mathrm{MEAN}}\right\} \leq m^2 \exp\left\{-\frac{\lambda^2\delta^2}{4(\lambda-\tau_q q)}\right\} \leq \epsilon.$$

In other words, for every similarity matrix $S \in \mathcal{S}(q)$, with probability at least $(1-\epsilon)$, the top $(k-t)$ papers will receive higher score than bottom $(m-k-t)$ papers. Thus, among accepted papers $\mathcal{T}_k\left(A_{1-h}^{\mathrm{PR4A}}, \widehat{\theta}^{\mathrm{MEAN}}\right)$, at most t papers will not belong to $\mathcal{T}_k^*$, thereby ensuring that

$$\mathcal{D}_H\left(\mathcal{T}_k\left(A_{1-h}^{\mathrm{PR4A}}, \widehat{\theta}^{\mathrm{MEAN}}\right), \mathcal{T}_k^*\right) \leq 2t$$

with probability at least $1-\epsilon$.

9.5.2 Proof of lower bound

To prove the lower bound, we follow similar path as we used when we derived a lower bound in Theorem 2. However, we now need more advanced technique to construct necessary set of instances.

As in the proof of Theorem 2(b), we assume that $k \leq \frac{m}{2}$. If the converse holds, than the result holds by the symmetry of the problem. Next, consider similarity matrix $\widetilde{S} = \left\{h^{-1}\left(1-\frac{q}{\lambda}\right)\right\}^{n\times m} \in \mathcal{S}(q)$. To apply Fano's inequality, it remains to construct a set $\mathcal{P} = \{1, 2, \ldots, L\}$ of suitable instances of paper accepting/rejecting problem: every problem instance in this set has the same similarity matrix $\widetilde{S}$, but differs in the set of top k papers $\mathcal{T}_k^*$. We note that in contrast to the proof of Theorem 2(b), it is not enough to create $(m-k+1)$ instances where the sets of top k papers differ only in a single paper. As we will see below, it suffices to construct instances such that for every $\ell_1, \ell_2 \in \mathcal{P}$, the sets of top k papers satisfy $\mathcal{D}_H\left(\mathcal{T}_k^*(\ell_1), \mathcal{T}_k^*(\ell_2)\right) > 4t$.

Note that requirements of Lemma 2 are satisfied by the conditions of Theorem 3. Let $\left\{b^1, b^2, \ldots, b^L\right\}$ be the corresponding binary strings. For every problem $\ell \in \mathcal{P}$, consider the following binary string:

$$\widetilde{b}^\ell = \overbrace{\underbrace{1,1,\ldots,1}_{k-2(1+\nu_2)t}, 0, 0, \ldots, 0}^{m/2}, b_1^\ell, b_2^\ell, \ldots, b_{m/2}^\ell. \tag{3.40}$$

First, note that $2t \leq \frac{1}{1+\nu_2}k$, and hence $k - 2(1+\nu_2)t \geq 0$, thereby ensuring that the construction (3.40) is not vacuous. Now let $\mathcal{T}_k^*(\ell)$ be the set of indices such that their corresponding elements in string $\widetilde{b}^\ell$ equal 1. By construction, the cardinality of $\mathcal{T}_k^*(\ell)$ is k so it is a valid set of top k papers. Finally, we need to set the scores of papers. Let for every paper $j \in [m]$:

$$\theta_j^*(\ell) = \begin{cases} \delta & \text{if} \quad \widetilde{b}_j^\ell = 1 \\ 0 & \text{if} \quad \widetilde{b}_j^\ell = 0, \end{cases}$$

which ensures that for every $\ell \in \mathcal{P}$, $(\theta_1^*(\ell), \theta_2^*(\ell), \ldots, \theta_m^*(\ell)) \in \mathcal{F}_k \subset \mathcal{F}_{k,t}$.

The strategy for the remaining part of the proof is the following. We first show that the problem instances defined above are well-separated in a sense that for any two of them, the corresponding sets of the top k papers differ in sufficiently many elements. We then assume that there exists an (assignment algorithm, estimator) pair which for every similarity matrix $S \in \mathcal{S}(q)$ recovers the set of top k papers with at most t errors with high probability. Then this pair must be able to determine with high probability the problem

instance ℓ, sampled uniformly at random from $\mathcal{P}$, by observing corresponding reviewers' scores. We then apply Fano's inequality to show the impossibility of the last implication.

Following the plan described above, we note that for every two distinct instances $\ell_1, \ell_2 \in \mathcal{P}$, we have

$$\mathcal{D}_H\left(\mathcal{T}_k^*(\ell_1), \mathcal{T}_k^*(\ell_2)\right) > 4t.$$

Consequently, for every set $\mathcal{T}_k^*$ of k papers, $\mathcal{D}_H\left(\mathcal{T}_k^*, \mathcal{T}_k^*(\ell)\right) \le 2t$ for at most one instance $\ell \in \mathcal{P}$. Now assume for the sake of contradiction that for every similarity matrix $S \in \mathcal{S}(q)$, there exists an assignment $\widetilde{A} = \widetilde{A}(S)$ and estimator $\widehat{\theta} = \widehat{\theta}(S)$ such that for arbitrarily large value of m

$$\sup_{\left(\theta_1^*,\dots,\theta_m^*\right)\in\mathcal{F}_k(\delta)} \mathbb{P}\left\{\mathcal{D}_H\left(\mathcal{T}_k\left(\widetilde{A},\widehat{\theta}\right),\mathcal{T}_k^*\right) > 2t\right\} < \frac{1}{2}. \tag{3.41}$$

This assumption implies that estimator $\widehat{\theta}(\widetilde{S})$ might be used to determine the problem $P = \ell$ sampled uniformly at random from $\mathcal{P}$ correctly with probability greater than $1/2$. Indeed, notice that similarity matrix $\widetilde{S}$ was constructed in a way that $\mathcal{T}_k\left(\widetilde{A},\widehat{\theta}\right)$ does not depend on assignment $\widetilde{A}$.

Given $P = \ell$, let $Y^{(\ell)}$ be the random matrix of reviewers' scores. The distribution $\mathbb{P}^{(\ell)}$ of components of $Y^{(\ell)}$ is defined in (3.29). To apply Fano's inequality (3.30), it remains to derive an upper bound on the quantity $\max\limits_{\ell_1\neq\ell_2\in\mathcal{P}} \mathrm{KL}\left[\mathbb{P}^{(\ell_1)}||\mathbb{P}^{(\ell_2)}\right]$.

First, note that entries of $Y^{(\ell)}$ are independent. Second, note that for every pair $\ell_1 \neq \ell_2 \in \mathcal{P}$ and for every $j \in [m/2]$, the distribution of the j^{th} column of $Y^{(A,\ell_1)}$ is identical to the distribution of the j^{th} column of $Y^{(A,\ell_2)}$. Among the last $m/2$ columns, the distributions of at most $4(1+\nu_2)t$ columns of $Y^{(A,\ell_1)}$ differ from the distributions of the corresponding columns in $Y^{(A,\ell_2)}$. Thus, for arbitrary $\ell_1 \neq \ell_2 \in \mathcal{P}$

$$\mathrm{KL}\left[\mathbb{P}^{(\ell_1)}||\mathbb{P}^{(\ell_2)}\right] \le 2(1+\nu_2)t\lambda\left\{\mathrm{KL}\left[\mathcal{N}\left(\delta, 1-\frac{q}{\lambda}\right)||\mathcal{N}\left(0,1-\frac{q}{\lambda}\right)\right] + \mathrm{KL}\left[\mathcal{N}\left(0,1-\frac{q}{\lambda}\right)||\mathcal{N}\left(\delta,1-\frac{q}{\lambda}\right)\right]\right\}.$$

Recalling (3.32), we deduce that

$$\max_{\ell_1\neq\ell_2\in\mathcal{P}} \mathrm{KL}\left[\mathbb{P}^{(\ell_1)}||\mathbb{P}^{(\ell_2)}\right] \le 4(1+\nu_2)t\lambda\frac{\lambda\delta^2}{2(\lambda-q)} = 2(1+\nu_2)t\frac{\lambda^2\delta^2}{\lambda-q} \le 4c^2\nu_1\nu_2 t\ln m.$$

Finally, Fano's inequality together with Lemma 2 ensures that for every estimator $\varphi : Y \to \mathcal{P}$

$$\mathbb{P}\left\{\varphi(Y) \neq P\right\} \ge 1 - \frac{4c^2\nu_1\nu_2 t\ln m + \log 2}{\frac{9}{10}\nu_1\nu_2 t\log m} \ge 1 - \frac{40}{9}c^2\frac{\ln m}{\log m} - \frac{1}{\frac{9}{10}\nu_1\nu_2 t\log m} \ge \frac{1}{2}$$

for m larger than some (ν_1,ν_2)-dependent constant and small enough universal constant c. This leads to a contradiction with (3.41), thus proving the theorem.

9.6 Proof of Corollary 3

The proof of the Corollary 3 is based on the ideas of the proofs of Theorem 3 and Corollary 2 and repeats them with minor changes.

9.6.1 Proof of upper bound

To show the required upper bound, we repeat the proof of Theorem 3(a) from Section 9.5.1 with the following changes. Equation (3.38) should be substituted with:

$$\sup_{S\in\mathcal{S}_\kappa(v)} \mathbb{P}\left\{\widehat{\theta}_{j_1}^{\mathrm{MLE}} - \widehat{\theta}_{j_2}^{\mathrm{MLE}} \le 0\right\} \le \exp\left\{-\left(\frac{\delta}{2\sup\limits_{S\in\mathcal{S}_\kappa(v)}\widetilde{\sigma}(A_{h^{-1}}^{\mathrm{PR4A}},\widehat{\theta}^{\mathrm{MLE}})}\right)^2\right\}.$$

Equation (3.39) should be substituted with:

$$\sup_{S\in\mathcal{S}_\kappa(v)} \widetilde{\sigma}^2(A^{\mathrm{PR4A}}, \widehat{\theta}^{\mathrm{MLE}}) \le \frac{h(v)h(0)}{\kappa h(0) + (\lambda-\kappa)h(v)}.$$

In the remaining part of the proof, pair $(A^{\mathrm{PR4A}}_{1-h}, \widehat{\theta}^{\mathrm{MEAN}})$ should be substituted with the pair $(A^{\mathrm{PR4A}}_{h^{-1}}, \widehat{\theta}^{\mathrm{MLE}})$.

9.6.2 Proof of lower bound

To prove the lower bound, we use the set of problems $\mathcal{P}$ constructed in Section 9.5.2 and the similarity matrix $\widetilde{S}$ as defined in (3.34).

Given $P=\ell$ and any feasible assignment $A\in\mathcal{A}$, let $Y^{(A,\ell)}$ be the random matrix of reviewers' scores. The distribution $\mathbb{P}^{(A,\ell)}$ of components of $Y^{(A,\ell)}$ is defined in (3.35). Since the distribution of reviewers' scores now depends on the assignment, to apply Fano's inequality (3.30), we need to derive an upper bound on the quantity $\sup_{A\in\mathcal{A}} \max_{\ell_1\neq\ell_2\in\mathcal{P}} \mathrm{KL}\left[\mathbb{P}^{(A,\ell_1)}||\mathbb{P}^{(A,\ell_2)}\right]$.

First, note that entries of $Y^{(A,\ell)}$ are mutually independent. Second, note that for every pair $\ell_1\neq\ell_2\in\mathcal{P}$ and for every $j\in[m/2]$, the distribution of the j^{th} column of $Y^{(A,\ell_1)}$ is identical to the distribution of the j^{th} column of $Y^{(A,\ell_2)}$. Among the last $m/2$ columns, the distributions of at most $4(1+\nu_2)t$ columns of $Y^{(A,\ell_1)}$ differ from the distributions of the corresponding columns in $Y^{(A,\ell_2)}$. Next, consider arbitrary feasible assignment $A\in\mathcal{A}$. Let $\gamma^{(r)}_{\ell_1}, r\in[2(1+\nu_2)t]$, denote the number of strong reviewers (with similarity v) assigned in A to paper $j^{(r)}_1\in\mathcal{T}^*_k(\ell_1)$, where paper $j^{(r)}_1$ corresponds to the the second part of the string $\widetilde{b}^{\ell_1}$ defined in (3.40). Recall now that there are at most $4(1+\nu_2)t$ papers that belong to exactly one of the sets $\mathcal{T}^*_k(\ell_1)$ and $\mathcal{T}^*_k(\ell_2)$. Hence, the equation for upper bound of the Kullback-Leibler divergence between $\mathbb{P}^{(A,\ell_1)}$ and $\mathbb{P}^{(A,\ell_2)}$ is obtained by assuming that all the papers that belong to the $\mathcal{T}^*_k(\ell_1)$ and correspond to the second half of the string $\widetilde{b}^{\ell}$ do not belong to $\mathcal{T}^*_k(\ell_2)$ and vice versa. Thus, similar to (3.36)-(3.37), for arbitrary $\ell_1\neq\ell_2\in\mathcal{P}$ and for arbitrary feasible assignment $A\in\mathcal{A}$, we have

$$\begin{aligned}
\mathrm{KL}\left[\mathbb{P}^{(A,\ell_1)}||\mathbb{P}^{(A,\ell_2)}\right] &\le \sum_{r=1}^{2(1+\nu_2)t}\left\{\gamma^{(r)}_{\ell_1}\mathrm{KL}\left[\mathcal{N}\big(\delta,h(v)\big)||\mathcal{N}\big(0,h(v)\big)\right] + \left(\lambda-\gamma^{(r)}_{\ell_1}\right)\mathrm{KL}\left[\mathcal{N}\big(\delta,h(0)\big)||\mathcal{N}\big(0,h(0)\big)\right]\right\}\\
&+ \sum_{r=1}^{2(1+\nu_2)t}\left\{\gamma^{(r)}_{\ell_2}\mathrm{KL}\left[\mathcal{N}\big(0,h(v)\big)||\mathcal{N}\big(\delta,h(v)\big)\right] + \left(\lambda-\gamma^{(r)}_{\ell_2}\right)\mathrm{KL}\left[\mathcal{N}\big(0,h(0)\big)||\mathcal{N}\big(\delta,h(0)\big)\right]\right\}\\
&= \left(\sum_{r=1}^{2(1+\nu_2)t}\left(\gamma^{(r)}_{\ell_1}+\gamma^{(r)}_{\ell_2}\right)\right)\frac{\delta^2}{2h(v)} + \left(4(1+\nu_2)t\lambda - \sum_{r=1}^{2(1+\nu_2)t}\left(\gamma^{(r)}_{\ell_1}+\gamma^{(r)}_{\ell_2}\right)\right)\frac{\delta^2}{2h(0)}.
\end{aligned}$$

Noting that $\frac{\delta^2}{2h(v)}\ge\frac{\delta^2}{2h(0)}$, we obtain

$$\begin{aligned}
\sup_{A\in\mathcal{A}}\max_{\ell_1\neq\ell_2\in\mathcal{P}}\mathrm{KL}\left[\mathbb{P}^{(A,\ell_1)}||\mathbb{P}^{(A,\ell_2)}\right] &\le 2(1+\nu_2)t\left(\frac{\kappa\delta^2}{h(v)} + \frac{(\lambda-\kappa)\,\delta^2}{h(0)}\right)\\
&= 2(1+\nu_2)t\delta^2\left(\frac{\kappa h(0)+(\lambda-\kappa)\,h(v)}{h(v)h(0)}\right)\\
&\le 4c^2\nu_1\nu_2 t\ln m.
\end{aligned}$$

Applying Fano's inequality (3.30), we obtain the desired lower bound.

9.7 Proof of Theorem 4

Note that Theorem 4 is similar in nature with Theorem 2, the only difference is that now we are trying to recover a ranking which is induced by the assignment.

9.7.1 Proof of upper bound

Given any feasible assignment A, the "ground truth" ranking that we try to recover is given by

$$\widetilde{\theta}^{\star}_j(A) = \frac{1}{\lambda} \sum_{i \in \mathcal{R}_A(j)} \widetilde{\theta}_{ij}. \tag{3.42}$$

Then the estimates $\widehat{\theta}^{\mathrm{MEAN}}_j, j \in [m]$, are distributed as

$$\widehat{\theta}^{\mathrm{MEAN}}_j \sim \mathcal{N}\left(\frac{1}{\lambda} \sum_{i \in \mathcal{R}_A(j)} \widetilde{\theta}_{ij}, \frac{1}{\lambda^2} \sum_{i \in \mathcal{R}_A(j)} \sigma^2_{ij} \right) = \mathcal{N}\left(\widetilde{\theta}^{\star}_j(A), \bar{\sigma}^2_j \right), \tag{3.43}$$

where $\bar{\sigma}^2_j = \frac{1}{\lambda^2} \sum\limits_{i \in \mathcal{R}_A(j)} \sigma^2_{ij}$. Now observe that Lemma 1, with $\mathcal{T}^{\star}_k\left(A, \widetilde{\theta}^{\star}(A)\right)$ substituted for $\mathcal{T}^{*}_k$, also holds for the subjective score model and the averaging estimator $\widehat{\theta}^{\mathrm{MEAN}}$. Thus, repeating the proof of the upper bound for averaging estimator in Theorem 2(a) and substituting $\mathcal{T}^{*}_k$ with $\mathcal{T}^{\star}_k\left(A^{\mathrm{PR4A}}, \widetilde{\theta}^{\star}(A^{\mathrm{PR4A}})\right)$ in (3.25), yields the claimed result.

9.7.2 Proof of lower bound

The lower bound directly follows from Theorem 2(b). To see this, consider the following matrix of reviewers' subjective scores: $\widetilde{\Theta} = \left\{\widetilde{\theta}_{ij}\right\}_{i \in [n], j \in [m]}$, where $\widetilde{\theta}_{ij} = \theta^*_j$. Under this assumption, the total ranking induced by assignment A does not depend on the assignment: $\widetilde{\theta}^{\star}_j(A) = \theta^*_j$. Now we can conclude that such choice of $\widetilde{\Theta}$ brings us to the objective model setup in which true underlying ranking exists and does not depend on the assignment. Thus, the lower bound of Theorem 2(b) transfers to the subjective score model.

9.8 Proof of Theorem 5

The proof of the Theorem 5 is based on the ideas of the proofs of Theorem 3 and Theorem 4 and repeats them with minor changes.

9.8.1 Proof of upper bound

Having equations (3.42) and (3.43), we note that the goal now mimics the goal we achieved when proved an upper bound for averaging estimator in Theorem 3.

9.8.2 Proof of lower bound

The argument from Section 9.7.2 ensures that the lower bound established in Theorem 3 directly transfers to the to the subjective score model.

10 Discussion

Researchers submit papers to conferences expecting a fair outcome from the peer-review process. This expectation is often not met, as is illustrated by the difficulties that non-mainstream or inter-disciplinary research faces in present peer-review systems. We design a reviewer-assignment algorithm PeerReview4All to address the crucial issues of fairness and accuracy. Our guarantees impart promise for deploying the algorithm in conference peer-reviews.

There are number of open problems suggested by our work. The first direction is associated with approximation algorithms and corresponding guarantees established in this work. One goal is to determine

whether there exists a polynomial-time algorithm with worst case approximation guarantees better than $1/\lambda$ established in this work (3.7b). It would also be useful to obtain a deeper understanding of the adaptive behavior of our algorithm with bounds more nuanced than (3.7a). Finally, we leave the task of improving the computational efficiency of our PeerReview4All algorithm out of the scope of this work. However, we suggest that optimal implementation of Subroutine 1 should not be based on the general max-flow algorithm and instead should rely on algorithms specifically designed to work fast on layered graphs.

The second direction is related to the statistical part of our work. In this chapter, we provide a minimax characterization of the simplified version of the paper acceptance problem. This simplified procedure may be considered as an initial estimate that can be used as a guideline for the final decisions. However, there remain a number of other factors, such as self-reported confidence of reviewers or inter-reviewer discussions, that may additionally be included in the model.

Finally, an important related problem is to improve the assessment of similarities between reviewers and papers. It will be interesting to see whether the problems of assessing similarities and assigning reviewers can be addressed jointly in an active manner possibly incorporating feedback from the previous iterations of the conference

Appendix

We provide supplementary materials and additional discussion.

A1 Discussion of approximation results

In this section we discuss the approximation-related results. In what follows we consider function $f(s) = s$ and for any value $c \in \mathbb{R}$, we denote the matrix all of whose entries are c as $\mathbf{c}$.

A1.1 Example for ILPR algorithm.

We begin by construction a series of similarity matrices for various λ such that $\Gamma^S\left(A^{\mathrm{ILPR}}\right) = 0$ while assignments A^{PR4A} and A^{HARD} have non-trivial fairness.

Proposition 1. *For every positive integer λ, there exists a similarity matrix S such that $\Gamma^S\left(A^{ILPR}\right) = 0$ and $\Gamma^S\left(A^{PR4A}\right) \geq \frac{1}{\lambda}\Gamma^S\left(A^{HARD}\right) > 0$.*

Proof. Given any positive integer $\lambda \in \mathbb{N}$, consider an instance of reviewer assignment problem with $m = n$, $\mu = \lambda$ and similarities given by the block matrix

$$S = \left[\begin{array}{c|c|c} \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & (\widetilde{s}-\varepsilon)\cdot\mathbf{1} \\ \hline \underbrace{(\widetilde{s}-\varepsilon)\cdot\mathbf{1}}_{m_1} & \underbrace{(\widetilde{s}-\varepsilon)\cdot\mathbf{1}}_{m_1} & \underbrace{\widetilde{s}\cdot\mathbf{1}}_{m_1} \end{array}\right] \begin{array}{l} \} n_1 \\ \} n_2 \\ \} n_3 \end{array} \tag{3.44}$$

Here $\widetilde{s} = \frac{n_1}{n_1+n_2}$, the value $\varepsilon > 0$ is some small constant strictly smaller than $\widetilde{s}$, and $n_r = m_r > 0$ for every $r \in \{1,2,3\}$. We also require $n_3 > \lambda$ and

$$n_2 = (\lambda - 1)\, n_1 + 1. \tag{3.45}$$

We refer to the first m_1 papers and n_1 reviewers as belonging to the first group, the second m_2 papers and n_2 reviewers as belonging to the second group, and so on.

	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 4$
$\Gamma^S\left(A^{\text{ILPR}}\right)$	0	0	0	0
$\Gamma^S\left(A^{\text{HARD}}\right)$	0.49	0.65	0.72	0.76
$\Gamma^S\left(A^{\text{PR4A}}\right)$	0.49	0.65	0.72	0.76

Table 6: Fairness of various assignment algorithms for the class of similarity matrices (3.44).

The ILPR algorithm involves two steps. The first step consists of solving a linear programming relaxation and finding the most fair fractional assignment. The second step then performs a rounding procedure in order to obtain integer assignments. Let us first see the output of the first step of the ILPR algorithm — the fractional assignment with the highest fairness — on the similarity matrix (3.44). Observe that for each of the m_3 papers in the third group, the sum of the similarities of any λ reviewers is at most $\lambda\widetilde{s}$, and furthermore, that this value is achieved with equality if and only if they are reviewed by λ reviewers from the third group. Next, the n_1 reviewers from the first group can together review λn_1 papers. Dividing this amount equally over the $m_1 + m_2$ papers in the first two groups (in any arbitrary manner) and complementing the assignment with reviewers from the second group, we see that each paper from the first and the second groups receives a sum similarity $\lambda\frac{n_1}{m_1+m_2} = \lambda\widetilde{s}$. It is not hard to see that any deviation from the assignment introduced above will lead to a strict decrease of the fairness.

The second step of the ILPR algorithm is a rounding procedure that constructs a feasible assignment from the fractional assignment (solution of linear programming relaxation) obtained in the previous step. The rounding procedure is guaranteed to assign λ reviewers to each paper, respecting the following condition: any reviewer assigned to any paper $j \in [m]$ in the resulting feasible assignment must have a non-zero fraction allocated to that paper in the fractional assignment.

Now notice that aforementioned condition ensures that all papers from the third group must be assigned to reviewers from the third group. Next, recall that on one hand, reviewers from the first group can together review at most λn_1 different papers. On the other hand, in each optimally fair fractional assignment, the first $m_1 + m_2$ papers are assigned to reviewers from the first two groups. Thus, in the resulting integral assignment these papers also must be assigned to reviewers from the first two groups. These two facts together with the inequality $\lambda n_1 < m_1 + m_2$ that we obtain from (3.45) ensure that at least one paper in the resulting integral assignment will be reviewed by λ reviewers with zero similarity. Hence, the assignment computed by the ILPR algorithm has zero fairness $\Gamma^S\left(A^{\text{ILPR}}\right) = 0$.

On the other hand, it is not hard to see that $\Gamma^S\left(A^{\text{HARD}}\right) \geq \widetilde{s} - \varepsilon$. Indeed, let us assign one reviewer to each paper by the following procedure: the m_1 papers from the first group and some $m_2 - 1$ papers from the second group are all assigned one arbitrary reviewer each from the first group of reviewers. Such an assignment is possible since $\lambda n_1 = m_1 + m_2 - 1$ due to (3.45). The remaining paper from the second group is assigned one arbitrary reviewer from the third group. At this point, there are m_3 papers (in the third group) which are not yet assigned to any reviewer, and $n_3 + n_2 - 1 \geq m_3$ reviewers who have not been assigned any paper and have similarity higher than $\widetilde{s} - \varepsilon$ with these m_3 papers in the third group. Assigning one reviewer each from this set to each of these m_3 papers, we obtain an assignment in which each paper is allocated to one reviewer with similarity at least $\widetilde{s} - \varepsilon$. Completing the remaining assignments in an arbitrary fashion, we conclude that $\Gamma^S\left(A^{\text{PR4A}}\right) \geq \frac{1}{\lambda}\Gamma^S\left(A^{\text{HARD}}\right) \geq \widetilde{s} - \varepsilon > 0$ where first inequality is due to Theorem 1. □

The results of simulations for $\lambda \in \{1, 2, 3, 4\}$, parameters $n_1 = 1, n_2 = \lambda, n_3 = \lambda + 1, \varepsilon = 0.01$ and similarity matrices $\widetilde{S}$ defined in (3.44) are depicted in Table 6. Interestingly, for these choices of parameters, our PeerReview4All algorithm is not only superior to ILPR , but is also able to exactly recover the fair assignment.

	Paper a	Paper b	Paper c	Paper d
Reviewer 1	$0.3+\epsilon$	1	1	0
Reviewer 2	$0.3-\epsilon$	0	1	1
Reviewer 3	0	0.1	0	0.3
Reviewer 4	0	0.1	0	0.3

Table 7: An example of similarities that yield $1/\lambda$ approximation factor of the PeerReview4All algorithm.

A1.2 Sub-optimality of TPMS

In this section we show that assignment obtained from optimizing the objective (3.2) can be highly sub-optimal with respect to the criterion (3.4) even when f is the identity function.

Proposition 2. *For any $\lambda \geq 1$, there exists a similarity matrix S such that $\Gamma^S\left(A^{PR4A}\right) = \Gamma^S\left(A^{HARD}\right) \geq \frac{\lambda}{4}$ and $\Gamma^S\left(A^{TPMS}\right) = 0$.*

Proof. Consider an instance of the problem with $m = n = 2\lambda$, and similarities given by the block matrix

$$S = \left[\begin{array}{c|c} \mathbf{1} & \mathbf{0.4} \\ \hline \underbrace{\mathbf{0.4}}_{\lambda} & \underbrace{\mathbf{0}}_{\lambda} \end{array}\right] \begin{matrix} \} \lambda \\ \} \lambda \end{matrix} \tag{3.46}$$

Then A^{TPMS} assigns the first λ reviewers to the first λ papers (in some arbitrary manner) and the remaining reviewers to the remaining papers, obtaining

$$\sum_{j\in[m]} \sum_{i\in\mathcal{R}_{A^{\text{TPMS}}}(j)} s_{ij} = \lambda^2 \text{ and}$$
$$\Gamma^S\left(A^{\text{TPMS}}\right) = 0$$

In contrast, assignments A^{PR4A} and A^{HARD} assign the first $\frac{1}{2}n$ reviewers to the second group of papers and the remaining reviewers to the remaining papers. This assignment yields

$$\sum_{j\in[m]} \sum_{i\in\mathcal{R}_{A^{\text{PR4A}}}(j)} s_{ij} = \sum_{j\in[m]} \sum_{i\in\mathcal{R}_{A^{\text{HARD}}}(j)} s_{ij} = 0.8\lambda^2 \text{ and}$$
$$\Gamma^S\left(A^{\text{PR4A}}\right) = \Gamma^S\left(A^{\text{HARD}}\right) = 0.4\lambda \geq \frac{\lambda}{4}.$$

This concludes the proof. □

A1.3 Example of $1/\lambda$ approximation factor for A^{PR4A}

Let us consider an instance of fair assignment problem with $m = n = 4$, $\lambda = \mu = 2$ and similarities represented in Table 7.

First, note that $\Gamma^S\left(A^{\text{HARD}}\right) \leq 0.6$. This is because in every feasible assignment $A \in \mathcal{A}$ paper 1 in the best case is assigned to reviewers 1 and 2. Moreover, there exists a feasible assignment represented as A^{HARD} in Table 8 which achieves a max-min fairness of 0.6 and hence we have $\Gamma^S\left(A^{\text{HARD}}\right) = 0.6$.

Let us now analyze the performance of PeerReview4All algorithm. Again, the fairness of the resulting assignment is determined in the first iteration of Step 2 to 7 of Algorithm 1, so we restrict our attention to that part of the algorithm. It is not hard to see that after Step 2 is executed, we have two candidates assignments, A_1 and A_2, represented in Table 8 (up to not important randomness in braking ties). Computing the fairness of these assignments, we obtain

$$\Gamma^S(A_1) = 0.3 + \varepsilon \quad \text{and} \quad \Gamma^S(A_2) = 0.2.$$

	A^{HARD}		A_1		A_2	
	1ST Reviewer	2ND Reviewer	1ST Reviewer	2ND Reviewer	1ST Reviewer	2ND Reviewer
Paper a	1	2	1	3	1	2
Paper b	1	3	1	3	3	4
Paper c	2	4	2	4	1	2
Paper d	3	4	2	4	3	4

Table 8: The optimal assignment as well as and PeerReview4All 's intermediate assignments for the similarities in Table 7.

which implies that

$$\frac{\Gamma^S\left(A^{\text{PR4A}}\right)}{\Gamma^S\left(A^{\text{HARD}}\right)} = \frac{\max\left\{\Gamma^S\left(A_1\right), \Gamma^S\left(A_2\right)\right\}}{\Gamma^S\left(A^{\text{HARD}}\right)} = \frac{1}{2} + \frac{\varepsilon}{0.6}.$$

Setting ϵ small enough, we can see that the approximation factor is very close to $1/2 = 1/\lambda$.

A2 Computational aspects

A naïve implementation of the PeerReview4All algorithm has a polynomial computational complexity (under either an arbitrary choice or one computable in polynomial-time in Step 6) and requires $\mathcal{O}\left(\lambda m^2 n\right)$ iterations of the max-flow algorithm. There are a number of additional ways that the algorithm may be optimized for improved computational complexity while retaining all the approximation and statistical guarantees.

One may use Orlin's method (Orlin, 2013; King et al., 1992) to compute the max-flow which yields a computational complexity of the entire algorithm at most $\mathcal{O}\left(\lambda(m+n)m^3n^2\right)$. Instead of adding edges is Step 3 of the subroutine one by one, a binary search may be implemented, reducing the number of max-flow iterations to $\mathcal{O}\left(\lambda m \log mn\right)$ and the total complexity to $\widetilde{\mathcal{O}}\left(\lambda(m+n)m^2n\right)$.

Finally, note that the max-min approximation guarantees (Theorem 1), as well as statistical results (Theorems 2 to 5 and corresponding corollaries) remain valid even for the assignment $\widetilde{A}$ computed in Step 3 of Algorithm 1 during the *first* iteration of the algorithm. The algorithm may thus be stopped at any time after the first iteration if there is a strict time-deadline to be met. However, the results of Corollary 1 on optimizing the assignment for papers beyond the most worst-off will not hold any more.[9] The computational complexity of each of the iterations is at most $\widetilde{\mathcal{O}}\left(\lambda(m+n)mn\right)$, and stopping the algorithm after a constant number of iterations makes it comparable to the complexity of TPMS algorithm which is successfully implemented in many large scale conferences.

Let us now briefly compare the computational cost of PeerReview4All and ILPR algorithms. The full version of ILPR algorithm requires $\mathcal{O}(m^2)$ solutions of linear programming problems. Given that finding a max-flow in a graph constructed by our subroutine can be casted as linear programming problem (with constraints similar to those in Garg et al. 2010), we conclude that slightly optimized implementation of our algorithm results in $\mathcal{O}(\lambda m \log mn)$ solutions of linear programming problems, which is asymptotically better. To be fair, the ILPR algorithm also can be terminated in an earlier stage with theoretical guarantees satisfied, which brings both algorithms on a similar footing with respect to the computational complexity.

A3 Topic coverage

In this section we discuss an additional benefit of "topic coverage" that can be gained from the special choice of heuristic in Step 6 of Subroutine 1 of our PeerReview4All algorithm.

[9] If the algorithm is terminated after p' iterations, then bound (3.8) from Corollary 1 holds for $r \in [p']$.

Research is now increasingly inter-disciplinary and consequently many papers submitted to modern conferences make contributions to multiple research fields and cannot be clearly attributed to any single research area. For instance, computer scientists often work in collaboration with physicists or medical researchers resulting in papers spanning different areas of research. Thus, it is important to maintain a broad topic coverage, that is, to ensure that such multidisciplinary papers are assigned to reviewers who not only have high similarities with the paper, but also represent the different research areas related to the paper. For example, if a paper proposes an algorithm to detect new particles in the CERN collider, then that paper should ideally be evaluated by competent physicists, computer scientists, and statisticians.

There are prior works both in peer-review (Long et al., 2013) and in text mining (Lin and Bilmes, 2011) which propose a submodular objective function to incentivize topic coverage. According to Long et al. (2013), the appropriate measure of coverage is a number of distinct topics of the paper covered, summed across the all papers. Let us introduce a piece of notation to formally describe the underlying optimization problem. For every paper $j \in [m]$, let $T(j) = \{t_1^{(j)}, \ldots, t_{r_j}^{(j)}\}$ be related research topics and for every reviewer $i \in [n]$, let $T(i) = \{t_1^{(i)}, \ldots, t_{r_i}^{(i)}\}$ be the topics of expertise of reviewer i. For every assignment A, we define $\omega(A)$ to be the total number of distinct topics of all papers covered by the assigned reviewers:

$$\omega(A) = \sum_{j \in [m]} \operatorname{card}\left(\bigcup_{i \in \mathcal{R}_A(j)} \Big(T(j) \bigcap T(i) \Big) \right), \tag{3.47}$$

where $\operatorname{card}(\mathcal{C})$ denotes the number of elements in the set $\mathcal{C}$. The goal in Long et al. (2013) is to find an assignment that maximizes $\omega(A)$ and respects the constraints on the paper/reviewer load. However, instead of the requirement that each paper is assigned to λ reviewers as in our work, Long et al. (2013) consider a relaxed version and require each paper to be reviewed by at most λ reviewers.

Using the submodular nature of the objective (3.47), Long et al. (2013) propose a greedy algorithm that is guaranteed to achieve a constant-factor approximation of the optimal coverage (3.47). This greedy algorithm, however, has the following two important drawbacks:

(i) Like the TPMS algorithm, the greedy algorithm aims at optimizing the global functional, and consequently may fare poorly in terms of fairness. Indeed, in order to optimize the global objective (3.47), the greedy algorithm may sacrifice the topic coverage for some of the papers, assigning relevant reviewers to other papers.

(ii) While guaranteed to achieve a constant factor approximation of the objective (3.47), the greedy algorithm may yield an assignment in which papers are reviewed by (much) less than λ reviewers. It is not even guaranteed that in the resulting assignment each paper has at least one reviewer.

Nevertheless, both the PeerReview4All algorithm and the algorithm of Long et al. (2013) can benefit from each other if the latter is used as a heuristic to choose a feasible assignment in Step 6 of the subroutine of the former. In what follows we detail the procedure to combine the two algorithms. The greedy algorithm of Long et al. (2013) picks (reviewer, paper) pairs one-by-one and adds them to the assignment. At each step, it picks the pair that yields the largest incremental gain to (3.47) while still meeting the paper/reviewer load constraints. In Step 6 of the subroutine of PeerReview4All, we may use the greedy algorithm, restricted to the (reviewer, paper) pairs added to the network in the previous steps, to find an assignment that approximately maximizes (3.47). Next, for every (reviewer, paper) pair that belongs to this assignment, we set the cost of the corresponding edge in the flow network to 1 and the costs of the remaining edges to 0. Finally, we compute the maximum flow with maximum cost in the resulting network and fix (reviewer, paper) pairs that correspond to edges employed in that flow in the final output of the subroutine.

Let us now discuss the benefits of this approach. First, in PeerReview4All we modify only the procedure of tie-breaking among max-flows, and hence all the guarantees established in the paper continue to hold. Second, the introduced procedure allows to overcome the issue (ii), because the max-flow guarantees that each paper is assigned with exactly requested number of reviewers. Third, by setting the cost of selected edges to 1, we encourage the topic coverage (although the pproximation guarantee of the greedy algorithm no

longer holds). Finally, we do not allow the algorithm of Long et al. (2013) to sacrifice some papers in order to maximize the global coverage (3.47), because the subroutine ensures that in the resulting assignment all the papers are assigned to pre-selected reviewers with high similarity, thereby overcoming (i).

Part II

Bias and Policies

Chapter 4

Identity-Related Biases in Single-Blind Peer Review

1 Introduction

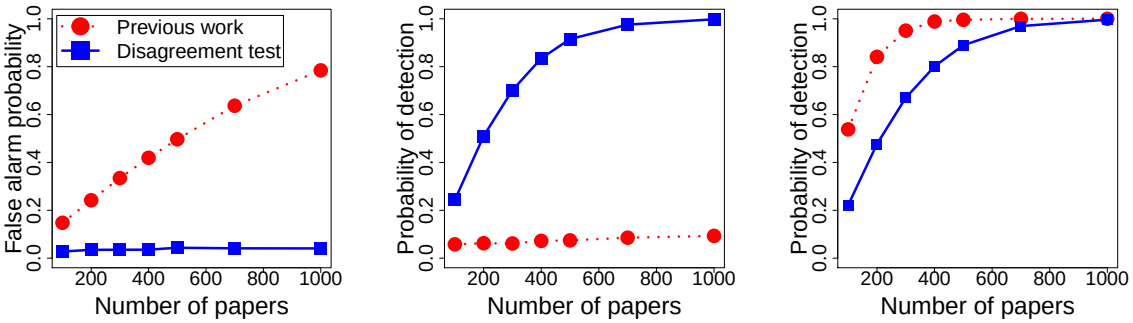
Past research in social sciences indicates that humans display various biases including gender, race and age biases in many critical domains such as hiring (Bertrand and Mullainathan, 2004), university admission (Thornhill, 2018), bail decisions (Arnold et al., 2018) and many others. This chapter considers the problem of identity-related biases in scientific peer review. Specifically, we follow the long-standing debate (Blank, 1991; Seeber and Bacchelli, 2017; Snodgrass, 2006; Largent and Snodgrass, 2016; Okike et al., 2016; Budden et al., 2008; Webb et al., 2008; Hill and J. Provost, 2003, and references therein) on whether the authors' identities should be hidden from reviewers or not. *In that, we focus on designing statistical tests to detect the presence of identity-related biases in single-blind peer review.*

In a remarkable piece of work, Tomkins et al. (2017) conducted a large scale (semi-) randomized controlled trial during the peer review for the ACM International Conference on Web Search and Data Mining (WSDM) 2017. In their experiment, the entire pool of reviewers was partitioned uniformly at random into two equal groups – single blind and double blind – and each paper was assigned to two reviewers from each of the groups. In this manner, the peer-review data contained both single-blind and double-blind reviews for each paper. The experiment allowed them to conduct a causal inference to test for biases, and conclude that the single-blind system induces a bias in favor of papers authored by (i) researchers from top-universities, (ii) researchers from top companies and (iii) famous authors. Interestingly, no bias against female-authored submissions was detected by their test, though a meta-analysis confirmed the presence of such bias. The conclusions of this experiment have had a significant impact. For instance, the WSDM conference itself completely switched to double-blind peer review starting 2018.

Testing for the presence of hypothesized phenomena is a common task in various branches of science including the biological, social, and physical sciences. The general approach therein is to impose a hard constraint on the probability of false alarm (claiming existence of the phenomenon when there is none; also called Type-I error) to some predefined threshold called significance level typically set as 0.05 or 0.01. The test would then aim to maximize the probability of detecting the phenomenon when it is actually present, while not violating the aforementioned hard constraint. The present work also follows this general approach, for the specific setting of testing for biases using single versus double blind reviewing.

Contributions. In this chapter, we study the problem of detecting bias in peer review, and present two sets of results.

(1) Detailed investigation into methodology of past work (Section 3) We first analyze the testing procedure used by Tomkins et al. (2017), and show that under plausible conditions the statistical test employed therein does not control for false alarm probability. In other words, *we show that under reasonable*



(a) Bias is absent. Any valid test must have false alarm probability below 0.05.

(b) Bias is present. Higher probability of detection is better.

(c) Bias is present. Higher probability of detection is better.

Figure 1: Synthetic simulations evaluating performance of the test in Tomkins et al. (2017) ("previous work") and the test proposed in this work ("DISAGREEMENT test"). Subfigures (a) and (b) are in presence of correlations and noisy estimates of true scores by double-blind reviewers; subfigure (c) has zero correlations and perfect estimate of true scores by double-blind reviewers. Details of the simulation setup are provided in Section 3. The error bars are too small to be visible.

conditions, the test used by Tomkins et al. (2017) can, with probability as large as 0.5 or higher, declare the presence of a bias when the bias is in fact absent (even when the test is tuned to have a false alarm error rate below 0.05). Specifically, we show that in presence of correlations that are reasonable to expect, any of the following factors breaks their false alarm probability guarantees: (a) measurement error caused by noise or subjectivity of reviewers, (b) model mismatch caused by violation of strong parametric assumptions on reviewers' behavior and (c) reviewer's calibration if she/he reviews more than one paper. Figures 1a and 1b illustrate the effect of measurement error on the false alarm probability and probability of detection of the test used by Tomkins et al. The issues we identify suggest that their test is at risk of committing Type-I error in declaring biases in their analysis.

Moving beyond the specific test used in Tomkins et al. (2017), we also study the effect of their experimental design, which is simply the standard peer-review procedure with an additional random partition of reviewers into single and double blind groups. We show that two factors – (d) asymmetrical bidding procedure and (e) non-random assignment of papers to referees – as is common in peer-review procedures today may introduce spurious correlations in the data, breaking some key independence assumptions and thereby violating the requisite guarantees on testing.

(2) Novel approach to testing for biases (Sections 4 - 6) We propose a general framework for the design of statistical tests to detect biases in this problem setting, that overcomes the aforementioned limitations. Specifically, our framework does not assume objectivity of reviewers and does not make any parametric assumptions on reviewers' behaviour. Conceptually, we propose to think of this problem as an instance of a two-sample testing problem where single-blind and double-blind reviews form two samples and the test operates on these samples. (In contrast, Tomkins et al. (2017) study the problem under one-sample testing paradigm, operating on reviews of single-blind reviewers and using double-blind reviews to estimate some parameters in their parametric model).

We then design computationally-efficient hypothesis testing procedures that under minimal assumptions guarantee a provable control over the false alarm probability under various conditions, including aforementioned conditions (a) - (c). We supplement these tests with an alternative design of the experimental setup which coupled with our tests mitigates issues (d) - (e) while not restricting the choice of assignment algorithm.

Our tests also have non-trivial power in that they have considerably higher probability of detection in hard cases where test used by Tomkins et al. fails, and a power comparable to that of Tomkins et al. when their

assumptions are exactly met. The performance of one of these tests is illustrated in Figure 1. Additionally, we show that assumptions required by our tests to control for the Type-I error rate are essentially minimal in that they cannot be further relaxed without making reliable testing impossible.

We note that while the discussion in this chapter focuses on testing for biases with respect to protected attributes, our experimental setup and statistical tests are not restricted to that alone. Instead of comparing the single versus double blind settings, our work can be used to test for effects of aspects of a submission exogenous to the manuscript's content, for instance, the effects of the reviewer questionnaire or that of asking authors to provide extraneous information (such as prior submission history). Our work enables conducting such semi-randomized controlled trials while retaining the no-bias and veracity conditions (Tomkins et al., 2017), not requiring additional reviews, and having rigorous guarantees on the tests.

Related work The problem of identifying biases in human decisions is commonly studied in social science and there are many works that design and conduct randomized field experiments in various settings, including resume screening (Bertrand and Mullainathan, 2004), hiring in academia (Moss-Racusin et al., 2012), and peer review (Blank, 1991; Okike et al., 2016). However, the conference peer review setup we consider in this work does not comprise a fully randomized control trial (i.e., the reviewers are not assigned to submissions at random) and past approaches fail due to idiosyncrasies of the peer-review process. For example, a popular approach (Bertrand and Mullainathan, 2004; Moss-Racusin et al., 2012) is to assign author identities to (fabricated) documents (resumes, application packages or papers) uniformly at random and compare the outcomes for different categories of authors. In our setup, *random assignment of author identities to real (i.e., non-fabricated) submissions* is problematic due to various logistical and ethical issues such as reviewers guessing actual authors thereby causing biases, and requrements of getting authors to agree to have their paper/name modified. Another approach (Okike et al., 2016) is to submit *the same paper* to multiple reviewers in both single-blind and double-blind conditions and test for the difference in the acceptance rates between conditions. However, such an approach necessitates a considerable additional reviewing load. Other approaches include observational studies, and we refer the interested readers to Tomkins et al. (2017) for a more in-depth literature review.

It is important to note that in this work, we do not aim to prove or disprove the existence of biases declared in the experiment by Tomkins et al. (2017). Instead, our focus is on the theoretical validity of the statistical procedures used to conduct such experiments and more generally on principled statistical approach towards designing such experiments.

Finally, the results and tests we discuss in this work are also applicable beyond peer review, and can be used to test for biases in other domains such as admissions and hiring.

The remainder of this chapter is organized as follows. In Section 2 we present the problem setting formally and describe the experimental setup of Tomkins et al. (2017). In Section 3 we uncover issues (a) - (e) with their test and setup and illustrate the detrimental effect of such issues through simulations. Next, in Sections 4 and 5 we present a novel non-parametric approach to testing for biases and corresponding statistical tests as well as the alternative design of the experimental procedure. The detailed analysis is given in Section 6. We conclude the chapter with a discussion in Section 7.

2 Preliminaries

The general peer-review setup we study for testing biases using single and double blind review is as considered in Tomkins et al. (2017). We study a conference peer-review setup where n papers are submitted at once and m independent reviewers are available to review submissions, where m is assumed to be an even number. With a goal to test whether single-blind reviewing induces a bias against or in favor of some groups of authors, we consider some pre-defined set of k binary mutually non-exclusive properties pertaining to the author(s) of any paper to be tested for bias. For example, a property could be "the first author is female" or "majority of authors are from the USA". Each paper $j \in [n]$ is then associated with k indicator variables $w_j^{(1)}, \ldots, w_j^{(k)}$, where $w_j^{(\ell)} = 1$ if paper j satisfies property ℓ and $w_j^{(\ell)} = -1$ otherwise. For each $\ell \in [k]$ we let $\mathcal{J}_\ell \subseteq [n]$ denote

the set of papers that satisfy property ℓ and $\overline{\mathcal{J}}_\ell = [n] \backslash \mathcal{J}_\ell$ denote its complement.[1]

For each property $\ell \in [k]$ we are interested in whether *single-blind peer review setup induces a bias* against or in favor of papers that satisfy this property. For example, if we consider property "the first author is female", then we aim at testing for the bias against or in favor of papers with female first author. Note that with respect to the properties, the study is observational in that we cannot assign author identities to papers at random. Hence, the effect of confounding is unavoidable and utmost care must be taken to address presence of confounding factors.

For brevity, in the main text we consider the case of a single property of interest ($k = 1$) which captures the complexity of our problem. For ease of notation we drop index ℓ from $w^{(\ell)}$ and $\mathcal{J}_\ell$. In Appendix A1 we generalize the results to $k > 1$. Let us now give details of the testing procedure used by Tomkins et al. (2017).

Experimental setup of Tomkins et al. The peer review process in their experiment is organized as follows. Reviewers are uniformly at random divided into two groups of equal sizes, corresponding to two conditions: (i) Double-Blind condition (DB) in which reviewers do not observe identities of papers' authors; and (ii) Single-Blind condition (SB) in which reviewers observe identities of the papers' authors. Next, each paper is assigned to λ reviewers from the SB group and λ reviewers from the DB group such that each reviewer reviews at most μ submissions, where λ and μ are predefined constants. In both conditions, if any reviewer $i \in [m]$ is assigned to any paper $j \in [n]$, then she/he returns a binary accept/reject recommendation and possibly a numeric score that estimates a quality of the paper as perceived by reviewer, accompanied by a textual review.

Model and test used by Tomkins et al. We begin by introducing an idealized version of their model. They assume a parametric, logistic model for the binary decisions made by SB reviewers. Specifically, for each paper $j \in [n]$, let $Y_{1j}, \ldots, Y_{\lambda j}$ denote the binary accept/reject decisions given by the λ reviewers assigned to paper j in the SB setup. It is assumed that $\{Y_{rj}\}_{r \in [\lambda]}$ are independent draws from a Bernoulli random variable with an expectation π_j satisfying

$$\log \frac{\pi_j}{1 - \pi_j} = \beta_0 + \beta_1 q_j^* + \beta_2 w_j, \tag{4.1}$$

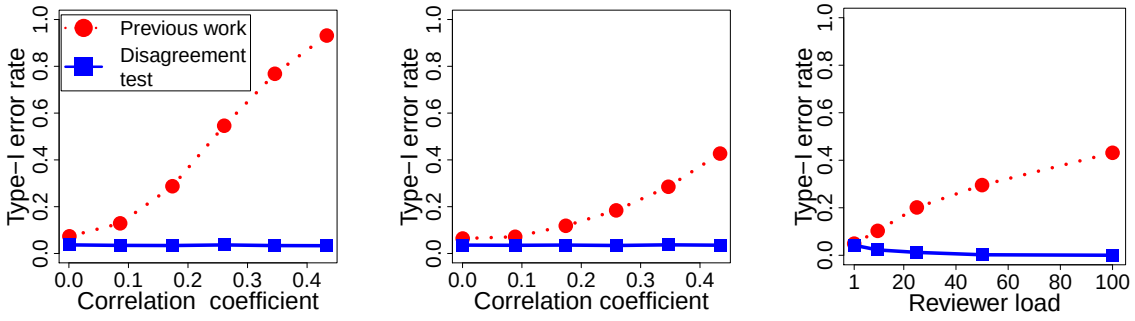
where q_j^* is a "true" underlying score of paper j, w_j is an indicator of property satisfaction and $\{\beta_0, \beta_1, \beta_2\}$ are unknown coefficients. In words, the model says that if there is a positive (respectively negative) bias with respect to a property of interest, then the fact that paper satisfies the property increases (respectively decreases) the log-odds of the probability of recommending acceptance by $2\beta_2$ as compared to the case if the same paper does not satisfy the property. The main difficulty with this model in the peer review setting lies in the fact that true scores $\{q_j^*, j \in [n]\}$ are unknown and hence standard tests for logistic regression model are not readily applicable.

In order to overcome the unavailability of true scores $\{q_j^*, j \in [n]\}$ in the model (4.1), Tomkins et al. (2017) use a plug-in estimate: they replace q_j^* with the mean $\widetilde{q}_j$ of scores given by the DB reviewers to paper j, for every $j \in [n]$. Under this approximation and using $\widetilde{q}_1, \ldots, \widetilde{q}_n$, they obtain maximum likelihood estimates of coefficients $\{\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2\}$ and then use the standard Wald test (Weisberg, 2005) to test for significance of the coefficient β_2. A bias is declared present if the coefficient β_2 is found significant; the direction of the bias is determined as the sign of $\widehat{\beta}_2$.

3 Problems with the past approach

In this section we identify several issues that should be taken into account when testing for biases in the setup we consider. Noting that the issues themselves are general, we motivate and discuss them in context of the prior work by Tomkins et al. (2017) and investigate possible consequences of these issues through synthetic simulations. In the simulations to follow, we juxtapose algorithm by Tomkins et al. (2017) to our DISAGREEMENT test introduced later in the paper. Complete details of all simulations are given in Appendix A5.

[1] Here, we adopt the standard notation $[\nu] = \{1, 2, \ldots, \nu\}$ for any positive integer ν.



(a) Measurement error (b) Model mismatch (c) Reviewer calibration

Figure 2: Type-I error of the test from previous work (Tomkins et al. 2017) blows up under three different setups: bias is absent in all simulations and the tests are designed to limit the Type-I error to at most 0.05. In contrast, our DISAGREEMENT test is robust to violations of modelling assumptions. Error bars are too small to be visible.

3.1 Testing procedure

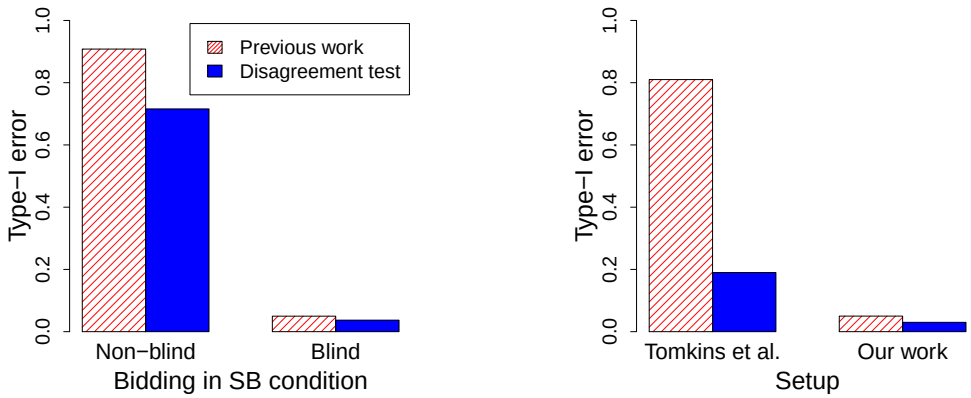
We begin from the issues that are pertinent to the testing procedure used by Tomkins et al. (2017). To this end, recall that with respect to the property of interest the experiment is observational. Hence we *cannot* assume independence between the indicator of property satisfaction w and the true score q^*. Moreover, a non-trivial amount of correlation between some properties is plausible. Consider for example a property "paper has author from top univeristy". For this property a non-trivial correlation between true scores and indicator of property satisfaction is natural to expect. While correlation itself does not cause issues, we identify three conditions which coupled with correlation can be significantly harmful.

(a) Measurement error Tomkins et al. (2017) report low interreviewer agreement between DB reviewers which means that the estimates $\widetilde{q}_1, \ldots, \widetilde{q}_n$ of the true scores by the DB reviewers are noisy. It is known (Stefanski and Carroll, 1985; Brunner and Austin, 2009) that noisy covariate measurement coupled with correlation between some covariates may inflate the Type-I error rate of the Wald test for logistic regression. We now investigate the impact of measurement error on the Type-I error rate of the Tomkins et al. test through simulations. We consider absence of any bias, and assume that model (4.1) with $\beta_2 = 0$ is correct for both DB and SB reviewers. We consider DB reviewers to report noisy estimates of true scores q_j^*, and vary the correlation between q^* and w. The level of noise was selected to keep correlation between the two DB reviewers assigned to each paper at the level of 0.6, which is much better than the actual interreviewer agreement observed by Tomkins et al. (2017) (correlation 0.37). We plot the Type-I error rates in Figure 2a for the test in Tomkins et al. (2017) and our proposed test, both tests are designed to restrict the Type-I error rate to 0.05.

Figure 2a indicates a strong detrimental effect of measurement error on the validity of the test by Tomkins et al. (2017). Given that interreviewer agreement in the actual WSDM conference experiment was low, the fact that some properties considered by Tomkins et al. may lead to correlations between q^* and w is concerning, because it could potentially undermine the validity of their findings.

The simulations in Section 1 follow the setup presented here: Figures 1a and 1b consider measurement error with correlation fixed at 0.4 (Figure 1a) and 0.6 (Figure 1b) and show that (a) the negative effect of measurement error on the Type-I error rate exacerbates as sample size grows and (b) measurement error may also hinder the power of the test. Figure 1c has zero correlation and no measurement error, satisfying all the assumptions of the test by Tomkins et al.

(b) Model mismatch Model (4.1) assumes a specific parametric relationship, which may not hold in practice. In order to check the effect of model mismatches, we consider a violation of the model (4.1) and suppose that



(a) Non-blind bidding

(b) Reviewer assignment

Figure 3: The experimental setup from previous work (Tomkins et al. 2017) violates Type-I error guarantees of testing procedures. Bias is absent in all simulations and the tests are designed to limit the Type-I error to at most 0.05. Note that the issues which pertain to the experimental setup rather than the modelling break guarantees of both tests (leftmost columns). In contrast, our proposed setup with fully blind bidding procedure and careful management of assignment ensures Type-I error guarantees for both tests (rightmost columns). Error bars are too small to be visible.

the correct model for both SB and DB reviewers is

$$\log \frac{\pi_j}{1-\pi_j} = \beta_0 + \beta_1 \left(q_j^*\right)^3 + \beta_2 w_j,$$

that is, instead of expected linear input, true scores of papers appear in the model raised to the power 3. To isolate the effect of model mismatch, we assume that true scores $q_j^*, j \in [n]$, are known exactly to the test of Tomkins et al. and hence abstract out the impact of the measurement error. We again consider an absence of any bias and set $\beta_2 = 0$ for both SB and DB reviewers. We then perform simulations similar to those in item (a). Figure 2b shows the results of the simulations.

(c) Reviewer calibration The test employed by Tomkins et al. (2017) treats reviews given by the same reviewer as independent. In practice this assumption may be violated due to correlations introduced by reviewer's calibration (Wang and Shah, 2019). While some easy calibrations such as harshness/leniency can be captured by simple parametric extensions of model (4.1), more subtle patterns are beyond the scope of this model. Suppose for example that the strength of reviewers' input depends on paper's clarity — the better the paper is written, the lower the contribution due to reviewers' calibration. Assume also that we are given a set of papers such that true score of each paper is proportional to the clarity of the paper (we formalize construction in Appendix A5.1.3). Coupled with the correlation between q^* and w, this pattern is sufficient to break Type-I error guarantees of the test of Tomkins et al. Again, to isolate the impact of reviewers' calibration, we assume that (i) true scores $q_j^*, j \in [n]$, are known to the test by Tomkins et al. and (ii) model (4.1) is marginally correct for each reviewer, that is, each reviewer follows model (4.1) for each paper she/he reviews, but her/his decisions for different papers are correlated in a specific way.

Figure 2c shows a result of simulations in which we vary the number of papers per reviewer, keeping correlation between q^* and w fixed at 0.75 and the total number of papers fixed at $n = 1000$. We simulate a wide range of reviewer load μ including small to medium loads of 5-15 papers typical in machine learning conferences like NeurIPS and larger loads of 40 or higher found in other smaller conferences.

3.2 Experimental setup

The issues discussed above pertain to the testing procedure and modelling assumptions made by Tomkins et al. (2017). We now issue a commentary regarding the experimental setup considered in their work which

comprises a random partition of reviewers into SB and DB groups within a standard peer review procedure. In particular, we show that the setup itself may create problems in controlling the Type-I error.

(d) Non-blind bidding In the experiment by Tomkins et al. papers are allocated to reviewers based on preferences ("bids") declared by reviewers (reviewers could indicate that they want to review some papers and do not want to review others). Importantly, the reviewers in the SB setup also get to see author identities in the bidding stage, which may act as a confounding factor in tests for bias in the acceptance/rejection of papers. This is indeed pointed out as a caveat by Tomkins et al. (2017) in their paper.

To illustrate the possible effect, consider a property of interest "paper has a famous author" and suppose that among all reviewers there is a subset of lenient reviewers who additionally want to read papers from top authors with the hope of reading better papers. Then in DB setup such reviewers cannot use author identity information and hence make their bidding decisions based on title and abstract only; in contrast, in SB setup these reviewers tend to bid on papers authored by top authors. Given that reviewers who preferentially bid on papers with top authors in SB condition are by coincidence lenient, the difference in bidding behavior may result in structurally different evaluations between conditions even when reviewers' evaluations are unbiased, leading to a blow-up of the Type-I error rate of any reasonable test. Figure 3a shows a result of simulations (formal setup is in Appendix A5.1.4) in which we compare non-blind and blind bidding conditions for SB reviewers and indicates a possible detrimental effect of non-blind bidding.

(e) Reviewer assignment One might imagine that a natural requirement to conduct the bidding in a double blind fashion for both DB and SB reviewers would fix the issues with the setup of Tomkins et al. However, perhaps surprisingly, we show that even if both groups bid in a double blind fashion (or even if the bidding process is eliminated entirely), and even if the reviewers are assigned to DB or SB groups uniformly at random, the non-random assignment using algorithms such as TPMS (Charlin and Zemel, 2013) that assigns reviewers to papers maximizing some notion of "similarities" can still lead to a violation of the Type-I error guarantees. We give a formal construction in Appendix A5.1.5; the intuition is as follows. Quoting Lamont (2009), "evaluators often define excellence as ¡¡what speaks to me¿¿ which is akin to ¡¡what is most like me¿¿", that is, a similarity between a paper and a reviewer may influence the decision. That said, we construct these similarities in a careful manner: our choice ensures that despite reviewers being allocated to DB or SB conditions at random, the popular TPMS assignment algorithm with high probability constructs assignments that are in some sense structurally different between SB and DB conditions, which in turn leads to structurally different evaluations. Our construction, along with correlation between q^* and w, introduces spurious correlations in the data, thereby violating some key independence assumptions and leading to the inflation of Type-I error.

Figure 3b shows a result of simulations in which we compare the setup of Tomkins et al. (2017) with our proposed experimental setup introduced in Section 5.2. Notably, under the setup of Tomkins et al. even the Disagreement test which is robust to various issues discussed in Section 3.1 is unable to control for the Type-I error. In contrast, observe that under our proposed experimental setup, both the Disagreement test and the test by Tomkins et al. control for the Type-I error rate at the desired level.

Importantly, we underscore that while our experimental procedure mitigates the issues with the experimental setup of Tomkins et al., their test is still *susceptible to the issues we discussed in Section 3.1* even under our experimental setup. Finally, under the setup of Tomkins et al. the phenomenon of the Type-I guarantee violation is not restricted to the TPMS assignment and can occur in a much broader class of reviewer assignment algorithms.

4 Novel framework to test for biases

In Section 3 we identified five key limitations of the approach taken by Tomkins et al. (2017). Three of these limitations pertain to the testing procedure and the two limitations relate to the design of the experiment itself. In the next sections we design a set of tests and experimental setup with strong guarantees, and which overcome the aforementioned limitations. In this section we begin from principled definition of a bias testing problem that generalizes one made by Tomkins et al. and does not make any restrictive assumptions.

At a high level, our approach to testing for biases is different from those proposed by Tomkins et al. in two ways. First, we relax two strict modelling assumptions: (i) instead of assuming existence of true qualities of submissions, we allow subjectivity in reviewer evaluations (Kerr et al., 1977; Ernst and Resch, 1994; Bakanic et al., 1987; Mahoney, 1977; Lamont, 2009; Noothigattu et al., 2020), and (ii) we do not assume any specific form of the relationship between a paper and its probability of acceptance by a reviewer. Instead, we allow these probabilities to be completely arbitrary and define the bias in terms of these probabilities. Second, we treat this problem conceptually differently from the work of Tomkins et al. The test therein treats the problem as that of one-sample testing and uses DB scores as a plugin estimate of true scores in SB model. In contrast, we approach this problem through the lenses of two-sample testing, where SB and DB reviews form the two samples, and the goal is to test whether they belong to the same distribution. This perspective helps us to avoid a number of issues discussed in Section 3.

Formally, let $\Pi^{\mathrm{db}} \in [0,1]^{m \times n}$ be a matrix whose $(i,j)^{\mathrm{th}}$ entry, denoted as $\pi_{ij}^{(\mathrm{db})}$, represents a probability that reviewer i would recommend acceptance of paper j if that paper is assigned to that reviewer in DB setup. Similarly, let matrix $\Pi^{\mathrm{sb}} \in [0,1]^{m \times n}$ be an analogous matrix in SB setup, and denote its $(i,j)^{\mathrm{th}}$ entry as $\pi_{ij}^{(\mathrm{sb})}$.

Let $\mathcal{R}_{\mathrm{SB}}$ be the set of reviewers allocated to the SB condition. Moreover, for each $i \in \mathcal{R}_{\mathrm{SB}}$, let $\mathcal{P}_{\mathrm{SB}}(i)$ denote the set of papers assigned to reviewer i and let $Y_{ij} \in \{0,1\}$ denote the accept/reject decision given by reviewer i for paper $j \in \mathcal{P}_{\mathrm{SB}}(i)$. We similarly define set of DB reviewers $\mathcal{R}_{\mathrm{DB}}$ and their decisions $\{X_{ij} : i \in \mathcal{R}_{\mathrm{DB}}, j \in \mathcal{P}_{\mathrm{DB}}(i)\}$. We are interested in testing for biases with respect to a property of interest. To this end, recall our notation $\mathcal{J} \subseteq [n]$ for the set of papers that satisfy a property of interest, and $\overline{\mathcal{J}}$ as its complement.

With this notation in place, we now define two formulations of the bias testing problem — "absolute" and "relative": The relative bias setting is strictly more general than the absolute bias setting, but also leads to more restrictive results.[2] Importantly, the tests we will introduce in Section 5.1 are applicable to both formulations without additional modifications.

4.1 Absolute bias problem

In the absence of bias, the knowledge of authors' identities does not induce any difference in reviewers' behaviour. In the biased hypothesis, there is a positive bias in favor of papers that satisfy a property of interest: reviewers in SB condition are more lenient towards papers from $\mathcal{J}$ and more harsh towards papers from $\overline{\mathcal{J}}$ than they would be in DB condition. The following problem formalizes this intuition.

Problem 1 (Absolute bias problem)**.** Given significance level $\alpha \in (0,1)$ and decisions of SB and DB reviewers, the goal is to test the following hypotheses:

$$
\begin{aligned}
H_0 &: \forall i \in [m]\ \forall j \in [n] \quad \pi_{ij}^{(\mathrm{sb})} = \pi_{ij}^{(\mathrm{db})} \\
H_1 &: \forall i \in [m]\ \forall j \in [n] \begin{cases} \pi_{ij}^{(\mathrm{sb})} \geq \pi_{ij}^{(\mathrm{db})} & \text{if } j \in \mathcal{J} \\ \pi_{ij}^{(\mathrm{sb})} \leq \pi_{ij}^{(\mathrm{db})} & \text{if } j \in \overline{\mathcal{J}}, \end{cases}
\end{aligned} \tag{4.2}
$$

where at least one inequality in the alternative hypothesis (4.2) is strict.

Note that one can define an alternative that represents a bias against papers from $\mathcal{J}$ simply by exchanging the sets $\mathcal{J}$ and $\overline{\mathcal{J}}$ in (4.2). Our goal is to design a testing procedure that controls for Type-I error and has non-trivial power for any pair of matrices $\Pi^{\mathrm{sb}}, \Pi^{\mathrm{db}}$ that fall under definition of Problem 1.

Non-trivial power. Informally, we say that the test has non-trivial power if for choices of Π^{sb} and Π^{db} for which the presence of bias is "obvious", the test is able to detect the bias with probability that goes to 1 as number of papers in both $\mathcal{J}$ and $\overline{\mathcal{J}}$ grows to infinity. Formally, we say that matrices Π^{sb} and Π^{db} satisfy alternative hypothesis (4.2) with margin δ, if all inequalities in equation (4.2) are satisfied with margin $\delta > 0$, that is, $|\pi_{ij}^{(\mathrm{sb})} - \pi_{ij}^{(\mathrm{db})}| > \delta \ \ \forall\ (i,j) \in [m] \times [n]$. Then we say that the testing procedure has non-trivial power

[2] An equivalent definition of the problem from the perspective of causal inference can be found in Appendix A4.

if for any $\varepsilon > 0$ and for any $\delta > 0$ there exists $n_0 = n_0(\varepsilon, \delta)$ such that if $\min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\} > n_0$, then for any Π^{sb} and Π^{db} that satisfy alternative hypothesis (4.2) with margin δ, the power of testing procedure is at least $1 - \varepsilon$.

For instance, if the logistic model (4.1) is correct for both SB and DB reviewers for some $\beta_0^{(\text{sb})} = \beta_0^{(\text{db})} = \beta_0$, $\beta_1^{(\text{sb})} = \beta_1^{(\text{db})} = \beta_1 > 0$, $\beta_2^{(\text{db})} = 0$ and $|\beta_2^{(\text{sb})}| > 0$, then the requirement of non-trivial power ensures that for any choice of true scores bounded in absolute value by a universal constant and any choice of property satisfaction indicators, the test has power growing to 1 as $\min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\}$ goes to infinity.

4.2 Relative bias problem

In Problem 1 we assumed that SB (or DB) condition itself does not cause any change in reviewers' behaviour. We now consider a generalization of Problem 1 which accommodates an additional confounding factor — a bias in the reviewer simply due to her/his assignment in the SB or the DB group (and independent of the paper or its characteristics). For example, reviewers may not have any bias with respect to the property of interest, but just being placed in the SB condition may induce more harsh opinions than the reviewers in DB. Formally, recall the null hypothesis $\pi_{ij}^{(\text{sb})} = \pi_{ij}^{(\text{db})} \quad \forall (i,j) \in [m] \times [n]$ in Problem 1. Instead, under the null, we now allow $\pi_{ij}^{(\text{sb})} = f_0(\pi_{ij}^{(\text{db})})$, for some non-decreasing function $f_0 : [0,1] \to [0,1]$. Of course, one may not know the function f_0 and the goal of this general problem is to design a test that is guaranteed to control over Type-I error and has non-trivial power uniformly for all functions f_0 that belong to some set of non-decreasing functions $\mathcal{F}$.

Problem 2 (Relative bias problem)**.** Given significance level $\alpha \in (0,1)$, class of functions $\mathcal{F}$ and decisions of SB and DB reviewers, the goal is to test the following hypotheses:

$$\begin{aligned} H_0 &: \forall i \in [m]\ \forall j \in [n] \quad \pi_{ij}^{(\text{sb})} = f_0(\pi_{ij}^{(\text{db})}) \\ H_1 &: \forall i \in [m]\ \forall j \in [n] \quad \begin{cases} \pi_{ij}^{(\text{sb})} \geq f_0(\pi_{ij}^{(\text{db})}) & \text{if } j \in \mathcal{J} \\ \pi_{ij}^{(\text{sb})} \leq f_0(\pi_{ij}^{(\text{db})}) & \text{if } j \notin \mathcal{J} \end{cases}, \end{aligned} \tag{4.3}$$

where f_0 is some unknown function from $\mathcal{F}$ and at least one inequality in the alternative hypothesis (4.3) is strict.

For example, if the logistic model (4.1) is correct for both SB and DB reviewers (with $\beta_2^{(\text{db})} = 0$), but intercepts β_0 in SB and in DB conditions are allowed to be different, then the corresponding matrices Π^{sb} and Π^{db} do not fall under the definition of Problem 1, but can be captured by Problem 2 with specific choice of $\mathcal{F}$ as we will discuss in Section 6.2.

The definition of non-trivial power transfers to the relative bias problem with the exception that all $\pi_{ij}^{(\text{db})}$ are substituted by $f_0(\pi_{ij}^{(\text{db})})$ for $f_0 \in \mathcal{F}$. Our goal is to design a testing procedure that controls for Type-I error and has non-trivial power for any pair of matrices $\Pi^{\text{sb}}, \Pi^{\text{db}}$ that fall under definition of Problem 2 for any function $f_0 \in \mathcal{F}$. Ideally, we would like to achieve this goal for a set of functions $\mathcal{F}$ that contains all non-decreasing functions $f : [0,1] \to [0,1]$.

5 Proposed solution

We now introduce the proposed experimental setup as well as statistical tests we study in this work. We subsequently analyze them in the context of Problems 1 and 2 in Section 6.

5.1 Testing procedures

In order to avoid correlations introduced by reviews given by the same reviewer, our tests use at most one decision per reviewer. As we discuss in Section 5.2, we do so by first matching reviewers into pairs, consisting of one SB and one DB reviewer who review a common paper. For the moment, assume that we are given a

set of tuples $\mathcal{T}$, where each tuple $t \in \mathcal{T}$ consists of a paper $j_t \in [n]$, decision of a SB reviewer for this paper Y_{j_t}, decision of a DB reviewer for this paper X_{j_t} and indicator of property satisfaction w_{j_t}, with a constraint that each reviewer contributes her/his decision to at most one tuple. With this notation, we now present two tests we consider in this work. As we show subsequently, either of these tests would suffice for the absolute bias problem, but for the relative bias problem they cater to different models of reviewers' behaviour with non-intersecting areas of applicability. To provide intuition behind the tests, we define them in context of the absolute bias problem (Problem 1) and discuss their applicability to the relative bias problem later.

Disagreement-based test A high-level idea of the test is as follows. Consider a pair of SB and DB reviewers who disagree in their decisions for some paper. Then under the null hypothesis, the events "SB accepts and DB rejects" and "SB rejects and DB accepts" are equally likely. In contrast, if the null hypothesis is violated, then depending on the property satisfaction and the direction of the bias, SB reviewer is more (or less) likely to vote for acceptance than her/his DB counterpart.

Test 1 DISAGREEMENT

Input: Significance level $\alpha \in (0, 1)$
Set of tuples $\mathcal{T}$, where each $t \in \mathcal{T}$ is of the form $(j_t, Y_{j_t}, X_{j_t}, w_{j_t})$ for some paper $j \in [n]$.

1. Initialize U and V to be empty arrays.
2. For each tuple $t \in \mathcal{T}$, if $Y_{j_t} \neq X_{j_t}$, append Y_{j_t} to $\begin{cases} U & \text{if } w_{j_t} = 1 \\ V & \text{if } w_{j_t} = -1 \end{cases}$.
3. Run a permutation test (Fisher, 1935) at the level α to test if entries of U and V are exchangeable random variables, using the test statistic:
$$\tau = \frac{1}{|U|} \sum_{r \in [|U|]} U_r - \frac{1}{|V|} \sum_{r \in [|V|]} V_r.$$
4. Reject the null if and only if the permutation test rejects the null. (If either of the arrays V and U is empty, the test keeps the null.)

We now formally present the DISAGREEMENT test as Test 1. In Step 1 two empty arrays U and V are initialized. Next, in Step 2 we focus on pairs of SB and DB reviewers disagreeing in their decisions for a paper they both review. For each of the corresponding tuples, we add the decision of SB reviewer to the array U if a paper satisfies the property of interest and to V otherwise. Finally, in Step 3 we define a test statistic τ. According to the aforementioned intuition, under the null hypothesis τ should be close to 0, but under the alternative it should be large in absolute value. Hence, to make a decision we run a permutation test and reject the null in Step 3 if this test suggests that $|\tau|$ is too large for a given significance level α.

Counting-based test The test is built on a simple intuition. Assume for the moment that SB setup induces a bias against papers from $\mathcal{J}$ and a bias in favor of papers from $\overline{\mathcal{J}}$. Then it is likely that papers from $\mathcal{J}$ will receive less number of positive recommendations in SB setup as compared to DB setup. Symmetrically, for papers from $\overline{\mathcal{J}}$ we expect reviewers in SB to be more lenient than their DB counterparts. In contrast, if there is no bias at all, then we expect the aforementioned differences to be small.

We now formally present the COUNTING test as Test 2. In Step 1 two empty arrays U and V are created which in Step 2 are populated with differences between decisions of SB and DB reviewers for papers from $\mathcal{J}$ and $\overline{\mathcal{J}}$ respectively. Importantly, in contrast to the DISAGREEMENT test, in the COUNTING test we do not condition on disagreeing pairs of reviewers. Noticing that mean value of entries of U (respectively V) measures the change of attitude towards papers from $\mathcal{J}$ (respectively $\overline{\mathcal{J}}$) between SB and DB conditions, in Step 3 we compute a test statistic γ which compares these changes. According to the aforementioned intuition, under correct null hypothesis the test statistic should be close to 0. Finally, in Step 4 we make a decision using concentration properties of the test statistic.

Test 2 COUNTING

Input: Significance level $\alpha \in (0,1)$
Set of tuples $\mathcal{T}$, where each $t \in \mathcal{T}$ is of the form $(j_t, Y_{j_t}, X_{j_t}, w_{j_t})$ for some paper $j \in [n]$.

1. Initialize U and V to be empty arrays.
2. For each tuple $t \in \mathcal{T}$, append $(Y_{j_t} - X_{j_t})$ to $\begin{cases} U & \text{if } w_{j_t} = 1 \\ V & \text{if } w_{j_t} = -1 \end{cases}$.
3. If either of the arrays V and U is empty, keep the null and terminate. Otherwise, set the test statistic γ as follows:

$$\gamma = \frac{1}{|U|} \sum_{r \in [|U|]} U_r - \frac{1}{|V|} \sum_{r \in [|V|]} V_r. \tag{4.4}$$

4. Reject the null hypothesis if and only if

$$|\gamma| > \sqrt{2\left(|U|^{-1} + |V|^{-1}\right) \log 2/\alpha}.$$

Effect size In Section 6 we will establish theoretical guarantees on Type-I error control for both DISAGREEMENT and COUNTING tests. In addition to these guarantees, both tests provide a natural measure of the effect size:

- COUNTING. The test statistic γ of the COUNTING test compares the within-subject differences in acceptance rates for papers from $\mathcal{J}$ and $\overline{\mathcal{J}}$. Indeed, the first term in equation (4.4) measures the difference between acceptance rates in SB and DB setups for papers from $\mathcal{J}$. Similarly, the second term measures the same difference for papers from $\overline{\mathcal{J}}$. A positive value of the test statistics then indicates that papers from $\mathcal{J}$ benefit from SB review more than papers from $\overline{\mathcal{J}}$.

- DISAGREEMENT. Slightly informally, the test statistic τ of the DISAGREEMENT test measures the difference in acceptance rates of "borderline" papers from $\mathcal{J}$ and $\overline{\mathcal{J}}$ in the SB setup. Indeed, by conditioning on pairs of disagreeing reviewers in Step 2 of Test 1, the test rules out "clear accept" and "clear reject" papers thus considering only the papers for which reviewers disagree (i.e., borderline papers).

Overall, absolute values of the test statistics τ and γ are reasonable estimates of the effect size and are in a similar vein to Cohen's d and other popular effect size measures (Cohen, 1992).

5.2 Setup of the experiment

We now propose the setup of the experiment to overcome the issues highlighted in Section 3.2 and discuss a construction of the set $\mathcal{T}$ used by the tests introduced above. At a higher level, the proposed setup has two main differences from one considered by Tomkins et al. (2017). First, bidding is performed in blind manner by both SB and DB reviewers (Step 1 below). Second and more importantly, to avoid issues caused by non-random reviewers' assignment, we perform paper assignment and reviewer allocation to conditions jointly in a carefully selected manner (Steps 2-4 below).

Procedure 1 Design of the experiment

Input: Paper load $\lambda \geq 1$
Reviewer load $\mu \geq 1$
Assignment algorithm $\mathcal{A}$

1. Reviewers bid on papers in blind manner
2. Depending on the relationship between number of papers (n) and reviewers (m):
   (a) If $m > 2n$, select $2n$ reviewers uniformly at random and use algorithm $\mathcal{A}$ to assign each paper to 2 reviewers from the selected pool such that each reviewer is assigned to one paper
   (b) If $m < 2n$, select $m/2$ papers such that proportions of papers from $\mathcal{J}$ and $\overline{\mathcal{J}}$ are as close to each other as possible. Use algorithm $\mathcal{A}$ to assign each selected paper to 2 reviewers such that each reviewer is assigned to one paper
   (c) If $m = 2n$, use algorithms $\mathcal{A}$ to assign each paper to 2 reviewers such that each reviewer is assigned to one paper

   Denote the corresponding assignment as A^*
3. For each paper in assignment A^*, allocate one assigned reviewer to DB condition and another assigned reviewer to SB condition uniformly at random. If at this point there are reviewers who are not allocated to conditions, allocate half of them to SB and half to DB uniformly at random
4. Using algorithm $\mathcal{A}$, complement assignment A^* such that each paper is assigned to λ SB and λ DB reviewers and each reviewer reviews at most μ papers. Denote the corresponding assignment as A and begin review process according to this assignment
5. When the review process is finished, construct a set $\mathcal{T}$ as follows. For every paper j from the assignment A^* and corresponding pair (i_1, i_2) of SB and DB reviewer, add tuple $(j, Y_{i_1 j}, X_{i_2 j}, w_j)$ to the set $\mathcal{T}$
6. Run statistical test on the set $\mathcal{T}$

We now formally present the experimental procedure as Procedure 1. It takes as input parameters of paper and reviewer loads together with any assignment algorithm that operates on similarities and/or bids. In Step 1 reviewers bid on the papers in blind manner, that is, using only title and abstract of submissions. Notice that in contrast to the Tomkins et al. setup, bidding happens even before the reviewers are allocated to SB or DB conditions. In Step 2 we find a partial assignment of papers to reviewers which satisfies $(\lambda = 2, \mu = 1)$-load constraints. Depending on the relationship between n and m, we may include only a subset of papers or reviewers in this assignment. For example, in case 2b we do not have enough reviewers to respect the one paper per reviewer constraint and hence we select subset of papers of appropriate size such that it includes approximately equal number of papers from $\mathcal{J}$ and $\overline{\mathcal{J}}$ and find the assignment for selected papers only. The constraint on the number of papers from $\mathcal{J}$ and $\overline{\mathcal{J}}$ is to ensure that the resulting set $\mathcal{T}$ is balanced which is necessary for non-trivial power. The corresponding assignment A^* is a building block for our tests which will use the reviews from this assignment only. Next, in Step 3 reviewers are allocated to conditions in a specific manner which is crucial for our statistical guarantees. In Step 4 we find a full assignment A that is a completion of the partial assignment A^*, meaning that if reviewer i was assigned to paper j in assignment A^*, she/he is also assigned to this paper in A. Finally, in Step 5 we construct a set of tuples $\mathcal{T}$ that is used by the Disagreement and Counting algorithms in Step 6. Importantly, by construction we ensure that each reviewer contributes at most one decision to the set $\mathcal{T}$.

As we show below, the experimental Procedure 1 overcomes the issues with the experimental setup we discussed in Section 3.2 and leads to provable control over Type-I error for our Disagreement and Counting algorithms. We underscore that (i) Disagreement and Counting tests are not tied to particular experimental procedure we introduce and can be applied under the setup of Tomkins et al. (2017) with caveats discussed in Section 3.2. For instance, in simulations (a)-(d) of Section 3 the Disagreement test was applied under the setup of Tomkins et al. More details on this remark are provided in Appendix A2; (ii) as requested by the Disagreement and Counting tests, the set of tuples $\mathcal{T}$ constructed in Step 5 contains at

most one decision of each reviewer. This requirement allows our tests to be agnostic to reviewer calibration which may otherwise undermine Type-I error guarantees as demonstrated in Section 3.1. However, if one treats reviews given by the same reviewer as independent, thereby ignoring issues with reviewer calibration, then in Step 5 of Procedure 1 one can construct a larger set $\mathcal{T}$ by using full assignment A and allowing each reviewer to contribute multiple decisions to the set $\mathcal{T}$.

Finally, in addition to facilitating the experiment, in the interest of fairness in the review (Tomkins et al., 2017) the experimental procedure should ensure that in the eventual assignment each paper is reviewed by equal number of SB and DB reviewers. By construction, Procedure 1 satisfies this requirement: Step 4 ensures that in the final assignment A each paper is assigned to λ SB reviewers and λ DB reviewers.

6 Analysis

We now present the analysis of the COUNTING and DISAGREEMENT tests in context of absolute and relative bias problems.

6.1 Absolute bias problem

We begin our analysis from the absolute bias problem and first formulate the main theorem of this section.

Theorem 1. *For any significance level $\alpha \in (0, 1)$, under the setup of the absolute bias problem (Problem 1), let the experiment be organized according to Procedure 1. Then the* DISAGREEMENT *and* COUNTING *tests are guaranteed to control for Type-I error at the level α, and also satisfy the requirement of non-trivial power.*

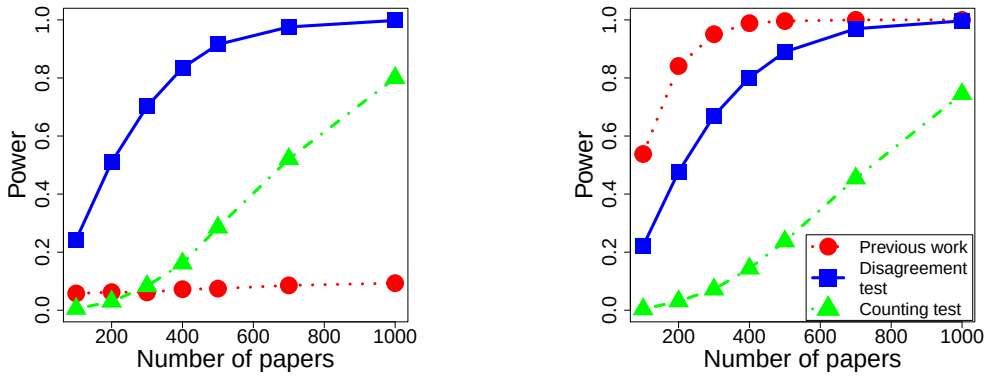
Remark. 1. As demonstrated in Figure 4, In practice, the DISAGREEMENT test has a higher power as compared to the COUNTING test and should be employed under conditions of the absolute bias problem.

2. Notice that the outcomes of the DISAGREEMENT and COUNTING tests depend on a set $\mathcal{T}$ provided to the tests as input. That is, for two different sets $\mathcal{T}_1$ and $\mathcal{T}_2$ (that for example correspond to different assignments A_1^* and A_2^* constructed in Step 2 of Procedure 1), the outcomes of the tests might be different. Hence, one should fix a set $\mathcal{T}$ before observing reviewers' decisions to avoid chasing statistical significance.

3. Finally, the DISAGREEMENT and COUNTING tests are also applicable to the experimental procedure used by Tomkins et al. (2017) and are guaranteed to be robust to issues (a)-(c) from Section 3.1. The formal statement is given in Appendix A2.

We now discuss the issues (a)-(e) considered in Section 3 in the context of our DISAGREEMENT and COUNTING tests.

- **Noise** The DISAGREEMENT and COUNTING tests do not rely on any estimation of papers' qualities made by reviewers. Moreover, we do not even assume that there exists some objective quantity that can be estimated. Hence, our tests do not suffer from issues caused by noisy estimates of scores given by DB reviewers as illustrated by Figure 2a in case of the DISAGREEMENT test.
- **Model mismatch** The only assumption we make is that under correct null hypothesis there is no difference in behavior of SB and DB reviewers. Hence, Theorem 1 guarantees that our tests are robust to violations of specific parametric model (4.1) as illustrated by Figure 2b.
- **Reviewer calibration** We circumvent the detrimental effect of correlations introduced by reviewers' calibration by requiring that each reviewer contributes at most one review to the test. See Figure 2c for an illustration. Of course, such robustness comes at the cost of some power, but we notice that our matching procedures guarantee the use of at least a constant fraction of available data, thereby limiting reduction in the power.
- **Non-blind bidding** The issue with bidding is straightforwardly resolved by requesting blind bidding from both SB and DB reviewers. As illustrated by Figure 3a, we abstract out possible confoundings due to difference in bidding behaviour and ensure that the observed difference in decisions (if any) is due to bias in evaluations and not in bidding.



(a) DB reviewers estimate true scores with some noise in presence of correlations.

(b) DB reviewers estimate true scores without noise and correlations are absent.

Figure 4: Synthetic simulations evaluating performance of the test in Tomkins et al. (2017) ("previous work") and the tests introduced in this section ("Disagreement and Counting tests") in presence of bias when logistic model (4.1) of reviewers is correct. Larger values are better. Details of the simulation setup are provided in Appendix A5.2. Error bars are too small to be visible.

- **Reviewer assignment** The experimental design proposed in Procedure 1 allows to execute any assignment algorithm without breaking the guarantees of the tests as demonstrated in case of TPMS assignment by Figure 3b. The key part of the procedure that ensures such robustness are Steps 2 and 3 where we first find triples of two reviewers and a paper they review and then randomly allocate one reviewer from each triple to SB and one to DB. In this manner, we ensure that parts of the final assignment A that are used for testing do not exhibit any structural difference caused by non-random assignment.

We conclude the section with brief discussion of the power of the tests we introduced in this section. To provide fair comparison of tests, the simulations are performed under the setup of the experiment by Tomkins et al. (2017) and the input to the Disagreement and Counting tests is computed according to procedure described in Appendix A2. Figure 4 contrasts the powers of the tests under the logistic model (4.1) in two cases. In Figure 4a, DB reviewers estimate the true scores of the papers with some noise, which in presence of correlation dramatically decreases the power of the test by Tomkins et al. As discussed above, the Disagreement and Counting tests are robust to issues caused by measurement error, as seen in Figure 4a.

In contrast, Figure 4b considers the case when DB reviewers estimate true scores without noise, that is, true scores are known to the Tomkins et al. test. Although in this case their test has the highest power among 3 tests under consideration, we notice that the margin between their test and the Disagreement test is not as large as in Figure 4a. Notice also that the test by Tomkins et al. gains power by overfitting to the strict model (4.1) which leads to higher power when the model is correct, but at the cost of not being able to control over Type-I error rate under reasonable violations of the modelling assumptions as discussed in Section 3.

Finally, notice that the Counting test relies on the sub-Gaussian approximation to define a threshold and consequently has lower power than the Disagreement test in both cases. While this suggests that for absolute bias problem the Disagreement test dominates the Counting test, we will show in the next section that under the relative bias problem these tests are incomparable.

6.2 Relative bias problem

In Section 6.1 we showed that if under the absence of bias the behaviour of reviewers in SB and DB conditions is the same, then the Disagreement and Counting tests control for Type-I error and have non-trivial

power thus leading to a reliable testing procedure. In this section we relax that assumption and consider a relative bias problem with two specific choices of $\mathcal{F}$ that correspond to popular linear and logistic models. We first show that for each of these choices either the DISAGREEMENT or the COUNTING test leads to reliable testing. We then conclude the analysis with a negative result saying that no test can control for Type-I error and have non-trivial power over both choices of $\mathcal{F}$.

Let us now introduce these two choices of $\mathcal{F}$ that we consider throughout the remaining part of this section. To this end, for each $j \in [n]$ we let $q_j \in \mathbb{R}$ denote an unknown "representation" of the paper j, where by representation we imply any function of a paper's content that defines reviewers' perception of a paper. For example, it could be that $q_j = q_j^*$, where q_j^* is a true score as defined by Tomkins et al.

Generalized linear model Under the generalized linear model, the SB condition itself induces a change in reviewers' evaluations, making them more harsh (or lenient). Moreover, under absence of bias the change in behaviour between SB and DB conditions is described by a constant shift in probability of acceptance for all papers, irrespective of whether they satisfy a property of interest or not. Formally, consider a fixed constant $\Delta \in (0, 0.5)$. The generalized linear model assumes that (i) for every $j \in [n]$ a corresponding representation q_j belongs to the interval $(\Delta, 1-\Delta)$ and (ii) under absence of bias for every $(i, j) \in [m] \times [n]$ the behavior of reviewer i if she/he reviews paper j is described by the following parametric equations:

$$\text{DB: } \pi_{ij}^{(\mathrm{db})} = q_j \tag{4.5a}$$

$$\text{SB: } \pi_{ij}^{(\mathrm{sb})} = \nu + q_j, \tag{4.5b}$$

for some unknown constant $\nu \in (-\Delta, \Delta)$. Provided that matrix Π^{db} was generated according to the model (4.5a) of DB reviewers, the generalized linear model corresponds to an instance of a relative bias problem with a set of functions $\mathcal{F}_\Delta$ associated to a fixed constant Δ and defined as

$$\mathcal{F}_\Delta = \left\{ h_\nu(t) : (\Delta, 1-\Delta) \to [0,1] \;\middle|\; \nu \in (-\Delta, \Delta) \right\}, \tag{4.6}$$

where

$$h_\nu(t) = t + \nu, \quad t \in (\Delta, 1-\Delta). \tag{4.7}$$

Indeed, observe that if for any $(i, j) \in [n] \times [m]$ the probability of acceptance $\pi_{ij}^{(\mathrm{db})}$ was generated according to the model (4.5a), then $\pi_{ij}^{(\mathrm{sb})}$ defined as $\pi_{ij}^{(\mathrm{sb})} = h_\nu(\pi_{ij}^{(\mathrm{db})})$ satisfies model (4.5b).

Generalized logistic model Similar to the generalized linear model, under the generalized logistic model the SB condition also induces a change in reviewers' evaluations, but the change now is described as a constant shift in space of log-odds of the acceptance probabilities. Formally, consider a fixed constant $\widetilde{\Delta} > 0$. The generalized logistic model assumes that (i) for every $j \in [n]$ a corresponding representation q_j belongs to the interval $(-\widetilde{\Delta}, \widetilde{\Delta})$ and (ii) under absence of bias for every $(i, j) \in [m] \times [n]$ behaviour of reviewer i if she/he reviews paper j is described by the following parametric equations:

$$\text{DB: } \log \frac{\pi_{ij}^{(\mathrm{db})}}{1 - \pi_{ij}^{(\mathrm{db})}} = \beta_0 + \beta_1 q_j \tag{4.8a}$$

$$\text{SB: } \log \frac{\pi_{ij}^{(\mathrm{sb})}}{1 - \pi_{ij}^{(\mathrm{sb})}} = \beta_0 + \widetilde{\nu} + \beta_1 q_j, \tag{4.8b}$$

for some unknown constant $\widetilde{\nu} \in (-\widetilde{\Delta}, \widetilde{\Delta})$, where unknown coefficients β_0 and β_1 are also bounded in absolute value by $\widetilde{\Delta}$ and $\beta_1 > 0$. Provided that matrix Π^{db} is generated according to the model (4.8a) of DB reviewers, one can verify that the generalized logistic model corresponds to an instance of a relative bias problem with a set of functions $\widetilde{\mathcal{F}}_{\widetilde{\Delta}}$ associated to a fixed constant $\widetilde{\Delta}$ and defined as

$$\widetilde{\mathcal{F}}_{\widetilde{\Delta}} = \left\{ g_{\widetilde{\nu}}(t) : [0,1] \to [0,1] \;\middle|\; \widetilde{\nu} \in (-\widetilde{\Delta}, \widetilde{\Delta}) \right\}, \tag{4.9}$$

where

$$g_{\widetilde{\nu}}(t) = \frac{te^{\widetilde{\nu}}}{1 - t + te^{\widetilde{\nu}}}, \quad t \in [0, 1]. \tag{4.10}$$

Indeed, observe that if for some $(i, j) \in [n] \times [m]$ the probability that reviewer i accepts paper j in DB setup $\pi_{ij}^{(\mathrm{db})}$ is generated according to the model (4.8a), then setting $\pi_{ij}^{(\mathrm{sb})} = g_{\widetilde{\nu}}(\pi_{ij}^{(\mathrm{db})})$ we ensure that $\pi_{ij}^{(\mathrm{sb})}$ satisfies model (4.8b).

The models we defined follow an objective parametric approach assumed by Tomkins et al. (2017) with two differences: (i) we do not assume that q_j has a known meaning or that it can be measured (for instance, it may be that $q_j = q_j^*$, or that $q_j = (q_j^*)^3$, or q_j may be a complex function of the content of the paper) and (ii) we do not assume that the bias is described by a linear shift in space of probabilities or log-odds, and instead consider a non-parametric definition of the bias as specified in the alternative hypothesis (4.3).

6.2.1 Positive results

We now show that the Counting and Disagreement tests lead to reliable testing under the generalized linear and generalized logistic models respectively. Before we formulate the main result of this section, let us provide some intuition behind the models and the corresponding tests.

Generalized linear model A natural strategy to test for biases under the generalized linear model is to estimate the shift in reviewers' behaviour on papers that belong to $\mathcal{J}$ and to $\overline{\mathcal{J}}$ separately and then compare the estimates. In fact, the Counting testing procedure introduced above as Test 2 follows this strategy and, as guaranteed by Theorem 2, leads to reliable testing under the generalized linear model.

Generalized logistic model Intuitively, estimating a constant shift in reviewers' behavior as done by the Counting test may not be the optimal strategy under the generalized logistic model, as under absence of bias the change of behaviour between SB and DB conditions is given by a constant shift in *log-odds* space and not in *probability* space. However, it turns out that the Disagreement test is able to capture the constant shift in *log-odds* space and hence we still can perform reliable testing under this model, using the Disagreement algorithm.

Theorem 2. *For any significance level $\alpha \in (0, 1)$, let the experiment be organized according to Procedure 1. Then*

(a) Under the generalized linear model with any $\Delta \in (0, 0.5)$, the Counting *test is guaranteed to control for Type-I error at the level α, and also satisfies the requirement of non-trivial power.*

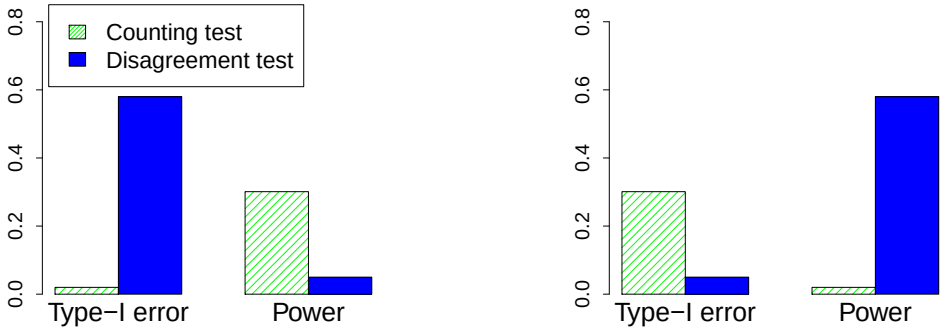
(b) Under the generalized logistic model with any $\widetilde{\Delta} > 0$, the Disagreement *test is guaranteed to control for Type-I error at the level α, and also satisfies the requirement of non-trivial power.*

Remark. 1. Result of Theorem 2(a) holds even for a subjective version of the generalized linear model in which for each $(i, j) \in [n] \times [m]$ we substitute q_j with q_{ij} (that is, different values across reviewers) in equations (4.5a) and (4.5b), thereby accounting for subjectivity of reviewers.

2. If the logistic model (4.1) assumed by Tomkins et al. is correct for both SB and DB reviewers with possibly different intercepts, then Theorem 2(b) ensures that the Disagreement test provably controls for the Type-I error and can detect a bias with probability that goes to 1 as sample size grows, without requiring knowledge (neither exact nor approximate) of papers' scores $q_1^*, \ldots, q_n^*$.

3. Notice that the Counting and Disagreement tests do not require the knowledge of Δ and $\widetilde{\Delta}$ parameters to control for Type-I error and satisfy the requirement of non-trivial power.

Figure 5 compares the performance of the Disagreement and Counting tests under specific instances of the generalized linear and logistic models, illustrating the results of Theorem 2. Figure 5a shows that the Counting test controls for Type-I error and has a non-trivial power under the generalized linear model. Notice that the Disagreement test does not control for Type-I error in this instance, implying that Theorem 2(b) cannot be extended to guarantee the Type-I error control under the generalized linear model



(a) Instance of generalized linear model.

(b) Instance of generalized logistic model.

Figure 5: Synthetic simulations evaluating performance of the DISAGREEMENT and COUNTING tests under setup of the relative bias problem when generalized linear (Figure 5a) or generalized logistic (Figure 5b) models is correct. For each model presence and absence of bias are simulated. Details of the simulation setup are provided in Appendix A5.3. Error bars are too small to be visible.

as well. Symmetrically, Figure 5b shows that while the DISAGREEMENT test leads to reliable testing under the generalized logistic model, the COUNTING test is unable to control for Type-I error under this model. Hence, we conclude that under the setup of the relative bias problem, none of the tests dominates another, each leading to reliable testing under the corresponding model.

We see in Figure 5 that neither the DISAGREEMENT nor the COUNTING test is suitable for *both* generalized linear and logistic models. In the next section we show that this is not a drawback of these specific tests, but rather a manifestation of a more general impossibility result.

6.2.2 Negative result

We conclude our analysis with a negative result that limits the complexity of the class $\mathcal{F}$ for which reliable testing in the relative bias problem is possible. Let us first state the main result of this section.

Theorem 3. *Suppose that there exist two functions $g, h \in \mathcal{F}$ and some $0 \leq x_1 < x_2 \leq 1$ such that $g(x_1) < h(x_1)$ and $g(x_2) > h(x_2)$. Suppose also that there exists a testing procedure ψ operating on decisions of SB and DB reviewers that for any given $\alpha \in (0, 1)$ keeps Type-I error below α for all matrices Π^{sb}, Π^{db} that satisfy the null hypothesis of Problem 2 specified by some function $f_0 \in \mathcal{F}$. Then the testing procedure ψ cannot satisfy the requirement of non-trivial power.*

The intuition behind Theorem 3 is as follows. If the class $\mathcal{F}$ contains "too many" functions, then some matrices Π^{sb} and Π^{db} satisfy the null hypothesis of Problem 2 defined by $f_0 \in \mathcal{F}$ and simultaneously satisfy the alternative hypothesis (4.3) defined by some other function $f_0' \in \mathcal{F}$ with margin $\delta > 0$. Hence, any testing procedure must either have high Type-I error rate or sacrifice the non-trivial power requirement over the class of functions $\mathcal{F}$, implying that reliable testing is impossible.

Using Theorem 3 we can deduce that one can hope to control for the Type-I error rate and simultaneously have a non-trivial power only when functions contained in $\mathcal{F}$ are pointwise totally ordered, that is, for any two functions $f, g \in \mathcal{F}$ it must be the case that either $f(x) \geq g(x)$ for all $x \in [0, 1]$ or $f(x) \leq g(x)$ for all $x \in [0, 1]$.

Let us now illustrate the consequences of Theorem 3 for the generalized linear and logistic models.

Corollary 1. *For any significance level $\alpha \in (0, 1)$, let ψ_1 and ψ_2 be any testing procedures which operate on decisions of SB and DB reviewers. Suppose that under the generalized linear model with any $\Delta \in (0, 0.5)$,*

procedure ψ_1 controls for the Type-I error at the level α and satisfies the non-trivial power requirement. Suppose also that under the generalized logistic model with any $\widetilde{\Delta} > 0$, procedure ψ_2 controls for the Type-I error at the level α and satisfies the non-trivial power requirement. Then

(a) Under the generalized logistic model with any $\widetilde{\Delta} > 0$, procedure ψ_1 incurs a Type-I error rate strictly greater than α.

(b) Under the generalized linear model with any $\Delta \in (0, 0.5)$, procedure ψ_2 incurs a Type-I error rate strictly greater than α.

Corollary 1 shows that there does not exist a testing procedure that controls over Type-I error and has non-trivial power under both generalized linear and generalized logistic models. As a result, one needs to design different procedures for these models as we did with the Disagreement and Counting tests. In case of the Disagreement and Counting tests, Corollary 1 is illustrated by Figure 5.

In Appendix A3 we discuss another application of Theorem 3 in context of the generalized logistic model that also suggests that generality of the Disagreement test cannot be further increased.

7 Discussion

Peer review is the backbone of academia but faces a number of challenges of unfairness, biases, and inefficiency. This work contributes to the growing literature (Shah et al., 2018; Kang et al., 2018; Gao et al., 2019; Wang and Shah, 2019; Stelmakh et al., 2021a; Kobren et al., 2019; Noothigattu et al., 2020; Balietti et al., 2016; Xu et al., 2019; Fiez et al., 2020) in the domain of addressing these challenges in peer review, by designing a principled method to test for biases. We show that under various conditions the approach used by the prior work of Tomkins et al. does not control the Type-I error rate. We underscore that we do not aim at confirming or disproving the presence of biases found in that work, but our focus is on the validity of testing methods. With this goal in mind, we propose a principled approach to testing for biases and design two statistical procedures that coupled with our novel experimental setup provably control for the Type-I error rate. Additionally, these procedures have non-trivial power under essentially a single assumption of no difference in the behavior of SB and DB reviewers when the bias is absent. We then show that this assumption cannot be relaxed in general and that to accommodate the aforementioned difference in behavior one needs to make some modelling assumptions, as we demonstrated with our tests and generalizations of popular linear and logistic models.

We presented the Disagreement and Counting tests in the context of peer review. However, we underscore that one can adapt our experimental setup (Procedure 1) to use our testing procedures (Test 1 and Test 2) in other applications. These applications include peer grading, university admission, and hiring where some protected attributes might be available to reviewers.

There are several open problems suggested by our work. The first direction is associated with the statistical power of the testing procedures we propose. In this work, we show that our tests have power that going to one under certain conditions on the alternative. It is of interest to establish a bound on the statistical power of our tests in a finite sample setting and compare it with an upper bound on the maximum power that can be achieved by any computationally-efficient testing procedure.

The second direction is related to the design of the experimental procedure. To accommodate tests for biases, one needs to deviate from the standard peer-review pipeline, thus introducing a trade-off between the quality of the peer-review process and the accuracy of the testing. Quantification of such a trade-off may help to design a better setup and understand the cost of the experiment in terms of the peer-review quality. In this work, we designed a procedure that leads to the desired accuracy, but is suboptimal in terms of the TPMS objective. In contrast, the optimal TPMS assignment would not allow to perform reliable testing. Hence, an open problem is to design an experimental procedure that accommodates our statistical tests and subject to this maximizes the quality of the assignment in terms of the TPMS objective.

Appendix

We provide supplementary materials and additional discussion.

A1 More than one property of interest

Throughout the main body of the paper, we considered the case of a single property of interest. We now generalize some of the results to the case of more than one property of interest. First of all, we recall some notation. Let k be the total number of properties, then for each paper $j \in [n]$, we let variables $w_j^{(1)}, w_j^{(2)}, \dots, w_j^{(k)}$ indicate whether or not the paper satisfies the corresponding property. For each property ℓ, the set $\mathcal{J}_\ell \subseteq [n]$ contains papers that satisfy a property ℓ with $\overline{\mathcal{J}}_\ell$ being its complement.

We argue that when $k > 1$, one needs to think about the bias testing problem as of an instance of the relative bias problem as defined in Section 4.2. Indeed, consider for example the case of two properties of interest ($k = 2$) and assume that we are interested in testing for biases with respect to the first property. Then even if there is no bias with respect to this property, the behavior of reviewers between SB and DB conditions might be different due to possible biases with respect to the second property.

The negative result of Theorem 3 we established in Section 6.2.2 also applies to the case of multiple properties, implying that reliable testing is possible only under some restrictions on the difference in reviewers' behavior between SB and DB conditions under the absence of bias. Following the relative bias problem defined as Problem 2, we now generalize it for $k > 1$. To this end, we consider a problem of testing for biases with respect to the property $\ell \in [k]$ and introduce an additional piece of notation. For each paper $j \in [n]$, let $\mathbf{w}_j$ denote a vector of indicators of property satisfaction: $\mathbf{w}_j = (w_j^{(1)}, w_j^{(2)}, \dots, w_j^{(k)})$ and let $\mathbf{w}_j^{(-\ell)}$ denote the same vector but with ℓ^{th} component omitted, that is, $\mathbf{w}_j^{(-\ell)} = (w_j^{(1)}, \dots, w_j^{(\ell-1)}, w_j^{(\ell+1)}, \dots, w_j^{(k)})$.

Following the definition of the relative bias problem (Problem 2), the set $\mathcal{F}_\ell$ contains functions that under the absence of bias with respect to the property ℓ, specify the difference in behavior between DB and SB conditions. In case of a single property of interest, $\mathcal{F}$ was a subset of all non-decreasing functions $f : [0,1] \to [0,1]$. However, when $k > 1$, even under the absence of the bias with respect to the property ℓ, the change in reviewers' behavior between DB and SB conditions may be influenced by whether the paper satisfies properties other than ℓ, due to possible biases with respect to these properties. Hence, under the absence of bias with respect to the property ℓ, the change of behavior between SB and DB conditions is described as follows:

$$\forall (i,j) \in [n] \times [m] : \ \pi_{ij}^{(\text{sb})} = f_0(\pi_{ij}^{(\text{db})}, \mathbf{w}_j^{(-\ell)}),$$

where function $f_0 \in \mathcal{F}_\ell$ is non-decreasing in its first argument. Thus, the set $\mathcal{F}_\ell$ is a subset of all functions with domain $[0,1] \times \{0,1\}^{k-1}$ which are non-decreasing in their first argument, that is,

$$\mathcal{F}_\ell \subseteq \left\{ f : [0,1] \times \{0,1\}^{k-1} \to [0,1] \,\middle|\, f \text{ is non-decreasing in its first argument} \right\}.$$

Having defined the necessary notation, we are ready to introduce the relative bias problem in case of multiple properties of interest.

Problem 3 (Relative bias problem for multiple properties.)**.** Given significance level $\alpha \in (0,1)$, the property of interest ℓ, the class of functions $\mathcal{F}_\ell$ and decisions of SB and DB reviewers, the goal is to test the following hypotheses:

$$\begin{aligned} H_0 &: \forall i \in [m] \ \forall j \in [n] \ \ \pi_{ij}^{(\text{sb})} = f_0(\pi_{ij}^{(\text{db})}, \mathbf{w}_j^{(-\ell)}) \\ H_1 &: \forall i \in [m] \ \forall j \in [n] \ \begin{cases} \pi_{ij}^{(\text{sb})} \geq f_0(\pi_{ij}^{(\text{db})}, \mathbf{w}_j^{(-\ell)}) & \text{if } j \in \mathcal{J}_\ell \\ \pi_{ij}^{(\text{sb})} \leq f_0(\pi_{ij}^{(\text{db})}, \mathbf{w}_j^{(-\ell)}) & \text{if } j \notin \mathcal{J}_\ell \end{cases}, \end{aligned} \tag{4.11}$$

for some unknown $f_0 \in \mathcal{F}_\ell$, and where at least one inequality in the alternative hypothesis (4.11) is strict.

Intuitively, under the null hypothesis of absence of bias with respect to the property ℓ, the change of reviewers' behaviour between SB and DB conditions is determined by (i) bias introduced by SB condition itself which is independent of papers' authorship information and (ii) bias with respect to properties other than the property ℓ. The generalized linear and logistic models can be formulated in case of multiple properties of interest as follows.

Generalized linear model. Given a fixed constant $\Delta \in (0, 0.5)$, we follow the case of a single property and assume that each paper $j \in [n]$ has some unknown representation $q_j \in (\Delta, 1-\Delta)$. The generalized linear model assumes that for each $(i,j) \in [m] \times [n]$, the behaviour of reviewer i if she/he reviews paper j is described by the following parametric equations:

$$\text{DB: } \pi_{ij}^{(\text{db})} = q_j \tag{4.12a}$$

$$\text{SB: } \pi_{ij}^{(\text{sb})} = q_j + \beta_0^{(\text{sb})} + \sum_{\ell \in [k]} \beta_\ell^{(\text{sb})} w_j^{(\ell)}, \tag{4.12b}$$

where unknown coefficients are such that $|\beta_0^{(\text{sb})}| + \sum_{\ell=1}^{k} |\beta_\ell^{(\text{sb})}| < \Delta$. Under the generalized linear model, a bias with respect to the property ℓ is present whenever $\beta_\ell^{(\text{sb})} \neq 0$.

Generalized logistic model. Given a fixed constant $\widetilde{\Delta} > 0$, the generalized logistic model assumes that (i) for every $j \in [n]$, a corresponding representation q_j belongs to the interval $(-\widetilde{\Delta}, \widetilde{\Delta})$ and (ii) for each $(i,j) \in [m] \times [n]$, the behaviour of reviewer i if she/he reviews paper j is described by the following parametric equations:

$$\text{DB: } \log \frac{\pi_{ij}^{(\text{db})}}{1 - \pi_{ij}^{(\text{db})}} = \beta_0^{(\text{db})} + \beta_1 q_j \tag{4.13a}$$

$$\text{SB: } \log \frac{\pi_{ij}^{(\text{sb})}}{1 - \pi_{ij}^{(\text{sb})}} = \beta_0^{(\text{sb})} + \beta_1 q_j + \sum_{\ell \in [k]} \beta_{\ell+1}^{(\text{sb})} w_j^{(\ell)}, \tag{4.13b}$$

where all coefficients are bounded in absolute value by $\widetilde{\Delta}$ and $\beta_1 > 0$. Under the generalized logistic model, a bias with respect to the property ℓ is present whenever $\beta_{\ell+1}^{(\text{sb})} \neq 0$.

Remark. 1. First, provided that matrix Π^{db} is generated according to the one of the introduced models, one can define a set of functions $\mathcal{F}_\ell$ that puts the corresponding model in the context of the relative bias problem for multiple properties of interest defined as Problem 3.

2. The goal under each of the models introduced above is to test the significance of the coefficient in equation describing the behavior of SB reviewer that corresponds to the indicator of the property of interest. For example, if we are interested in testing for biases with respect to the property ℓ under the generalized logistic model, then we want to test the significance of the coefficient $\beta_{\ell+1}^{(\text{sb})}$ in equation (4.13b).

3. Notice that in case of multiple properties, the models we introduced above describe reviewers' behaviour both under the absence of bias and under the presence of bias. In this way they allow simultaneous testing for biases with respect to many properties of interest.

Observe that the relationships that describe the behaviour of SB reviewers in models (4.12b) and (4.13b) are reminiscent of the linear regression and logistic regression models respectively. As mentioned above, one cannot fit decisions of SB reviewers to these models using existing methods, because one of the covariates (paper representation q) is unknown. In their work, Tomkins et al. employed DB reviewers to estimate this unknown covariate and used these estimates to fit the logistic model. As we discussed in Section 3, this approach leads to an unreliable testing procedure under various realistic conditions.

We now show that using ideas of the Disagreement and Counting tests, one can use decisions of both SB and DB reviewers to eliminate the unknown covariate from the model, thereby enabling standard tools without the need to estimate any covariate.

Proposition 1. *Let reviewers i and i' be assigned to paper j in SB and DB setups correspondingly, then*

(a) Under the generalized linear model the expectation of the quantity $Y_{ij} - X_{i'j}$ follows the linear model (4.12b) *with $q_j = 0$:*

$$\mathbb{E}\left[Y_{ij} - X_{i'j}\right] = \beta_0^{(sb)} + \sum_{\ell \in [k]} \beta_\ell^{(sb)} w_j^{(\ell)}. \tag{4.14a}$$

(b) Under the generalized logistic model the expectation of the quantity $Y_{ij} | \left(Y_{ij} \neq X_{i'j}\right)$ follows the logistic model (4.13b) *with $q_j = 0$:*

$$\log \frac{\mathbb{E}\left[Y_{ij} | \left(Y_{ij} \neq X_{i'j}\right)\right]}{1 - \mathbb{E}\left[Y_{ij} | \left(Y_{ij} \neq X_{i'j}\right)\right]} = \beta_0 + \sum_{\ell \in [k]} \beta_{\ell+1}^{(sb)} w_j^{(\ell)}, \tag{4.14b}$$

where $\beta_0 = \beta_0^{(sb)} - \beta_0^{(db)}$.

Proposition 1 provides a mean to eliminate the unknown covariate from the models of SB decisions (4.12b) and (4.13b) using decisions of DB reviewers. For example, in case of the generalized logistic model it relies on the core idea of the Disagreement test and suggests conditioning on pairs (SB reviewer, DB reviewer) such that reviewers disagree in their decisions for some paper. After conditioning, decisions of SB reviewers follow model (4.14b) with all covariates known and hence standard test for logistic regression can be applied to evaluate significance of coefficients.

Proposition 1 also allows to avoid using noisy measurements and hence any test for significance of the coefficients applied to the models (4.14a) and (4.14b) will not be susceptible to issues caused by the use of noisy measurements and misspecification of the meaning of q (issues (a) and (b) from Section 3). If one restricts each reviewer to input at most one decision to the testing procedure, then issue (c) will also be mitigated.

A2 Our tests under the setup of Tomkins et al.

In this section we give additional comments on the applicability of our testing procedures to the setup of Tomkins et al. (2017). To this end, recall that our tests take as input the set of tuples $\mathcal{T}$ such that (i) each tuple $t \in \mathcal{T}$ is of the form $t = (j_t, Y_{j_t}, X_{j_t}, w_{j_t})$, where j_t is a corresponding paper, Y_{j_t}, X_{j_t} are decisions of SB and DB reviewers for this paper and w_{j_t} equals 1 if $j_t \in \mathcal{J}$ and -1 otherwise; (ii) each reviewer contributes at most one decision to the set $\mathcal{T}$. Potentially our tests can be coupled with any experimental procedure as long as this procedure enables a construction of such a set $\mathcal{T}$. However, one needs to understand that while Procedure 1 is robust to issues we discussed in Section 3.2, other experimental setups may lead to an inflation of the Type-I error.

In the experiment conducted by Tomkins et al. (2017), reviewers were split into two groups (SB and DB) uniformly at random at the very beginning of the experiment. Then two assignments A_{SB} and A_{DB} were computed separately for each group of reviewers. As discussed in Section 3.2, even if both groups of reviewers bid in a blind manner, the design of Tomkins et al. may lead to inflated Type-I error. Nonetheless, we now show that even under the setup of Tomkins et al., one can employ the Disagreement and Counting tests to fix issues (a)-(c) with the testing procedure discussed in Section 3.1.

A2.1 Matching algorithms

Let us first introduce two matching procedures that construct an input for our tests under the setup of Tomkins et al. (2017). Given assignment of papers to reviewers in both SB and DB conditions, we discuss two choices of matching algorithms depending on the relationship between parameters λ (required number of reviewers per paper in each condition) and μ (maximum number of papers per reviewer). Notice that our goal is not to

maximize the size of $\mathcal{T}$, but instead to maximize the minimum of the number of papers from $\mathcal{J}$ included in the set $\mathcal{T}$ and the number of papers from $\overline{\mathcal{J}}$ included in the set $\mathcal{T}$. This is because our statistical tests need decisions for papers from both $\mathcal{J}$ and $\overline{\mathcal{J}}$ to maximize their power. Depending on the relationship between λ and μ we can solve this problem either exactly or approximately.

Case 1 ($\lambda \geq \mu$). In this case, each paper can be matched to 1 SB reviewer and 1 DB reviewer by finding two separate maximum matchings (papers to SB reviewers and papers to DB reviewers) using the Hungarian matching algorithm. We formally present the matching procedure as Algorithm 1.

Algorithm 1 Exact matching algorithm

Input: Assignments A_{SB}, A_{DB} of SB and DB reviewers to papers, respectively.

1. Construct a graph G that consists of 3 layers:

   - **Layer 1.** One node for each SB reviewer
   - **Layer 2.** One node for each paper
   - **Layer 3.** One node for each DB reviewer

   and add edges between reviewers and papers according to assignments A_{SB} and A_{DB}. Set $\mathcal{T} = \emptyset$.

2. Using the Hungarian matching algorithm with uniform tie-breaking find matchings $\mathcal{M}_{\text{SB}}$ and $\mathcal{M}_{\text{DB}}$ where $\mathcal{M}_{\text{SB}}$ (respectively $\mathcal{M}_{\text{DB}}$) is a maximum 1-1 matching between SB (respectively DB) reviewers and papers (each reviewer is matched to at most 1 paper and each paper is matched to at most 1 reviewer).

3. Leave in graph G only those edges that correspond to matched pairs in $\mathcal{M}_{\text{SB}}$ and $\mathcal{M}_{\text{DB}}$.

4. For any triple of (SB reviewer i_1, paper j, DB reviewer i_2) such that there is a path from a node that corresponds to reviewer i_1 to a node that corresponds to reviewer i_2 through a node that corresponds to paper j, add $t = (j, Y_{i_1 j}, X_{i_2 j}, w_j)$ to $\mathcal{T}$.

5. Return $\mathcal{T}$.

Lemma 1. *For any assignments of SB and DB referees to papers that satisfy (λ, μ)-load constraints with $\lambda \geq \mu$, the matching procedure in Algorithm 1 is guaranteed to construct a set of tuples $\mathcal{T}$ such that for each paper $j \in [n]$ there is one tuple that corresponds to this paper.*

Case 2 ($\lambda < \mu$). In this case we cannot use the above idea, because there does not exist a matching such that each paper is matched to one SB and one DB reviewer, subject to a constraint that each reviewer is matched with at most one paper. While solving the exact optimization problem in this case might be hard, a simple greedy procedure constructs a sufficiently large matching for the DISAGREEMENT and COUNTING tests to satisfy the non-trivial power requirement. The iterative greedy procedure in each iteration matches one paper from $\overline{\mathcal{J}}$ and one paper from $\mathcal{J}$ to 1 SB and 1 DB reviewer and removes those reviewers from subsequent interations to maintain the constraint that each reviewer contributes at most one decision to the set $\mathcal{T}$. We formally introduce the greedy procedure as Algorithm 2.

Lemma 2. *For any assignments of SB and DB referees to papers that satisfy (λ, μ)-load constraints, the matching procedure in Algorithm 2 is guaranteed to construct a set of tuples $\mathcal{T}$ that for large enough $\min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\}$ contains at least $c \min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\}$ tuples corresponding to papers from $\mathcal{J}$ and at least $c \min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\}$ tuples corresponding to papers from $\overline{\mathcal{J}}$, where c is a constant that may depend only on λ and μ.*

Remark. 1. If the set $\mathcal{T}$ constructed by the Algorithm 1 is such that there exist reviewers who do not contribute any of their decisions to this set, then one can run Algorithm 2 on assignments of these reviewers to papers and obtain the set $\mathcal{T}'$. Next, consider the updated set $\mathcal{T}^* = \mathcal{T} \cup \mathcal{T}'$ and observe that each reviewer contributes at most one decision to this set.

2. By construction both matching algorithms introduced in this section include at most one decision per reviewer in a set of tuples $\mathcal{T}$.

Algorithm 2 Greedy matching algorithm

Input: Assignments A_{SB}, A_{DB} of SB and DB reviewers to papers, respectively.

1. Construct a graph G that cosists of 3 layers:

   - **Layer 1.** One node for each SB reviewer
   - **Layer 2.** One node for each paper
   - **Layer 3.** One node for each DB reviewer

   and add edges between reviewers and papers according to assignments A_{SB} and A_{DB}. Set $\mathcal{T} = \emptyset$.

2. Find a triple (SB reviewer i_1, paper $j \in \mathcal{J}$, DB reviewer i_2) such that there is a path in graph G from a node corresponding to SB reviewer to a node corresponding to DB reviewer through a node corresponding to a paper. If there are many such triples, break ties uniformly at random. If such a triple exists, define $t_1 = (j, Y_{i_1 j}, X_{i_2 j}, w_j)$, otherwise set $t_1 = \emptyset$.

3. Find a triple (SB reviewer $i_1' \neq i_1$, paper $j' \in \overline{\mathcal{J}}$, DB reviewer $i_2' \neq i_2$) such that there is a path in graph G from a node corresponding to the SB reviewer to a node corresponding to the DB reviewer through a node corresponding to the paper. If there are many such triples, break ties uniformly at random. If such a triple exists, define $t_2 = (j', Y_{i_1' j'}, X_{i_2' j'}, w_{j'})$, otherwise set $t_2 = \emptyset$.

4. Update $\mathcal{T} = \mathcal{T} \cup \{t_1, t_2\}$. If both t_1 and t_2 are empty, return $\mathcal{T}$. Otherwise delete reviewers i_1, i_1', i_2, i_2' from the graph G together with the corresponding edges and go to Step 2.

Overall, let $\mathcal{A}$ denote a procedure that takes assignments A_{SB} and A_{DB} as input and depending on the relationship between λ and μ calls Algorithm 1 or Algorithm 2 to construct the set $\mathcal{T}$.

A2.2 Guarantees

Having defined a procedure to construct input for the Disagreement and Counting tests, we are now ready to formulate corresponding theoretical guarantees. Recalling that the experimental setup of Tomkins et al. itself breaks the Type-I error guarantees, we abstract out these issues by assuming that instead of TPMS assignment algorithm or any other algorithm that computes assignments of papers to SB and DB reviewers, the assignment is selected uniformly at random from the set of all assignments satisfying $(\lambda, \mu)-$constraints. For brevity, we only show the result for the absolute bias problem, but analogue of Theorem 2 also holds.

Proposition 2. *For any given significance level $\alpha \in (0, 1)$, under the setup of the absolute bias problem (Problem 1), let experiment be organized according to the procedure of Tomkins et al. with random assignment. Then the* Disagreement *and* Counting *tests coupled with procedure $\mathcal{A}$ (Algorithm 1 and Algorithm 2) are guaranteed to control for the Type-I error rate at the level α and also satisfy the requirement of non-trivial power.*

Remark. 1. From theoretical standpoint, the requirement of random assignment can be substituted with the following conditions, under which any assignment algorithm can be used: (i) reviewers in both conditions bid blindly and (ii) reviewers' evaluations are independent of similarities. That is, rows of matrices Π^{sb} and Π^{db} are assigned to reviewers uniformly at random after the assignment is computed.

2. Proposition 2 ensures that the Counting and Disagreement tests, coupled with matching algorithms we introduced above, are robust to issues (a)-(c) discussed in Section 3.1. However, in practice even our robust tests may still be susceptible to issues caused by the experimental setup of Tomkins et al.

A3 Additional impossibility result for the generalized logistic model

In this section we formulate an additional impossibility result that highlights the generality of the Disagreement test.

As we demonstrated in Theorem 2, the DISAGREEMENT test leads to reliable testing under the generalized logistic model. Recalling the definition of papers representations $q_j, j \in [n]$, let us now consider an extended version of the generalized logistic model which is given by the following parametric equations describing the behaviour of DB and SB reviewers under the absence of bias

$$\text{DB: } \log \frac{\pi_{ij}^{(\text{db})}}{1-\pi_{ij}^{(\text{db})}} = \beta_0^{(\text{db})} + \beta_1^{(\text{db})} q_j \tag{4.15a}$$

$$\text{SB: } \log \frac{\pi_{ij}^{(\text{sb})}}{1-\pi_{ij}^{(\text{sb})}} = \beta_0^{(\text{sb})} + \beta_1^{(\text{sb})} q_j, \tag{4.15b}$$

where parameters $\beta_0^{(\text{sb})}, \beta_1^{(\text{sb})} > 0, \beta_0^{(\text{db})}, \beta_1^{(\text{db})} > 0$ as well as papers' scores $q_j, j \in [n]$, are bounded in absolute value but otherwise are allowed to be arbitrary. This extended version specializes to the standard generalized logistic model if $\beta_1^{(\text{sb})} = \beta_1^{(\text{db})}$. In words, under the extended version, the reviewers in SB and DB conditions under the absence of bias may have not only different intercepts, but also different coefficients in front of q. The presence of bias is then defined as a violation of model (4.15b) where the direction of violation is different for papers from $\mathcal{J}$ and $\overline{\mathcal{J}}$.

Unfortunately, as we show in Corollary 2, reliable testing under this extension of the generalized logistic model is not possible.

Corollary 2. *For any significance level $\alpha \in (0,1)$, consider the extension of the generalized logistic model given by equations* (4.15a) *and* (4.15b)*. Then no test operating on decisions of SB and DB reviewers can control for the Type-I error rate at the level α and simultaneously satisfy the non-trivial power requirement.*

The negative result of Corollary 2 applies to the DISAGREEMENT test, implying that even our testing procedure which is robust to various issues discussed in Section 3 cannot handle the extension of the generalized logistic model specified by equations (4.15a) and (4.15b).

A4 Causal inference viewpoint

When testing for biases in peer review, we aim at discovering a causal relationship between paper's authorship information and reviewers' perception of the paper. In this section, we describe a causal model under which we approach the problem, and provide an equivalent formulation of the problem from the causal inference viewpoint.

Recall that a decision of reviewer i for paper j if this reviewer is assigned to this paper in SB setup is denoted as Y_{ij} and is a Bernoulli random variable with expectation $\pi_{ij}^{(\text{sb})}$. Our ultimate goal is to evaluate whether the indicator variable $w_j \in \{-1, 1\}$ which encodes the property satisfaction has *causal impact* on the decisions of SB reviewers. To this end, we assume that for each reviewer $i \in [m]$ and for each paper $j \in [n]$, probability $\pi_{ij}^{(\text{sb})}$ can be expressed as:

$$\pi_{ij}^{(\text{sb})} = \xi(r_i, q_j, w_j), \tag{4.16}$$

for some unknown function ξ with co-domain $[0,1]$, where q_j is an anonymized content of a paper and r_i is an arbitrary complex representation of a reviewer. That is, we assume that decisions of SB reviewers are determined by the paper content, reviewer identity and, possibly, authorship information.

In this notation, we can state a canonical formulation of the bias testing problem:

Problem $1'$ (Canonical formulation of the bias testing problem)**.** Given significance level $\alpha \in (0,1)$, and decisions of SB reviewers that are distributed according to equation (4.16), the goal is to test the following hypotheses:

$$\begin{aligned} H_0 &: \forall i \in [m]\ \forall j \in [n]\ \ \xi(r_i, q_j, 1) = \xi(r_i, q_j, -1) \\ H_1 &: \forall i \in [m]\ \forall j \in [n]\ \ \xi(r_i, q_j, 1) \geq \xi(r_i, q_j, -1) \end{aligned} \tag{4.17}$$

where at least for one pair $(i,j) \in [m] \times [n]$ the inequality in the alternative hypothesis (4.17) is strict.

Unfortunately, the formulation of Problem 1′ is too challenging and cannot be tested without further assumptions in the peer-review setup for the following reasons:

1. Fully randomized controlled experiments cannot be performed in peer-review settings, because we cannot randomize indicators $w_j, j \in [n]$, that is, we cannot randomize authors of the papers
2. To facilitate an observational study without further assumptions on function ξ, we need to have many papers with the same content but with different authors which is also impossible

Tomkins et al. (2017) attempted to circumvent the aforementioned challenges by making the following assumptions: (i) for each paper $j \in [n]$, representation q_j is simply a true score of a paper $q_j^* \in \mathbb{R}$; (ii) function ξ follows logistic model (4.1); (iii) function ξ is independent of reviewer identity r; and (iv) double blind reviewers can estimate true scores of submissions with reasonable accuracy. As we discussed in Section 3, even if assumptions (i)–(iv) are satisfied, the test used by Tomkins et al. (2017) is at risk of violating its Type-I error guarantees unless DB reviewers estimate true scores of submissions without any noise (which is not the case in conference settings). Of course, further violations of the assumptions exacerbate the issues.

In contrast, in our work we attempt the problem making fundamentally different assumptions. Without loss of generality, we denote the probability of reviewer i recommending acceptance for paper j in DB condition as:

$$\pi_{ij}^{(\mathrm{db})} = \xi(r_i, q_j, 0), \tag{4.18}$$

where the last argument of the function ξ is censored, indicating that DB reviewers do not have access to the authorship information. In this notation, our assumption is formulated as follows:

Assumption 1. Under absence of a bias, the behaviour of reviewers does not change between SB and DB conditions, that is, for any reviewer representation r, for any paper representation q and for any value of the indicator $w \in \{-1, 1\}$, we have

$$\xi(r, q, w) = \xi(r, q, 0).$$

Under Assumption 1, the presence of bias is defined as a deviation of reviewers in SB condition from their behavior in DB condition such that the direction of the deviation is determined by the value of the indicator w. Given that, the canonical formulation of the bias testing problem (Problem 1′) corresponds to the absolute bias problem (Problem 1).

Observe that Assumption 1 does not restrict the generality of representations q and r and also does not make strong parametric assumptions about function ξ. Instead, it essentially postulates that the condition in which a reviewer is put does not serve as a confounder, that is, under the absence of bias, the probability that reviewer i votes to accept paper j is independent of whether reviewer i reviews paper j in the SB or DB condition.

To accommodate an additional confounding factor — a distributional shift due to assignment of a reviewer in the SB or DB condition which is independent of papers' characteristics — we substitute Assumption 1 with its less restrictive version.

Assumption 2. Under the absence of bias, the behaviour of any reviewer i in SB condition is connected to the behaviour of that reviewer in DB condition through a linking function f_0, that is, for any reviewer representation r, for any paper representation q and for any value of the indicator $w \in \{-1, 1\}$, we have

$$\xi(r, q, w) = f_0(\xi(r, q, 0)), \tag{4.19}$$

where f_0 is an (unknown) member of a (known) family $\mathcal{F}$ of monotonic functions acting from $[0, 1]$ to $[0, 1]$.

First, observe that if we restrict $\mathcal{F}$ to be a singleton containing only the identity function, then Assumption 2 reduces to Assumption 1. However, richer choices of family $\mathcal{F}$ allow to incorporate various models of confoundings due to the setup. Second, if we again define the presence of bias as a deviation from (4.19), where the direction of the deviation is determined by indicator w, then the canonical formulation of the bias testing problem (Problem 1′) reduces to the relative bias problem (Problem 2).

In this section we have formulated two assumptions that allow us to perform causal inference and lead to the absolute and relative bias testing problems defined in Section 4. With this formulation, in Section 5 we introduce two statistical procedures to test for biases in the peer-review setup and provide their theoretical analysis in Section 6.

A5 Setup for simulations

In this section we describe setup for simulations we conducted in this work. Notice that in contrast to the test of Tomkins et al. (2017) which operates on accept/reject decisions of SB reviewers and scores provided by DB reviewers, the tests we introduce in this work operate on decisions of both SB and DB reviewers. Hence, to compare tests we need to specify (i) models of DB/SB reviewers' decisions and (ii) models of DB reviewers' scores. All simulations are run for 5000 iterations.

A5.1 Simulations in Section 3

We now provide necessary details for the simulations in Section 3.

A5.1.1 Measurement error (Figure 2a)

For this simulation we consider the following model of SB and DB reviewers:

$$\text{DB:} \quad \log \frac{\pi_j^{(\text{db})}}{1-\pi_j^{(\text{db})}} = \beta_0 + \beta_1 q_j^* \tag{4.20a}$$

$$\text{SB:} \quad \log \frac{\pi_{ij}^{(\text{sb})}}{1-\pi_{ij}^{(\text{sb})}} = \beta_0 + \beta_1 q_j^* + \beta_2 w_j, \tag{4.20b}$$

that is, model (4.1) is correct and reviews given by the same reviewer for different papers are independent. Notice that under this model all reviewers are identical and hence issues with the setup do not manifest in this case.

We set $m = 2n = 1000$ and $\mu = \lambda = 2$. At each iteration we independently sample true scores of papers $q_j^*, j \in [n]$, from uniform distribution $\mathcal{U}[-2, 2]$ and assume that mean scores by two DB reviewers assigned to a paper $j \in [n]$ estimates true score q_j^* with some Gaussian noise ($\sigma = 0.7$). We then sample values of $w_j, j \in [n]$, such that correlation between q^* and w equals φ for values of φ between 0 and 0.5. To this end, we let each paper $j \in [n]$ with the score $q_j^* < 0$ have $w_j = 1$ with probability $0.5 - \gamma$ and $w_j = -1$ otherwise. Similarly, each paper $j \in [n]$ with the score $q_j^* \geq 0$ has $w_j = 1$ with probability $0.5 + \gamma$ and $w_j = -1$ otherwise. We then vary the value of $\gamma \in (0, 0.5)$ to achieve the necessary correlation. Finally, using models (4.20a) and (4.20b) with $\beta_0 = 1$, $\beta_1 = 2$ and $\beta_2 = 0$ (no bias condition) we sample decisions of SB and DB reviewers and run the Disagreement test and the test used by Tomkins et al. (2017), setting the significance level to be $\alpha = 0.05$. We then compute a Type-I error as a fraction of iterations in which the null hypothesis ($\beta_2 = 0$) was rejected.

A5.1.2 Model mismatch (Figure 2b)

For this simulation we consider a violation of model (4.1) and the following model of SB and DB reviewers with $\beta_2 = 0$ (no-bias condition):

$$\text{DB:} \quad \log \frac{\pi_j^{(\text{db})}}{1-\pi_j^{(\text{db})}} = \beta_0 + \beta_1 (q_j^*)^3$$

$$\text{SB:} \quad \log \frac{\pi_{ij}^{(\text{sb})}}{1-\pi_{ij}^{(\text{sb})}} = \beta_0 + \beta_1 (q_j^*)^3 + \beta_2 w_j.$$

To abstract out the effect of measurement error, in this section we assume that the true scores $q^*_j, j \in [n]$, are known, but the test used by Tomkins et al. (2017) fits the model defined by equation (4.20b). Besides the change of correct model and availability of true scores $\{q^*_j, j \in [n]\}$, the simulations follow scenario we described in Appendix A5.1.1.

A5.1.3 Reviewer calibration (Figure 2c)

In this simulation we model the effect of correlations introduced by reviewer calibration. More concretely, we construct a model of reviewer calibration under which the test by Tomkins et al. (2017) fails to control for the Type-I error rate. In this section we assume that true scores of submissions are proportional to the clarity of the writing. We then sample clarity scores $\zeta_j, j \in [n]$, from uniform distribution $\mathcal{U}[-1, 1]$ and define $q^*_j = \zeta_j$ for each $j \in [n]$. Eventually, we consider the following model of reviewer. For each $i \in [m]$ and for each $j \in [n]$:

$$\begin{aligned} \text{DB:} \quad & \pi^{(\text{db})}_{ij} = \pi^{(\text{db})}_j + \ell_i \times \mathbb{I}\left[\zeta_j < 0.5\right] \\ \text{SB:} \quad & \pi^{(\text{sb})}_{ij} = \pi^{(\text{sb})}_j + \ell_i \times \mathbb{I}\left[\zeta_j < 0.5\right], \end{aligned}$$

where ℓ_i is reviewers' leniency which equals 0.4 with probability 0.5 and -0.4 otherwise and $\pi^{(\text{db})}_j, \pi^{(\text{sb})}_j$ are defined by equations (4.20a) and (4.20b) with $\beta_0 = 0, \beta_1 = 0.25$ and $\beta_2 = 0$ (no bias condition). Parameters are selected to ensure that $0 \le \pi^{(\text{db})}_{ij}, \pi^{(\text{sb})}_{ij} \le 1$.

In words, the above model says that for papers with high quality of writing ($\zeta > 0.5$) reviewers understand their content well and follow models (4.20a) and (4.20b) exactly, but for papers with lower writing quality their leniency parameter influences their decision. Notice that under this model it is natural to expect that estimates of the true scores provided by DB reviewers are also influenced by their leniency and hence are noisy. However, to isolate the effect of reviewer identity we assume that the test used by Tomkins et al. (2017) knows true scores $q^*_j, j \in [n]$, exactly. Additionally, notice that marginally each reviewer follows the model defined by equations (4.20a) and (4.20b), and hence when $\mu = 1$, the test by Tomkins et al. (2017) has control over the Type-I error for any correlation between q^* and w.

In this section we consider an extreme pattern of correlations between q^* and w. Concretely, we assume that for any paper $j \in [n]$, we have $w_j = 1$ if and only if $q^*_j > 0.5$ and $w_j = -1$ otherwise. Notice that in practice such strong dependence is unlikely to happen, but we underscore that in practice the test by Tomkins et al. (2017) also does not have access to noiseless true scores which will cause measurement errors and hence will exacerbate the issue.

We then perform simulations as discussed above having $n = 1000$ and $\lambda = 1$ fixed and varying the number of papers per reviewer and using the modification of the Wald test with factor variable for each reviewer added (reviewer-depedent intercept).

A5.1.4 Non-blind bidding (Figure 3a)

Formalizing the intuition we mentioned in Section 3.2, we consider a setting with $n = 1000, m = 2000, \lambda = \mu = 1$ and consider a property of interest "paper has a famous author". Suppose that during the bidding procedure each reviewer $i \in [m]$ gives a score $b_{ij} \in \{-1, 0, 1\}$ to each paper $j \in [n]$, where $b_{ij} = 1$ means that reviewer wants to review the paper, $b_{ij} = -1$ means that reviewer does not want to review the paper and $b_{ij} = 0$ is an intermediate between $b_{ij} = 1$ and $b_{ij} = -1$. Given the bids, the assignment is computed maximizing the total sum of the bids. Namely, for all $(i, j) \in [m] \times [n]$ let a binary indicator A_{ij} equal 1 if reviewer i is assigned to paper j and 0 otherwise and let $\mathcal{R}_{\text{SB}} \subset [m]$ be the set of reviewers allocated to SB condition. Then the assignment of SB reviewers to papers is computed maximizing the following objective subject to the standard (λ, μ)-load constraints.

$$\sum_{i \in \mathcal{R}_{\text{SB}}} \sum_{j \in [n]} A_{ij} b_{ij}.$$

The same objective is used to assign DB reviewers to papers. Next, we suppose that for each paper $j \in [n]$ there is a true score $q_j^* \in [0, 0.9]$ and that all reviewers belong to one of the following personality types:

- **Type A:** Lenient reviewers who accept each paper $j \in [n]$ assigned to them with probability $q_j^* + 0.1$ and want to read papers from top authors. If bidding is blind, they do not have any information about author identity and bid 0 on each paper, but if bidding is non-blind, then for each paper $j \in \mathcal{J}$ reviewer i of type A places a bid $b_{ij} = 1$ and for each paper $j \in \overline{\mathcal{J}}$ she/he places a bid $b_{ij} = -1$.
- **Type B:** Accurate reviewers who accept each paper $j \in [n]$ assigned to them with probability q_j^* and do not mind reviewing any paper. Independent of whether bidding is blind or not, reviewer i of type B places a bid $b_{ij} = 0$ on each paper $j \in [n]$.

Notice that evaluations of reviewers of both types are unbiased — the probability of acceptance is not determined by author identities. The type of each reviewer is determined independently: reviewer $i \in [m]$ is of type A with probability 0.3 and of type B with probability 0.7. Independently, each paper $j \in [n]$ belongs to $\mathcal{J}$ with probability 0.3 and to $\overline{\mathcal{J}}$ with probability 0.7.

Having defined the setup, in each iteration we independently sample true scores of submissions from $\mathcal{U}[0, 0.9]$ (no correlation with indicator w) and compute two bidding matrices: (i) when SB reviewers observe author identities during bidding and (ii) when bidding is blind for both SB and DB reviewers. For each bidding matrix we compute assignments of SB and DB reviewers to papers and pass observed decisions to the Disagreement test and the test used by Tomkins et al. (2017). For the test of Tomkins et al., we assume that true scores $q_j^*, j \in [n]$, are known exactly.

A5.1.5 Non-random assignment (Figure 3b)

In this section we construct a similarity matrix S and formalize the dependence of reviewer's perception of a paper on similarity between paper and reviewer that leads to the effect demonstrated in Figure 3b. We notice that the construction we provide here is artificial and serves as a proof of concept for our claim that non-random assignment may violate some key independence assumptions of statistical tests even if it is not based on reviewers' bids. While in practice we do not expect to observe such specific similarity matrices, we can still observe some more subtle manifestations of issues caused by non-randomness of the assignment.

First, in this section we assume that assignment is performed using the TPMS algorithms (Charlin and Zemel, 2013), that is, given similarity matrix S between reviewers and papers, each paper is assigned to λ reviewers in a way that each reviewer is assigned to at most μ papers such that total sum similarity of the assignment is maximized.

Second, consider a similarity matrix S, defined as follows. For each reviewer $i \in [m]$ and for each paper $j \in [n]$:

$$S_{ij} = (m + 1 - i) \times (n + 1 - j). \tag{4.21}$$

Given that reviewers are allocated to conditions at random, similarity matrices S_{SB} (SB condition) and S_{DB} (DB condition) are constructed by random division of rows of S into two groups of equal size and stacking them into S_{SB} and S_{DB} correspondingly.

Third, we assume that each reviewer $i \in [m]$ has some value of threshold z_i such that if reviewer i is assigned to paper $j \in [n]$ in either of setups, reviewer accepts the paper with probability π_{ij} given by:

$$\pi_{ij} = \begin{cases} 0.9 & \text{if } S_{ij} \geq z_i \\ q_j^* & \text{if } S_{ij} < z_i, \end{cases} \tag{4.22}$$

where $q_j^* \in [0, 0.9]$ is a true score of paper j. We also assume that reviewer i in DB condition returns π_{ij} as an estimate of q_j^*.

Fourth, for every reviewer i we set a value of threshold as follows:

$$z_i = (m + 1 - i) \times (n - [(i-1)/2]), \tag{4.23}$$

where $[x]$ is the integral part of x.

Fifth and finally, we assume that true scores q^* are independently sampled from $\mathcal{U}[0, 0.9]$ and sample indicators w such that they are correlated with q^*, fixing the value of correlation $\varphi = 0.45$. We also set $\mu = \lambda = 1$ and $m = 2n = 1000$. Now we allocate half of reviewers to SB condition and half to DB condition uniformly at random. We then compare the performance of the DISAGREEMENT test and the test by Tomkins et al. under (i) experimental setup of Tomkins et al. and (ii) our experimental setup.

The intuition behind our construction of matrix S in equation (4.21) is that for any square submatrix of S, the TPMS algorithm with parameters $\mu = \lambda = 1$ will compute an assignment that corresponds to the diagonal of this submatrix. Coupled with specific choice of thresholds (4.23), probabilities of acceptance (4.22) and correlation between q^* and w at the level of 0.45, this choice of similarity matrix ensures that under the setup of Tomkins et al., with non-zero probability most of SB reviewers will receive papers with similarities above the corresponding threshold and most of DB reviewers will receive papers with similarities below the corresponding threshold or vice versa. Hence, the assignments will be structurally different and, as demonstrated by Figure 3b, this difference will be confused with bias by both Tomkins et al. (2017) and DISAGREEMENT tests. In contrast, under our proposed setup the assignments of SB and DB reviewers to papers do not exhibit any structural difference and hence do not break the Type-I error guarantees of the tests.

A5.2 Simulations in Section 1 and Section 6.1

The simulations in Section 1 and Section 6.1 were performed under the model of reviewers in (4.20a) and (4.20b) following the setup described in Appendix A5.1.1 with small differences. Instead of varying the value of correlation φ between q^* and w, we fix the value of φ and vary the number of papers n. Moreover, we independently assign papers to the sets $\mathcal{J}$ and $\overline{\mathcal{J}}$ as follows: each paper j such that $q^*_j < 0$ belongs to the set $\mathcal{J}$ with probability $0.5 - \gamma$ and otherwise belongs to the set $\overline{\mathcal{J}}$, similarly, each paper j with $q^*_j > 0$ belongs to the set $\mathcal{J}$ with probability $0.5 + \gamma$ and otherwise belongs to the set $\overline{\mathcal{J}}$. The value of γ is selected to achieve the required level of correlation φ between q^* and w.

- For Figure 1a we set $\varphi = 0.4$ and perform simulations under $\beta_0 = 1, \beta_1 = 2, \beta_2 = 0$ (no bias), $\lambda = 2, \mu = 1$, where true scores are sampled from $\mathcal{U}[-1, 1]$. We see that for the test used by Tomkins et al. (2017) a violation of Type-I error guarantees caused by measurement error coupled with correlations (see Appendix A5.1.1 for details) exacerbates as sample size grows.
- For Figure 1b we set $\varphi = 0.6$ and perform simulations under $\beta_0 = 1, \beta_1 = 2, \beta_2 = -0.35$ (bias against papers that satisfy the property), $\lambda = 2, \mu = 1$, where true scores are sampled from $\mathcal{U}[-0.5, 0.5]$. We see that in this case measurement error has strong harmful impact on the power of the test used by Tomkins et al. (2017).
- For Figure 1c we set $\varphi = 0$ and additionally assume that DB reviewers estimate true scores with no noise. In this case all parametric assumptions made by Tomkins et al. (2017) are satisfied. We then perform simulations under $\beta_0 = 1, \beta_1 = 2, \beta_2 = 0.35$ (bias in favour of papers that satisfy the property), $\lambda = 2, \mu = 1$, where true scores are sampled from $\mathcal{U}[-1, 1]$.
- Simulations in Section 6.1 follow the simulations in Figure 1b and Figure 1c with the exception that the COUNTING test is added for comparison.

A5.3 Simulations in Section 6.2

In this section we illustrate that the COUNTING test designed to control for Type-I error under the generalized linear model does not lead to reliable testing under the generalized logistic model under which the DISAGREEMENT test is suitable, and vice versa. To this end, we design two instances of the relative bias problem under the generalized *linear* model — instance (i) with presence of bias and instance (ii) with absence of bias. Our construction ensures that the resulting matrices Π^{sb} and Π^{db} *simultaneously also* fall in the relative bias problem under the generalized *logistic* model with the exception that instance (i) corresponds to absence of

bias under the generalized logistic model and instance (ii) corresponds to the presence of bias under this model.

Instance (i) Under the generalized linear model, for each paper $j \in \mathcal{J}$ let $q_j^* = 0.7$ and for each paper $j \in \overline{\mathcal{J}}$ let $q_j^* = 0.5$. Additionally, let $\nu = 0.175$. This choice of parameters defines matrices Π_1^{sb} and Π_1^{db} that are generated according to the equations (4.5a) and (4.5b) and fall under the null hypothesis of no bias.

Instance (ii) Under the generalized linear model, for each paper $j \in \mathcal{J}$ let $q_j^* = 0.65$ and for each paper $j \in \overline{\mathcal{J}}$ let $q_j^* = 0.25$. Now let matrix Π_2^{db} be defined according to the model (4.5a) and matrix Π_2^{sb} be defined as follows:

$$\pi_{ij}^{(\text{sb})} = \begin{cases} q_j^* + \nu_1 & \text{if } w_j = 1 \\ q_j^* + \nu_2 & \text{if } w_j = -1 \end{cases}, \tag{4.24}$$

where we carefully select $\nu_2 > \nu_1$ as explained below. This choice of parameters leads to a correct alternative hypothesis of presence of bias against papers that satisfy the property of interest.

We now simulate reviewers decisions with $\lambda = 2, \mu = 1, n = 1000, m = 4000$, independently allocating each paper to $\mathcal{J}$ with probability 0.5 and to $\overline{\mathcal{J}}$ otherwise. We then apply the Counting and Disagreement tests for each of these instances, and present the results in Figure 5a. Instance (i) allows to compare Type-I error rates, and instance (ii) allows to compare powers of the tests under the generalized linear model.

One can verify that the instances we constructed above under the generalized *linear* model also fall under the generalized *logistic* model for some specific choice of parameters. Indeed, consider an instance of the generalized logistic model specified by parameters $\beta_0 = -2.5 \log 7/3, \beta_1 = 5 \log 7/3$ and $\widetilde{\nu} = 1$. Then a straightforward verification shows that matrix Π_1^{db} satisfies equation (4.8a) which specifies the behavior of DB reviewers under the generalized logistic model. Next, observe that for each reviewer $i \in [m]$ and for each paper $j \in \mathcal{J}$ the corresponding entry of the matrix Π_1^{sb} is *larger* than prescribed by the model of SB reviewers under the absence of bias (4.8b). Similarly, for each paper $j \in \overline{\mathcal{J}}$ the corresponding entry of the matrix Π_1^{sb} is *smaller* than it should be if the bias is absent (4.8b). Hence, the pair of matrices $\Pi_1^{\text{db}}, \Pi_1^{\text{sb}}$ satisfies the *alternative* hypothesis under the generalized logistic model.

Conversely, consider an instance of the generalized logistic model specified by parameters $\beta_0 = \log 1/3 - 0.625 \log 39/7, \beta_1 = 2.5 \log 39/7$ and $\widetilde{\nu} = 1.5$. Then a straightforward verification shows that matrix Π_2^{db} satisfies equation (4.8a) which specifies the behavior of DB reviewer under the generalized logistic model. Recall that at this point we didn't specify how we selected values ν_1, ν_2 in equation (4.24). In fact, we selected these values such that entries of the matrix Π_2^{sb} satisfy equation (4.8b) which specifies the behavior of SB reviewers under the generalized logistic model when the bias is absent. Namely, we set

$$\nu_1 = -0.65 + \left(1 + \exp\left\{-\beta_0 - \widetilde{\nu} - 0.65\beta_1\right\}\right)^{-1}$$
$$\nu_2 = -0.25 + \left(1 + \exp\left\{-\beta_0 - \widetilde{\nu} - 0.25\beta_1\right\}\right)^{-1}$$

As a result, the pair of matrices $\Pi_2^{\text{db}}, \Pi_2^{\text{sb}}$ satisfies the *null* hypothesis under the generalized logistic model.

Finally, the power of the Counting test in Figure 5a becomes the Type-I error rate under the instance of the generalized logistic model with $\beta_0 = \log 1/3 - 0.625 \log 39/7, \beta_1 = 2.5 \log 39/7$ and $\widetilde{\nu} = 1$. Similarly, the Type-I error of the Counting test in Figure 5a becomes its power under the instance of the generalized logistic model with $\beta_0 = -2.5 \log 7/3, \beta_1 = 5 \log 7/3$ and $\widetilde{\nu} = 1$. The same applies to the Disagreement test and eventually we obtain Figure 5b by simply exchanging the bars in Figure 5a.

A6 Proofs of main results

In this section we give proofs of our main results.

A6.1 Proof of Theorem 1

We prove Theorem 1 in two steps. First, we show the result for the DISAGREEMENT test and then for the COUNTING test. Before we delve into proofs, let us make two observations that we use in this section.

Observations:

A For every paper $j \in [n]$, if in assignment A^* (Step 2 of Procedure 1) paper j is attributed to reviewers i_1 and i_2, then the events "reviewer i_1 is allocated to SB and reviewer i_2 is allocated to DB" and "reviewer i_1 is allocated to DB and reviewer i_2 is allocated to SB" are mutually exclusive and happen with probability 0.5 each. This is ensured by Step 3 of Procedure 1 where reviewers are allocated to conditions.

B By construction of Procedure 1, at least $c\min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\}$ papers each from sets $\mathcal{J}$ and $\overline{\mathcal{J}}$ appear in assignment A^* for some constant c that depends only on parameters λ and μ. Indeed, in cases (a) and (c) of Step 2, all papers are included into assignment A^* and hence our claim holds with $c = 1$. In case (b) $\frac{m}{2} \geq \frac{\lambda}{\mu} n \geq 2\frac{\lambda}{\mu}\min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\}$ papers are selected and hence our claim holds with $c = \frac{\lambda}{\mu}$.

A6.1.1 Proof for DISAGREEMENT test

The proof of Theorem 1 for the DISAGREEMENT test consists of two parts. First, we show that under the null hypothesis defined in Problem 1, for any matrices Π^{db} and $\Pi^{\text{sb}}(= \Pi^{\text{db}})$ and for any assignment A^* constructed by Procedure 1 in Step 2, the test rejects the null with probability at most α. Second, we show that if the number of papers in both $\mathcal{J}$ and $\overline{\mathcal{J}}$ is large enough, then the DISAGREEMENT test satisfies the requirement of non-trivial power.

We prove both parts conditioned on the assignment A^*. The unconditional statement of the theorem then follows from the law of total probability.

Control over Type-I error

Let Π^{db} and $\Pi^{\text{sb}}(= \Pi^{\text{db}})$ be arbitrary matrices that fall under the definition of null hypothesis in Problem 1. Consider arrays U and V constructed in Step 2 of the DISAGREEMENT test from the set of tuples $\mathcal{T}$ passed to the test by Procedure 1. If any of them is empty, the test keeps the null and hence does not commit the Type-I error. Now without loss of generality assume that both U and V are non-empty.

The idea of the proof is to show that under the null hypothesis, entries of arrays U and V are mutually independent and identically distributed. Assume for the moment that it is indeed the case. Then entries of arrays U and V are exchangeable random variables and hence the permutation test with statistic τ defined in Step 3 of Test 1 is guaranteed to provide control over the Type-I error rate for any given significance level $\alpha \in (0, 1)$ and hence the result for Type-I error control follows.

Consider any entry u of array U. Then u is a decision of SB reviewer for some paper $j_t \in \mathcal{J}$, where t is a tuple that corresponds to u. Corresponding SB and DB reviewers disagree in their decisions, that is, $Y_{j_t} \neq X_{j_t}$. Recalling Observation A, we deduce that conditioned on assignment A^*, the symmetry of the null hypothesis guarantees that

$$Y_{j_t} \,|\, (Y_{j_t} \neq X_{j_t}) \;\sim\; \text{Bernoulli}(0.5). \tag{4.25}$$

Indeed, given that both Y_{j_t} and X_{j_t} are Bernoulli random variables, one can verify that

$$\mathbb{P}\left[Y_{j_t} = 1, X_{j_t} = 0\right] = \mathbb{P}\left[Y_{j_t} = 0, X_{j_t} = 1\right],$$

which coupled with the definition of condition probability implies (4.25).

Hence, entries of array U are Bernoulli random variables with expectation 0.5. Provided that each reviewer contributes at most one decision to $\mathcal{T}$, entries of U are also independent. The same argument applies to entries of array V and hence we have shown that under the null hypothesis entries of U and V are independent Bernoulli random variables with probability of success 0.5 and thus are exchangeable.

Non-trivial power

Consider any fixed choice of $\delta > 0$ and $\varepsilon > 0$ in the definition of non-trivial power. The goal now is to show that there exists $n_0 = n_0(\varepsilon, \delta)$ such that if $\min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\} > n_0$, then for any matrices Π^{db} and Π^{sb}

that satisfy the alternative hypothesis in Problem 1 with margin δ, the DISAGREEMENT test coupled with Procedure 1 is guaranteed to reject the null hypothesis with probability at least $1-\varepsilon$. Throughout the proof we use c to denote a universal constant and allow its value to change from line to line due to multiplications by some other universal constants. Recall that problem parameters λ, μ and α are treated as constants. For concreteness, throughout the proof we assume that the bias is in favor of papers from $\mathcal{J}$. The same argument can be repeated in case of bias against papers from $\mathcal{J}$.

Step 1. Cardinality of U and V.

Let us first show that arrays U and V will with high probability contain order n_0 elements. To this end, recall that for tuple $t \in \mathcal{T}$ we add Y_{j_t} to U if (i) $w_{j_t} = 1$ and (ii) $Y_{j_t} \neq X_{j_t}$. Observation B ensures that $\mathcal{T}$ will contain at least cn_0 tuples that correspond to papers from $\mathcal{J}$. Consider any such tuple, and let (j_t, i_1, i_2) be a corresponding paper and two reviewers assigned to this paper in assignment A^*. Then conditioned on assignment A^*, $\mathbb{P}[Y_{j_t} \neq X_{j_t}]$ is lower bounded by:

$$\begin{aligned}
\mathbb{P}[Y_{j_t} \neq X_{j_t}] &= \frac{1}{2}\left(\pi^{(\mathrm{sb})}_{i_1 j_t}(1-\pi^{(\mathrm{db})}_{i_2 j_t}) + \pi^{(\mathrm{db})}_{i_2 j_t}(1-\pi^{(\mathrm{sb})}_{i_1 j_t})\right) + \frac{1}{2}\left(\pi^{(\mathrm{sb})}_{i_2 j_t}(1-\pi^{(\mathrm{db})}_{i_1 j_t}) + \pi^{(\mathrm{db})}_{i_1 j_t}(1-\pi^{(\mathrm{sb})}_{i_2 j_t})\right) \\
&\geq \frac{1}{2}\left(\pi^{(\mathrm{sb})}_{i_1 j_t}(1-\pi^{(\mathrm{db})}_{i_2 j_t})\right) + \frac{1}{2}\left(\pi^{(\mathrm{sb})}_{i_2 j_t}(1-\pi^{(\mathrm{db})}_{i_1 j_t})\right) \\
&\overset{(i)}{\geq} \frac{1}{2}\left(\delta^2 + \delta^2\right) = \delta^2,
\end{aligned}$$

where inequality (i) follows from the fact that for any reviewer $i \in [m]$ and for any paper $j \in [n]$ we have $\delta \leq \pi^{(\mathrm{sb})}_{ij} \leq 1$ and $0 \leq \pi^{(\mathrm{db})}_{ij} \leq 1-\delta$ by the definition of non-trivial power requirement.

The same argument applies to tuples $t \in \mathcal{T}$ that correspond to papers from $\overline{\mathcal{J}}$. Hence, we conclude that for any tuple $t \in \mathcal{T}$ we are guaranteed that $Y_{j_t} \neq X_{j_t}$ with probability at least δ^2.

Now notice that $|U| = \sum_{t \in \mathcal{T}:\ w_{j_t}=1} \mathbb{I}[Y_{j_t} \neq X_{j_t}]$ and hence $\mathbb{E}[|U|] \geq cn_0\delta^2$. Applying Hoeffding's inequality, we can also derive that for large enough n_0 with probability at least $1 - \frac{\varepsilon}{4}$ we have

$$|U| > cn_0\delta^2.$$

The same argument applies to V and hence we conclude that with probability at least $1 - \frac{\varepsilon}{2}$ we have

$$|U| > cn_0\delta^2 \quad \text{and} \quad |V| > cn_0\delta^2. \tag{4.26}$$

Step 2. Distribution.

Now we describe the distribution of components of U and V. By construction, the entries of these arrays are independent, so it suffices to study a single component. Consider an entry u of array U and let (j, i_1, i_2) be a corresponding paper and two reviewers assigned to this paper in assignment A^*. For brevity, denote $p = \pi^{(\mathrm{sb})}_{i_1 j} \in (\delta, 1]$, $q = \pi^{(\mathrm{db})}_{i_2 j} \in [0, 1-\delta)$, $\gamma_1 = p - \pi^{(\mathrm{db})}_{i_1 j}$ and $\gamma_2 = \pi^{(\mathrm{sb})}_{i_2 j} - q$, where $\gamma_1 > \delta$ and $\gamma_2 > \delta$ by definition of non-trivial power requirement. Then, we can derive the following chain of bounds:

$$\begin{aligned}
2\mathbb{P}[u=1] - 1 &= 2\mathbb{P}[Y_j = 1 | Y_j \neq X_j] - 1 \\
&= \frac{p(1-q)}{p(1-q)+q(1-p)} + \frac{(q+\gamma_2)(1-p+\gamma_1)}{(q+\gamma_2)(1-p+\gamma_1)+(p-\gamma_1)(1-q-\gamma_2)} - 1 \\
&\overset{(i)}{\geq} \frac{p(1-q)}{p(1-q)+q(1-p)} + \frac{(q+\delta)(1-p+\delta)}{(q+\delta)(1-p+\delta)+(p-\delta)(1-q-\delta)} - 1 \\
&= \frac{1}{2}\left(\frac{p-q}{p+q-2pq} - \frac{p-q-2\delta}{p+q-2pq+2\delta(\delta+q-p)}\right)
\end{aligned}$$

where inequality (i) holds due to monotonicity of the expression over γ_1 and γ_2 and lower bounds $\gamma_1 > \delta$, $\gamma_2 > \delta$.

Optimizing the last expression over $p \in (\delta, 1]$ and $q \in [0, 1-\delta)$, we obtain

$$2\mathbb{P}[u=1] - 1 \geq \frac{\delta^2}{\delta^2 + (1-\delta)^2},$$

and hence $\mathbb{P}[u=1] \geq \frac{1}{2} + \frac{1}{2}\frac{\delta^2}{\delta^2+(1-\delta)^2} = \frac{1}{2} + \gamma$. Similarly, we can show that $\mathbb{P}[v=1] \leq \frac{1}{2} - \frac{1}{2}\frac{\delta^2}{\delta^2+(1-\delta)^2} = \frac{1}{2} - \gamma$, where $\gamma > 0$ is a constant that depends on δ.

Step 3. Permutation.

At this point we are guaranteed that vectors V and U constructed in Step 2 of the DISAGREEMENT test, with probability $1 - \frac{\varepsilon}{2}$, contain at least $cn_0\delta^2$ elements and their entries are independent Bernoulli random variables. Moreover, the entries of U have expectations larger than $1/2 + \gamma$ and entries of V have expectations smaller than $1/2 - \gamma$, where γ is independent of n_0.

Conditioned on $\min\{|V|, |U|\} > cn_0\delta^2$, notice that as n_0 grows, the permutation test for exchangeablility of entries of V and U has power growing to 1. Hence, there exists n_0^* such that if $n_0 > n_0^*$, then the permutation test rejects the null with probability at least $1 - \frac{\varepsilon}{2}$.

Finally, taking union bound over (i) probability that either of U and V has cardinality smaller than $cn_0\delta^2$ and (ii) probability that the permutation test fails to reject the null given $\min\{|V|, |U|\} > cn_0\delta^2$, we deduce that conditioned on A^*, the requirement of non-trivial power is satisfied. It now remains to notice that the established fact holds for any A^* that is constructed by Procedure 1 and hence Theorem 1(a) holds.

A6.1.2 Proof for COUNTING test

Similar to the proof for theDISAGREEMENT test, the proof for the COUNTING test consists of two parts — control over Type-I error and non-trivial power. As in the proof for the DISAGREEMENT test, we prove both parts conditioned on the assignment A^* computed in Step 2 of Procedure 1. The unconditional statement of the theorem then follows from the law of total probability.

Control over Type-I error

Let Π^{db} and $\Pi^{\text{sb}} (= \Pi^{\text{db}})$ be arbitrary matrices that fall under the definition of the null hypothesis in Problem 1. Consider arrays U and V constructed in Step 2 of the COUNTING test. If any of them is empty, the test keeps the null and hence does not commit the Type-I error. Now without loss of generality assume that both U and V are non-empty. By construction, conditioned on the assignment A^*, entries of arrays U and V are mutually independent and bounded by 1 in absolute value. Moreover, conditioned on A^* the size of arrays U and V is fixed and is not a random variable. Next, we can show that expectation of any entry of arrays U and V is zero. Indeed, consider any arbitrary entry $u \in U$ and let (j, i_1, i_2) be a corresponding paper and reviewers assigned to this paper in assignment A^*. Then:

$$\mathbb{E}[u] = \frac{1}{2}\left(\pi^{(\text{sb})}_{i_1 j} - \pi^{(\text{db})}_{i_2 j}\right) + \frac{1}{2}\left(\pi^{(\text{sb})}_{i_2 j} - \pi^{(\text{db})}_{i_1 j}\right) = 0,$$

where two terms correspond to two equiprobable allocations of reviewers i_1 and i_2 to conditions and the last equality follows from the fact that under the null hypothesis $\Pi^{\text{sb}} = \Pi^{\text{db}}$. Hence, we conclude that the expectation of test statistic γ equals 0. Independence and boundedness of entries of arrays V and U ensure that the test statistic γ is sub-Gaussian random variable with noise parameter σ given by

$$\sigma^2 = |U|^{-1} + |V|^{-1}.$$

Finally, applying Hoeffding's inequality we deduce that

$$\mathbb{P}\left[|\gamma| > \sqrt{2\left(|U|^{-1} + |V|^{-1}\right)\log 2/\alpha}\right] \leq 2\exp\left\{-\frac{2\left(|U|^{-1} + |V|^{-1}\right)\log 2/\alpha}{2\left(|U|^{-1} + |V|^{-1}\right)}\right\} = \alpha,$$

which concludes the proof.

Non-trivial power
Consider any fixed choice of $\delta > 0$ and $\varepsilon > 0$ in the definition of non-trivial power. The goal now is to show that there exists $n_0 > 0$ such that if $\min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\} > n_0$, then for any matrices Π^{db} and Π^{sb} that satisfy the alternative hypothesis in Problem 1 with margin δ, the COUNTING test coupled with Procedure 1 rejects the null hypothesis with probability at least $1 - \varepsilon$. Throughout the proof we use c to denote a universal constant and allow its value to change from line to line due to multiplications by some other universal constants. Recall that problem parameters λ, μ and α are treated as constants. For concreteness, suppose that there is a bias in favor of papers that satisfy the property of interest.

We now consider an arbitrary instance of the bias testing problem with matrices Π^{sb} and Π^{db} that fall under the definition of non-trivial power. First, Observation B ensures that the set $\mathcal{T}$ passed to the COUNTING algorithm is such that the resulting vectors U and V contain at least cn_0 elements each. Next, let $\gamma_1 = \frac{1}{|U|} \sum\limits_{u \in U} u$ and $\gamma_2 = \frac{1}{|V|} \sum\limits_{v \in V} v$, in this notation the test statistic is defined as $\gamma = \gamma_1 - \gamma_2$. Conditioned on the assignment A^*, we have:

$$\mathbb{E}[\gamma_1] = \frac{1}{|U|} \sum_{u \in U} \mathbb{E}[u] \geq \delta.$$

Indeed, for any arbitrary entry u of array U let (j, i_1, i_2) be corresponding paper and reviewers assigned to this paper in assignment A^*. Then requirement of non-trivial power guarantees that

$$\mathbb{E}[u] = \frac{1}{2}\left(\pi^{(\text{sb})}_{i_1 j} - \pi^{(\text{db})}_{i_2 j}\right) + \frac{1}{2}\left(\pi^{(\text{sb})}_{i_2 j} - \pi^{(\text{db})}_{i_1 j}\right) \geq \delta + \frac{1}{2}\left(\pi^{(\text{db})}_{i_1 j} - \pi^{(\text{db})}_{i_2 j}\right) + \frac{1}{2}\left(\pi^{(\text{db})}_{i_2 j} - \pi^{(\text{db})}_{i_1 j}\right) = \delta.$$

Similarly,

$$\mathbb{E}[\gamma_2] = \frac{1}{|V|} \sum_{v \in V} \mathbb{E}[v] \leq -\delta.$$

Applying Hoeffding's inequality we obtain:

$$\mathbb{P}[\gamma_1 - \gamma_2 < \delta] \leq \mathbb{P}[\gamma_1 - \gamma_2 < \mathbb{E}[\gamma_1 - \gamma_2] - \delta] \leq \exp\left(-\frac{\delta^2}{2\left(|V|^{-1} + |U|^{-1}\right)}\right) \leq \exp\left(-c\delta^2 n_0\right).$$

On the other hand, the threshold for rejecting the null is such that

$$\sqrt{2\left(|U|^{-1} + |V|^{-1}\right)\log 2/\alpha} \leq c\sqrt{\frac{1}{n_0}}.$$

Finally, setting $n_0 = c\frac{\log 1/\varepsilon}{\delta^2}$, we ensure that if $\min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\} > n_0$, then the COUNTING algorithm with probability at least $1 - \varepsilon$ rejects the null for any matrices Π^{sb}, Π^{db} that satisfy alternative hypothesis with margin δ.

A6.2 Proof of Theorem 2

We prove Theorem 2 separately for the DISAGREEMENT test and for the COUNTING tests.

A6.2.1 Proof for DISAGREEMENT test

Again, the proof is presented in two parts: control over Type-I error and non-trivial power. The conceptual difference from the proof of the corresponding result for absolute bias problem is that now the parametric relationships (4.8a) and (4.8b) allow us to avoid conditioning on the assignment A^*.

Control over Type-I error
Let Π^{db} and Π^{sb} be arbitrary matrices generated from the generalized logistic model under the absence of bias. Consider arrays U and V constructed in Step 2 of the DISAGREEMENT test from the set of tuples

$\mathcal{T}$ passed to the test by Procedure 1. If any of them is empty, the test keeps the null and hence does not commit the Type-I error. Now without loss of generality assume that both arrays U and V are non-empty. Following the idea of the proof of Theorem 1, we need to show that entries of arrays U and V are exchangeable random variables. First, the mutual independence follows from construction of the set $\mathcal{T}$. Second, using equations (4.8a) and (4.8b), we deduce that for any paper $j \in [n]$ and for any reviewer $i \in [m]$:

$$\log \frac{\pi_{ij}^{(\mathrm{sb})}(1-\pi_{ij}^{(\mathrm{db})})}{\pi_{ij}^{(\mathrm{db})}(1-\pi_{ij}^{(\mathrm{sb})})} = \widetilde{\nu}.$$

Noticing that $\pi_{ij}^{(\mathrm{sb})}$ and $\pi_{ij}^{(\mathrm{db})}$ under the generalized logistic model are independent of reviewer's identity, we drop index i from the above equation. Now we consider any entry u of array U together with a corresponding tuple $t = (j_t, Y_{j_t}, X_{j_t}, w_{j_t})$ and conclude that:

$$\begin{aligned}
\mathbb{P}[u=1] &= \mathbb{P}[Y_{j_t} = 1 | Y_{j_t} \neq X_{j_t}] \\
&= \frac{\pi_{j_t}^{(\mathrm{sb})}(1-\pi_{j_t}^{(\mathrm{db})})}{\pi_{j_t}^{(\mathrm{sb})}(1-\pi_{j_t}^{(\mathrm{db})}) + \pi_{j_t}^{(\mathrm{db})}(1-\pi_{j_t}^{(\mathrm{sb})})} \\
&= \frac{1}{1 + \frac{\pi_{j_t}^{(\mathrm{db})}(1-\pi_{j_t}^{(\mathrm{sb})})}{\pi_{j_t}^{(\mathrm{sb})}(1-\pi_{j_t}^{(\mathrm{db})})}} \\
&= \frac{1}{1+e^{-\widetilde{\nu}}}.
\end{aligned} \tag{4.27}$$

Importantly, the value of the paper representation q_j does not appear in equation (4.27), implying that entries of array U are identically distributed. Applying the same argument to entries of array V we deduce that entries of arrays U and V are exchangeable random variables and hence the permutation test with the test statistic τ defined in Step 3 of Test 1 is guaranteed to control for the Type-I error rate at any given significance level $\alpha \in (0,1)$ which concludes the proof.

Non-trivial power

Consider any fixed choice of $\delta > 0$ and $\varepsilon > 0$ in the definition of non-trivial power. The goal now is to show that there exists $n_0 = n_0(\varepsilon, \delta)$ such that if $\min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\} > n_0$, then for any matrices Π^{db} and Π^{sb} generated from the generalized logistic model that satisfy the alternative hypothesis in Problem 2 with margin δ, the Disagreement test coupled with Procedure 1 is guaranteed to reject the null hypothesis with probability at least $1-\varepsilon$. Throughout the proof we use c to denote a universal constant and allow its value to change from line to line due to multiplications by some other universal constants. Recall that problem parameters λ, μ and α are treated as constants. For concreteness, throughout the proof we assume that the bias is in favor of papers from $\mathcal{J}$. The same argument can be repeated in case of bias against papers from $\mathcal{J}$.

Step 1. Cardinality of U and V.

Consider any matrices Π^{sb} and Π^{db} generated from the generalized logistic model that satisfy the alternative hypothesis in Problem 2 with margin δ. First, we notice that scores $q_j, j \in [n]$, and coefficients β_0, β_1 are bounded in absolute value by some constant $\widetilde{\Delta}$, and hence using equation (4.8a) we conclude that for all $(i,j) \in [n] \times [m]$

$$\pi_{ij}^{(\mathrm{db})} \in (\ell, b) \quad \forall j \in [n], \tag{4.28}$$

where $0 < \ell < b < 1$ and values of ℓ and b are determined by $\widetilde{\Delta}$. Now consider any tuple $t = (j_t, Y_{i_1 j_t}, X_{i_2 j_t}, w_{j_t})$

from the set of tuples $\mathcal{T}$. Then

$$\begin{aligned}\mathbb{P}\left[Y_{j_t} \neq X_{j_t}\right] &= \pi_{i_1 j_t}^{(\mathrm{sb})}(1-\pi_{i_2 j_t}^{(\mathrm{db})}) + \pi_{i_2 j_t}^{(\mathrm{db})}(1-\pi_{i_1 j_t}^{(\mathrm{sb})}) \\ &\geq \min\{\pi_{i_2 j_t}^{(\mathrm{db})}, 1-\pi_{i_2 j_t}^{(\mathrm{db})}\}\left(\pi_{i_1 j_t}^{(\mathrm{sb})} + 1 - \pi_{i_1 j_t}^{(\mathrm{sb})}\right) \\ &= \min\{\pi_{i_2 j_t}^{(\mathrm{db})}, 1-\pi_{i_2 j_t}^{(\mathrm{db})}\} \\ &\geq \min\{\ell, 1-b\},\end{aligned}$$

where the last inequality follows from equation (4.28). Applying Hoeffding's inequality in the same way as we did in the proof of Theorem 1 to get the bound (4.26), we deduce that with probability at least $1-\frac{\varepsilon}{2}$, cardinalities of arrays U and V are at least cn_0 for some constant c that may depend on δ and $\widetilde{\Delta}$.

Step 2. Distribution

By definition of non-trivial power requirement, it must be the case that for all $(i,j) \in [m] \times [n]$ we have $|\pi_{ij}^{(\mathrm{sb})} - f_0(\pi_{ij}^{(\mathrm{db})})| > \delta$, where function f_0 belongs to class $\widetilde{\mathcal{F}}_{\widetilde{\Delta}}$ defined in (4.9) and for all $(i,j) \in [m] \times [n]$ satisfies:

$$\log \frac{f_0(\pi_{ij}^{(\mathrm{db})})}{1-f_0(\pi_{ij}^{(\mathrm{db})})} = \beta_0 + \widetilde{\nu} + \beta_1 q_j \tag{4.29a}$$

$$= \log \frac{\pi_{ij}^{(\mathrm{db})}}{1-\pi_{ij}^{(\mathrm{db})}} + \widetilde{\nu}, \tag{4.29b}$$

for some value of $\widetilde{\nu} \in (-\widetilde{\Delta}, \widetilde{\Delta})$. Observe that values $\beta_0, \widetilde{\nu}, \beta_1, q_j$ in the RHS of equation (4.29a) are bounded in absolute value by constant $\widetilde{\Delta}$. Next, recall that the definition of the non-trivial power requirement ensures that for each reviewer $i \in [m]$ it must be the case that (a) for each paper $j \in \mathcal{J}$ we have $f_0(\pi_{ij}^{(\mathrm{db})}) < 1-\delta$ and (b) for each paper $j \in \overline{\mathcal{J}}$ we have $f_0(\pi_{ij}^{(\mathrm{db})}) > \delta$. Finally, we are guaranteed that for any pair of reviewer $i \in [m]$ and paper $j \in [n]$ we have

$$f_0(\pi_{ij}^{(\mathrm{db})}) \in \begin{cases}(\ell', \min\{b', 1-\delta\}) & \text{if } j \in \mathcal{J} \\ (\max\{\ell', \delta\}, b') & \text{if } j \in \overline{\mathcal{J}}.\end{cases}$$

Notice that constants ℓ' and b' are such that $0 < \ell' < b' < 1$ and may be different from ℓ and b, because in equation (4.29a) we have additional term $\widetilde{\nu}$ which is absent in (4.8a).

Let us now define two quantities d_1 and d_2 as

$$d_1 = \inf_{t \in (\ell', \min\{b', 1-\delta\})} \left(\log \frac{t+\delta}{1-(t+\delta)} - \log \frac{t}{1-t}\right) \tag{4.30a}$$

$$d_2 = \inf_{t \in (\max\{\ell', \delta\}, b')} \left(\log \frac{t}{1-t} - \log \frac{t-\delta}{1-(t-\delta)}\right). \tag{4.30b}$$

Notice that both quantities d_1 and d_2 are some functions of δ and $\widetilde{\Delta}$ and are strictly positive, because function $\log \frac{x}{1-x}$ is strictly increasing on the interval $(0,1)$ with its derivative being lower bounded by $c > 0$, where c is independent of problem parameters.

Putting together equations (4.29a) - (4.30b), we now show that for each reviewer $i \in [m]$ the definition of non-trivial power requirement ensures that for each paper $j \in \mathcal{J}$

$$\log \frac{\pi_{ij}^{(\mathrm{sb})}}{1-\pi_{ij}^{(\mathrm{sb})}} \geq \log \frac{\pi_{ij}^{(\mathrm{db})}}{1-\pi_{ij}^{(\mathrm{db})}} + \widetilde{\nu} + d_1, \tag{4.31}$$

and for each paper $j \in \overline{\mathcal{J}}$

$$\log \frac{\pi_{ij}^{(\mathrm{sb})}}{1-\pi_{ij}^{(\mathrm{sb})}} \leq \frac{\pi_{ij}^{(\mathrm{db})}}{1-\pi_{ij}^{(\mathrm{db})}} + \widetilde{\nu} - d_2. \tag{4.32}$$

Consider any arbitrary entry u of array U and the corresponding tuple $(j_t, Y_{i_1 j_t}, X_{i_2 j_t}, w_{j_t})$. Then,

$$\begin{aligned}
\mathbb{P}[u=1] &= \mathbb{P}[Y_{i_1 j_t} = 1 | Y_{i_1 j_t} \neq X_{i_2 j_t}] \\
&= \frac{\pi_{i_1 j_t}^{(\mathrm{sb})}(1-\pi_{i_2 j_t}^{(\mathrm{db})})}{\pi_{i_1 j_t}^{(\mathrm{sb})}(1-\pi_{i_2 j_t}^{(\mathrm{db})}) + \pi_{i_2 j_t}^{(\mathrm{db})}(1-\pi_{i_1 j_t}^{(\mathrm{sb})})} \\
&= \frac{1}{1+\frac{\pi_{i_2 j_t}^{(\mathrm{db})}(1-\pi_{i_1 j_t}^{(\mathrm{sb})})}{\pi_{i_1 j_t}^{(\mathrm{sb})}(1-\pi_{i_2 j_t}^{(\mathrm{db})})}} \\
&\geq \frac{1}{1+e^{-\widetilde{\nu}-d_1}},
\end{aligned}$$

where the last inequality follows from (4.31). Similarly, using (4.32) we show that for each entry v of array V

$$\mathbb{P}[v=1] \leq \frac{1}{1+e^{-\widetilde{\nu}+d_2}}.$$

Step 3. Permutation.

At this point we are guaranteed that vectors V and U constructed in Step 2 of the DISAGREEMENT test, with probability $1-\frac{\varepsilon}{2}$, contain at least cn_0 elements and their entries are independent Bernoulli random variables. Moreover, the entries of U have expectations larger than $\frac{1}{1+e^{-\widetilde{\nu}}}+\gamma$ and entries of V have expectations smaller than $\frac{1}{1+e^{-\widetilde{\nu}}}-\gamma$, where γ is independent of n_0, but depends on δ and $\widetilde{\Delta}$.

Conditioned on $\min\{|V|,|U|\} > cn_0$, notice that as n_0 grows, the permutation test for exchangeablility of entries of V and U has power growing to 1. Hence, there exists n_0^* such that if $n_0 > n_0^*$, then the permutation test rejects the null with probability at least $1-\frac{\varepsilon}{2}$.

Finally, taking union bound over (i) probability that either of U and V has cardinality smaller than cn_0 and (ii) probability that the permutation test fails to reject the null given $\min\{|V|,|U|\} > cn_0$, we deduce that the requirement of non-trivial power is satisfied.

A6.2.2 Proof for COUNTING test

We give a proof for an extended version of the generalized linear model in which for each $(i,j) \in [n] \times [m]$ we substitute q_j with q_{ij}, thus allowing subjectivity of reviewers. In the proof we will be using two observations we made in the beginning of Appendix A6.1. As in the proof of Theorem 1, we prove the result conditioned on the assignment A^* constructed in Step 2 of Procedure 1. The unconditional statement of the theorem then follows from the law of total probability.

Control over Type-I error

Let Π^{db} and Π^{sb} be arbitrary matrices generated under the generalized linear model that fall under the null hypothesis in Problem 2. Consider arrays U and V constructed in Step 2 of the COUNTING test. If any of them is empty, the test keeps the null and hence does not commit the Type-I error. Now without loss of generality assume that both U and V are non-empty. By construction, conditioned on the assignment A^*, entries of arrays U and V are mutually independent and bounded by 1 in absolute value. Moreover, conditioned on A^* the size of arrays U and V is fixed and is not a random variable. Next, for any arbitrary entry $u \in U$ let (j, i_1, i_2) be a corresponding paper and reviewers assigned to this paper in assignment A^*.

Then,

$$\begin{aligned}\mathbb{E}[u] &= \frac{1}{2}\left(\pi^{(\mathrm{sb})}_{i_1 j} - \pi^{(\mathrm{db})}_{i_2 j}\right) + \frac{1}{2}\left(\pi^{(\mathrm{sb})}_{i_2 j} - \pi^{(\mathrm{db})}_{i_1 j}\right) \\ &= \frac{1}{2}\left(q_{i_1 j} + \nu - q_{i_2 j}\right) + \frac{1}{2}\left(q_{i_2 j} + \nu - q_{i_1 j}\right) \\ &= \nu.\end{aligned}$$

Similarly, it follows that for any arbitrary entry $v \in V$:

$$\mathbb{E}[v] = \nu.$$

Hence, we conclude that the expectation of the test statistic γ equals 0. Independence and boundedness of entries of arrays V and U ensure that the test statistic γ is sub-Gaussian random variable with noise parameter σ given by

$$\sigma^2 = |U|^{-1} + |V|^{-1}.$$

Finally, applying Hoeffding's inequality we deduce that

$$\mathbb{P}\left[|\gamma| > \sqrt{2\left(|U|^{-1} + |V|^{-1}\right)\log 2/\alpha}\right] \leq 2\exp\left\{-\frac{2\left(|U|^{-1} + |V|^{-1}\right)\log 2/\alpha}{2\left(|U|^{-1} + |V|^{-1}\right)}\right\} = \alpha,$$

which concludes the proof.

Non-trivial power

Consider any fixed choice of $\delta > 0$ and $\varepsilon > 0$ in the definition of non-trivial power. The goal now is to show that there exists $n_0 > 0$ such that if $\min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\} > n_0$, then for any matrices Π^{db} and Π^{sb} generated under the generalized linear model that satisfy the alternative hypothesis in Problem 2 with margin δ, the Counting test coupled with Procedure 1 rejects the null hypothesis with probability at least $1 - \varepsilon$. Throughout the proof we use c to denote a universal constant and allow its value to change from line to line due to multiplications by some other universal constants. Recall that problem parameters λ, μ and α are treated as constants. For concreteness, suppose that there is a bias in favor of papers that satisfy the property of interest.

We now consider an arbitrary instance of the bias testing problem with matrices Π^{sb} and Π^{db} that fall under the definition of non-trivial power. First, Observation B ensures that the set $\mathcal{T}$ passed to the Counting algorithm is such that resulting vectors U and V contain at least cn_0 elements each. Next, let $\gamma_1 = \frac{1}{|U|}\sum\limits_{u \in U} u$ and $\gamma_2 = \frac{1}{|V|}\sum\limits_{v \in V} v$, in this notation the test statistic is defined as $\gamma = \gamma_1 - \gamma_2$. Conditioned on the assignment A^*, we have:

$$\mathbb{E}[\gamma_1] = \frac{1}{|U|}\sum_{u \in U}\mathbb{E}[u] \geq \nu + \delta.$$

Indeed, for any arbitrary entry u of array U let (j, i_1, i_2) be corresponding paper and reviewers assigned to this paper in assignment A^*. Then the definition of the non-trivial power guarantees that

$$\begin{aligned}\mathbb{E}[u] &= \frac{1}{2}\left(\pi^{(\mathrm{sb})}_{i_1 j} - \pi^{(\mathrm{db})}_{i_2 j}\right) + \frac{1}{2}\left(\pi^{(\mathrm{sb})}_{i_2 j} - \pi^{(\mathrm{db})}_{i_1 j}\right) \\ &\geq \frac{1}{2}\left(q_{i_1 j} + \nu + \delta - q_{i_2 j}\right) + \frac{1}{2}\left(q_{i_2 j} + \nu + \delta - q_{i_1 j}\right) \\ &= \nu + \delta.\end{aligned}$$

Similarly,

$$\mathbb{E}[\gamma_2] = \frac{1}{|V|}\sum_{v \in V}\mathbb{E}[v] \leq \nu - \delta.$$

Applying Hoeffding's inequality we obtain:

$$\mathbb{P}\left[\gamma_1 - \gamma_2 < \delta\right] \leq \mathbb{P}\left[\gamma_1 - \gamma_2 < \mathbb{E}\left[\gamma_1 - \gamma_2\right] - \delta\right] \leq \exp\left(-\frac{\delta^2}{2\left(|V|^{-1} + |U|^{-1}\right)}\right) \leq \exp\left(-c\delta^2 n_0\right).$$

On the other hand, the threshold for acceptance is such that

$$\sqrt{2\left(|U|^{-1} + |V|^{-1}\right)\log 2/\alpha} \leq c\sqrt{\frac{1}{n_0}}.$$

Finally, setting $n_0 = c\frac{\log 1/\varepsilon}{\delta^2}$, we ensure that if $\min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\} > n_0$, then the COUNTING algorithm with probability at least $1-\varepsilon$ rejects the null for any matrices Π^{sb}, Π^{db} that satisfy the alternative hypothesis with margin δ.

A6.3 Proof of Theorem 3

Assume that the premises of Theorem 3 are satisfied, that is, there exist functions $g, h \in \mathcal{F}$ and values $0 \leq x_1 < x_2 \leq 1$ such that $g(x_1) < h(x_1)$ and $g(x_2) > h(x_2)$.

The high-level idea of the proof is to construct matrices Π^{db} and Π^{sb} which simultaneously satisfy the null hypothesis of Problem 2 specified by some function $f_0 \in \mathcal{F}$ and the alternative hypothesis of Problem 2 specified by another function $f_0' \in \mathcal{F}$ with margin $\delta > 0$. If such matrices exist, then there exist two instances of a bias testing problem — one with presence of bias and the other with absence of bias — such that the distributions of the reviewers' decisions for these two instances coincide. Hence, any test that uniformly controls for the Type-I error rate at the level α for every $f_0 \in \mathcal{F}$ must under the second instance have power upper bounded by α and thus violate the requirement of non-trivial power over the class of functions $\mathcal{F}$.

We begin with building a matrix Π^{db}. For any reviewer $i \in [m]$ and for any paper $j \in [n]$ we let

$$\pi_{ij}^{(\text{db})} = \begin{cases} x_1 & \text{if } w_j = 1 \\ x_2 & \text{if } w_j = -1. \end{cases}$$

Next, we define Π^{sb} as follows. For any reviewer $i \in [m]$ and for any paper $j \in [n]$

$$\pi_{ij}^{(\text{sb})} = h(\pi_{ij}^{(\text{db})}).$$

By construction matrices Π^{sb} and Π^{db} satisfy the null hypothesis specified by function $h \in \mathcal{F}$. On the other hand, notice that for each paper $j \in \mathcal{J}$ we have

$$\pi_{ij}^{(\text{sb})} = h(x_1) > g(x_1) = g(\pi_{ij}^{(\text{db})}),$$

and for each paper $j \in \overline{\mathcal{J}}$ we have

$$\pi_{ij}^{(\text{sb})} = h(x_2) < g(x_2) = g(\pi_{ij}^{(\text{db})}).$$

Hence, matrices Π^{db} and Π^{sb} also satisfy the alternative hypothesis specified by function g. Moreover, Π^{db} and Π^{sb} satisfy this alternative with margin $\delta = \min\{|h(x_1) - g(x_1)|, |h(x_2) - g(x_2)|\} > 0$. We now conclude the proof by noting that our construction holds for any choice of parameters λ, μ, n, m and hence the requirement of non-trivial power must be violated by any testing algorithm that controls for Type-I error at the level $\alpha \in (0, 1)$.

A6.4 Proof of Corollary 1

To prove Corollary 1, we consider any choice of parameters $\Delta \in (0, 0.5)$ and $\widetilde{\Delta} > 0$ and construct two functions f_0 and $\widetilde{f_0}$ together with two numbers $0 \leq x_1 < x_2 \leq 1$ such that

(i) Functions f_0 and $\widetilde{f}_0$ describe the behavior of reviewers under the absence of bias under the generalized linear and generalized logistic models respectively, that is, $f_0 \in \mathcal{F}_\Delta$, $\widetilde{f}_0 \in \widetilde{\mathcal{F}}_{\widetilde{\Delta}}$, where class $\mathcal{F}_\Delta$ is defined by equation (4.6) and class $\widetilde{\mathcal{F}}_{\widetilde{\Delta}}$ is specified in equation (4.9)

(ii) Values x_1 and x_2 are such that:

(a) One can select parameters $q_j^*, j \in [n]$, that fall under the definition of the generalized linear model such that matrix Π^{db} generated according to the equation (4.5a) satisfies the following equation:

$$\pi_{ij}^{(\text{db})} = \begin{cases} x_1 & \text{if } j \in \mathcal{J} \\ x_2 & \text{if } j \in \overline{\mathcal{J}}. \end{cases} \tag{4.33}$$

(b) One can select parameters $q_j^*, j \in [n]$, and β_0, β_1 that fall under the definition of the generalized logistic model such that matrix Π^{db} generated according to the equation (4.8a) satisfies the equation (4.33).

(iii) Functions f_0 and $\widetilde{f}_0$ are such that

$$\text{sign}\left(f_0(x_1) - \widetilde{f}_0(x_1)\right) \times \text{sign}\left(f_0(x_2) - \widetilde{f}_0(x_2)\right) = -1,$$

where $\text{sign}(\cdot)$ is the sign function. That is, at x_1 the function f_0 is strictly larger than $\widetilde{f}_0$ and at x_2 the function f_0 is strictly smaller than $\widetilde{f}_0$, or vice versa.

Assume for the moment that conditions (i)-(iii) are satisfied and consider the matrix Π^{db} whose entries are given by equation (4.33). Then one can select values of papers' representations $q_j, j \in [n]$, such that Π^{db} satisfies the model of DB reviewers in the generalized linear model (4.5a). Similarly, there exists another choice of papers' representations $q'_j, j \in [n]$, and parameters β_0, β_1, such that the same matrix Π^{db} satisfies the model of DB reviewers in the generalized logistic model (4.8a). Now define matrix Π_1^{sb} whose entries for each $(i, j) \in [m] \times [n]$ are given by:

$$\pi_{ij}^{(\text{sb})} = f_0(\pi_{ij}^{(\text{db})})$$

and matrix Π_2^{sb} whose entries for each $(i, j) \in [m] \times [n]$ are given by:

$$\pi_{ij}^{(\text{sb})} = \widetilde{f}_0(\pi_{ij}^{(\text{db})}).$$

Matrices $\Pi^{\text{db}}, \Pi_1^{\text{sb}}$ satisfy the null hypothesis under the generalized *linear* model specified by the function f_0. Moreover, condition (iii) ensures that they *simultaneously* satisfy the alternative hypothesis under the generalized *logistic* model specified by the function $\widetilde{f}_0$ with margin $\delta = \min\{|f_0(x_1) - \widetilde{f}_0(x_1)|, |f_0(x_2) - \widetilde{f}_0(x_2)|\}$. Hence, if the testing procedure ψ_2 (which has a non-trivial power under the generalized logistic model) is given decisions of SB and DB reviewers sampled according to the pair of matrices $\Pi^{\text{db}}, \Pi_1^{\text{sb}}$, then it will reject the null hypothesis with probability that goes to 1 as the minimum of $|\mathcal{J}|$ and $|\overline{\mathcal{J}}|$ grows. Finally, given that matrices Π^{db} and Π^{sb} solely determine the distribution of observed reviewers' decisions, our construction implies that under the generalized linear model procedure ψ_2 does not control for the Type-I error rate at any level $\alpha < 1$.

A similar argument applies to the pair of matrices $\Pi^{\text{db}}, \Pi_2^{\text{sb}}$, and it follows that under the generalized logistic model the procedure ψ_1 does not control for the Type-I error rate at any level $\alpha < 1$.

To conclude the proof it remains to find $x_1, x_2, f_0, \widetilde{f}_0$ that satisfy aforementioned conditions (i)-(iii). To this end, let us define quantities γ_1, γ_2:

$$\gamma_1 = \max\left\{\Delta, \left(1 + \exp\{\widetilde{\Delta} + \widetilde{\Delta}^2\}\right)^{-1}\right\}$$
$$\gamma_2 = \min\left\{1 - \Delta, \left(1 + \exp\{-\widetilde{\Delta} - \widetilde{\Delta}^2\}\right)^{-1}\right\}.$$

Notice that the value of γ_1 is by definition smaller than 0.5. Moreover, for each pair $(i,j) \in [m] \times [n]$ it gives a lower bound on the value $\pi_{ij}^{(\mathrm{db})}$ that can be generated from both the generalized linear (with parameter Δ) and generalized logistic (with parameter $\widetilde{\Delta}$) models. Likewise, the value of γ_2 is at least 0.5 and gives the corresponding upper bound. Hence, any values of x_1, x_2 such that $\gamma_1 < x_1 < x_2 < \gamma_2$ satisfy the condition (ii).

Next, find values $\nu \in (0, \Delta)$ and $\widetilde{\nu} \in (0, \widetilde{\Delta})$ such that for functions $h_\nu \in \mathcal{F}_\Delta$ and $g_{\widetilde{\nu}} \in \widetilde{\mathcal{F}}_{\widetilde{\Delta}}$ defined in equations (4.6) and (4.9) respectively the following equality holds:

$$h_\nu(0.5) = g_{\widetilde{\nu}}(0.5).$$

Observe that such values must exist because h_ν and $g_{\widetilde{\nu}}$ are continuous functions of ν and $\widetilde{\nu}$ respectively and

$$\lim_{\nu \to +0} h_\nu(0.5) = \lim_{\widetilde{\nu} \to +0} g_{\widetilde{\nu}}(0.5) = 0.5.$$

Consider now two possible cases:

Case 1. Functions h_ν and $g_{\widetilde{\nu}}$ are such that there exist two points $y \in (\gamma_1, 0.5)$ and $z \in (0.5, \gamma_2)$ for which the following equation holds:

$$\operatorname{sign}\Big(h_\nu(y) - g_{\widetilde{\nu}}(y)\Big) \times \operatorname{sign}\Big(h_\nu(z) - g_{\widetilde{\nu}}(z)\Big) = -1, \tag{4.34}$$

Observe that in this case conditions (i)-(iii) are satisfied by the choice $f_0 = h_\nu, \widetilde{f}_0 = g_{\widetilde{\nu}}, x_1 = y, x_2 = z$ and hence the result of the theorem follows.

Case 2. Functions h_ν and $g_{\widetilde{\nu}}$ are such that h_ν is a tangent line to $g_{\widetilde{\nu}}$ at 0.5. This case reduces to the Case 1 by setting $\nu' = \nu - \varepsilon$ for a sufficiently small $\varepsilon \in (0, \nu)$. Indeed, if ε is sufficiently small, then due to strict concavity and differentiability of the function $g_{\widetilde{\nu}}$, by shifting the tangent line down we ensure that there exist points $y = 0.5$ and $z \in (0.5, \gamma_2)$ such that

$$\operatorname{sign}\Big(h_{\nu'}(y) - g_{\widetilde{\nu}}(y)\Big) \times \operatorname{sign}\Big(h_{\nu'}(z) - g_{\widetilde{\nu}}(z)\Big) = -1.$$

Hence, we can satisfy conditions (i)-(iii) by setting $f_0 = h_{\nu'}, \widetilde{f}_0 = g_{\widetilde{\nu}}, x_1 = 0.5, x_2 = z$.

To conclude the proof, we notice that Cases 1 and 2 are complementary, because function $g_{\widetilde{\nu}}$ is differentiable and strictly concave on the interval (0, 1) and function h_ν is a linear function.

A6.5 Proofs of auxiliary results

In this section we give proofs for auxiliary results stated in appendix.

A6.5.1 Proof of Proposition 1

We prove Lemma 1 by straightforward verification. First, let Y_{ij} be generated from model (4.12b) and $X_{i'j}$ be generated from model (4.12a). Then

$$\mathbb{E}\left[Y_{ij} - X_{i'j}\right] = q_j + \beta_0^{(\mathrm{sb})} + \sum_{\ell \in [k]} \beta_\ell^{(\mathrm{sb})} w_j^{(\ell)} - q_j = \beta_0^{(\mathrm{sb})} + \sum_{\ell \in [k]} \beta_\ell^{(\mathrm{sb})} w_j^{(\ell)}.$$

Similarly, let Y_{ij} be generated from model (4.13b) and $X_{i'j}$ be generated from model (4.13a). Then

$$\begin{aligned}
\mathbb{E}\left[Y_{ij} | Y_{ij} \neq X_{i'j}\right] &= \frac{\pi_{ij}^{(\mathrm{sb})}(1 - \pi_{i'j}^{(\mathrm{db})})}{\pi_{ij}^{(\mathrm{sb})}(1 - \pi_{i'j}^{(\mathrm{db})}) + \pi_{i'j}^{(\mathrm{db})}(1 - \pi_{ij}^{(\mathrm{sb})})} \\
&= \left[1 + \frac{\pi_{i'j}^{(\mathrm{db})}(1 - \pi_{ij}^{(\mathrm{sb})})}{\pi_{ij}^{(\mathrm{sb})}(1 - \pi_{i'j}^{(\mathrm{db})})}\right]^{-1} \\
&= \left[1 + \exp\left\{-\left(\beta_0^{(\mathrm{sb})} + \sum_{\ell \in [k]} \beta_{\ell+1}^{(\mathrm{sb})} w_j^{(\ell)} - \beta_0^{(\mathrm{db})}\right)\right\}\right]^{-1},
\end{aligned}$$

and hence

$$\log \frac{\mathbb{E}\left[Y_{ij} \middle| \left(Y_{ij} \neq X_{i'j}\right)\right]}{1 - \mathbb{E}\left[Y_{ij} \middle| \left(Y_{ij} \neq X_{i'j}\right)\right]} = \beta_0^{(\mathrm{sb})} - \beta_0^{(\mathrm{db})} + \sum_{\ell \in [k]} \beta_{\ell+1}^{(\mathrm{sb})} w_j^{(\ell)}.$$

A6.5.2 Proof of Lemma 1

Consider any assignment of papers to SB reviewers that satisfy $(\lambda, \mu)-$constraint with $\lambda > \mu$. Then pick any subset of papers $\mathcal{P} \subseteq [n]$ and denote a set of SB reviewers who are assigned to at least one paper from $\mathcal{P}$ as $\mathcal{R}_{\mathrm{SB}}$. Then one can notice that

$$|\mathcal{R}_{\mathrm{SB}}| \geq \frac{\lambda |\mathcal{P}|}{\mu} \geq |\mathcal{P}|,$$

and hence by Hall's theorem there exists a matching that maps each paper to one reviewer such that each reviewer is matched to at most one paper. This matching is computed in Step 2 of Algorithm 1.

The same argument applies to DB reviewers and hence, joining these two matchings, the algorithm in Step 4 constructs a set of tuples $\mathcal{T}$ where for each paper $j \in [n]$ there exists a tuple that corresponds to this paper.

A6.5.3 Proof of Lemma 2

Consider any assignments of papers to SB and DB reviewers that satisfy $(\lambda, \mu)-$constranints. Let γ be a maximum integer that satisfies inequality

$$\gamma \leq \min\left\{ \frac{|\mathcal{J}|}{4\mu}, \frac{|\overline{\mathcal{J}}|}{4\mu} \right\}.$$

Without loss of generality, assume that $\gamma > 1$. Given that μ and λ are treated as constants and that we only need to proof the result for large enough $\min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\}$, we ignore the cases when $\min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\}$ is small.

Consider a graph G before the first iteration of Steps 2 - 4 of Algorithm 2. Each paper in this graph is connected to λ SB and λ DB reviewers such that each reviewer is connected to at most μ papers.

Now let (i_1, j, i_2) and (i_1', j', i_2') be triples found in the first iteration of the algorithm. These triples exists provided that $\gamma > 1$. Then in Step 4 we remove reviewers i_1, i_1', i_2, i_2' and corresponding edges from graph G. One can see that these reviewers are connected to at most 4μ papers in total and hence before the second iteration of Steps 2 - 4 graph G will have at least $|\mathcal{J}| - 4\mu \geq 4\mu(\gamma - 1)$ papers from $\mathcal{J}$ and $|\overline{\mathcal{J}}| - 4\mu \geq 4\mu(\gamma - 1)$ papers from $\overline{\mathcal{J}}$ that are connected to λ SB and λ DB remaining reviewers and each of the remaining reviewers (there must be at least $8\lambda(\gamma - 1)$ SB and $8\lambda(\gamma - 1)$ DB reviewers) will be connected to at most μ papers.

By induction we can show that in the first γ iterations of Steps 2 - 4 the greedy algorithm will be able to find non-empty triples in Steps 2 and 3. Hence the resulting set of tuples $\mathcal{T}$ will contain at least γ tuples that correspond to papers from $\mathcal{J}$ and at least γ tuples that correspond to papers from $\overline{\mathcal{J}}$. We then conclude the proof noticing that $\gamma = c \min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\}$, where c is a constant that depends only on μ.

A6.5.4 Proof of Proposition 2

The proof of Proposition 2 follows the idea of the proof of Theorem 1 with some changes which we now discuss. Consider any set of triples $\mathcal{C}$ such that (i) each triple $c \in \mathcal{C}$ is of the form (j, i_1, i_2) (one paper and two reviewers) and (ii) each reviewer $i \in [m]$ appears in at most one triple. Let $\mathbb{C}$ denote a collection of all such sets of triples. Then any set of tuples $\mathcal{T}$ passed to the Disagreement or Counting tests as input corresponds to one member of $\mathbb{C}$ which is constructed as follows: for each $t \in \mathcal{T}$ let (j_t, i_t, i_t') be a corresponding paper, SB reviewer and DB reviewer assigned to this paper, then $\mathcal{C} = \bigcup_{t \in \mathcal{T}} (j_t, i_t, i_t')$. Conversely, each member $\mathcal{C} \in \mathbb{C}$ gives rise to a family of sets of tuples $\mathbb{T}(\mathcal{C})$ which contains $2^{|\mathcal{C}|}$ elements and each element

corresponds to a different allocation of reviewers in each triple $(j, i_1, i_2) \in \mathcal{C}$ to SB and DB conditions. For example, let $\mathcal{C} = \{(j, i_1, i_2), (j', i_1', i_2')\}$, then the family $\mathbb{T}(\mathcal{C})$ consists of four sets of tuples:

$$\mathcal{T}_1 = \{(j, Y_{i_1 j}, X_{i_2 j}, w_j), (j', Y_{i_1 j'}, X_{i_2 j'}, w_{j'})\}$$
$$\mathcal{T}_2 = \{(j, Y_{i_2 j}, X_{i_1 j}, w_j), (j', Y_{i_1 j'}, X_{i_2 j'}, w_{j'})\}$$
$$\mathcal{T}_3 = \{(j, Y_{i_1 j}, X_{i_2 j}, w_j), (j', Y_{i_2 j'}, X_{i_1 j'}, w_{j'})\}$$
$$\mathcal{T}_4 = \{(j, Y_{i_2 j}, X_{i_1 j}, w_j), (j', Y_{i_2 j'}, X_{i_1 j'}, w_{j'})\}$$

Next, for concreteness assume that $\lambda \geq \mu$, that is, Algorithm 1 is used to construct a set $\mathcal{T}$. Then conditioned on the fact that the set of tuples $\mathcal{T}$ constructed by the algorithm belongs to $\mathbb{T}(\mathcal{C})$, the randomness of the allocation of reviewers to conditions, the random assignment procedure used to assign reviewers to papers in each condition and randomness in the tie-breaking in the matching algorithm ensure that $\mathcal{T} \in \mathcal{U}[\mathbb{T}(\mathcal{C})]$, that is, all elements of $\mathbb{T}(\mathcal{C})$ are equally likely to be constructed and no other set of tuples can be constructed.

For each member $\mathcal{C} \in \mathbb{C}$, let $\mathbb{P}[\mathcal{C}]$ be probability that Algorithm 1 constructs a set of tuples that belongs to $\mathbb{T}(\mathcal{C})$. Notice that for some $\mathcal{C} \in \mathbb{C}$ we have $\mathbb{P}[\mathcal{C}] = 0$ which happens for example when $|\mathcal{C}| < n$, because Lemma 1 ensures that $|\mathcal{T}| = n$. Now, conditioning on any set $\mathcal{C}$ with $\mathbb{P}[\mathcal{C}] > 0$ (instead of conditioning on A^*) and using Lemma 1 (instead of Observation B), we repeat the proof of Theorem 1 for both DISAGREEMENT and COUNTING tests. The unconditional result then follows from the law of total probability. The same argument applies to the case when $\lambda < \mu$ and hence we conclude the proof.

A6.5.5 Proof of Corollary 2

The high-level idea of the proof is to construct matrices Π^{db} and Π^{sb} that simultaneously (for different choices of $\beta_0^{(\text{sb})}$ and $\beta_1^{(\text{sb})}$ coefficients) satisfy the null and the alternative hypotheses under the extended model given by equations (4.15a) and (4.15b).

We begin our construction from specifying values of $q_j, j \in [n]$. For each paper $j \in [n]$, let

$$q_j = \begin{cases} -1 & \text{if } \ w_j = 1 \\ 0 & \text{if } \ w_j = -1. \end{cases}$$

Then Π^{db} is generated from model (4.15a) with $\beta_0^{(\text{db})} = 0$ and $\beta_1^{(\text{db})} = 1$. In this way, for any reviewer $i \in [m]$ and for any paper $j \in [n]$, probability of acceprance $\pi_{ij}^{(\text{db})}$ satisfies:

$$M_0: \ \log \frac{\pi_{ij}^{(\text{db})}}{1 - \pi_{ij}^{(\text{db})}} = q_j.$$

That is, for any reviewer $i \in [m]$ and for any paper $j \in [n]$ we have

$$\pi_{ij}^{(\text{db})} = \begin{cases} \frac{1}{1+e} & \text{if } w_j = 1 \\ 0.5 & \text{if } w_j = -1. \end{cases}$$

We now consider two different choices of coefficients for SB reviewers which result into two different models of behaviour of SB reviewers under the absence of bias:

$$M_1 \ (\beta_0^{(\text{sb})} = 1, \beta_1^{(\text{sb})} = 1): \ \ \log \frac{\pi_{ij}^{(\text{sb})}}{1 - \pi_{ij}^{(\text{sb})}} = 1 + q_j$$

$$M_2 \ (\beta_0^{(\text{sb})} = 3/2, \beta_1^{(\text{sb})} = 2): \ \ \log \frac{\pi_{ij}^{(\text{sb})}}{1 - \pi_{ij}^{(\text{sb})}} = \frac{3}{2} + 2q_j$$

Consider a matrix Π^{sb} whose components for each $i \in [m]$ and $j \in [n]$ are defined as follows:

$$\pi_{ij}^{(\text{sb})} = \begin{cases} 0.5 & \text{if } w_j = 1 \\ \frac{1}{1+e^{-1}} & \text{if } w_j = -1, \end{cases}$$

it is not hard to see that (i) entries of matrix Π^{sb} satisfy the model M_1 and (ii) for each paper $j \in \mathcal{J}$ corresponding entries of matrix Π^{sb} are larger than prescribed by model M_2 by $\delta > 0$ and for each paper $j \in \overline{\mathcal{J}}$ corresponding entries are smaller than those prescribed by M_2 by $\delta > 0$, where δ is some universal constant. Hence, depending on which model of SB reviewer under the absence of bias (M_1 or M_2) is correct, pair of matrices $(\Pi^{\text{db}}, \Pi^{\text{sb}})$ corresponds to the absence or presence of bias.

Given that matrices Π^{db} and Π^{sb} solely determine a distribution of reviewers' decisions, we have shown that reviewers' decisions are identically distributed under both null and alternative hypotheses under the extended version of the generalized logistic model. Hence, we conclude the proof by declaring that any algorithm that operates on reviewers' decision and keeps Type-I error below α must have power at most α under the alternative specified by models M_0, M_2 and matrices $\Pi^{\text{sb}}, \Pi^{\text{db}}$ for all values of $\min\{|\mathcal{J}|, |\overline{\mathcal{J}}|\}$ and hence violates the non-trivial power requirement.

Chapter 5

Identity-Related Biases in Double-Blind Peer Review

1 Introduction

While some venues still show author identities to reviewers and debate whether they should move to the double-blind review process or not, many venues have already implemented this change. Ideally, in a double-blind review process, neither the authors nor the reviewers of any papers are aware of each others' identity. However, a challenge for ensuring that reviews are truly double-blind is the exponential growth in the trend of posting papers online before review (Xie et al., 2021). Increasingly, authors post their preprints on online publishing websites such as arXiv and SSRN and publicize their work on social media platforms such as Twitter. The conventional publication route via peer review is infamously long and time-consuming. On the other hand, online preprint-publishing venues provide a platform for sharing research with the community usually without delays. Not only does this help science move ahead faster, but it also helps researchers avoid being "scooped". However, the increase in popularity of making papers publicly available—with author identities—before or during the review process, has led to the dilution of double-blinding in peer review. For instance, the American Economic Association, the flagship journal in economics, dropped double-blinding in their reviewing process citing its limited effectiveness in maintaining anonymity. The availability of preprints online presents a challenge in double-blind reviewing, which could lead to biased evaluations for papers based on their authors' identities, similar to single-blind reviewing.

This dilution has led several double-blind peer-review venues to debate whether authors should be allowed to post their submissions on the Internet, before or during the review process. For instance, top-tier machine learning conferences such as NeurIPS and ICML do not prohibit posting online. On the other hand, the Association of Computational Linguistics (ACL) recently introduced a policy for its conferences in which authors are prohibited from posting their papers on the Internet starting a month before the paper submission deadline till the end of the review process. The Conference on Computer Vision and Pattern Recognition (CVPR) has banned the advertisement of submissions on social media platforms for such a time period. Some venues are stricter, for example, the IEEE Communication Letters and IEEE International Conference on Computer Communications (INFOCOMM) disallows posting preprints to online publishing venues before acceptance.

Independently, authors who perceive they may be at a disadvantage in the review process if their identity is revealed face a dilemma regarding posting their work online. They stand to either hurt their paper's chances of acceptance by revealing their identity online or lose out on publicity for their paper by refraining from posting.

It is thus important to quantify the consequences of posting preprints online to (i) enable an evidence-based debate over conference policies, and (ii) help authors make informed decisions about posting preprints online. In this chapter, we conduct a large-scale survey-based study in conjunction with the review process of two

top-tier publication venues in computer science that have double-blind reviewing: the 2021 International Conference on Machine Learning (ICML 2021) and the 2021 ACM Conference on Economics and Computation (EC 2021).[1] Specifically, we design and conduct experiments aimed at answering the following research questions:

(Q1) What fraction of reviewers, who had not seen the paper they were reviewing before the review process, deliberately search for the paper on the Internet during the review process?

(Q2) What is the relation between the rank of the authors' affiliations and the visibility of a preprint to its target audience?

By addressing these research questions, we aim to measure some of the effects of posting preprints online, and help quantify their associated risks and benefits for authors from different institutions.

2 Related work

Surveys of reviewers. Several studies survey reviewers to obtain insights into reviewer perceptions and practices. Nobarany et al. (2016) surveyed reviewers in the field of human-computer interaction to gain a better understanding of their motivations for reviewing. They found that encouraging high-quality research, giving back to the research community, and finding out about new research were the top general motivations for reviewing. Along similar lines, Tite and Schroter (2007) surveyed reviewers in biomedical journals to understand why peer reviewers decline to review. Among the respondents, they found the most important factor to be conflict with other workload.

Resnik et al. (2008) conducted an anonymous survey of researchers at a government research institution concerning their perceptions about ethical problems with journal peer review. They found that the most common ethical problem experienced by the respondents was incompetent review. Additionally, 6.8% respondents mentioned that a reviewer breached the confidentiality of their article without permission. This survey focused on the respondents' perception, and not on the actual frequency of breach of confidentiality. In another survey, by Martinson et al. (2005), 4.7% authors self-reported publishing the same data or results in more than one publication. Fanelli (2009) provides a systematic review and meta analysis of surveys on scientific misconduct including falsification and fabrication of data and other questionable research practices.

Goues et al. (2018) surveyed reviewers in three double-blind conferences to investigate the effectiveness of anonymization of submitted papers. In their experiment, reviewers were asked to guess the authors of the papers assigned to them. Out of all reviews, 70%-86% of the reviews did not have any author guess. Here, absence of a guess could imply that the reviewer did not have a guess or they did not wish to answer the question. Among the reviews containing guesses, 72%-85% guessed at least one author correctly.

Analyzing papers posted versus not posted on arXiv. Bharadhwaj et al. (2020) aim to analyse the risk of selective de-anonymization through an observational study based on open review data from the International Conference on Learning Representations (ICLR). The analysis quantifies the risk of de-anonymization by computing the correlation between papers' acceptance rates and their authors' reputations separately for papers posted and not posted online during the review process. This approach however is hindered by the confounder that the outcomes of the analysis may not necessarily be due to de-anonymization of papers posted on arXiv, but could be a result of higher quality papers being selectively posted on arXiv by famous authors. Moreover, it is not clear how the paper draws conclusions based on the analysis presented therein. Our supporting analysis overlaps with the investigation of Bharadhwaj et al. (2020): we also investigate the correlation between papers' acceptance rates and their authors' associated ranking in order to support our main analysis and to account for confounding by selective posting by higher-ranked authors.

Aman (2014) also investigate possible benefits of publishing preprints on arXiv in *Quantitative Biology*, wherein they measure and compare the citations received by papers posted on arXiv and those received by

[1] In Computer Science, conferences are typically the terminal publication venue and are typically ranked at par or higher than journals. Full papers are reviewed in CS conferences, and their publication has archival value.

papers not posted on arXiv. A similar confounder arises here that a positive result could be a false alarm due to higher quality papers being selectively posted on arXiv by authors.

In our work, we quantify the risk of de-anonymization by directly studying reviewer behaviour regarding searching online for their assigned papers. We quantify the effects of publishing preprints online by measuring their visibility using a survey-based experiment querying reviewers whether they had seen a paper before.

Studies on peer review in computer science. Our study is conducted in two top-tier computer science conferences and contributes to a growing list of studies on peer review in computer science. Lawrence and Cortes (2014); Beygelzimer et al. (2021) quantify the (in)consistencies of acceptance decisions on papers. Several studies (Madden and DeWitt, 2006; Tung, 2006; Tomkins et al., 2017; Manzoor and Shah, 2020) study biases due to single-blind reviewing. Shah et al. (2018) study several aspects of the NeurIPS 2016 peer-review process. Stelmakh et al. (2021d) study biases arising if reviewers know that a paper was previously rejected. Stelmakh et al. (2021c) study a pipeline for getting new reviewers into the review pool. Stelmakh et al. (2020) study herding in discussions. A number of recent works (Charlin and Zemel, 2013; Stelmakh et al., 2021a; Kobren et al., 2019; Jecmen et al., 2020; Noothigattu et al., 2020) have designed algorithms that are used in the peer-review process of various computer science conferences. See Shah (2022) for an overview of such studies and computational tools to improve peer review.

3 Methods

We now outline the design of the experiment that we conducted to investigate the research questions in this work. First, in Section 3.1 we introduce the two computer science conferences ICML 2021 and EC 2021 that formed the venues for our investigation, and describe research questions Q1 and Q2 in the context of these two conferences. Second, in Section 3.2 we describe the experimental procedure. Finally, in Section 3.3 we provide the details of our analysis methods.

3.1 Preliminaries

Experiment setting The study was conducted in the peer-review process of two conferences:

- **ICML 2021** International Conference on Machine Learning is a flagship machine learning conference. ICML is a large conference with 5361 submissions and 4699 reviewers in its 2021 edition.
- **EC 2021** ACM Conference on Economics and Computation is the top conference at the intersection of Computer Science and Economics. EC is a relatively smaller conference with 498 submissions and 190 reviewers in its 2021 edition.

Importantly, the peer-review process in both conferences, ICML and EC, is organized in a double-blind manner, defined as follows. In a **double-blind peer-review process**, the identity of all the authors is removed from the submitted papers. No part of the authors' identity, including their names, affiliations, and seniority, is available to the reviewers through the review process. At the same time, no part of the reviewers' identity is made available to the authors through the review process.

We now formally define some terminology used in the research questions Q1 and Q2. The first research question, Q1, focuses on the fraction of reviewers who deliberately search for their assigned paper on the Internet. The second research question, Q2, focuses on the correlation between the visibility to a target audience of papers available on the Internet before the review process, and the rank of the authors' affiliations. In what follows, we explicitly define the terms used in Q2 in the context of our experiments—target audience, visibility, preprint, and rank associated with a paper.

Paper's target audience For any paper, we define its target audience as members of the research community that share similar research interests as that of the paper. In each conference, a 'similarity score' is computed between each paper-reviewer pair, which is then used to assign papers to reviewers. We used the same similarity score to determine the target audience of a paper (among the set of reviewers in the conference). We provide more details in Appendix A1.

Paper's visibility We define the visibility of a paper to a member of its target audience as a binary variable which is 1 if that person has seen this paper outside of reviewing contexts, and 0 otherwise. Visibility, as defined here, includes reviewers becoming aware of a paper through preprint servers or other platforms such as social media, research seminars and workshops. On the other hand, visibility does *not* include reviewers finding a paper during the review process (e.g., visibility does not include a reviewer discovering an assigned paper by deliberate search or accidentally while searching for references).

Preprint To study the visibility of papers released on the Internet before publication, we checked whether each of the papers submitted to the conference was available online. Specifically, for EC, we manually searched for all submitted papers to establish their presence online. On the other hand, for ICML, owing to its large size, we checked whether a submitted paper was available on arXiv (`arxiv.org`). ArXiv is the predominant platform for pre-prints in machine learning; hence we used availability on arXiv as a proxy indicator of a paper's availability on the Internet.

Rank associated with a paper In this work, the rank of an author's affiliation is a measure of author's prestige that, in turn, is transferred to the author's paper. We determine the rank of affiliations in ICML and EC based on widely available rankings of institutions in the respective research communities. Specifically, in ICML, we rank (with ties) each institution based on the number of papers published in the ICML conference in the preceding year (2020) with at least one author from that institution (Ivanov, 2020). On the other hand, since EC is at the intersection of two fields, economics and computation, we merge three rankings—the QS ranking for computer science (QS, 2021a), the QS ranking for economics and econometrics (QS, 2021b), and the CS ranking for economics and computation (CSRankings, 2021)—by taking the best available rank for each institution to get our ranking of institutions submitting to EC. By convention, better ranks, representing more renowned institutions, are represented by lower numbers; the top-ranked institution for each conference has rank 1. Finally, we define the rank of a paper as the rank of the best-ranked affiliation among the authors of that paper. Due to ties in rankings, we have 37 unique rank values across all the papers in ICML 2021, and 66 unique rank values across all the papers in EC 2021.

3.2 Experiment design

To address Q1 and Q2, we designed survey-based experiments for EC 2021 and ICML 2021, described next.

Design for Q1 To find the fraction of reviewers that deliberately search for their assigned paper on the Internet, we surveyed the reviewers. Importantly, as reviewers may not be comfortable answering questions about deliberately breaking the double-blindness of the review process, we designed the survey to be anonymous. We used the Condorcet Internet Voting Service (CIVS) (Myers, 2003), a widely used service to conduct secure and anonymous surveys. Further, we took some steps to prevent our survey from spurious responses (e.g., multiple responses from the same reviewer). For this, in EC, we generated a unique link for each reviewer that accepted only one response. In ICML we generated a link that allowed only one response per IP address and shared it with reviewers asking them to avoid sharing this link with anyone.[2] The survey form was sent out to the reviewers via CIVS after the initial reviews were submitted. In the e-mail, the reviewers were invited to participate in a one-question survey on the consequences of publishing preprints online. The survey form contained the following question:

> "During the review process, did you search for any of your assigned papers on the Internet?"

with two possible options: *Yes* and *No*. The respondents had to choose exactly one of the two options. To ensure that the survey focused on reviewers deliberately searching for their assigned papers, right after the question text, we provided additional text: "Accidental discovery of a paper on the Internet (e.g., through

[2]The difference in procedures between EC and ICML is due to a change in the CIVS policy that was implemented between the two surveys.

searching for related works) does not count as a positive case for this question. Answer *Yes* only if you tried to find an assigned paper itself on the Internet."

Following the conclusion of the survey, CIVS combined the individual responses, while maintaining anonymity, and provided the total number of *Yes* and *No* responses received.

Design for Q2 Recall that for Q2 we want to find the correlation between preprints' visibility to a target audience and its associated rank. Following the definitions provided in Section 3.1, we designed a survey-based experiment as follows. We conducted a survey to query reviewers about some papers for which they are considered a target audience. Specifically, we asked reviewers if they had seen these papers before outside of reviewing contexts. We provide more details about the survey, including the phrasing of the survey question, in Appendix A1. We queried multiple reviewers about each paper, and depending on their response, we considered the corresponding visibility to be 1 if the reviewer said they had seen the paper before outside of reviewing contexts and 0 otherwise. We note that in ICML reviewers were queried about the papers they were assigned to review using the reviewer response form, in which a response to the question of visibility was required. Meanwhile, in EC, reviewers were queried about a set of papers that they were not assigned to review, using a separate optional survey form that was emailed to them by the program chairs after the rebuttal phase and before the announcement of paper decisions. The survey designed for Q2 had a response rate of 100% in ICML, while EC had a response rate of 55.78%.

3.3 Analysis

We now describe the analysis for the data collected to address Q1 and Q2. Importantly, our analysis is the same for the data collected from ICML 2021 and EC 2021. For Q1, we directly report the numbers obtained from CIVS regarding the fraction of reviewers who searched for their assigned papers online in the respective conference. In this section, we describe our analysis for Q2, where we want to analyse the effect of papers' ranking on visibility. Recall that for Q2, we collected survey responses and observational data about which papers submitted to ICML or EC were posted online before the corresponding review process. Since the latter data is observational, we describe two possible confounding factors in our setting.

3.3.1 Confounding factors

For a paper posted online, the amount of time for which it has been available on the Internet can affect the visibility of the paper. For instance, papers posted online well before the deadline may have higher visibility as compared to papers posted near the deadline. Moreover, the time of posting a paper online could vary across institutions ranked differently. Thus, time of posting can be a confounding factor. In order to control for this factor, for the papers that were posted online before the review process, we incorporate the time gap between posting of the papers and submission of reviews in our analysis, which is described next in Section 3.3.2.

Second, one should ideally study the visibility of all papers submitted to the conference. However, in our experiment, we are naturally limited to studying the visibility of the papers that were released on the internet before the conclusion of the review process. The choice of posting online before or after the review process could vary depending on the quality of the paper as well as on the rank of the authors' affiliations. This can form a confounding factor in our analysis. To understand whether papers posted online before the review process had significantly different quality and rank profile from papers not posted online, we provide supporting analysis in Section 3.3.3.

3.3.2 Analysis procedure

We now describe our analysis to compute the relation between a paper's visibility and associated rank. In the following analysis procedure, we consider each response obtained in the survey for Q2 as one unit. Each response corresponds to a paper-reviewer pair, wherein the reviewer was queried about seeing the considered paper. In case of no response from reviewer, we do not consider the corresponding paper-reviewer pairs in

our data. We thus have two variables associated to each response: the visibility of the paper to the reviewer (in $\{0, 1\}$), and the rank associated with the paper. Recall that we define the rank of a paper as the rank of the best-ranked affiliation associated with that paper.

We first describe the approach to control for confounding due to time of posting. There is ample variation in the time of posting papers online within the papers submitted to ICML and EC: some papers were posted right before the review process began while some papers were posted two years prior. To account for the causal effect of time of posting on visibility, we divide the responses into bins based on the number of days between the paper being posted online and the deadline for submitting responses to the Q2 survey. Since similar conference deadlines arrive every three months roughly and the same conference appears every one year, we binned the responses accordingly into three bins. Specifically, if the number of days between the paper being posted online and the survey response is less than 90, it is assigned to the first bin, if the number of days is between 90 and 365, the response is assigned to the second bin, and otherwise, the response is assigned to the third bin. Following this binning, we assume that time of posting does not affect the visibility of papers within the same bin. Consequently, we analyse the correlation between papers' visibility and associated rank separately within each bin and then combine them to get the overall effect in two steps:

Step 1. We compute the correlation coefficient between papers' visibility and associated rank within each bin. For this we use Kendall's Tau-b statistic, which is closely related to the widely used Kendall's Tau rank correlation coefficient (Kendall, 1938). Kendall's Tau statistic provides a measure of the strength and direction of association between two variables measured on an ordinal scale. It is a non-parametric measure that does not make any assumptions about the data. However, it does not account for ties and our data has a considerable number of ties, since visibility is a binary variable and the rankings used contain ties. Therefore, we use a variant of the statistic, Kendall's Tau-b statistic, that accounts for ties in the data.

Within each bin we consider all the responses obtained and their corresponding visibility and rank value, and compute Kendall's Tau-b correlation coefficient between visibility and rank. The procedure for computing Kendall's Tau-b correlation coefficient between two real-valued vectors (of the same length) is described in Appendix A2.1. We now make a brief remark of a notational convention we use in this chapter, in order to address ambiguity between the terminology "high-rank institutions" as well as "rank 1, 2,... institutions", both of which colloquially refers to better-rank institutions. It is intuitive to interpret a positive correlation between visibility and rank as the visibility increasing with an *improvement* in the rank. Consequently, we flip the sign of all correlation coefficients computed with respect to the rank variable.

Step 2. With the correlation computed within each bin, we compute the overall correlation using a sample-weighted average (Corey et al., 1998). Formally, let N_1, N_2 and N_3 denote the number of responses obtained in the first, second and third bin respectively. Denote Kendall's Tau-b correlation coefficients within the three bins as τ_1, τ_2 and τ_3. Then the correlation T between papers' visibility and rank over all the time bins is computed as

$$T = \frac{N_1\,\tau_1 + N_2\,\tau_2 + N_3\,\tau_3}{N_1 + N_2 + N_3}. \tag{5.1}$$

The statistic T gives us the effect size for our research question Q2. Finally, to analyse the statistical significance of the effect, we conduct a permutation test, wherein we permute our data within each bin and recompute the test statistic T to obtain a p-value for our test. We provide the complete algorithm for the permutation test in Appendix A2.2.

3.3.3 Supporting analysis

As mentioned earlier in Section 3.3.1, we have to account for the papers not posted online before the survey for Q2 was conducted, to have a complete understanding of the effect of rank on paper visibility. We are unable to measure the visibility of these papers, as they were not posted online yet. Thus, we investigate whether the pool of papers posted online before the review process is significantly different, in terms of their rank profile, from the rest of the papers submitted to the conference. In the following analysis, we consider all papers submitted to the conference, and we consider each submitted paper as one unit of analysis.

		EC 2021	ICML 2021
1	# Reviewers	190	4699
2	# Survey Respondents	97	753
3	# Survey respondents who said they searched for their assigned paper online	41	269
4	% Survey respondents who said they searched for their assigned paper online	42%	36%

Table 1: Outcome of survey for research question Q1.

First, we analyse the relationship between a binary value indicating whether a submitted paper was posted online before the Q2 survey, and the paper's associated rank. For this, we compute Kendall's Tau-b statistic between the two values for all papers submitted to the conference, and flip the sign of the statistic with respect to the rank variable.

Second, we investigate whether there is a significant difference between the papers posted and not posted online before the review process, in terms of their quality and rank profile. Here we measure the quality of a paper as a binary variable based on its final decision in the conference (`accept` or `reject`). We give an example to understand the motivation for this supporting analysis. Suppose the double-blind process works perfectly, and assume that among the papers with better-ranked affiliations, only the high-quality papers are posted online, while there is no quality-based selection among other papers. Then, for better-ranked papers, the difference in acceptance rates for papers posted and those not posted online would be higher than the difference for the other papers. Suppose there is no causal effect of ranking on papers' visibility. Assuming higher-quality papers would enjoy higher visibility, such self-selection could lead to false discovery of effect.

We conduct this analysis by computing three statistics. First, for all papers posted online before the Q2 survey, we compute Kendall's Tau-b statistic between their rank and their final decision. Second, for all papers *not* posted online, we compute Kendall's Tau-b statistic between their rank and their final decision. Third, for each unique rank value, for the corresponding papers with that rank, we compute the difference between the average acceptance rate for papers posted online and those not posted online. Then, we compute Kendall's Tau-b statistic between the rankings and the difference in acceptance rate. Finally, we flip the sign of all correlation coefficients computed with respect to the rank variable. Hence, a positive correlation would imply that the (difference in) acceptance rate increases as the rank improves.

4 Main results

We now discuss the results from the experiments conducted in ICML 2021 and EC 2021.

4.1 Q1 results

Table 1 provides the results of the survey for research question Q1. The percentage of reviewers that responded to the anonymous survey for Q1 is 16% (753 out of 4699) in ICML and 51% (97 out of 190) in EC. While the coverage of the pool of reviewers is small in ICML (16%), the number of responses obtained is large (753). As shown in Table 1, the main observation is that, in both conferences, at least a third of the Q1 survey respondents self-report deliberately searching for their assigned paper on the Internet. There is substantial difference between ICML and EC in terms of the response rate as well as the fraction of *Yes* responses received, however, the current data cannot provide explanations for these differences.

4.2 Q2 results

We discuss the results of the survey conducted for Q2 in ICML 2021 and EC 2021. First we discuss the results of the main analysis described in Section 3.3.2. Then we discuss the results of the supporting analysis described in Section 3.3.3.

Main analysis Table 2 depicts the results of the survey for research question Q2. We received 7594 responses and 449 responses for the survey for Q2 in ICML and EC respectively (Row 1). Based on our binning rule based on time of posting described in Section 3.3.2, we see more papers in bin 1 and bin 2, compared to bin 3. This suggests that majority of preprints were posted online within one year of the review process (Row 2).

As shown in Table 2, for papers submitted to the respective conference and posted online before the review process, we observe that there is a weak positive correlation between the papers' visibility and its associated rank. The weak positive correlation implies that the visibility increases slightly as the rank improves.

To provide some interpretation of the correlation coefficient values in Row 4, we compare the mean visibility within and without responses obtained for papers with at least one affiliation ranked 10 or better (Row 8 and 9). There are 10 and 23 institutions among the top-10 ranks in ICML and EC respectively. We see that there is more than 3 percentage points decrease in mean visibility across these two sets of responses in both ICML and EC. Figure 1 displays additional visualization that helps to interpret the strength of the effect of the papers' rank on visibility. The data suggests that top-ranked institutions enjoy higher visibility than lower-ranked institutions in both venues ICML 2021 and EC 2021.

In summary, in ICML the analysis supports a small but statistically significant effect of paper ranking on its visibility. In EC the effect size is comparable, but the effect does not reach statistical significance. Without further data, for EC the results are only suggestive.

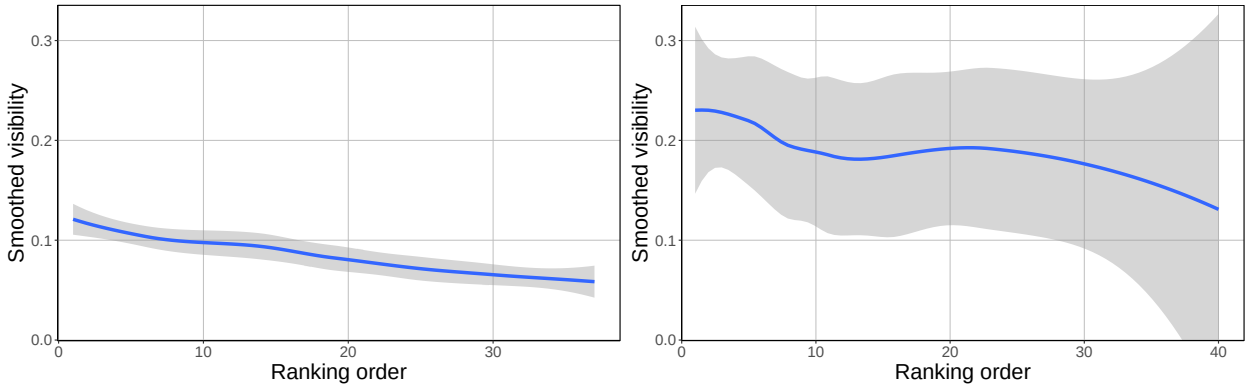
As an aside, we note that the mean visibility in ICML 2021 (8.36%) is much lower than that in EC 2021 (20.5%). This may be attributed to the following reason: The research community in EC is smaller and more tight-knit, meaning that there is higher overlap in research interests within the members of the community (reviewers). On the other hand, ICML is a large publication venue with a more diverse and spread-out research community.

Supporting analysis We provide the results for the supporting analysis described in Section 3.3.3 in Table 3. There were a total of 5361 and 498 papers submitted to ICML 2021 and EC 2021 respectively, out of which 1934 and 183 were posted online before the end of the review process respectively (Row 1 and 2). Thus, we see that more than a third of the papers submitted were available online. Among all the papers submitted, we observe that there is a positive correlation (Kendall's Tau-b) between paper's rank and whether it was posted online before the review process in both ICML and EC of 0.12 ($p < 10^{-5}$) and 0.09 ($p = 0.01$) respectively (Row 3). This implies that the authors from higher-ranked institutions are more likely to post their papers online before the review process. In Figure 2 we provide visualization to interpret the correlation between ranking and uploading behaviour.

Next, to understand if there is significant difference in the quality of papers uploaded online by authors from institutions with different ranks, we compare the final decision of the pool of papers posted online before

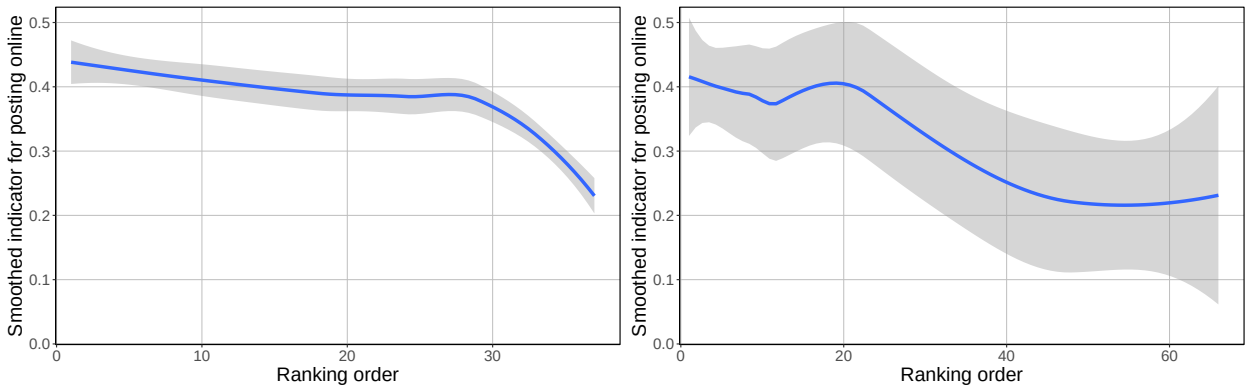
		EC 2021	ICML 2021
1	# Responses overall	449	7594
2	# papers in bins 1, 2, 3	$63, 82, 38$	$968, 820, 146$
3	# Responses in bins 1, 2, 3	$159, 233, 57$	$3799, 3228, 567$
4	Correlation between rank and visibility $[-1, 1]$	0.05 ($p = 0.11$)	0.06 ($p < 10^{-5}$)
5	Correlation between rank and visibility in bins 1, 2, 3	$0.06, 0.04, 0.04$	$0.04, 0.10, 0.03$
6	p-value associated with correlations in Row 5	$0.36, 0.46, 0.66$	$0.004, < 10^{-5}, 0.19$
7	% Visibility overall $[0 - 100]$	20.5% (92 out of 449)	8.36% (635 out of 7594)
8	% Visibility for papers with top 10 ranks $[0 - 100]$	21.93% (59 out of 269)	10.91% (253 out of 2319)
9	% Visibility for papers below top 10 ranks $[0 - 100]$	18.33% (33 out of 180)	7.24% (382 out of 5275)

Table 2: Outcome of main analysis for research question Q2. A positive correlation in Row 4 and Row 5 implies that the visibility increases as the rank of the paper improves. Recall that for ICML, we consider the set of responses obtained for submissions that were available as preprints on arXiv. There were 1934 such submissions.



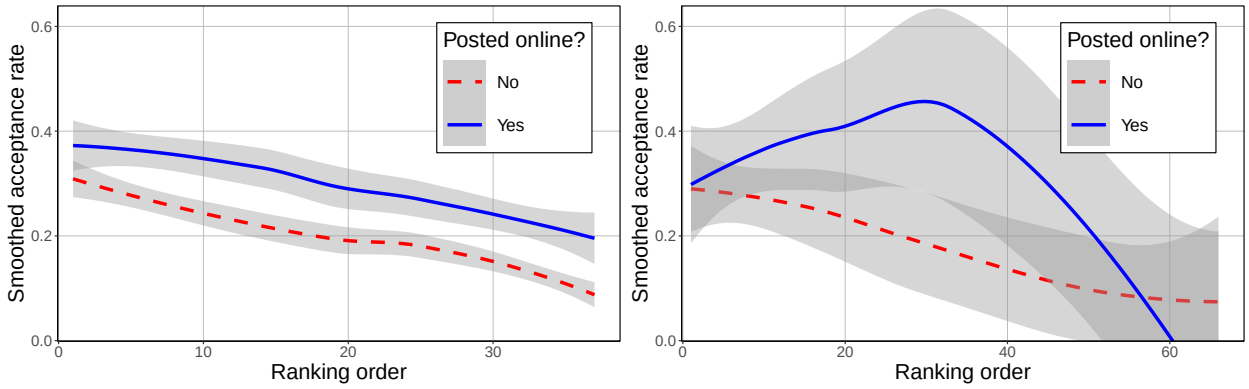
(a) ICML 2021 (b) EC 2021

Figure 1: Using responses obtained in Q2 survey, we plot the papers' visibility against papers' associated rank with smoothing. On the x-axis, we order papers by their ranks (i.e., paper with the best rank gets order 1, paper with the second best rank gets order 2, and so on). The range of x-axis is given by the number of unique ranks in the visibility analysis, which may be smaller than the total number of unique ranks associated with the papers in the respective conferences. The x-axis range is 37 in Figure 1a and 40 in Figure 1b due to ties in rankings used. On the y-axis, smoothed visibility lies in $[0, 1]$. We use local linear regression for smoothing (Cleveland and Loader, 1996). The solid line gives the smoothed visibility, and the grey region around the line gives the 95% confidence interval.



(a) ICML 2021 (b) EC 2021

Figure 2: For papers submitted to the respective conferences, we plot the indicator for paper being posted online before the end of the review process against papers' associated rank, with smoothing. On the x-axis, we have the ranking order as described in Figure 1. On the y-axis, smoothed indicator for posting online lies in $[0, 1]$. We use locally estimated smoothing to get the smoothed indicator for posting online across ranks, shown by the solid line, and a 95% confidence interval, shown by the grey region.



(a) ICML 2021

(b) EC 2021

Figure 3: For papers submitted to the respective conferences and (not) posted online before the review process, we plot the papers' final decision against papers' associated rank, with smoothing. On the x-axis, we have the ranking order as described in Figure 1. On the y-axis, smoothed acceptance rate lies in $[0, 1]$. We use locally estimated smoothing to get the smoothed acceptance rate across ranks, shown by the lines, and a 95% confidence interval, shown by the grey region. Note that in Figure 3b, the number of papers corresponding to the ranks on the right of the plot is very small.

the review process and the pool of papers that was not, across ranks. Here, we use the final decision as a proxy indicator for the quality of paper. Note that this proxy is not a perfect indicator, it could be affected by the ranking of the paper in case of de-anonymization of the paper, as we discussed in Section 1. Now, for the pool of papers posted online, we see that Kendall's Tau-b correlation between papers' rank and final decision is 0.11 ($p < 10^{-5}$) in ICML and 0.03 ($p = 0.58$) in EC (Row 4). Recall that a positive correlation implies that the acceptance rate increases as the rank improves. For the pool of papers *not* posted online, we see that Kendall's Tau-b correlation between papers' rank and final decision, 0.16 ($p < 10^{-5}$) in ICML and 0.13 ($p = 0.006$) in EC (Row 5). Further, the correlation between the rank values and the corresponding difference (between papers posted and not posted online) in mean acceptance rates is 0.01 ($p = 0.92$) in ICML and 0.12 ($p = 0.18$) in EC (Row 6).

To interpret these values, we provide visualization of the variation of mean acceptance rate as rank varies for the two pools of papers in Figure 3. In ICML, we see that the difference in acceptance rates between the two pools of papers roughly remains the same as the rank changes. Meanwhile, in EC this difference in acceptance rates does not suggest a clear trend across ranks. Now, recall the earlier discussion regarding self-selection of papers being posted online, that authors from high-rank institutions may upload only high quality papers, while authors from low-rank institutions may not select the papers to be uploaded online

		EC 2021	ICML 2021
1	# Papers	498	5361
2	# Papers posted online before the end of review process	183	1934
3	Correlation between papers' rank and whether they were posted online $[-1, 1]$	0.09	0.12
4	Correlation for papers posted online between their rank and decision $[-1, 1]$	0.03	0.11
5	Correlation for papers not posted online between their rank and decision $[-1, 1]$	0.13	0.16
6	Correlation between ranking and corresponding difference, between papers posted and not posted online, in mean acceptance rate $[-1, 1]$	0.12	0.01

Table 3: Outcome of supporting analysis for research question Q2. A positive correlation in row 3, 4 and 5 implies that the value of the variable considered increases as the rank of the paper improves. For instance, in row 3, the rate of posting online increases as the rank improves.

based on quality. Based on the results displayed in Table 3 and Figure 3, we do not see strong evidence for self-selection of papers being posted online by authors from high-rank institutions compared to those from low-rank institutions.

5 Discussion

To improve peer review and scientific publishing in a principled manner, it is important to understand the quantitative effects of the policies in place, and design policies in turn based on these quantitative measurements.

We find that more than a third of survey respondents self-report deliberately searching for their assigned papers online, thereby weakening the effectiveness of author anonymization in double-blind peer review. Further, the observed value of fraction of reviewers that searched for their assigned paper online in Table 1 might be an underestimate due of two reasons: (i) Reviewers who deliberately broke the double-blindedness of the review process may be more reluctant to respond to our survey for Q1. (ii) As we saw in Section 4.2, roughly 8% of reviewers in ICML 2021 had already seen their assigned paper before the review process began (Table 2 row 5). If these reviewers were not already familiar with their assigned paper, they may have searched for them online during the review process.

For Q2, the effect size is statistically significant in ICML, but not in EC. A possible explanation for the difference is in the method of assigning rankings to institutions, described in Section 3.1. For ICML, the rankings used are directly related to past representation of the institutions at ICML (Ivanov, 2020). In EC, we used popular rankings of institutions such as QS rankings and CS rankings. In this regard, we observe that there is no clear single objective measure for ranking institutions in a research area. This leads to many ranking lists that may not agree with each other. Our analysis also suffers from this limitation.

Next, while we try to carefully account for confounding factors based on time of posting in our analysis for Q2, our study remains dependent on observational data. Thus, the usual caveat of unaccounted for confounding factors applies to our work. For instance, the topic of research may be a confounding factor in the effect of papers' rank on visibility: If authors from better-ranked affiliations work more on cutting-edge topics compared to others, then their papers would be read more widely. This could potentially increase the observed effect.

Policy implications Double-blind venues now adopt various policies for authors regarding posting or advertising their work online before and during the review process. A notable example is a recent policy change by the Association for Computational Linguistics in their conference review process, which includes multiple conferences: ACL, NAACL (North American Chapter of the ACL) and EMNLP (Empirical Methods in Natural Language Processing). ACL introduced an anonymity period for authors, starting a month before the paper submission deadline and extending till the end of the review process. According to their policy, within the anonymity period authors are not allowed to post or discuss their submitted work anywhere on the Internet (or make updates to existing preprints online). In this manner, the conference aims to limit the de-anonymization of papers from posting preprints online. A similar policy change has been instituted by the CVPR computer vision conference. We provide some quantitative insights on this front using the data we collected from the Q2 survey in ICML 2021 and EC 2021. There were 918 (out of 5361 submitted) and 74 (out of 498 submitted) papers posted online *during* the one month period right before the submission deadline in ICML and EC respectively. These papers enjoyed a visibility of 8.11% (292 out of 3600) and 23.81% (45 out of 189) respectively. Meanwhile, there were 1016 (out of 5361) and 109 (out of 498) papers posted online *prior to* the one month period right before the submission deadline in ICML and EC, and these papers enjoyed a visibility of 8.59% (343 out of 3994) and 18.08% (47 out of 260) respectively. These measurements may help inform subsequent policy decisions.

While our work finds dilution of anonymization in double-blind reviewing, any prohibition on posting preprints online comes with its own downsides. For instance, consider fields such as Economics where journal publication is the norm, which can often imply several years of lag between paper submission and publication. Double-blind venues must grapple with the associated trade-offs, and we conclude with a couple of suggestions

for a better trade-off. First, many conferences, including but not limited to EC 2021 and ICML 2021, do not have clearly stated policies for reviewers regarding searching for papers online, and can clearly state as well as communicate these policies to the reviewers. Second, venues may consider policies requiring authors to use a different title and reword the abstract during the review process as compared to the versions available online, which may reduce the chances of reviewers discovering the paper or at least introduce some ambiguity if a reviewer discovers (a different version of) the paper online.

Appendix

A1 Survey details for Q2.

Target audience selection Recall that our objective in target audience selection is to find reviewers for each paper whose research interests intersect with the paper, so that we can survey these reviewers about having seen the corresponding papers outside of reviewing contexts. We describe the exact process for target audience selection in EC and ICML.

In EC, the number of papers posted online before the end of the review process was small. To increase the total number of paper-reviewer pairs where the paper was posted online and the reviewer shared similar research interests with the paper, we created a new paper-reviewer assignment. For the new paper-reviewer assignment, for each paper we considered at most 8 members of the reviewing committee that satisfied the following constraints as its target audience—(1) they submitted a positive bid for the paper indicating shared interest, (2) they are not reviewing the given paper.

In ICML, a large number of papers were posted online before the end of the review process. So, we did not create a separate paper-reviewer assignment for surveying reviewers. Instead, in ICML, we consider a paper's reviewers as its target audience and queried the reviewers about having seen it, directly through the reviewer response form.

Survey question For research question Q2, we conducted a survey to measure the visibility of papers submitted to the conference and posted online before or during the review process. We describe the details of the survey for EC 2021 and ICML 2021 separately. In EC 2021, we created a specialised reviewer-specific survey form shared with all the reviewers. Each reviewer was shown the title of five papers and asked to answer the following question for each paper:

"Have you come across this paper earlier, outside of reviewing contexts?"

In the survey form, we provided examples of reviewing contexts as "reviewing the paper in any venue, or seeing it in the bidding phase, or finding it during a literature search regarding another paper you were reviewing." The question had multiple choices as enumerated in Table 4, and the reviewer could select more than one choice. If they selected one or more options from (b), (c), (d) and (e), we set the visibility to 1, and if they selected option (a), we set the visibility to 0. We did not use the response in our analysis, if the reviewer did not respond or only chose option (f). In Table 4, we also provide the number of times each choice was selected in the set of responses obtained.

In ICML 2021, we added a two-part question corresponding to the research question Q2 in the reviewer response. Each reviewer was asked the following question for the paper they were reviewing:

"Do you believe you know the identities of the paper authors? If yes, please tell us how."

Each reviewer responded either *Yes* or *No* to the first part of the question. For the second part of the question, table 5 lists the set of choices provided for the question, and a reviewer could select more than one choice. If they responded *Yes* to the first part, and selected one or more options from (a), (d), (e) and (f) for

the second part, then we set the visibility to 1, otherwise to 0. In Table 5, we also provide the number of times each choice was selected in the set of responses that indicated a visibility of 1.

A2 Analysis procedure details

In this section we provide some more details of the analysis procedure.

A2.1 Kendall's Tau-b statistic

We describe the procedure for computing Kendall's Tau-b statistic between two vectors. Let n denote the length of each vector. Let us denote the two vectors as $[x_1, x_2, \ldots, x_n] \in \mathbb{R}^n$ and $[y_1, y_2, \ldots, y_n] \in \mathbb{R}^n$. Let P denote the number of concordant pairs in the two vectors, defined formally as

$$P = \sum_{\substack{(i,k) \in [n]^2 \\ i<k}} \left(\mathbb{I}\left(x_i > x_k\right) \mathbb{I}\left(y_i > y_k\right) + \mathbb{I}\left(x_i < x_k\right) \mathbb{I}\left(y_i < y_k\right) \right).$$

Following this, we let the number of discordant pairs in the two vectors be denoted by Q, defined as

$$Q = \sum_{\substack{(i,k) \in [n]^2 \\ i<k}} \left(\mathbb{I}\left(x_i > x_k\right) \mathbb{I}\left(y_i < y_k\right) + \mathbb{I}\left(x_i < x_k\right) \mathbb{I}\left(y_i > y_k\right) \right).$$

Observe that the concordant and discordant pairs do not consider pairs with ties in either of the two vectors. In our data, we have a considerable number of ties. To account for ties, we additionally compute the following statistics. Let A_x and A_y denote the number of pairs in the two vectors tied in exactly one of the two vectors as

$$A_x = \sum_{\substack{(i,k) \in [n]^2 \\ i<k}} \mathbb{I}\left(x_i = x_k\right) \mathbb{I}\left(y_i \neq y_k\right) \qquad \text{and} \qquad A_y = \sum_{\substack{(i,k) \in [n]^2 \\ i<k}} \mathbb{I}\left(x_i \neq x_k\right) \mathbb{I}\left(y_i = y_k\right).$$

Finally, let A_{xy} denote the number of pairs in the two vectors tied in both vectors, as

$$A_{xy} = \sum_{\substack{(i,k) \in [n]^2 \\ i<k}} \mathbb{I}\left(x_i = x_k\right) \mathbb{I}\left(y_i = y_k\right).$$

Observe that the five statistics mentioned above give a mutually exclusive and exhaustive count of pairs of indices, with $P + Q + A_x + A_y + A_{xy} = 0.5n(n-1)$. With this setup in place, we have the Kendall's Tau-b

List of choices for question in Q2 survey	Count
(a) I have NOT seen this paper before / I have only seen the paper in reviewing contexts	359
(b) I saw it on a preprint server like arXiv or SSRN	51
(c) I saw a talk/poster announcement or attended a talk/poster on it	22
(d) I saw it on social media (e.g., Twitter)	4
(e) I have seen it previously outside of reviewing contexts (but somewhere else or don't remember where)	29
(f) I'm not sure	24

Table 4: Set of choices provided to reviewers in EC in Q2 survey and the number of times each choice was selected in the responses obtained. There were 449 responses in total, out of which 92 responses indicated a visibility of 1.

List of choices for question in Q2 survey	Count
(a) I was aware of this work before I was assigned to review it.	373
(b) I discovered the authors unintentionally while searching web for related work during reviewing of this paper	47
(c) I guessed rather than discovered whose submission it is because I am very familiar with ongoing work in this area.	28
(d) I first became aware of this work from a seminar announcement, Archiv announcement or another institutional source	259
(e) I first became aware of this work from a social media or press posting by the authors	61
(f) I first became aware of this work from a social media or press posting by other researchers or groups (e.g. a ML blog or twitter stream)	52

Table 5: Set of choices provided to reviewers in ICML in Q2 survey question and the number of times each choice was selected in the set of responses considered that self-reported knowing the identities of the paper authors outside of reviewing contexts. There were a total of 635 such responses that indicated a visibility of 1. Recall that for ICML, we consider the set of responses obtained for submissions that were available as preprints on arXiv. There were 1934 such submissions.

statistic between $[x_1, x_2, \ldots, x_n] \in \mathbb{R}^n$ and $[y_1, y_2, \ldots, y_n] \in \mathbb{R}^n$ denoted by τ as

$$\tau = \frac{P - Q}{\sqrt{(P + Q + A_x)\,(P + Q + A_y)}}. \tag{5.2}$$

This statistic captures the correlation between the two vectors.

A2.2 Permutation test

The test statistic T in (5.1) gives us the effect size for our test. Recall from (5.1) that the test statistic T is defined as:

$$T = \frac{N_1\,\tau_1 + N_2\,\tau_2 + N_3\,\tau_3}{N_1 + N_2 + N_3},$$

where for each bin value $b \in \{1, 2, 3\}$, we have N_b as the number of responses obtained in that bin, and τ_b represents the Kendall Tau-b correlation between visibility and rank in the responses obtained in that bin. To analyse the statistical significance of the effect, we define some notation for our data. Let N denote the total number of responses. For each response $i \in [N]$ we denote the visibility of the paper to the reviewer as $v_i \in \{0, 1\}$ and the rank associated with response i as $\alpha_i \in \mathbb{N}_{<0}$. Finally, we denote the bin associated with response i as $\widetilde{t}_i \in \{1, 2, 3\}$. With this, we have the following algorithm for permutation testing.

Algorithm 1 Permutation test for correlation between papers' visibility and rank.

Input : Samples $v_i, \alpha_i, \widetilde{t}_i$ for $i \in [N]$, iteration count γ.
(1) Compute the test statistic T defined in (5.1).
(2) For $z \leftarrow 1$ to γ:
(i) For all $b \in \{1, 2, 3\}$: Let V_b denote the number of responses with $v_i = 1$ in bin b. Take all the responses in bin b and reassign each response's visibility to 0 or 1 uniformly at random such that the total number of responses with a visibility of 1 remains the same as V_b.
(ii) Using the new values of visibility in all bins, recompute the test statistic in (5.1). Denote the computed test statistic as T_z.
Output : P value $= \frac{1}{\gamma} \sum_{z=1}^{\gamma} \mathbb{I}(T_z - T > 0)$.

Chapter 6

The Novice Reviewers' Bias against Resubmissions in Conference Peer Review

1 Introduction

In contrast to many other fields of science, where journals are the only established venues for research publication, in machine learning (ML) and computer science (CS), conferences are considered to be equally or even more attractive (Vrettas and Sanderson, 2015). While being as selective as top journals, leading conferences ensure much shorter turnaround time thereby facilitating timely research dissemination and allowing authors to quickly resubmit their work to the next conference if it gets rejected. However, the explosion in the number of submissions received by top conferences has considerably challenged the sustainability of the conference peer-review process since the number of qualified reviewers is growing at a much slower rate (Sculley et al., 2019; Shah, 2022).

In an attempt to decrease the load on reviewers by discouraging resubmissions without substantial changes, several leading ML and CS conferences have started requesting or requiring authors to declare if a previous version of their submission was rejected at other peer-reviewed venues. For example, a top-tier conference in natural language processing EMNLP 2019 allowed authors to decide whether they want to disclose the past submission history and provide the summary of changes they made, making this information available only to senior committee members. A similar opportunity was offered at a leading conference in artificial intelligence and statistics (AISTATS 2017) with the exception that the information about past rejections was also available to regular reviewers. Additionally, the AISTATS 2017 conference implemented an automated review sharing with some past conferences, making these reviews visible to senior committee members after the initial reviewing was completed.

Other conferences make it mandatory for authors to disclose the past submission history: the NeurIPS conference — one of the most popular conferences in ML, which in 2019 received more than six thousand submissions — in 2020 required authors of previously rejected submissions to declare the changes they made to the current version of the paper. Another top conference in artificial intelligence (IJCAI 2020) went even further and made full reviews from past venues available to reviewers by requiring authors to include them in the submission file *before* the actual paper.

In addition to changes implemented by specific venues, the `openreview` platform — a growing conference management system that hosts the leading deep-learning conference ICLR and other forums — offers a novel approach towards managing conferences in a transparent manner by allowing organizers to make both accepted and rejected submissions accompanied with full reviews publicly available. This option has been used by the ICLR conference since 2017 and all reviews for papers submitted to the conference since then are now publicly available, allowing subsequent reviewers (even at different venues) of rejected papers to consult

them at any time.

All these steps are supposed to facilitate improving the quality of peer review as reviews from past venues may reduce the burden on reviewers by allowing them to "quickly focus in on what previous issues were and how they may or may not have been fixed" (Francis, 2008; Lin et al., 2020a). While being especially actual for the fields of ML and CS, similar discussion is also happening in other areas. For example, a survey conducted by Cals et al. (2013) among editors of general medical journals showed that despite there being a concern that reviews from previous venues may bias subsequent reviewers, 45% of the participating editors prefer authors to indicate whether a paper has been previously rejected (the survey does not indicate how this information is supposed to be used in the review process). Only 24% of editors oppose this idea, and the rest are indecisive.

Despite the interest of journal editors and conference program committees in reusing reviews from past iterations of the review cycle, authors are less enthusiastic: ICFP — the ACM-sponsored conference on functional programming — in 2020 removed the option to upload the past reviews since no authors took this advantage in the previous edition of the conference. This skepticism suggests that authors do not believe that revealing old reviews will increase the acceptance chances of their submission.

Authors' concerns are in fact supported by a long line of research in psychology that establishes susceptibility of human judgements to various biases (see Tversky and Kahneman, 1974; Kahneman, 2011; Gilovich et al., 2002, for overview), some of which are caused by additional (and sometimes irrelevant) information available to the decision-maker (Baron and Hershey, 1988; Tversky and Kahneman, 1974; Fischhoff and Beyth, 1975; Ross et al., 1975; Carretta and Moreland, 1983). Projecting this evidence on the peer-review context, we hypothesize that the knowledge that a paper was rejected at a previous venue may negatively bias reviewers' evaluations, leading to what in this study we call a "resubmission bias". In other words, reviewers may judge the paper differently depending on whether they are notified about this paper being a resubmission or not.

To highlight the potential impact of the resubmission bias, we note that the peer-review process of the major machine learning conferences is far from being absolutely consistent and objective. Indeed, several controlled experiments (Lawrence and Cortes, 2014; Price, 2014; Pier et al., 2018) found very low degree of agreement between reviewers evaluating the same manuscript. This implies that the outcome of the review process heavily depends on the reviewers to whom the submission is assigned, introducing significant randomness in the final decisions. Hence, even a strong paper that deserves acceptance has a nontrivial chance of being rejected due to this randomness. The resubmission bias can *amplify this unfairness by putting a previously rejected paper at disadvantage in the subsequent conferences.* Given that success in academia is largely determined by the publication profile of a researcher, the resubmission bias can have far reaching consequences not just for a particular paper, but more generally also for career trajectories of researchers due to the widespread prevalence of the Matthew effect ("rich get richer") in academia (Merton, 1968; Squazzoni and Gandelli, 2012).

Therefore, in this chapter, we aim at testing for the presence of the resubmission bias in peer review. Focusing on the population of novice reviewers, in conjunction with a reviewer-recruiting process of ICML 2020 (a top ML conference), we design and conduct a randomized controlled trial to test the following research hypothesis:

> *Research Hypothesis:* The signal about rejection from the same or similar venue in the past, received by novice reviewers before they read the paper, induces a bias in reviewers' evaluations.

In our hypothesis, we do not specify the direction of the effect and note that while authors are concerned about a negative bias, the bias can hypothetically be positive. For example, a reviewer may think that previously rejected papers have gone through another iteration of revisions and improvements and might be better on average than papers that have not previously been peer reviewed. Therefore, in this work we also aim to confirm the direction of the effect (if it exists).

Importantly, we caveat that in this study we target the population of novice reviewers and the findings we report in this work must not be overgeneralized to the whole reviewer population. While the choice of the study participants is mostly justified by the difficulty of engaging senior reviewers in the experiment, we discuss below (Section 3) that novice reviewers constitute a large fraction of the leading ML and CS conferences reviewer pool.

In this work, we do not aim to support or undermine the idea of reusing the past reviews. Instead, the answer to our research question will inform conference organizers and journal editors about potential side-effects of reusing reviews, and will help them to carefully select the point during the peer-review pipeline (if at all) at which previous reviews become available to reviewers. For example, if the resubmission bias in peer review is present, then editors or program chairs may prefer to keep the past submission history hidden from reviewers in the initial stages of the process to ensure that the reviewers form an unbiased opinion about the paper first. Additionally, the results of our experiment can inform the current practices of novice reviewer training where more emphasis could be made on how to avoid the potential resubmission bias.

Past research on human decision making decisively establishes the presence of various cognitive biases in human judgement that are relevant to our research hypothesis. For example, the famous anchoring effect (Tversky and Kahneman, 1974) manifests in human evaluations being dependent on an (irrelevant) piece of information received at the beginning of the decision task. We defer the discussion of the general literature on decision making and cognition to Section 5, but underscore that these findings do not necessarily transfer to our setting. Indeed, from the perspective of the dual-process model of cognition (Kahneman and Frederick, 2002; Stanovich, 1999; Stanovich and West, 2008), biases are properties of an autonomous heuristic system (System 1) whereas the demanding and rational reviewing task invokes the analytic system (System 2) that can potentially override such biases.

2 Related literature

The work described in this chapter contributes to a long line of empirical works studying various aspects of the academic peer-review process and in this section we discuss the most relevant papers.

Resnik and Smith (2020), in the chapter of their book dedicated to peer review, discuss the presence of groupthink — a strong desire of group consensus that results in all deviating ideas being rejected — in peer review. One potential contribution to the overall groupthink effect is susceptibility of experts to social influence, as demonstrated by Teplitskiy et al. (2019). In their experiment, reviewers (faculty at US medical schools) first evaluated and rated submissions assigned to them, and then were exposed to scores presumably given to these submissions by other anonymous reviewers. Unbeknown to participants, these scores were randomly sampled to be either above or below their scores. The experiment demonstrated that 47% of reviewers decided to update their scores, and in all but one case (out of approximately 185) the update was in the direction of the external scores. This finding indicates a strong impact of social influence on experts' evaluations. The experiment of Teplitskiy et al. (2019) was conducted alongside a review process of grant applications in which real awards were distributed; in this setup, reviewers are generally expected to reach a consensus, hence, updates of review scores do not necessarily indicate a bias in reviewers' judgements of the proposal quality and instead can be seen as attempts to decrease the inconsistency of evaluations. In our work, we design the experiment to remove the group deliberation component of peer review by eliminating the discussion stage of the process, and measure the bias in reviewers' attitude towards the submission.

Another widely documented bias (Rosenthal, 1979; Moscati et al., 1994; Callaham et al., 1998; Emerson et al., 2010, and others) that manifests in peer review is the *positive*-outcome bias also known as the *file drawer problem*. This bias results in reviewers judging soundness of the experimental works depending on the outcome of the study: works in which the null hypothesis is rejected are rated more favourably than otherwise equal papers that show nonsignificant results.

One difference between the social influence, file drawer problem and resubmission bias which may result in different cognitive heuristics responsible for the effects is the stage of the review process at which the stimulus is received by reviewers. In case of the social influence, reviewers get biased by signals they receive *after* the review process and this bias can be attributed to the desire of consensus, attempt to improve the accuracy of a review or aim to achieve some other social goals (Resnik and Smith, 2020; Teplitskiy et al., 2019; Cialdini and Goldstein, 2004). The file drawer problem is induced by information observed *during* reviewing and may be attributed to the desire of accepting "newsworthy" (i.e., positive) studies for publications (Callaham et al., 1998; Lynch et al., 2007). Finally, the incarnation of the resubmission bias we study in this work completes the picture and is related to the information reviewers receive *prior* to reviewing submissions.

Confirmatory bias — a tendency of people to emphasize evidence that support their views and ignore or misinterpret those that do not — is another relevant effect connected to prior beliefs of reviewers. It was identified in context of peer review by Mahoney (1977) who conducted an experiment in which reviewers were exposed to different versions of the manuscript with identical experimental procedure but different directions of the obtained results (either confirming or disproving the beliefs of reviewers). It turned out that reviewers who received the version contradicting their theoretical perspective were significantly harsher in their evaluations than those who received the version supporting their views. In Section 5 we draw further connections between the resubmission and confirmatory biases.

A separate line of work (Tomkins et al., 2017; Blank, 1991; Okike et al., 2016, and others) studies the presence of various biases, including gender, affiliation and fame biases in single-blind peer review (i.e., when author identities are visible to reviewers). The work of Tomkins et al. (2017) is of particular interest as it identifies strong fame and affiliation biases[1] in reviewers' evaluations. As a result of this work, WSDM — a premier conference in web-inspired research — switched to double-blind reviewing (i.e., author identities are hidden from reviewers), demonstrating the potential impact of empirical research on peer-review practices.

Finally, our work contributes to the growing literature in computer science, both theoretical and empirical, that aims to understand and improve the conference peer-review process. These works develop methodologies to address various biases and other issues in peer review such as miscalibration (Roos et al., 2011; Ge et al., 2013; Wang and Shah, 2019), commensuration bias (Noothigattu et al., 2020), strategic or dishonest behavior (Aziz et al., 2019; Xu et al., 2019; Stelmakh et al., 2021b; Jecmen et al., 2020), biases with respect to author demographics (Tomkins et al., 2017; Stelmakh et al., 2019; Manzoor and Shah, 2020), and methods for better assignments of reviewers to papers (Garg et al., 2010; Charlin and Zemel, 2013; Kobren et al., 2019; Fiez et al., 2020; Stelmakh et al., 2021a). We envisage that the biases discussed in the current paper may be mitigated by a combination of policy guidelines and such computational techniques.

3 Experimental setup

To test our research hypothesis of presence of the resubmission bias, we design a randomized controlled experiment that replicates the relevant components of the peer-review pipeline of machine learning conferences while giving us more control over the resubmission signal received by reviewers. Two main components of the peer-review process are pools of papers and reviewers, and we now describe how these pools were constructed.

Papers We solicited $m = 19$ anonymized preprints in various sub-areas of machine learning. To ensure that participants of the experiment cannot obtain a signal about a paper being a resubmission from anywhere outside of the experimental context, we restricted the pool of papers to works that had not yet been accepted to any conference or journal and had never been submitted to conferences hosted by the openreview platform or any other venue that makes the list of rejected submissions publicly available. We note that in machine learning it is common to publish preprints on arXiv (`arxiv.org`) and we allow papers available on arXiv to be used in the experiment. Additionally, we allow papers that were previously presented at workshops — less formal and prestigious venues without proceedings — because it is also common to present a preliminary version of the work in a workshop and then submit the full version to the conference.

The final pool of papers consisted of working papers, papers under review, workshop publications and unpublished manuscripts. The papers were 6–12 pages long excluding references and appendices (a standard range for many ML and CS conferences) and were formatted in various popular journals' and conferences' templates with all explicit venue identifiers removed.

Reviewers The reviewer pool of a typical machine learning conference consists of researchers at various seniority levels, working in areas covered by the conference. Perhaps unique to ML, a recent surge in the number of papers submitted to leading conferences has forced organizers to expand the reviewer pool by relaxing the qualification bar, that is, by introducing rather junior researchers to the reviewer pool: for example, 33% of the NeurIPS 2016 reviewers (1082 out of 3233) were graduate students (Shah et al., 2018). Using data on the structure of the reviewer pool of the ICML 2020 conference, we estimate that approximately

[1] Meta-analysis they perform also reveals a bias against female-authored submissions.

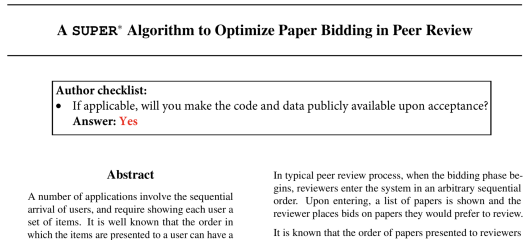
35% of the reviewers are rather junior individuals who self-nominated and satisfied the screening requirements of having at least two papers published in some top ML venues and being a reviewer for at least one top ML conference in the past. Overall, we conclude that a significant fraction of ML reviewers are novices and in this study we concentrate on the population of novice and junior reviewers. We admit that in this work we do not approximate the general reviewer population by leaving out more experienced reviewers, but notice that at the ICML 2020 conference each submission was assigned to at least one reviewer from the aforementioned subset, making it a significant part of the reviewer community.

To recruit participants, we messaged mailing lists of five large large, top US universities (CMU, MIT, UMD, UC Berkeley and Stanford) and targeted master's and junior PhD students working in ML-related fields. The invitation also propagated to a small number of students outside of these schools through the word of mouth. The recruiting materials contained an invitation to participate in the ICML reviewer-selection experiment (selection of reviewers was indeed a key goal of the experiment and we studied the resubmission bias on top of it). Specifically, we asked participants to write a review for one paper and promised that those who provide a high-quality review will be invited to join the ICML 2020 reviewer pool. Being a reviewer at the flagship ML conference is a recognition of one's expertise and we expected that this potential benefit will motivate students to join our experiment. As a result, we received responses from $n = 200$ candidates, more than 90% of whom were master's and PhD students or recent graduates of the aforementioned universities, and all of them were added to the pool of participants without further screening. The research hypothesis we study in this work may be sensitive to awareness of participants; therefore, we employed deception and did not reveal the dual goal of this work to participants, that is, subjects were unaware that in parallel with selecting reviewers for ICML, we also want to measure the impact of the resubmission signal on the reviewers' attitude towards papers they review.

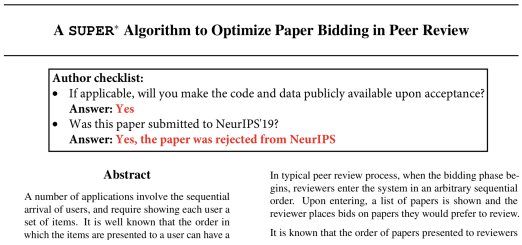
The experimental procedure closely followed the initial stages of the standard ML conference peer-review pipeline and was hosted using Microsoft Conference Management Toolkit (`https://cmt3.research.microsoft.com`). First, we asked participants to express their interest in reviewing specific papers by entering bids that take the following values: "Not Willing", "In a Pinch", "Willing" and "Eager". Thirteen participants did not enter any bids and were removed from the pool. The remaining participants were active in bidding (mean number of "Willing" and "Eager" bids is 4.7) and we assigned all of them to 1 paper each, where we tried to satisfy reviewer bids, subject to a constraint that each paper is assigned to at least 8 reviewers. As a result, 186 participants were assigned to a paper they bid either "Willing" or "Eager" and 1 participant was assigned to a paper they bid "In a Pinch" (this participant did not bid "Eager" or "Willing" on any paper).

Finally, we instructed participants that they should review the paper as if it was submitted to the real ICML conference with the exception that the relevance to ICML and the formatting issues (e.g., page limit, margins) should not be considered as criteria. To help participants in writing their reviews, we provided reviewer guidelines adapted from NeurIPS instructions (see supplementary materials on the website of the author of this thesis) that discuss the best practices of reviewing. We gave participants 15 days to complete the review and then extended the deadline for 16 more days to accommodate late reviews as our original deadline interfered with the final exams at various US universities and the US holiday period.

Unbeknown to participants, we allocated half of them to the test condition and half to the control condition uniformly at random. Participants in the control condition did not receive the resubmission signal while subjects in the test condition were notified that the paper they are reviewing was rejected at the NeurIPS 2019 conference. Since the goal of the experiment declared to participants was to test out a new approach towards recruiting reviewers, the presence of the resubmission signal could have been confusing for the participants as this information is irrelevant to the task. To ensure that the presence of the signal does not look odd to reviewers, we incorporated it in a small author checklist placed on the first page of each submission as shown in Figure 1. Participants in the control condition received submissions with a single-item checklist asking about the code submission, while participants in the test condition were additionally informed that the paper is a resubmission.



(a) Control condition.



(b) Test condition.

Figure 1: Different versions of checklists shown to reviewers in the test and control conditions. The answer to all questions were "Yes" for all papers in both conditions.

4 Analysis of the experiment

Out of 187 participants who received a paper for review, 134 handed in the reviews (response rate of 71.7%). The remaining reviewers are excluded from subsequent analysis. Additionally, one of the participants misread the instructions and wrote a one-line review, rejecting the paper based on a minor violation of the standard ICML page limit. This participant was also excluded from the analysis as formatting violation was not considered as a review criterion in the experiment. Table 1 compares populations of reviewers between the test and control conditions using demographic information available to us.

The review form offered to participants (included in supplementary materials on the website of the author of this thesis) was adapted from the NeurIPS 2019 form and contained 3 fields for open-ended feedback and 6 multiple choice questions in which reviewers were asked to give the overall score to the submission (10-point Likert item), evaluate the submission on 4 criteria (Originality, Quality, Clarity and Significance, 5-point Likert item for each criteria) and finally self-access their confidence in the scores assigned (5-point Likert item). Options offered for multiple choice questions were ordered from the most negative to the most positive and for the sake of analysis we associate these options to the numeric scales: 10-point scale for the overall score and 5-point scale for the other questions (larger numbers indicate more positive evaluations).[2] In the sequel of this section, we compare these values between the test and control reviewers to see whether the resubmission signal significantly impacted the behaviour of reviewers in the test condition.

4.1 Testing procedure and effect size

In this work, we employ a modification of the permutation test (Fisher, 1935) that accounts for the fact that each paper is reviewed by different number of the test/control reviewers and is defined as follows. Recall that m stands for the number of papers in the experiment, for each paper $i \in [m]$ we let $\mathcal{R}_i^{\text{cntr}}$ (respectively $\mathcal{R}_i^{\text{test}}$) be a set of reviewers assigned to this paper in the control (respectively test) condition. Next, for any reviewer $j \in \mathcal{R}_i^{\text{cntr}}$ we use Y_i^j to denote a numeric evaluation given by reviewer j to paper i. Similarly, for any test reviewer $j \in \mathcal{R}_i^{\text{test}}$ we let X_i^j denote their evaluation of paper i. With this notation, we define the

	# Participants	#With Prior Review Experience	#With Publications
Control	68	20	47
Test	65	28	49

Table 1: Comparison of demographics of reviewers across the test and control conditions. All differences are not significant at the level $\alpha = 0.1$.

[2]We release numeric evaluations reported by reviewers and the dataset is available on the website of the author of this thesis.

test statistic:

$$\tau = \frac{1}{m} \sum_{i \in [m]} \Bigg[\underbrace{\frac{1}{|\mathcal{R}_i^{\text{test}}|} \sum_{j \in \mathcal{R}_i^{\text{test}}} X_i^j}_{M_i^{\text{test}}} - \underbrace{\frac{1}{|\mathcal{R}_i^{\text{cntr}}|} \sum_{j \in \mathcal{R}_i^{\text{cntr}}} Y_i^j}_{M_i^{\text{cntr}}} \Bigg].$$

Our test statistic compares mean scores M_i^{test} and M_i^{cntr} received by each paper $i \in [m]$ from the test and control reviewers, respectively, and then averages differences of these scores across submissions. A negative value of the test statistic implies that reviewers in the test condition are more harsh than those in the control condition, whereas a positive value indicates that the test reviewers are more lenient.

Having defined the test statistic, we employ the permutation test to quantify the significance of its value. To this end, we permute reviewers within each paper between the test and control conditions uniformly at random, ensuring that the number of reviewers in each condition remains the same. We then recompute the value of the test statistic for 10,000 permutations and compare these values with the original value of the test statistic to obtain P-values.

In addition to reporting P-values of the test, we also provide the following measures of the effect size that capture different aspects of the effect:

- **Simple Effect Size** First, we note that our test statistic measures the between-conditions shift of the distributions of scores, that is, it represents the mean difference (in points of the corresponding Likert item) between scores given by two groups of reviewers. Hence, the value of the test statistic τ serves as a natural measure of the unscaled (simple) effect size (Baguley, 2008). Note that the larger the absolute value of the test statistic, the stronger the effect.

- **Scaled Effect Size (adaptation of Cohen's *d*)** Following recommendations of Cohen (Cohen, 1992), we supplement the aforementioned simple effect size (which is expressed in the original units of analysis) with its scaled version, making it independent of the corresponding units. While the scaling inevitably obscures the interpretation of the resulting value, it helps to compare effect sizes between evaluations of the overall score (measured on the 10-point Likert item) and criteria scores (measured on 5-point Likert items). To report the scaled effect size, we use an adaptation of the Cohen's d (Cohen, 1992):

  $$d = \frac{\tau}{\sqrt{v}},$$

  where v is an upper bound on the variance of the terms M_i^{cntr} and M_i^{test}, $i \in [m]$, which is obtained by using the boundedness of numeric evaluations Y and X. Formally, $v = {}^{(k-1)^2}/_4$, where k is the number of points in the corresponding Likert item (10 for overall score and 5 for criteria scores). Note that we rely on the upper bound to avoid inflation of the effect size as terms corresponding to different papers have different variances, determined by the number of reviews written for this paper. The effect size computed in this manner is conservative because it assumes the largest possible sample variance. Similar to the simple effect size, the larger the absolute value of the scaled effect size, the stronger the effect.

- **Relative Effect Size (stochastic superiority)** To reduce the impact of the extreme numeric evaluations on the effect size, we also consider the measure of stochastic superiority that is computed taking into account relative rather than absolute differences (Vargha and Delaney, 2000):

  $$A = \frac{1}{\sum\limits_{i \in [m]} |\mathcal{R}_i^{\text{test}}| \times |\mathcal{R}_i^{\text{cntr}}|} \sum_{i \in [m]} \left(\sum_{j_1 \in \mathcal{R}_i^{\text{test}}} \sum_{j_2 \in \mathcal{R}_i^{\text{cntr}}} \left[\mathbb{I}\left(X_i^{j_1} > Y_i^{j_2} \right) + 0.5 \cdot \mathbb{I}\left(X_i^{j_1} = Y_i^{j_2} \right) \right] \right),$$

  where $\mathbb{I}(\cdot)$ is an indicator function that equals 1 if its argument is correct and 0 otherwise. The denominator of this equation is the total number of (test, control) pairs of reviews written for the same paper. Each of these pairs contributes 1 to the numerator if the test review is more positive than the control review and

		Simple ES		Scaled ES		Relative ES	
	P-value	Size	95% CI	Size	95% CI	Size	95% CI
Overall score	.036*	-0.78	$[-1.30, -0.24]$	-0.17	$[-0.29, -0.05]$	0.42	$[0.32, 0.52]$
Quality	.005*	-0.46	$[-0.69, -0.23]$	-0.23	$[-0.35, -0.11]$	0.37	$[0.27, 0.46]$
Clarity	.022*	-0.44	$[-0.68, -0.19]$	-0.22	$[-0.34, -0.10]$	0.43	$[0.34, 0.52]$
Significance	.037*	-0.36	$[-0.61, -0.10]$	-0.18	$[-0.30, -0.05]$	0.43	$[0.35, 0.50]$
Originality	.105	-0.21	$[-0.40, -0.03]$	-0.11	$[-0.20, -0.02]$	0.41	$[0.32, 0.50]$
Confidence	.902	-0.01	$[-0.20, 0.17]$	-0.01	$[-0.10, 0.09]$	0.50	$[0.42, 0.59]$

Table 2: Comparison of evaluations given by participants in the test and control conditions to papers assigned to them for review. All P-values are two-sided and are computed by the permutation test with 10,000 permutations. Asterisk indicates significance at the level $\alpha = 0.05$. Confidence intervals for effect sizes are computed using 5,000 bootstrapped samples.

0.5 if the reviews contain the same numeric evaluation. Intuitively, this measure of the effect size equals to the empirical probability of a randomly sampled pair of (test, control) reviews written for the same paper having the test review more positive than the control (with ties broken uniformly at random). In contrast to the previous measures of the effect size, the further the value of A from 0.5, the stronger the effect.

In addition to reporting the point estimates of the effect sizes, we also report bootstrapped 95% confidence intervals computed over 5,000 iterations, where bootstrapping is performed at the level of papers. With all the preliminaries introduced, we are now ready to present the results of our experiment.

4.2 Results

Table 2 compares evaluations given by reviewers from the test and control conditions. The results indicate that reviewers in the test condition were significantly more strict in evaluating submissions, in average reporting almost 1 point lower overall score on the 10-point Likert item ($\Delta = -0.78$, 95% CI $= [-1.30, -0.24]$) than reviewers in the control condition. The difference appears to be small but considerable and we provide more discussion of the strength of the effect in the next section. Looking at the criteria score, we observe the similar trend of reviewers in the test conditions being stricter. Comparing unit-independent effect sizes (adaptation of Cohen's d and stochastic superiority), we note that the bias manifests the most in the evaluations of the submissions' quality which is known to be of high importance for the overall evaluation of the submission (Noothigattu et al., 2020).

Overall, the data presented in Table 2 supports our research hypothesis, suggesting that junior reviewers are indeed susceptible to the resubmission bias. Notably, while the resubmission bias makes reviewers harsher, it does not seem to impact the confidence of reviewers. However, we qualify that self-evaluations of the confidence may not be a reliable measure of actual confidences, due to other biases manifesting in parallel. Indeed, as reviewers participate in the experiment to receive the invitation to join the pool of ICML reviewers, they may be reluctant to report too high or too low confidence as it may hypothetically hurt their chances. This hypothesis agrees with the observed data as 122 out of 133 reviewers reported confidence level 3 or 4 on the 5-point Likert item (recall that larger values indicate higher confidence).

As a final remark, we note that reported bootstrapped confidence intervals may be slightly more inaccurate than one would expect for the given sample size due to specific nature of data. Indeed, 133 subjects of the experiment were broken into two groups (test and control) and distributed across 19 papers. The bootstrapping was performed at the level of papers, that is, for each paper we bootstrapped test and control reviewers from the actual test and control reviewers assigned to this paper. The sample size for each paper is small and hence the resulting intervals could capture the excessive variance or underestimate the actual variance. Nevertheless, the combination of different measures of the effect size and the results of the permutation test

(which is guaranteed to control the false alarm probability even under the small sample size) suggest that the effect is present in the data.

5 Discussion

The experiment we conduct in this work identifies the presence of the resubmission bias that manifests in junior reviewers being significantly harsher in evaluating submissions that they know were previously rejected from similar venues. The design of our experiment ensures the absence of unobserved confounders that could drive the result, as we obtain reviews for the *same* submission in both test and control conditions. In this section, we discuss some aspects of our experiment and suggest directions for future work.

Strength of the effect The size of the impact of the resubmission bias on the overall score received by submissions appears to be small according to the measures of the effect size we use in this work. We note, however, that top machine learning and computer science conferences are highly competitive and even small changes in reviewers' scores may have significant impact on the outcome of a submission and more generally on the researchers' career (Thurner and Hanel, 2011; Squazzoni and Gandelli, 2012). As a concrete example, data from the ICML 2012 conference (Langford, 2012b) demonstrates that papers with mean reviewer score 2.67 (on a 4-point Likert item) were 6 times more likely to be accepted than papers with mean score 2.33: the difference between these scores is a single point decrease in a single review. While this data is observational and does not account for potential unobserved "true quality" of submissions, it suggests that even small effects may result in large changes in the outcomes.

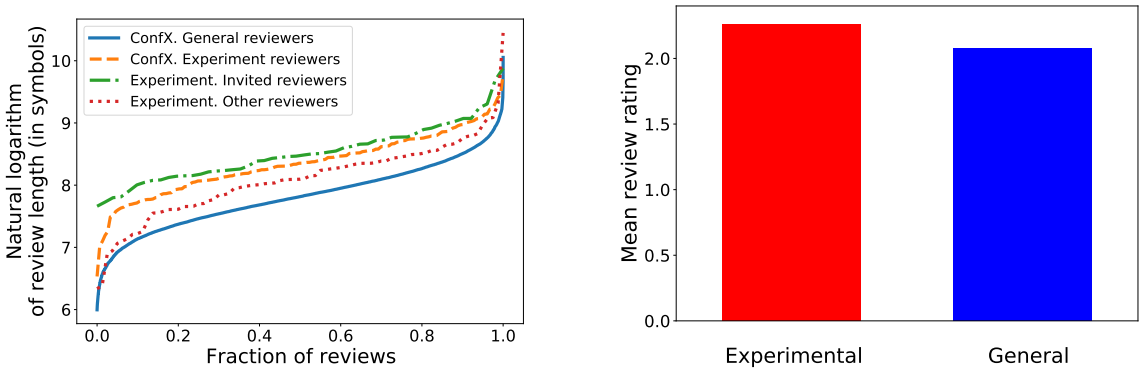
Next, we note that reviewers' evaluations are known to be subjective and have a high variance (Kerr et al., 1977; Mahoney, 1977; Ernst and Resch, 1994; Bakanic et al., 1987; Lamont, 2009). As a result, the effect that would be perceived as "strong" in a hypothetical within-subject experiment may be much less prominent in the between-subject design due to additional variance in evaluations due to subjectivity.

Finally, the pool of papers we have is diverse as it contains both papers that are likely to be accepted to the top conferences and papers that are not. In absence of the ground-truth ranking of papers and due to a limited sample size, we cannot look closer at the strength of the bias as a function of a paper quality, but past work suggests that some biases may be especially prominent in a subset of borderline papers (Blank, 1991). As a direction for future work, it would be interesting to understand the extent of the resubmission bias with a breakdown by the quality of the submission and some characteristics of reviewers (e.g., experience).

Population of participants The data obtained in this experiment unfortunately does not allow us to decisively align the population of the experiment participants with the general population of top-tier machine learning conference reviewers. Moreover, most of participants of the experiment would not be invited to join the reviewer pool of the ICML 2020 conferences through the standard ways of reviewer recruiting. However, as a result of the experiment, 52 participants whose reviews were found to be strong enough were invited to join the ICML reviewer pool. Therefore, these participants allow us to make some indirect comparisons to the general reviewer pool and we now present some relevant results. For a detailed analysis of performance of reviewers recruited through our experiment in the real ICML conference, we refer the reader to the companion paper (Stelmakh et al., 2021c).

Figure 2a juxtaposes the lengths of reviews written by different populations of reviewers at different venues. If we treat the length of a review as a measure of diligence, then we can see that reviewers selected in our experiment are not only more diligent than other participants of the experiment, but are also more diligent than the general reviewer population at the ICML conference. Perhaps surprisingly, even if we consider the participants of our experiment who did not get invited to review for the main conference, the length of their reviews stochastically dominates the length of the reviews written by general ICML reviewers with a significant margin, suggesting that they put a non-trivial effort in writing reviews.

Next, Figure 2b compares independent evaluations of review quality between the general population of ICML reviewers and reviewers recruited through our experiment. In the ICML review process each paper is assigned to several reviewers and one area chair. The area chairs are in charge of overseeing the reviewer activity, and at the end of the process, evaluate each reviewer on a 3-point Likert item: "Failed to Meet Expectations" (score 1), "Met Expectations" (score 2) and "Exceeded Expectations" (score 3). As



(a) Distribution of the natural logarithm of the review length (in symbols). For clarity, we remove reviews shorter than 400 symbols from the set of reviews written by general ICML reviews and plot linearly-interpolated curves instead of discrete points.

(b) Mean rating of reviews given by area chairs across two groups of reviewers.

Figure 2: Comparison of participants of the experiment with the general population of ICML reviewers.

seen in Figure 2b, experimental reviewers appear to produce reviews of high quality according to the area chair's ratings, being on average better than general population of reviewers ($N_{\text{experimental}} = 111,\ N_{\text{general}} = 11624,\ \Delta = 0.18,\ P < .001$).[3]

As a word of caution, we note that comparisons we make in this section are based on observational data and may be influenced by confounding variables. For example, our experiment and the real conference employ different reviewer forms that could impact the length of the reviews. Additionally, experimental reviewers received at most 3 papers for review at ICML and only 1 paper in the experiment while general ICML reviewers received up to 6 submissions. There were also some other subtle differences between experimental and general reviewers in the assignment procedure of ICML that could contribute to the observed result.

Subject to the aforementioned caveats, indirect evidence presented in Figure 2 suggests that the participants of the experiment who got invited to join the ICML 2020 reviewer pool are comparable to the general population of top ML conference reviewers in terms of the length and quality of their reviews.

However, we underscore again that the findings of this work must be interpreted with care and must not be overgeneralized to a more senior population of reviewers. Indeed, a past study of reviewers' behaviour (Teplitskiy et al., 2019) demonstrated that seniority and expertise of reviewers impact their behaviour and hence serve as confounding factors. Therefore, we do not know whether the results reported in this chapter extend to the whole reviewer pool, and an interesting direction for future work is to understand the presence of the resubmission bias in the general population of reviewers.

Impact of this study on the reviewer-selection process Recall that we conducted this study on top of the main experiment whose primary goal was to select participants to join the ICML reviewer pool. To minimize the effect of this study on the main experiment, prior to the selection process we removed numeric evaluations given by participants from the reviews, so the selection was based solely on the textual part of reviews. We then manually analyzed reviews that fall in the study team members' area of expertise, asked authors to comment on the review quality, and crowdsourced expert opinions for reviews that we were unable to evaluate ourselves. Combining feedback from these sources, we eventually invited 52 reviewers to join the ICML reviewer pool: 20 reviewers from the test condition and 32 reviewers from the control condition. The difference between the invitation rates appears to be insignificant at the level $\alpha = 0.05$ ($\Delta = 0.16, P = 0.075$).

Cognitive mechanism of the resubmission bias While in this work we do not aim to identify the cognitive mechanism of the resubmission bias, we now briefly comment on its relationship to several relevant

[3]To evaluate significance, we use a permutation test treating the rating of each review as an independent random variable.

cognitive biases known from general psychological literature.

- *Anchoring (Tversky and Kahneman, 1974; Strack and Mussweiler, 1997; Mussweiler and Strack, 2001)* Human judgements and evaluations are known to depend heavily on an initial piece of information (anchor), even when the anchor is obviously irrelevant (e.g., a number generated by the wheel of fortune). The signal about the previous outcome received by reviewers may be seen as an anchor that prevents them from adjusting to a positive opinion about the paper; therefore, anchoring may be responsible for the resubmission bias.
- *Social proof (Asch, 1951; Baron et al., 1996; Resnik and Smith, 2020; Cialdini and Goldstein, 2004)* When being a part of a group, individuals are known to be susceptible to social influence, often exhibiting conformity to the opinion of the group. While in our experiment we removed the group deliberation component of the review process and participants were acting individually, they could consider the rejection from the previous venue as a decision made by a group of more experienced reviewers and decide to comply, deferring to the authority.
- *Confirmatory bias (Mahoney, 1977; Lord et al., 1979; Rabin and Schrag, 1999; Garcia et al., 2020)* Once having a belief about the state of the world, people tend to be selective in accepting new evidence: evidence that supports their views gets accepted more easily while contradicting evidence may be ignored or misinterpreted. In this paradigm, the resubmission bias can contribute to creating an initial belief that the paper under review is of a low quality that can later be exacerbated by a biased interpretation of strength and weaknesses of the submission.
- *Hindsight bias (Fischhoff and Beyth, 1975; Hawkins and Hastie, 1990; Roese and Vohs, 2012) and Outcome bias (Baron and Hershey, 1988; Allison et al., 1996; Marshall and Mowen, 1993; Gino et al., 2008)* Hindsight bias ("I knew it all along") is a tendency of people to overestimate the predictability of the event after it becomes observed. Outcome bias is a slightly different effect that manifests in distortion of human judgements of decisions. When a stochastic outcome of the decision is observed, even having all the information available to the decision-maker at the time of the decision, people tend to judge the quality of the decision differently depending on whether they view the outcome as good or bad. While these effects do not directly correspond to the resubmission bias we study in this work, they suggest that availability of outcome information can adversely impact the evaluations.

Despite all of the above biases are known to be present in human judgements, their presence in peer review is not obvious because of the nature of the task performed by reviewers. Indeed, the reviewing task is analytical and requires rational thinking, thereby potentially reducing the reliance on heuristics responsible for cognitive biases (Stanovich, 1999; Kahneman and Frederick, 2002). For example, Gino et al. (2008), in context of the outcome bias, show that the outcome information biases participants less when they use rational mindset. Given that the task conducted by participants of the experiment of Gino et al. (2008) is much less demanding than reviewing a paper, one could hypothesize that in peer review the cognitive heuristics may be overridden as reviewers naturally engage in rational and analytical thinking. However, our study identifies the presence of the resubmission bias in reviewers' judgements, demonstrating that the resubmission bias is strong enough to manifest even if reviewers put a non-trivial amount of cognitive effort into their work, as evidenced by the length of the reviews and expert judgements.

On general idea of reusing reviews In this work we do not aim to justify or oppose the idea of reusing reviews and do not say whether the resubmission bias is desirable or not. On the one hand, top-tier conferences are overloaded with substandard-quality papers making multiple rounds of submissions without major changes in a hope that they will get accepted at some point. Availability of past reviews could aid in identifying such submissions and optimizing the review efforts with respect to them, thereby reserving reviewers' effort for stronger papers. Additionally, when authors are aware of the resubmission bias, they may evaluate their work against higher standards and refrain from submitting works that are not yet in perfect shape to avoid the adverse effect on subsequent resubmissions, thereby further decreasing the load on reviewers.

On the other hand, the amount of randomness in the review process observed in several experiments (Lawrence and Cortes, 2014; Price, 2014; Pier et al., 2018) and in a multitude of anecdotal evidence

(for example, see a survey of Nobel Prize laureates in economics by Gans and Shepherd 1994) suggests that even strong papers that do not require revisions can be rejected by pure chance. It is also known (Travis and Collins, 1991; Lamont, 2009) that works that are novel or not mainstream, particularly those interdisciplinary in nature, face significantly higher difficulty in gaining acceptance. All these works can be put at significant disadvantage if reviewers are exposed to the resubmission signal as they may not have major aspects to be improved.

While in this work we do not resolve the aforementioned dichotomy, we identify the presence of the resubmission bias in reviewers' decisions and measure its effect size, letting the community and conference organizers decide on its desirability. Specifically, to perform the randomized controlled trial we choose perhaps the simplest format of exposing the previous outcome to reviewers, omitting the content of reviews received by submissions at past venues and withholding author statements that could explain changes (if any) made in the manuscript after the previous rejection.

In its simplicity, our stimulus can perhaps be seen as a point on a spectrum between two ways of exposure to the resubmission signal employed in the real world: the openreview system allows subsequent reviewers to see that the paper was rejected and exposes any damning or hypercritical past reviews without letting authors present their view beyond what is stated in the discussion. In contrast, some other venues allow authors to present their case more clearly. Exploring the wider range of the aforementioned spectrum and identifying the desirable point on it (if any) is an interesting direction for future work.

Overall, the questions of whether the resubmission bias is desirable and how to reuse the reviews from the past venues in a fair and efficient way is an open question that requires a discussion in the community. In this work, we provide some concrete evidence that can help conference organizers to make informed decisions when designing the peer-review system. A promising direction for the future work is to evaluate the proposal of reusing reviews in a holistic manner, studying the interactions between different aforementioned positive and negative effects both theoretically and empirically.

Chapter 7

A Large Scale Randomized Controlled Trial on Herding in Peer-Review Discussions

1 Introduction

In this chapter, we continue the effort on scrutinizing various parts of the review system and report the results of a large-scale randomized experiment on the discussion stage of scientific peer review.

Discussion stage in scientific peer review In many journals (e.g., Nature and PNAS) reviewers do not communicate with each other, and decisions are made by editors based on independent opinions of reviewers. In contrast, many conferences (which in computer science are considered to be a final destination for research and typically ranked higher than journals) and grant committees (e.g., NSF) implement an additional discussion stage that takes place after reviewers submit their initial independent reviews. The purpose of the discussion stage is to allow reviewers to exchange their opinions about the submission and correct each other's misconceptions. As a result of the discussion, reviewers are supposed to reach a consensus on the submission or boil their disagreement down to concrete arguments that can later be evaluated by chairs of the selection committee.

Given that independent opinions of reviewers often demonstrate a substantial amount of disagreement (Hofer et al., 2000; Obrecht et al., 2007; Fogelholm et al., 2011; Pier et al., 2017), the discussion stage may seem to be an appealing opportunity to reduce the load on editors and chairs by letting reviewers resolve their disagreements themselves. However, the aforementioned studies also demonstrate that while discussion increases the within-group agreement, the agreement between two groups of reviewers participating in parallel discussions of the same submission does *not* improve (and often the agreement across groups decreases after within-group discussions). This finding hints that reviewers within the group reach a consensus not because they identify the "correct" decision, but due to some other artifacts of group discussion. More generally, this observation agrees with a long line of research in psychology (Asch, 1951; Baron et al., 1996; Lorenz et al., 2011; Janis, 1982; Cialdini and Goldstein, 2004) which demonstrates that decisions stemming from a *group discussion* are susceptible to various biases related to social influence.

Importantly, lack of reliability in peer-review discussions may have far-reaching consequences not just for a particular submission, but also for career trajectories of researchers due to the widespread prevalence of the Matthew effect ("rich get richer") in academia (Merton, 1968; Squazzoni and Gandelli, 2012). Thus, in this work, we focus on investigating a cause of lack of reliability in peer review discussions and specifically focus on examining the presence of herding behaviour.

Herding behaviour We consider a specific manifestation of social influence that results in *"herding behaviour"* — an effect when agents are doing what others are doing rather than choosing the course of actions based on the information available to them (Banerjee, 1992; Bikhchandani and Sharma, 2000) — in peer-review discussions. A long line of work on human decision-making establishes the presence of various biases related to the first piece of information received by an individual (Tversky and Kahneman, 1974; Strack and Mussweiler, 1997; Mussweiler and Strack, 2001). In particular, these works show that an initial signal received by a decision-maker (even when being clearly unrelated to the underlying task) often has a disproportionately strong influence on the final decision. Projecting this observation on the group discussion setting, we study whether the first argument made in the discussion may exert an undue influence on the opinions of subsequent contributors, thereby leading to herding behaviour.

Past literature on group decision making (McGuire et al., 1987; Dubrovsky et al., 1991; Weisband, 1992) indeed suggests that the herding behaviour (titled "first advocacy" effect in these works) may be present in group discussions.[1] *With this motivation, in this chapter we analyze the presence of the herding effect in the discussion stage of peer review.*

To formalize the research question, we note that in the current review system it is often the job of the discussion chair (area chair in conferences, committee chair in grant proposal review) to maintain the order in which reviewers speak up in the discussion. In the absence of a standardized approach, some chairs may call upon the reviewer whose initial opinion is the most extreme to initiate the discussion, others may request the most positive or most negative reviewer to start. Another option for the discussion chair is to initiate the discussion themselves or to choose an initiator based on their seniority or expertise. In the presence of herding, the uncertainty in the choice of the strategy may impact the outcome of a paper or a grant proposal (which becomes dependent on the essentially arbitrary choice of the discussion initiator by the chair), thereby increasing the undesirable arbitrariness of the process. With the above motivation, in this work we aim at testing the presence of the herding effect in peer-review discussions, investigating the following research question:

> *Research Question:* Given a set of reviewers who participate in the discussion of a submission, does the final decision for the submission causally depend on the choice of the discussion initiator made by the discussion chair?

2 Methods

In this section, we outline the design of the experiment we conducted to investigate the research question of this chapter.

Setting of the experiment The experiment was conducted in the peer-review process of ICML 2020 — a flagship machine learning conference that receives thousands of paper submissions and manages a pool of

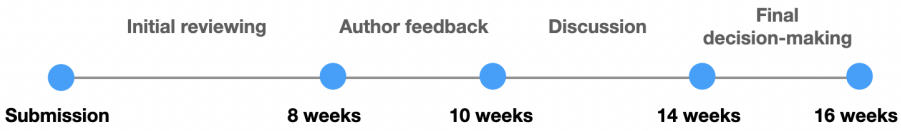


Figure 1: Timeline of the peer-review process of typical machine learning and artificial intelligence conferences. Upon the release of initial reviews, authors of papers have several days to write a response to reviewers, followed by the discussion stage. Finally, program chairs aggregate the results of the review process into final decisions. The duration of each stage varies across conferences, and this figure corresponds to the ICML 2020 review process with the duration of each stage rounded to weeks.

[1] We discuss these past works in more detail in Appendix A4.

thousands of reviewers. The ICML peer review is organized in a double-blind manner and, similar to most other top machine learning and artificial intelligence conferences, follows the timeline outlined in Figure 1.

Given our focus on the discussion dynamics, we describe the discussion process of ICML 2020 in more detail. During the discussion, reviewers (typically three or four per paper) and area chairs (one per paper; the role of area chairs is equivalent to that of associate editors in journal peer review) have access to the author feedback and are able to asynchronously communicate with each other (but not with authors) via a special online interface. The discussion between reviewers is anonymous (i.e., reviewers do not see each other's names), but area chairs know identities of reviewers and vice versa. For the papers assigned to them, each reviewer is expected to carefully read the author rebuttal as well as the reviews written by the other reviewers, and participate in the discussion.

Idea of the experiment To investigate a *causal* relationship between the choice of the discussion initiator and the outcome of a paper, the experiment we design in this work follows an A/B testing pipeline. Specifically, the set of papers involved in the experiment is split into two groups that receive different treatments, where the treatments are designed such that the difference in some observable outcomes of papers across groups is indicative of the presence of herding. For the reasons that are clarified below, we refer to the two groups of papers as $\mathcal{P}_+$ and $\mathcal{P}_-$.

The key challenge of this work is to design appropriate treatments and we begin by specifying requirements that the treatment scheme must satisfy. First, we intuitively expect herding (if present) to move the outcome of a discussion towards the opinion of the discussion initiator. Thus, to achieve a high detection power, we induce the following requirement:

> **Requirement 1:** The treatment scheme should induce a difference across two groups of papers in terms of the initial opinion of reviewers who initiate discussions within these groups. That is, in $\mathcal{P}_+$, reviewers with a positive initial opinion should initiate discussion more often while in $\mathcal{P}_-$, reviewers with a negative initial opinion should be more active in initiating discussion.

Moving on to the second requirement, we note that not every treatment scheme that satisfies Requirement 1 is valid for testing our research question. Indeed, one idiosyncrasy of conference peer review (at least in the machine learning and artificial intelligence areas) is that some reviewers may choose to ignore the discussion. In fact, the analysis of the review process of another leading machine learning conference NeurIPS 2016 (Shah et al., 2018) revealed that only 30% of 13,674 paper-reviewer pairs had a message posted by the reviewer in the associated discussion, showing that the set of discussion participants is generally a strict (and somewhat small) subset of reviewers assigned to the paper. Thus, in the conference peer-review setting, any intervention that impacts *the order* in which reviewers join the discussion may also impact *the population* of reviewers who participate in the discussion, thereby introducing a confounding factor in our analysis. To mitigate this issue, we introduce another requirement that must be satisfied by the treatments:

> **Requirement 2:** The treatment scheme should not introduce any difference across two groups of papers other than in the opinion of the discussion initiator. That is, distributions of the participating reviewers and other characteristics of the discussion should be the same across $\mathcal{P}_+$ and $\mathcal{P}_-$.

We note that in some settings (e.g., panel discussion in grant proposal review where all reviewers are required to participate) Requirement 2 (at least its part about the distribution of participating reviewers) is naturally met. However, as our experiment is implemented in conference peer review, we impose this requirement to remove associated confounding factors.

Treatment scheme Keeping Requirements 1 and 2 in mind, we now introduce the treatment scheme that we use in the experiment. The key component of our treatment scheme is a proxy towards the opinions of discussion participants. For this, recall that discussions begin *after* the initial reviews are submitted. Thus, we use overall scores given in the initial reviews to identify reviewers with the most positive and the most negative initial opinions about the paper. With these preliminaries, we execute the following treatments:

- Treatment for the positive group of papers $\mathcal{P}_+$
  - *Step 1* We ask a reviewer with the most positive initial opinion about the paper to *initiate* the discussion.
  - *Step 2* We then ask a reviewer with the most negative initial opinion to *contribute* to the discussion.
- Treatment for the negative group of papers $\mathcal{P}_-$
  - *Step 1* We ask a reviewer with the most negative initial opinion about the paper to *initiate* the discussion.
  - *Step 2* We then ask a reviewer with the most positive initial opinion to *contribute* to the discussion.

Both treatments consist of two steps: in the first step, a reviewer with an extreme opinion about the paper is asked to initiate the discussion. In the second step, the reviewer whose opinion is on the another extreme of the spectrum is asked to contribute to the discussion. Importantly, both treatments proceed to Step 2 (after some waiting time) even if the reviewer requested to initiate the discussion in the first step fails to fulfil the request.

Let us now discuss the design of the treatment scheme in light of the requirements we formulated earlier. First, observe that the order in which reviewers with the most positive and the most negative initial opinions about the paper are approached is different across treatments. Provided that a sufficiently large fraction of reviewers comply with our requests, we should expect to see a difference in opinions of discussion initiators across $\mathcal{P}_+$ and $\mathcal{P}_-$ as stated in Requirement 1. Second, we note that Step 1 of the treatments alone may induce a difference in the population of reviewers who *participate* in the discussion across $\mathcal{P}_+$ and $\mathcal{P}_-$. To remove this confounding factor, both treatments implement Step 2 that is designed to "balance" the impact of Step 1, thereby aiming to satisfy Requirement 2.

Overall, we designed the intervention with a goal of satisfying the aforementioned requirements. However, a priori we cannot guarantee that these requirements are indeed satisfied in the real experiment. For example, Requirement 2 is not satisfied if the balancing part (Step 2) of the treatments fails to equalize the populations of participating reviewers across conditions. To further support our experimental methodology, in Section 3 we empirically check for satisfaction of the stated requirements.

As a final remark, we note that ideally, we would like to fully control the choice of the discussion initiator for each paper, thereby implementing the conventional randomized controlled experiment. However, in the conference peer-review setup, organizers have only limited ability to impact the behaviour of reviewers. Thus, our treatment scheme follows the concept of the randomized encouragement design (West et al., 2008), in which the treatments are not enforced, but encouraged.

Details of the experiment Let us now clarify some important aspects of the experiment.

- **Sample size** In 2020, ICML received more than 5,000 paper submissions out of which 4,625 papers remained in the process (i.e., were not withdrawn) at the time when the discussion period began. While we would like to run the experiment using all these papers, we note that some reviewers may be the most positive or the most negative reviewers for multiple papers, being overburdened with requests to initiate (contribute to) the discussion of these papers.

  To limit the additional load on reviewers induced by our experiment, for each reviewer we limit the number of papers the reviewer is asked to initiate the discussion or contribute to the discussion to one each (in total, each reviewer may receive at most two requests). This condition puts the limit on the number of papers we can use in the experiment. Consequently, to compensate for the potential decrease of power, we focus the scope of the experiment on the *borderline* papers with some disagreement between reviewers as we expect the effect (if any) to be the most prominent in these papers. As a result, we end up with a pool $\mathcal{P}$ of 1,544 papers used in the experiment that we split into $\mathcal{P}_+$ and $\mathcal{P}_-$ uniformly at random subject to the aforementioned constraint on the additional reviewer load (see criteria for borderline papers and other details in Appendix A2). The experiment also involved 2,797 reviewers who participated in the discussion of at least one paper from the experimental pool.

- **Implementation of the treatment scheme** To ensure that the behaviour of area chairs and reviewers is not altered by awareness of the experiment, we implement the treatment scheme at the level of program chairs and do not notify other committee members about the experiment. Specifically, the requests to initiate or contribute to the discussion were sent over email on behalf of the program chairs. To further maximize the power of our test, we first open the discussion interface without notifying the general pool of reviewers, and implement Step 1 of both treatments, giving reviewers more time to fulfil our request.

- **Statistical analysis** We employ two-sided permutation test (Fisher, 1935) to test for difference across groups $\mathcal{P}_+$ and $\mathcal{P}_-$. Specifically, the analysis is conducted at the level of papers and we compute p values over 10,000 permutations of papers across conditions.

- **Data and code availability** We note that the release of experimental data would compromise the reviewers' confidentiality. Thus, following prior works that empirically analyze the conference peer-review process (Tomkins et al., 2017; Shah et al., 2018; Lawrence and Cortes, 2014), and complying with the conference's policy, we are unable to release the data and code from the experiment.

- **Avoiding conflict of interests** Three members of the study team were involved in the ICML decision-making process. Nihar Shah served as an area chair and Aarti Singh and Hal Daumé III were program chairs. To avoid the conflict of interests, Nihar Shah, Aarti Singh and Hal Daumé III were not aware of what papers were used in the experiment. Moreover, we excluded papers chaired by Nihar Shah from the analysis.

- **Ethics statement** Finally, we note that if the herding effect is present in ICML discussions, our intervention may place some papers ($\mathcal{P}_-$) at a disadvantage, while giving an unfair advantage to other papers ($\mathcal{P}_+$). We carefully considered this risk when designing the experiment and concluded that it does not exceed the risk that is otherwise present in the review process. Indeed, currently, there is no standardized approach towards choosing discussion initiators and it is often the job of the discussion chair to maintain the order in which reviewers speak up in the discussion. Different discussion chairs implement different strategies and, under the presence of herding, this variability results in randomness in decisions. Hence, even if herding is present, the impact of our intervention would not go beyond the unfairness that is otherwise present in the review system. This study was analyzed by Carnegie Mellon University's Institutional Review Board that agreed with our reasoning and approved the study.

  Additionally, to avoid the Hawthorne effect, we employ deception and do not notify reviewers about the experiment. The deception was approved by Carnegie Mellon University's Institutional Review Board and we debriefed all participants after the end of the review process.

More details on the experiment design and exact schedule of our intervention are given in Appendix A2.

3 Results of the experiment

In this section, we present the results of the experiment. First, we empirically check whether our treatment scheme satisfies requirements formulated in Section 2. In that, we begin with some general statistics on the discussion process to check satisfaction of Requirement 2 (Section 3.1). We then discuss the efficacy of the intervention we employed (Section 3.2) and confirm that Requirement 1 is well-satisfied. Finally, we conclude with the analysis of the research question we study in this work (Section 3.3). For brevity, for any paper, we use R_+ (respectively, R_-) to refer to the reviewer with the most positive (respectively, negative) initial opinion about the paper as determined by the overall scores given in the initial reviews.

3.1 Preliminary analysis (data to check for satisfaction of Requirement 2)

We begin with providing data-based evidence which lets us verify whether our treatment scheme satisfies Requirement 2, that is, does not introduce differences across $\mathcal{P}_+$ and $\mathcal{P}_-$ in characteristics other than the

Comparative Statistics on the Discussion Process

	$\mathcal{P}_+$	$\mathcal{P}_-$
1. Number of papers	755	789
2. Mean initial score (all revs)	3.52	3.52
3. Standard deviation of initial scores (all revs)	1.12	1.11
4. Mean initial score (revs in discussion)	3.44	3.46
5. Percentage of papers with active discussion	97%	97%
6. Mean number of discussion participants (revs + area chairs)	3.14	3.06
7. Mean discussion length (# messages)	4.41	4.24
8. Percentage of papers with R_+ active in discussion	79%	79%
9. Percentage of papers with R_- active in discussion	87%	84%

Table 1: Comparison of some discussion statistics between two groups of papers ($\mathcal{P}_+$ and $\mathcal{P}_-$) receiving different treatments. Except Row 4, all values are computed using all papers including those with no discussion. Permutation test at the level 0.05 (two-sided; before multiple-testing adjustment) with 10,000 iterations does not reveal significant differences between conditions in any of the criteria.

opinion of the discussion initiator. For this, Table 1 provides some comparative statistics on the discussion process for the papers involved in the experiment. Overall, we note that many important parameters of the discussion are similar across the two conditions. This observation provides quantitative evidence that the randomization of papers to conditions occurred successfully and Requirement 2 is satisfied.

Looking closer at the most relevant parameters, we first focus on Rows 2 and 4 of that compare mean overall scores (the overall score takes integer values from 1 to 6 where larger values indicate higher quality) given by reviewers in the initial reviews, that is, before reviewers got to see the other reviews and the author feedback. Interestingly, mean initial scores given by reviewers who participate in the discussion (Row 4) appear to be lower than mean scores computed over all reviewers (Row 2), suggesting that those who give lower scores are more active in discussing papers (see also Rows 8 and 9). However, there is no significant difference between two groups of papers ($\mathcal{P}_+$ and $\mathcal{P}_-$) in these values. Hence, the data indicates that this trend is independent of the choice of the treatment as requested by Requirement 2.

Next, the activity of reviewers with the most positive (respectively, negative) initial opinion in the discussion (Rows 8 and 9) is similar across the two groups of papers. This observation supports the intuition that our treatment scheme does not introduce a difference across conditions in the distributions of reviewers who participate in the discussion. Finally, we note that most of the papers used in the experiment had some discussion and the length of the discussion is similar across conditions (Rows 5 and 7). With this observation, we conclude this section and note that data reported in Table 1 supports our treatment scheme in light of Requirement 2.

3.2 Efficacy of the intervention (data to check for satisfaction of Requirement 1)

In the previous section we confirmed that our intervention did not introduce a difference across conditions in metrics such as intensity of discussions and the population of participating reviewers. This observation suggests that our treatment scheme satisfied Requirement 2 and indicates the appropriateness of our intervention. However, in order for the experiment to have sufficient power to detect the effect, the intervention needs to satisfy Requirement 1 and introduce a difference across conditions in the order in which reviewers join the discussion of the papers. Indeed, if all the emails we sent to reviewers were ignored (i.e., our attempt to impact the order failed), the subsequent analysis will not detect the phenomena even when the phenomena is present.

Table 2 reports relevant statistics and indicates a large difference between positive and negative groups of papers, suggesting that our intervention did impact the order in which reviewers joined the discussion. Indeed, Rows 2 and 3 show that more than half of discussions in the positive group $\mathcal{P}_+$ were initiated by reviewers

Does the Intervention Affect who Initiates the Discussion?

	$\mathcal{P}_+$	$\mathcal{P}_-$	Δ	Δ 95% CI	p value
1. Mean initial score (initiator)	4.03	2.76	1.27	$[1.15, 1.39]$	$< .001$
2. Percentage of discussions initiated by R_+	53%	9%	0.44	$[0.39, 0.48]$	$< .001$
3. Percentage of discussions initiated by R_-	15%	59%	-0.44	$[-0.48, -0.39]$	$< .001$

Table 2: The impact of the intervention on who initiates the discussion. To compute values for Row 1, we use 1,140 papers for which (i) the discussion was initiated, and (ii) the discussion initiator was a reviewer (and not the area chair). For the last two rows, we use all papers including those with no discussion. Bootstrapped confidence intervals are constructed for the difference of the relevant quantities between conditions. All p values for the difference between $\mathcal{P}_+$ and $\mathcal{P}_-$ are two-sided and computed using the permutation test with 10,000 iterations.

with the most positive initial opinion and only 15% were initiated by reviewers with the most negative initial opinion. Conversely, in the negative group $\mathcal{P}_-$, reviewers with the most negative initial opinion initiated the discussion a lot more often than reviewers with the positive initial opinion. This asymmetry results in a considerable difference of scores given by discussion initiators in their initial reviews (Row 1). Overall, Table 2 suggests that our treatment scheme satisfied Requirement 1. Coupled with a large sample size, this observation ensures that our experiment has a strong detection power.

3.3 Main causal analysis

Having confirmed that the intervention we implemented in the experiment reasonably satisfies the necessary requirements, we now continue to the analysis directly related to the research question we study in this work. Specifically, as we explained in the introduction and in Section 2, if herding behaviour exists, we expect it to manifest in the final decisions being disproportionately influenced by the opinion of the discussion initiator. Hence, given that for the positive group of papers $\mathcal{P}_+$ the initial opinion of the discussion initiator was on average significantly more positive than that of initiators of discussions for the negative group of papers $\mathcal{P}_-$, we expect (if herding exists) to observe a disparity in the eventual acceptance rates between conditions.

Table 3 formalizes the intuition and performs the comparison of acceptance rates across papers that received different treatments (Row 1). Additionally, Table 3 displays the updates of the scores made by reviewers (Rows 2–5). Based on the presented data, we make two observations:

- First, the data does not indicate a statistically significant difference between acceptance rates in the

Does the Intervention Affect the Outcome of Papers?

	$\mathcal{P}_+$	$\mathcal{P}_-$	Δ	Δ 95% CI	p value
1. Acceptance rate	0.21	0.25	-0.04	$[-0.08, 0.01]$	.122
2. Change in mean score (initiator)	-0.10	0.20	-0.30	$[-0.37, -0.23]$	$< .001$
3. Change in mean score (all revs)	0.01	0.01	0.00	$[-0.03, 0.04]$	.949
4. Change in mean score (revs in discussion)	0.03	0.02	0.01	$[-0.04, 0.06]$	.697
5. Change in standard deviation of scores (all revs)	-0.23	-0.21	-0.02	$[-0.05, 0.02]$	.296

Table 3: The impact of the intervention on the final outcome of papers. For Row 2, we use 1,140 papers for which (i) the discussion was initiated, and (ii) the discussion initiator was a reviewer (and not the area chair). For Row 4, we use papers with discussion. For all other rows, we use all papers including those with no discussion. Bootstrapped confidence intervals are constructed for the difference of the relevant quantities between conditions. All p values for the difference between $\mathcal{P}_+$ and $\mathcal{P}_-$ are two-sided and computed using the permutation test with 10,000 iterations.

two groups of papers ($\mathcal{P}_+$ versus $\mathcal{P}_-$). Thus, despite discussion initiators had considerably different initial opinions about papers from $\mathcal{P}_+$ and $\mathcal{P}_-$, the outcomes of the discussion were distributed similarly across conditions.

- Second, the data on the score updates suggests that in their final evaluations, reviewers tend to converge to the mean of initial independent opinions irrespective of the treatment we applied to a paper. Indeed, Row 2 demonstrates that the initiators of discussions update the scores towards the mean of all initial scores. Next, Rows 3 and 4 show that a significant update made by the discussion initiators is compensated by the update made by other reviewers, such that the overall amount of change in the mean scores is negligible. As expected, the outlined dynamics result in a significant decrease in the variance of scores per paper, but the effect is the same for both groups of papers (Row 5).

Overall, we find no evidence of herding in the discussion phase of ICML 2020 peer review.

4 Discussion

The experiment we conducted in the present work aims at identifying the herding behaviour in the discussions of the ICML 2020 conference. The results presented in Section 3 show that while we managed to achieve an imbalance in the opinion of the discussion initiators across conditions, and despite past work having documented an undue influence of the first piece of information on the final decision (Tversky and Kahneman, 1974; Strack and Mussweiler, 1997; Mussweiler and Strack, 2001) in various other settings and applications, the difference in the acceptance rates is not significant and hence there is no evidence of herding in peer review. The absence of the effect suggests that the absence of a unified approach towards discussion management does not result in an increased arbitrariness of the resulting decisions. Thus, the question of identifying the source of the spurious agreement between peer reviewers observed in past works (Hofer et al., 2000; Obrecht et al., 2007; Fogelholm et al., 2011; Pier et al., 2017) remains open.

Finally, we urge the reader to be aware of the caveats that we listed in Appendix A1. While we do not believe that any of these caveats affected the outcome of this experiment in a significant way, they may be important for the design of follow-up studies.

Appendix

We provide supplementary materials and additional discussion. In Appendix A1, we discuss several caveats that should be taken into account when interpreting the results of this work. In Appendix A2, we give additional details on the experiment, including its timeline and selection criteria for participating papers $\mathcal{P}$. Appendix A3 provides additional analysis of the collected data (see Caveat 4 in Appendix A1 for motivation). Finally, in Appendix A4 we discuss the relationship between the present study and past works on group discussion.

A1 Caveats regarding the design and analysis of the experiment

Given that our intervention induced the strong difference in the order in which reviewers joined the discussions (see Table 2), the absence of difference in score updates (see Table 3) allows us to conclude with a high degree of confidence that the choice of the discussion-management strategy does not impact the way reviewers update their scores. Of course, herding (if present) does not necessarily need to manifest in how reviewers change their scores after the discussion. Instead, it can change some other characteristics such as what reviewers write in the textual messages which are later analyzed by the area and program chairs who make the final decisions. To account for these potential manifestations, we compared acceptance rates between the groups

of papers (see Row 1 of Table 3) and observed some difference in this quantity. However, this difference does not appear significant despite the large sample size we had in the experiment, suggesting that even if present, the effect has at most small size. That being said, we urge the reader to be aware of the following caveats when interpreting the results of this work.

Caveat 1. The design of the intervention. Recall that our research question defines herding as a conditional dependence of the outcome of a paper on the choice of the discussion initiator. The test and the intervention we designed attempt to compare the outcomes of papers when the discussion is initiated by the most positive versus the most negative reviewers with the motivation that this difference is expected to be the largest in the presence of herding. Strictly speaking, the absence of a difference between these choices of the initiators does not imply the absence of the difference between any other choices of the initiators: for example, it is possible that the outcome of a paper would be impacted differently if we asked the reviewer with a non-extreme score to initiate the discussion.

Caveat 2. The choice of papers. As noted in Section 2, in this experiment we tried to identify a set of *borderline* papers as these papers are more susceptible to the impact of the herding effect if it is present. However, our choice of the borderline papers was based on some indirect indicators and hence we could potentially fail to uncover the set of true borderline papers which could reduce the power of our test.

To evaluate our choice of borderline papers, we use a rough classification of submissions into clear and borderline cases made by the area chairs. Note that this classification was performed after the discussion stage which could resolve the uncertainty present before the discussion stage when we selected the participating papers. Hence, the fraction of borderline papers in the area chairs' classification is a conservative estimate of the pre-discussion fraction of borderline papers. Nonetheless, 30% of submissions used in the experiment were classified by the area chairs as borderline cases in contrast to 18% of those not involved in the experiment ($\Delta = 0.12, p = .002$). Hence, our choice of the borderline papers was better than random and the set of the participating papers $\mathcal{P}$ contained a large fraction of papers for which the decisions were not clear before the discussion.

Caveat 3. Satisfaction of Requirement 2 The validity of the conclusions we make is based on the assumption that our treatment scheme satisfied requirements formulated in Section 2. Note that a violation of these requirements not only could increase the false alarm probability, but could also reduce the power of the test. The data we analyzed in Section 3.1 and Section 3.2 strongly supports the satisfaction of Requirements 1 and 2. However, as a note of caution, we remark that there is some space for potential violations of Requirement 2. Indeed, in Table 1 we establish that the *marginal* values of relevant indicators of discussion activity are similar across groups. However, this observation does not imply that the value of these indicators for each individual paper would not change if that paper was placed in the other condition. Hence, the the outcome of this study should be considered together with this opportunity for the violation of Requirement 2.

Caveat 4. Spurious correlations induced by reviewer identity. In peer review, each reviewer participates in the discussion of multiple papers. Similarly, each area chair manages several papers. Hence, strictly speaking, the outcomes of two papers that have at least one reviewer in common (are managed by the same area chair) may not be statistically independent due to correlations introduced by the reviewer (area chair) identities. Additionally, the limit on the additional burden on reviewers introduced by our experiment (see last subsection of Section 2) makes allocation of papers $\mathcal{P}$ into groups $\mathcal{P}_+$ and $\mathcal{P}_-$ not fully uniform random (some pairs of papers may be required to be placed in the same group to not exceed that limit). These issues put a strain on the testing procedure because in contrast to the vanilla A/B testing framework which assumes that samples are independent of each other, in our case we receive correlated samples. In the domain of empirical studies of the peer-review procedure (Shah et al., 2018; Tomkins et al., 2017; Lawrence and Cortes, 2014) such spurious correlations are usually tolerated, because otherwise the sample size would be negligible. Additionally, simulations performed by Stelmakh et al. (2019) demonstrate that unless reviewers are involved in the discussion of dozens of submissions, the impact of such spurious correlations is limited.

Nevertheless, in this work we take some additional steps to minimize the impact of these spurious correlations. To this end, we simultaneously also perform the analysis on a subset of 937 papers $\mathcal{P}^* = \mathcal{P}^*_+ \cup \mathcal{P}^*_-$, where $\mathcal{P}^*_+ \subset \mathcal{P}_+$ and $\mathcal{P}^*_- \subset \mathcal{P}_-$ are constructed such that each reviewer is requested to initiate the discussion

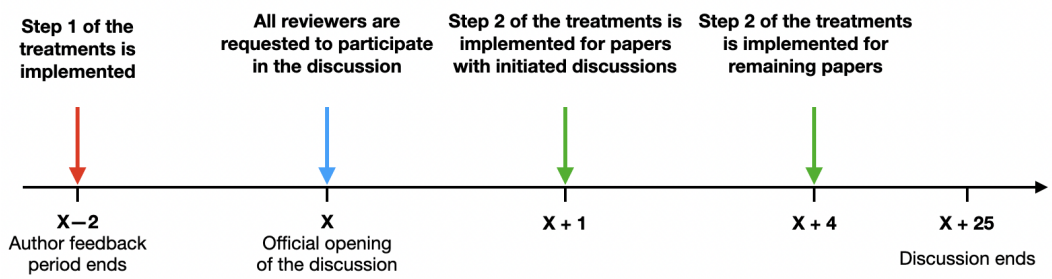


Figure 2: Timeline of the experiment. Day X is the day of the official discussion opening.

or contribute to the discussion of at most one paper from $\mathcal{P}^*$ in total.[2] This additional reduction of the sample size allows us to limit the impact of the reviewer identity on the outcome of submissions. Of course, by doing so we do not guarantee that there is no reviewer who participates in the discussion of more than one paper from the set $\mathcal{P}^*$, but we guarantee that the discussion participants who are targeted by our treatments are unique. Appendix A3 gives more details on how sets $\mathcal{P}^*_-$ and $\mathcal{P}^*_+$ were constructed and presents the results of additional analysis on this subset of the papers. Importantly, we note that this additional analysis leads to the same conclusions as in Section 3.

Caveat 5. Opinion of the discussion initiator. In this work we used the scores given by reviewers in the initial reviews to infer the pre-discussion opinion of reviewers and assumed that reviewers begin the discussion from advocating these opinions. However, the fact that a reviewer has some pre-discussion opinion does not guarantee that they advocate the same position in the discussion because the latter is also influenced by other reviewers' reviews and the author feedback. Indeed, past research (Teplitskiy et al., 2019) suggests that reviewers do listen to each other and may update their initial independent opinions in light of opinions expressed in initial reviews of other reviewers. The data we obtained in the experiment suggests that while such updates take place, their magnitude is small enough and does not break our intervention. Indeed, Row 2 of Table 3 and Row 1 of Table 2 indicate that reviewers with extreme pre-discussion opinions remain on the different sides of the mean pre-discussion group opinion (Row 2 of Table 1) even according to the final scores. Thus, we conclude that our intervention succeeded in creating a difference in opinions of discussion initiators across groups.

Caveat 6. Alternative model of herding. In this work we assume that the herding behaviour in peer review manifests in final decisions being moved towards the position of the reviewer who initiates the discussion. However, the data presented in Table 3 shows that initiators of the discussion tend to slightly update their scores towards the mean of initial scores given by all reviewers. Hence, an alternative model of the herding behaviour is that the sentiment demonstrated by the initiating reviewer carries over to other reviewers who could change their behaviour accordingly. For example, the positive score update of initiators with a negative initial opinion may demonstrate a positive sentiment, which could affect opinions of other reviewers in a positive way. Under this alternative model of herding, we would expect papers from the negative group $\mathcal{P}_-$ to enjoy a higher acceptance rate than their counterparts from the positive group $\mathcal{P}_+$. While this agrees with the observed acceptance rates reported in Table 3, we reiterate that the difference between the acceptance rates is not significant and the effect size is small, so our test does not provide evidence in support of this alternative model either.

A2 Additional details on the experiment

In this section, we provide additional details on the experimental procedure: timeline of the intervention and selection criteria for papers that were used in the experiment.

[2]Compare to $\mathcal{P} = \mathcal{P}_+ \cup \mathcal{P}_-$ which is constructed such that each reviewer is asked to initiate the discussion for at most one paper from $\mathcal{P}$ and contribute to the discussion of at most one paper from $\mathcal{P}$ (at most two requests in total).

Timeline of the experiment The experiment was conducted over the course of 4 weeks of the ICML 2020 discussion process. Figure 2 depicts the pipeline implemented in the experiment To further increase the power of the experiment, we unofficially opened the discussion portal two days before the scheduled beginning of the discussion and executed Step 1 of both treatments by sending emails to corresponding reviewers. Step 2 of the treatments was then executed in two stages: first, one day after the official opening of the discussion, we executed Step 2 for papers with already initiated discussion. We then waited for three more days before executing Step 2 for all the remaining papers, irrespective of whether the discussion was initiated or not.

Through the first ten days of the experiment, we sent reminders to the reviewers who have not fulfilled our request to initiate or contribute to the discussion of the corresponding papers. In order to avoid a disproportional impact on the discussion participants across two groups of papers, we ensured that the total number of reminders is the same for reviewers who were asked to initiate the discussion and reviewers who were asked to contribute to the discussion.

Construction of the sets $\mathcal{P}_+$ and $\mathcal{P}_-$ We now specify how the set $\mathcal{P} = \mathcal{P}_+ \cup \mathcal{P}_-$ of participating papers (introduced in Section 2) was constructed from the set of all $m = 4,625$ papers not withdrawn from ICML 2020 by the beginning of the discussion period. For this, recall that in Section 2 we mentioned that in order to limit the additional load on reviewers, we require that each reviewer is asked to initiate the discussion for at most one paper from $\mathcal{P}$ and contribute to the discussion of at most one paper from $\mathcal{P}$ (at most two requests in total). To meet this requirement, we construct the target set of papers $\mathcal{P}$ such that each reviewer is the most positive reviewer for at most one paper from $\mathcal{P}$ and the most negative reviewer for at most one paper from $\mathcal{P}$ (the most extreme reviewer for at most two papers from $\mathcal{P}$). To compensate for the associated decrease in the sample size, we design the selection procedure such that $\mathcal{P}$ consists of borderline papers for which the herding effect (if present) is expected to be the most prominent. Having $\mathcal{P}$ constructed, we split it into $\mathcal{P}_+$ and $\mathcal{P}_-$ uniformly at random subject to the aforementioned requirement on the additional burden on reviewers introduced by our intervention. More formally, sets $\mathcal{P}_+$ and $\mathcal{P}_-$ were constructed using the following three-step procedure:

Step 1. First, we identify the set of borderline papers as follows. The overall scores given in the initial reviews were in the set $\{1, 2, \dots, 6\}$ so for each paper $i \in [m]$, we let λ_i to denote the number of reviewers assigned to the paper (typically λ_i equals 3 or 4) and let $(\theta_1, \theta_2, \dots, \theta_{\lambda_i}) \in \{1, 2, \dots, 6\}^{\lambda_i}$ to denote the collection of overall scores given to paper i in initial reviews.[3] With this notation, using acceptance statistics of the ICML 2019 conference, we construct a set of borderline papers $\mathcal{T}$ by identifying submissions that satisfy the following criteria:

C1 The mean overall score is such that in ICML 2019 the paper is in the borderline category:

$$\frac{1}{\lambda_i} \sum_{j=1}^{\lambda_i} \theta_j \in [2.7, 4.5].$$

C2 The minimum and maximum overall scores are on the different sides of the decision spectrum:

$$\max(\theta_1, \theta_2, \dots, \theta_{\lambda_i}) \geq 4 \quad \text{and} \quad \min(\theta_1, \theta_2, \dots, \theta_{\lambda_i}) \leq 3.$$

Note that for each borderline paper $i \in \mathcal{T}$ we are guaranteed that there is some disagreement between reviewers.

Step 2. Having the set of borderline papers $\mathcal{T}$ defined, we construct $\mathcal{P}$ by greedily finding a subset of $\mathcal{T}$ that satisfies the requirement of each reviewer being the most positive reviewer for at most one paper from this subset and the most negative reviewer for at most one paper from this subset.

[3]Here, we adopt the standard notation $[\nu] = \{1, 2, \dots, \nu\}$ for any positive integer ν.

	$\mathcal{P}^*_+$	$\mathcal{P}^*_-$
1. Number of papers	460	477
2. Mean initial score (all revs)	3.51	3.52
3. Standard deviation of initial scores (all revs)	1.22	1.20
4. Mean initial score (revs in discussion)	3.42	3.45
5. Percentage of papers with active discussion	96%	98%
6. Mean number of discussion participants (revs + area chairs)	3.16	3.10
7. Mean discussion length (# messages)	4.52	4.25
8. Percentage of papers with R_+ active in discussion	78%	79%
9. Percentage of papers with R_- active in discussion	87%	85%

Table 4: Comparison of some discussion statistics between two groups of papers ($\mathcal{P}^*_+$ and $\mathcal{P}^*_-$) receiving different treatments. Except Row 4, all values are computed using all papers including those with no discussion. Permutation test at the level 0.05 (two-sided; before multiple-testing adjustment) with 10,000 iterations does not reveal significant differences between conditions in any of the criteria.

Step 3. Finally, we split $\mathcal{P}$ into $\mathcal{P}_+$ and $\mathcal{P}_-$ uniformly at random subject to the constraint that each reviewer is requested to initiate the discussion of at most one paper and contribute to the discussion of at most one paper (in total, each reviewer receives at most two requests).

A3 Additional analysis

In this section, we report additional analysis that aims to alleviate confounding factors mentioned in Caveat 4 of Appendix A1. Specifically, we replicate the analysis presented in Section 3, conditioning on a subset 937 of papers $\mathcal{P}^* = \mathcal{P}^*_+ \cup \mathcal{P}^*_-$ (see Caveat 4 in Appendix A1 for motivation and specification of the set $\mathcal{P}^*$).

Construction of the sets $\mathcal{P}^*_+$ and $\mathcal{P}^*_-$ Recall that $\mathcal{P}^*$ is a subset of the original set of participating papers $\mathcal{P}$ greedily selected such that each reviewer is only approached once by our treatments (that is, is asked to initiate the discussion or contribute to the discussion of at most one paper from $\mathcal{P}^*$ in total). Given that there exist many such subsets of approximately the same size, we facilitate tie-breaking by additionally requesting that for each paper $i \in \mathcal{P}^*$ the most positive and most negative reviewers disagree in the initial reviews by at least 2 points:

$$\max(\theta_1, \theta_2, \ldots, \theta_{\lambda_i}) - \min(\theta_1, \theta_2, \ldots, \theta_{\lambda_i}) \geq 2.$$

Hence, set $\mathcal{P}^*$ has an additional property of containing papers with high disagreement between reviewers in the initial reviews.

Additional analysis on $\mathcal{P}^*$ We now replicate the analysis described in Section 3, conditioning on the set of papers $\mathcal{P}^*$. First, mirroring the analysis on the full set of participating papers (Table 1), Table 4 indicates that various parameters of the discussion are similar across the two conditions even after we condition on the target set of papers $\mathcal{P}^*$. Hence, we also conclude that data supports our treatment scheme in light of Requirement 2 and the intervention did not result in a difference across conditions in the distributions of reviewers who participate in the discussion.

Next, we investigate the efficacy of our intervention and proceed to Table 5 that compares relevant statistics. Observe that the values in Table 5 are very similar to those reported in Table 2, suggesting that the intervention continues to introduce the required difference in opinions of discussion initiators between the groups of papers even when we zoom in on the target subset of papers $\mathcal{P}^*$. Thus, we conclude that Requirement 1 remains satisfied and out test continues to posses strong power.

Having confirmed the efficacy of the intervention, we proceed to the comparison of the outcomes of submissions across $\mathcal{P}^*_-$ and $\mathcal{P}^*_+$. The results presented in Table 6 mimic those reported in Table 3, suggesting that conditioning on the set of papers $\mathcal{P}^*$ does not qualitative change the findings. The most notable distinction between Table 6 and Table 3 is that the significance of the difference in acceptance rates (Row 1) becomes closer to the threshold of 0.05 after we condition on $\mathcal{P}^*$, but still does not cross it. Given that the number of submissions involved in the experiment is large, we conclude that we do not observe strong evidence of the herding behaviour even after conditioning on the set of papers $\mathcal{P}^*$.

A4 Relation to past work on group discussion

Past literature suggests the presence of herding behaviour in group discussions. McGuire et al. (1987) observe that the first solution proposed to a group predicts the group decision better than an aggregate of initial opinions independently expressed in a pre-discussion survey. Dubrovsky et al. (1991) document an impact of the interplay between the status of discussion participants and the opinion of the group member who proposed the first concrete solution on the final group decision. Closest to the present work, Weisband (1992) further investigates the herding effect in a semi-randomized controlled trial and declares that the initiators of discussion manage to influence the group opinion when they step in after an initial general discussion of the problem, that is, when they have some understanding of the general opinions of other discussants, but no concrete decisions have been proposed.

In contrast, in this experiment, we did not detect herding in the peer-review discussion. Let us now discuss the relationship of the current experiment to these past works. First, we note that the papers of McGuire et al. (1987) and Dubrovsky et al. (1991) study the herding effect when the discussion initiators are self-selected. The difference between the self-selected and assigned initiators appears to be significant, because the former may be associated with other personal qualities such as assertiveness and energy. Hence, our work is not directly comparable to these studies as we attempt to randomize the identity of the discussion initiator.

The experiment of Weisband (1992) employs randomization and in addition to the self-selection scenario considers the setup in which the first person to propose the solution to the group is chosen uniformly at random. This work finds that the randomly assigned initiator exerts much smaller influence on the group decision than the self-selected initiator. We caveat however, that in the experiment of Weisband (1992), the fact that the initiator is selected at random was known to the whole group before the beginning of the discussion. Hence, it is plausible that other group members did not perceive the initiator as a leader and could adjust their behaviour accordingly (Weisband, 1992). In contrast, in the present experiment the non-initiating reviewers were not aware of the intervention and hence from their point of view the assigned initiator of the discussion possessed all the properties of the self-selected initiator (Hollander, 1978).

Finally, there is a subtle difference between the definition of the herding effect made by Weisband (1992) and the definition we use in this work. According to Weisband (1992), the herding is present when the first

Does the Intervention Affect who Initiates the Discussion?

	$\mathcal{P}^*_+$	$\mathcal{P}^*_-$	Δ	Δ 95% CI	p value
1. Mean initial score (initiator)	4.09	2.69	1.40	$[1.24, 1.55]$	$< .001$
2. Percentage of discussions initiated by R_+	53%	11%	0.42	$[0.36, 0.47]$	$< .001$
3. Percentage of discussions initiated by R_-	16%	60%	-0.44	$[-0.49, -0.38]$	$< .001$

Table 5: The impact of the intervention on who initiates the discussion, conditioned on the subset of papers $\mathcal{P}^*$. To compute values for Row 1, we use 698 papers for which (i) the discussion was initiated, and (ii) the discussion initiator was a reviewer (and not the area chair). For the last two rows, we use all papers including those with no discussion. Bootstrapped confidence intervals are constructed for the difference of the relevant quantities between conditions. All p values for the difference between $\mathcal{P}^*_+$ and $\mathcal{P}^*_-$ are two-sided and computed using the permutation test with 10,000 iterations.

	$\mathcal{P}^*_+$	$\mathcal{P}^*_-$	Δ	Δ 95% CI	p value
1. Acceptance rate	0.21	0.26	-0.05	$[-0.11, 0.00]$	.079
2. Change in mean score (initiator)	-0.11	0.21	-0.32	$[-0.42, -0.22]$	$< .001$
3. Change in mean score (all revs)	0.01	0.01	0.00	$[-0.05, 0.05]$	.925
4. Change in mean score (revs in discussion)	0.02	0.02	0.00	$[-0.06, 0.07]$	.867
5. Change in standard deviation of scores (all revs)	-0.26	-0.25	-0.01	$[-0.06, 0.03]$	.560

Table 6: The impact of the intervention on the final outcome of papers, conditioned on the subset of papers $\mathcal{P}^*$. For Row 2, we use 698 papers for which (i) the discussion was initiated, and (ii) the discussion initiator was a reviewer (and not the area chair). For Row 4, we use papers with discussion. For all other rows, we use all papers including those with no discussion. Bootstrapped confidence intervals are constructed for the difference of the relevant quantities between conditions. All p values for the difference between $\mathcal{P}^*_+$ and $\mathcal{P}^*_-$ are two-sided and computed using the permutation test with 10,000 iterations.

solution formulated in the group discussion predicts the group final decision better than the mean of the pre-discussion independent opinions. Note that according to this definition, the herding may be present even if the first solution proposed to the group is independent of who is selected to formulate this opinion, that is, even when all discussants would propose the same solution should they be selected to start the discussion. In contrast, in our settings it is natural to define herding to be present only when the opinion of the discussion initiator is different depending on who is selected to initiate the discussion, because the goal of the present work is to inform the discussion chairs about the potential consequences of their discussion initiating strategy.

In addition to the aforementioned distinctions from the past work, we note that in the present experiment reviewers are engaged in a much more analytical task as compared to the previous works in which some toy problems were used to study the discussion dynamics. Hence, the absence of the herding behaviour in peer review may be due to the fact that reviewers have a rational mindset which is hypothesized to reduce a reliance on heuristics responsible for various cognitive biases (Stanovich, 1999; Kahneman and Frederick, 2002).

Beyond testing for herding, in this chapter we also document effects predicted by past works on discussion in peer review (Teplitskiy et al., 2019; Hofer et al., 2000; Obrecht et al., 2007; Fogelholm et al., 2011; Pier et al., 2017): reviewers tend to update their scores towards the consensus pre-discussion opinion, and the discussion increases the agreement among reviewers. Coupled with the observation made in these past works that an increased agreement does not necessarily result in an increased reliability of the decision, our findings highlight an importance of additional research on the discussion dynamics in peer review.

Part III

Incentives and Reviewing

Chapter 8

Detecting Strategic Behaviour in Peer Assessment

1 Introduction

Ranking a set of items submitted by a group of people (or ranking the people themselves) is a ubiquitous task that is faced in many applications, including education, hiring, employee evaluation and promotion, and scientific peer review. Many of these applications have a large number of submissions, which makes obtaining an evaluation of each item by a set of independent experts prohibitively expensive or slow. Peer-assessment techniques offer an appealing alternative: instead of relying on independent judges, they distribute the evaluation task across the fellow applicants and then aggregate the received reviews into the final ranking of items. This paradigm has become popular for employee evaluation (Edwards and Ewen, 1996) and grading students' homeworks (Topping, 1998), and is now expanding to more novel applications of massive open online courses (Kulkarni et al., 2013; Piech et al., 2013) and hiring at scale (Kotturi et al., 2020).

The downside of such methods, however, is that reviewers are incentivized to evaluate their counterparts strategically to ensure a better outcome of their own item (Huang et al., 2019; Balietti et al., 2016; Hassidim et al., 2018). Deviations from the truthful behaviour decrease the overall quality of the resulted ranking and undermine fairness of the process. This issue has led to a long line of work (Alon et al., 2009; Aziz et al., 2016; Kurokawa et al., 2015; Kahng et al., 2018; Xu et al., 2019) on designing "impartial" aggregation rules that can eliminate the impact of the ranking returned by a reviewer on the final position of their item.

While impartial methods remove the benefits of manipulations, such robustness may come at the cost of some accuracy loss when reviewers do not engage in strategic behaviour. This loss is caused by less efficient data usage (Kahng et al., 2018; Xu et al., 2019) and reduction of efforts put by reviewers (Kotturi et al., 2020). Implementation of such methods also introduces some additional logistical burden on the system designers; as a result, in many critical applications (e.g., conference peer review) the non-impartial mechanisms are employed. An important barrier that prevents stakeholders from making an informed choice to implement an impartial mechanism is a lack of tools to detect strategic behaviour. Indeed, to evaluate the trade off between the loss of accuracy due to manipulations and the loss of accuracy due to impartiality, one needs to be able to evaluate the extent of strategic behaviour in the system. With this motivation, in this chapter we *focus on detecting strategic manipulations in peer-assessment processes.* [1]

Specifically, in this work we consider a setting where each reviewer is asked to evaluate a subset of works submitted by their counterparts. In a carefully designed randomized study of strategic behaviour when evaluations take the form of *ratings*, Balietti et al. (2016) were able to detect manipulations by comparing the distribution of scores given by target reviewers to some truthful reference. However, other works (Huang et al., 2019; Barroga, 2014) suggest that in more practical settings reviewers may strategically decrease

[1] This chapter is framed slightly more generally and operates with a broader class of peer-assessment problems that includes scientific peer review as a special case.

some scores and increase others in attempt to mask their manipulations or intentionally promote weaker submissions, thereby keeping the distribution of output scores unchanged and making the distribution-based detection inapplicable. Inspired by this observation, we aim to design tools to detect manipulations when the distributions of scores output by reviewers are fixed, that is, we assume that evaluations are collected in the form of *rankings*. Ranking-based evaluation is used in practice (Hazelrigg, 2013) and has some theoretical properties that make it appealing for peer grading (Shah et al., 2013; Caragiannis et al., 2014) which provides additional motivation for our work.

Contributions In this work we present two sets of results.

- **Theoretical** First, we propose a non-parametric test for detection of strategic manipulations in peer-assessment setup with rankings. Second, we prove that our test has a reliable control over the false alarm probability (probability of claiming existence of the effect when there is none). Conceptually, we avoid difficulties associated to dealing with rankings as covariates by carefully accounting for the fact that each reviewer is "connected" to their submission(s); therefore, the manipulation they employ is naturally not an arbitrary deviation from the truthful strategy, but instead the deviation that potentially improves the outcome of their works.

- **Empirical** On the empirical front, we first design and conduct an experiment that incentivizes strategic behaviour of participants. This experiment yields a novel dataset of patterns of strategic behaviour that we make publicly available and that can be useful for other researchers (the dataset is available on the website of the author of this thesis). Second, we use the experimental data to evaluate the detection power of our test on answers of real participants and in a series of semi-synthetic simulations. These evaluations demonstrate that our testing procedure has a non-trivial detection power, while not making strong modelling assumptions on the manipulations employed by strategic agents.

The remainder of this chapter is organized as follows. We give a brief overview of related literature in Section 2. In Section 3 we formally present the problem setting and demonstrate an important difference between the ranking and rating setups. Next, in Section 4 we design a novel approach to testing for strategic behaviour and prove the false alarm guarantees for our test. Section 5 is dedicated to the discussion of the experiment that we designed and executed to collect a dataset of patterns of strategic behaviour. In Section 6 we use this dataset to evaluate the detection ability of our test. Finally, we conclude the chapter with a discussion in Section 7.

2 Related literature

Our work falls at the intersection of crowdsourcing, statistical testing, and a recent line of computer science research on the peer-assessment process. We now give an overview of relevant literature from these areas.

Crowdsourcing work on manipulations Despite motivation for this work comes from studies of Balietti et al. (2016) and Huang et al. (2019), an important difference between rankings and ratings that we highlight in Section 3.2 makes the models considered in these works inapplicable to our setup. Several other papers (Thurner and Hanel, 2011; Cabotà et al., 2013; Paolucci and Grimaldo, 2014) specialize on the problem of strategic behaviour in peer review and perform simulations to explore its detrimental impact on the quality of published works. These works are orthogonal to the present work because they do not aim to detect the manipulations.

Another relevant paper (Perez-Diaz et al., 2018) considers a problem of strategic behaviour in context of the relationships between electric vehicle aggregators in the electricity market. In that setup, each agent is supposed to solve a part of a certain distributed optimization problem and self-interested agents may be incentivized to misreport their solutions. Perez-Diaz et al. (2018) offer a heuristic procedure to identify manipulating agents, but the proposed method relies on the specific nature of the optimization problem and does not directly extend to our setup.

Finally, a long line of work (Akoglu et al., 2013; Kaghazgaran et al., 2018; Jindal and Liu, 2008) aims at detecting fraud in online consumer reviews. In contrast to our setting, these works try to identify malicious reviews without having a direct and known connection between the reviewers and the items being reviewed

that is present in our setup. Instead, these works often rely on additional information (e.g., product-specific features) which is irrelevant or unavailable in our problem.

Statistical testing In this work, we formulate the test for strategic behaviour as a test for independence of rankings returned by reviewers from their own items. Classical statistical works (Lehmann and Romano, 2005) for independence testing are not directly applicable to this problem due to the absence of low-dimensional representations of items. To avoid dealing with unstructured items, one could alternatively formulate the problem as a two-sample test and obtain a control sample of rankings from non-strategic reviewers. This approach, however, has two limitations. First, past work suggests that the test and control rankings may have different distributions even under the absence of manipulations due to misalignment of incentives (Kotturi et al., 2020). Second, existing works (Mania et al., 2018; Gretton et al., 2012; Jiao and Vert, 2018; Rastogi et al., 2020) on two-sample testing with rankings ignore the authorship information that is crucial in our case as we show in the sequel (Section 3.2).

Research on peer assessment This work also falls in the line of several recent works in computer science on the peer-evaluation process that includes both empirical (Tomkins et al., 2017; Sajjadi et al., 2016; Kotturi et al., 2020) and theoretical (Wang and Shah, 2019; Stelmakh et al., 2021a; Noothigattu et al., 2020; Fiez et al., 2020) studies. Particularly relevant works are recent papers (Tomkins et al., 2017; Stelmakh et al., 2019) that consider the problem of detecting biases (e.g., gender bias) in single-blind peer review. Biases studied therein manifest in reviewers being harsher to some subset of submissions (e.g., papers authored by females), making the methods designed in these works inapplicable to the problem we study. Indeed, in our case there does not exist a fixed subset of works that reviewers need to put at the bottom of their rankings to improve the outcome of their own submissions. However, these works share a conceptual approach of detecting the effect on the aggregate level of all agents rather than in each agent individually.

As discussed earlier, research on peer review also aims to prevent or mitigate strategic behavior, where reviewers may manipulate their reviews to help their own papers (Alon et al., 2009; Aziz et al., 2016; Kurokawa et al., 2015; Kahng et al., 2018; Xu et al., 2019) or manipulate reviews to maliciously influence the outcomes of other papers (Jecmen et al., 2020).

3 Problem formulation

In this section we present our formulation of the manipulation-testing problem.

3.1 Preliminaries

In this chapter, we operate in the peer-assessment setup in which reviewers first conduct some work (e.g., homework assignments) and then judge the performance of each other. We consider a setting where reviewers are asked to provide a total ranking of the set of works they are assigned to review.

We let $\mathcal{R} = \{1, 2, \dots, m\}$ and $\mathcal{W} = \{1, 2, \dots, n\}$ denote the set of reviewers and works submitted for review, respectively. We let matrix $C \in \{0,1\}^{m \times n}$ represent conflicts of interests between reviewers and submissions, that is, $(i,j)^{\text{th}}$ entry of C equals 1 if reviewer i is in conflict with work j and 0 otherwise. Matrix C captures all kinds of conflicts of interest, including authorship, affiliation and others, and many of them can be irrelevant from the manipulation standpoint (e.g., affiliation may put a reviewer at conflict with dozens of submissions they are not even aware of). We use $A \in \{0,1\}^{m \times n}$ to denote a subset of "relevant" conflicts — those that reviewers may be incentivized to manipulate for — identified by stakeholders. For the ease of presentation, we assume that A represents the authorship conflicts, as reviewers are naturally interested in improving the final standing of their own works, but in general it can capture any subset of conflicts. For each reviewer $i \in \mathcal{R}$, non-zero entries of the corresponding row of matrix A indicate submissions that are (co-)authored by reviewer i. We let $C(i)$ and $A(i) \subseteq C(i)$ denote possibly empty sets of works conflicted with and authored by reviewer i, respectively.

Each work submitted for review is assigned to λ non-conflicting reviewers subject to a constraint that each reviewer gets assigned μ works. For brevity, we assume that parameters n, m, μ, λ are such that $n\lambda = m\mu$ so we can assign exactly μ works to each reviewer. The assignment is represented by a binary matrix

$M \in \{0,1\}^{m \times n}$ whose $(i,j)^{\text{th}}$ entry equals 1 if reviewer i is assigned to work j and 0 otherwise. We call an assignment valid if it respects the (submission, reviewer)-loads and does not assign a reviewer to a conflicting work. Given a valid assignment M of works $\mathcal{W}$ to reviewers $\mathcal{R}$, for each $i \in \mathcal{R}$, we use $M(i)$ to denote a set of works assigned to reviewer i. $\Pi[M(i)]$ denotes a set of all $|M(i)|!$ total rankings of these works and reviewer i returns a ranking $\pi_i \in \Pi[M(i)]$. The rankings from all reviewers are aggregated to obtain a final ordering $\Lambda(\pi_1, \pi_2, \ldots, \pi_m)$ that matches each work $j \in \mathcal{W}$ to its position $\Lambda_j(\pi_1, \pi_2, \ldots, \pi_m)$, using some aggregation rule Λ known to all reviewers. The grades or other rewards are then distributed according to the final ordering $\Lambda(\pi_1, \pi_2, \ldots, \pi_m)$ with authors of higher-ranked works receiving better grades or rewards.

In this setting, reviewers may be incentivized to behave strategically because the ranking they output may impact the outcome of *their own* works. The focus of this work is on designing tools to detect strategic behaviour of reviewers when a non-impartial aggregation rule Λ (e.g., a rule that theoretically allows reviewers to impact the final standing of their own submissions) is employed.

3.2 Motivating example

To set the stage, we start from highlighting an important difference between rankings and ratings in the peer-assessment setup. To this end, let us consider the experiment conducted by Balietti et al. (2016) in which reviewers are asked to give a score to each work assigned to them for review and the final ranking is computed based on the mean score received by each submission. It is not hard to see that in their setting, the dominant strategy for each rational reviewer who wants to maximize the positions of their own works in the final ranking is to give the lowest possible score to all submissions assigned to them. Observe that this strategy is fixed, that is, it does not depend on the quality of reviewer's work — irrespective of position of their work in the underlying ordering, each reviewer benefits from assigning the lowest score to all submissions they review. Similarly, Huang et al. (2019) in their work also operate with ratings and consider a fixed model of manipulations in which strategic agents increase the scores of low-quality submissions and decrease the scores of high-quality submissions, irrespective of the quality of reviewers' works.

In contrast, when reviewers are asked to output *rankings* of submissions, the situation is different and reviewers can no longer rely on fixed strategies to gain the most for their own submission. To highlight this difference, let us consider a toy example of the problem with 5 reviewers and 5 submissions ($m = n = 5$), authorship and conflict matrix given by an identity matrix ($C = A = I$), and three works (reviewers) assigned to each reviewer (work), that is, $\lambda = \mu = 3$. In this example, we additionally assume that: (i) assignment of reviewers to works is selected uniformly at random from the set of all valid assignments, (ii) aggregation rule Λ is the Borda count, that is, the positional scoring rule with weights equal to positions in the ranking,[2] (iii) reviewers are able to reconstruct the ground-truth ranking of submissions assigned to them without noise, and (iv) all but one reviewers are truthful.

Under this simple formulation, we qualitatively analyze the strategies available to the strategic reviewer, say reviewer i^*. Specifically, following the rating setup, we consider the fixed deterministic strategies that do not depend on the work created by reviewer i^*. Such strategies are limited to permutations of the ground-truth ranking of submissions in $M(i^*)$. Figure 1 represents an expected gain of each strategy (measured in positions in the aggregated ordering) as compared to the truthful strategy for positions 2–4 of the work authored by reviewer i^* in the ground-truth ranking. To make this figure, for each target position of the strategic reviewer i^* in the underlying total ordering, we first compute their expected position in the final ranking (expectation is taken over the randomness in the assignment) if they use the truthful strategy. We then compute the same expectations for each manipulating strategy and plot the differences. The main observation is that there does not exist a fixed strategy that dominates the truthful strategy for every possible position of the reviewer's work. Therefore, in setup with rankings strategic reviewers need to consider how their own works compare to the works they rank in order to improve the outcome of their submissions.

[2]We use the variant without tie-breaking — tied submissions share the same position in the final ordering.

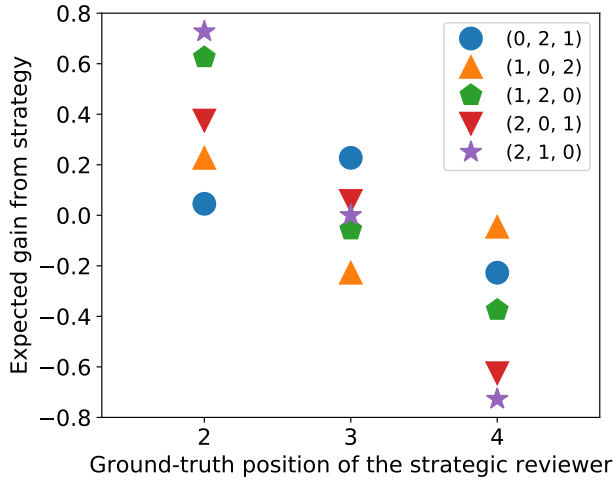


Figure 1: Comparison of fixed deterministic strategies available to a single strategic reviewer depending on position of their work in the true underlying ranking. A positive value of the expected gain indicates that the manipulation strategy in expectation delivers better position in the final ordering than the truthful strategy.

3.3 Problem setting

With the motivation given in Section 3.2, we are ready to present the formal hypothesis-testing problem we consider in this work. When deciding on how to rank the works, the information available to reviewers is the content of the works they review and the content of their own works. Observe that while a truthful reviewer does not take into account their own submissions when ranking works of others, the aforementioned intuition suggests that the ranking output by a strategic agent should depend on their own works. Our formulation of the test for manipulations as an independence test captures this motivation.

Problem 1 (Testing for strategic behaviour)**.** Given a non-impartial aggregation rule Λ, assignment of works to reviewers M, rankings returned by reviewers $\{\pi_i, i \in \mathcal{R}\}$, conflict matrix C, authorship matrix A and set of works submitted for review $\mathcal{W}$, the goal is to test the following hypotheses:

Null (H_0)

$$\forall i \in \mathcal{R} \text{ s.t. } A(i) \neq \emptyset \quad \pi_i \perp A(i).$$

Alternative (H_1)

$$\exists i \in \mathcal{R} \text{ s.t. } A(i) \neq \emptyset \quad \pi_i \not\perp A(i).$$

In words, under the null hypothesis reviewers who have their submissions under review do not take into account their own works when evaluating works of others and hence are not engaged in manipulations that can improve the outcome of their own submissions. In contrast, under the alternative hypothesis some reviewers choose the ranking depending on how their own works compare to works they rank, suggesting that they are engaged in manipulations.

Assumptions Our formulation of the testing problem makes two assumptions about the data-generation process to ensure that association between works authored by reviewer i and ranking π_i may be caused only by strategic manipulations and not by some intermediate mediator variables.

(A1) **Random assignment** We assume that the assignment of works to reviewers is selected uniformly at random from the set of all assignments that respect the conflict matrix C. This assumption ensures that the works authored by a reviewer do not impact the set of works assigned to them for review. The assumption of random assignment holds in many applications, including peer grading (Freeman and Parks, 2010; Kulkarni et al., 2013) and NSF review of proposals (Hazelrigg, 2013).

(A2) **Independence of ranking noise** We assume that under the null hypothesis of absence of strategic behaviour, the reviewer identity is independent of the works they author, that is, the noise in reviewers' evaluations (e.g., the noise due to subjectivity of the truthful opinion) is not correlated with their submissions. This assumption is satisfied by various popular models for generation of rankings, including Plackett-Luce model (Luce, 1959; Plackett, 1975) and more general location family of random utility models (Soufiani et al., 2012).

In the sequel, we show that Assumption A1 can be relaxed to allow for assignments of any fixed topology (Appendix A2) and that our test can control the false alarm probability under some practical violations of Assumption A2 (Appendix A1). More generally, we note that one needs to carefully analyze the assumptions in the specific application and carefully interpret the results of the test, keeping in mind that its interpretation depends heavily on whether the assumptions are satisfied or not.

4 Testing procedure

In this section, we introduce our testing procedure. Before we delve into details, we highlight the main intuition that determines our approach to the testing problem. Observe that when a reviewer engages in strategic behaviour, they tweak their ranking to ensure that *their own* works experience better outcome when all rankings are aggregated by the rule Λ. Hence, when *successful* strategic behaviour is present, we may expect to see that the ranking returned by a reviewer influences position of *their own* works under aggregation rule Λ in a more positive way than other works not reviewed by this reviewer. Therefore, the test we present in this work attempts to identify whether rankings returned by reviewers have a more positive impact on the final standing of their own works than what would happen by chance.

For any reviewer $i \in \mathcal{R}$, let $\mathcal{U}_i$ be a uniform distribution over rankings $\Pi[M(i)]$ of works assigned to them for review. With this notation, we formally present our test as Test 1 below. Among other arguments, our test accepts the optional set of rankings $\{\pi_i^*, i \in \mathcal{R}\}$, where for each $i \in \mathcal{R}$, π_i^* is a ranking of works $M(i)$ assigned to reviewer i, but is constructed by an impartial agent (e.g., an outsider reviewer who does not have their own works in submission). For the ease of exposition, let us first discuss the test in the case when the optional set of rankings is *not* provided (i.e., the test has no supervision) and then we will make a case for usefulness of this set.

In Step 1, the test statistic is computed as follows: for each reviewer $i \in \mathcal{R}$ and for each work $j \in A(i)$ authored by this reviewer, we compute the impact of the ranking returned by the reviewer on the final standing of this work. To this end, we compare the position actually taken by the work (first term in the inner difference in Equation 8.1) to the expected position it would take if the reviewer would sample the ranking of works $M(i)$ uniformly at random (second term in the inner difference in Equation 8.1). To get the motivation behind this choice of the test statistic, note that if a reviewer i is truthful then the ranking they return may be either better or worse for *their own* submissions than a random ranking, depending on how their submissions compare to works they review. In contrast, a strategic reviewer may choose the ranking that delivers a better final standing for their submissions, biasing the test statistic to the negative side.

Having defined the test statistic, we now understand its behaviour under the null hypothesis to quantify when its value is too large to be observed under the absence of manipulations for a given significance level α. To this end, we note that for a given assignment matrix M, there are many pairs of conflict and authorship matrices (C', A') that (i) are equal to the actual matrices C and A up to permutations of rows and columns and (ii) do not violate the assignment M, that is, do not declare a conflict between any pair of reviewer i and submission j such that submission j is assigned to reviewer i in M. Next, note that under the null hypothesis of absence of manipulations, the behaviour of reviewers would not change if matrix A was substituted by another matrix A', that is, a ranking returned by any reviewer i would not change if that reviewer was an author of works $A'(i)$ instead of $A(i)$. Given that the structure of the alternative matrices C' and A' is the same as that of the actual matrices C and A, under the null hypothesis of absence of manipulations we expect the actual test statistic to have a similar value as compared to that under C' and A'.

The aforementioned idea drives Steps 2-4 of the test. In Step 2 we construct the set of all pairs of conflict and authorship matrices of the fixed structure that do not violate the assignment M. We then compute the

Test 1 Test for strategic behaviour

Input: Reviewers' rankings $\{\pi_i, i \in \mathcal{R}\}$
Assignment M of works to reviewers
Conflict and authorship matrices (C, A)
Significance level α, aggregation rule Λ
Optional Argument: Impartial rankings $\{\pi_i^*, i \in \mathcal{R}\}$

1. Compute the test statistic τ as

$$\tau = \sum_{i \in \mathcal{R}} \sum_{j \in A(i)} \Big(\Lambda_j(\pi'_1, \pi'_2, \dots, \pi_i, \dots, \pi'_m) - \mathbb{E}_{\tilde{\pi} \sim \mathcal{U}_i} \left[\Lambda_j(\pi'_1, \pi'_2, \dots, \tilde{\pi}, \dots, \pi'_m) \right] \Big), \tag{8.1}$$

   where $\pi'_i, i \in \mathcal{R}$, equals π_i^* if the optional argument is provided and equals π_i otherwise.

2. Compute a multiset $\mathcal{P}(M)$ as follows. For each pair (p_m, p_n) of permutations of m and n items, respectively, apply permutation p_m to rows of matrices C and A and permutation p_n to columns of matrices C and A. Include the obtained matrix A' to $\mathcal{P}(M)$ if it holds that for each $i \in \mathcal{R}$:

$$A'(i) \subseteq C'(i) \subset \mathcal{W} \backslash M(i).$$

3. For each matrix $A' \in \mathcal{P}(M)$ define $\varphi(A')$ to be the value of the test statistic (8.1) if we substitute A with A', that is, $\varphi(A')$ is the value of the test statistic if the authorship relationship was represented by A' instead of A. Let

$$\Phi = \{\varphi(A'), A' \in \mathcal{P}(M)\} \tag{8.2}$$

   denote the multiset that contains all these values.

4. Reject the null if τ is strictly smaller than the $(\lfloor \alpha |\Phi| \rfloor + 1)^{\text{th}}$ order statistic of Φ.

value of the test statistic for each of these authorship matrices in Step 3 and finally reject the null hypothesis in Step 4 if the actual value of the test statistic τ appears to be too extreme against values computed in Step 3 for the given significance level α.

If additional information in the form of impartial rankings is available (i.e., the test has a supervision), then our test can detect manipulations better. The idea of supervision is based on the following intuition. In order to manipulate successfully, strategic reviewers need to have some information about the behaviour of others. In absence of such information, it is natural (and this idea is supported by data we obtain in the experiment in Section 5) to choose a manipulation targeted against the truthful reviewers, assuming that a non-trivial fraction of agents behave honestly. The optional impartial rankings allow the test to use this intuition: for each reviewer $i \in \mathcal{R}$ the test measures the impact of reviewer's ranking on their submissions as if this reviewer was the only manipulating agent, by complementing the ranking π_i with impartial rankings $\{\pi_1^*, \dots, \pi_{i-1}^*, \pi_{i+1}^* \dots, \pi_m^*\}$. As we show in Section 6, availability of supervision can significantly aid the detection power of the test.

The following theorem combines the above intuitions and ensures a reliable control over the false alarm probability for our test (the proof is given in Appendix A3).

Theorem 1. *Suppose that Assumptions A1 and A2 specified in Section 3.3 hold. Then, under the null hypothesis of absence of manipulations, for any significance level $\alpha \in (0, 1)$ and for any aggregation rule Λ, Test 1 (both with and without supervision) is guaranteed to reject the null with probability at most α. Therefore, Test 1 controls the false alarm probability at the level α.*

Remark. 1. In Section 6 we complement the statement of the theorem by demonstrating that our test has a non-trivial detection power.

2. In practice, the multiset $\mathcal{P}(M)$ may take $\mathcal{O}(m!n!)$ time to construct which is prohibitively expensive even for small values of m and n. The theorem holds if instead of using the full multiset $\mathcal{P}(M)$, when defining Φ, we only sample some k authorship matrices uniformly at random from the multiset $\mathcal{P}(M)$. The value of k should be chosen large enough to ensure that $(\lfloor \alpha |\Phi| \rfloor + 1)$ is greater than 1. The sampling can be performed by generating random permutations using the shuffling algorithm of Fisher (1935) and rejecting samples that lead to matrices $A' \notin \mathcal{P}(M)$.

3. The impartial set of rankings $\{\pi_i^*, i \in \mathcal{R}\}$ need not necessarily be constructed by a separate set of m reviewers. For example, if one has access to the (noisy) ground-truth (for example, to the ranking of homework assignments constructed by an instructor), then for each $i \in \mathcal{R}$ the ranking π_i^* can be a ranking of $M(i)$ that agrees with the ground-truth.

Effect size In addition to controlling for the false alarm probability, our test offers a measure of the effect size defined as $\Delta = \tau \cdot \left[\sum_{i \in \mathcal{R}} |A(i)|\right]^{-1}$. Each term in the test statistic τ defined in (8.1) captures the impact of the ranking returned by a reviewer on the final standing of the corresponding submission and the the mean impact is a natural measure of the effect size. Negative values of the effect size demonstrate that reviewers in average benefit from the rankings they return as compared to rankings sampled uniformly at random.

5 Experiment to elicit strategic behaviour

In this section we describe the experiment we designed and executed to collect a dataset of patterns of strategic behaviour that we will use in Section 6 to empirically evaluate the detection power of our test. The experiment was offered to attendees of a graduate-level AI course at Carnegie Mellon University and $N = 55$ students completed the experimental procedure described in Section 5.1. Exploratory analysis of the collected data is given in Section 5.2 and the dataset is available on the website of the author of this thesis.

5.1 Design of experiment

The goal of our experiment is to understand what strategies people use when manipulating their rankings of others. A real peer grading setup (i.e., homework grading) possesses an ethical barrier against cheating and hence many subjects of the hypothetical experiment would behave truthfully, reducing the efficiency of the process. To overcome this issue, we use gamification and organize the experiment as follows (game interface can be found in supplementary materials on the the website of the author of this thesis).

We design a game for $m = 20$ players and $n = 20$ hypothetical submissions. First, a one-to-one authorship relationship A is sampled uniformly at random from the set of permutations of 20 items and each player becomes an "author" of one of the submissions. Each submission is associated to a unique value $v \in \{1, 2, \ldots, 20\}$ and this value is privately communicated to the respective player; therefore, players are associated to values and in the sequel we do not distinguish between a player's value and their "submission". We then communicate values of some $\mu = 4$ other contestants to each player subject to the constraint that a value of each player becomes known to $\lambda = 4$ counterparts. To do so, we sample an assignment M from the set of assignments respecting the conflict matrix $C = A$ uniformly at random. Note that players do not get to see the full assignment and only observe the values of players assigned to them. The rest of the game replicates the peer grading setup: participants are asked to rank their peers (the truthful strategy is to rank by values in decreasing order) and the rankings are aggregated using the Borda count aggregation rule (tied submissions share the position in the final ordering).

For the experiment, we create 5 rounds of the game, sampling a separate authorship matrix A_k and assignment M_k for each round $k \in \{1, 2, \ldots, 5\}$. Each of the $N = 55$ subjects then participates in all 5 rounds, impersonating one (the same for all rounds) of the 20 game players.[3] Importantly, subjects are instructed that their goal is to *manipulate their ranking to improve their final standing.* Additionally, we inform participants that in the first 4 rounds of the game their competitors are truthful bots who always rank

[3]We sample a separate authorship matrix for each round so participants get different values between rounds.

players by their values. In the last round, participants are informed that they play against other subjects who also engage in manipulations.

To help participants better understand the rules of the game and properties of the aggregation mechanism, after each of the first four rounds, participants are given feedback on whether their strategy improves their position in the aggregated ordering. Note that the position of the player in the final ordering depends on the complex interplay between (i) the strategy they employ, (ii) the strategy employed by others, and (iii) the configuration of the assignment. In the first four rounds of the game, participants have the information about (ii), but do not get to see the third component. To make feedback independent of (iii), we average it out by computing the mean position of the player over the randomness in the part of the assignment unobserved by the player and give positive feedback if their strategy is in expectation better than the ranking sampled uniformly at random. Finally, after the second round of the game, we give a hint that additionally explains some details of the game mechanics.

The data we collect in the first four rounds of the game allows us to understand what strategies people use when they manipulate in the setup when (most) other reviewers are truthful. In the last round, we remove the information about the behaviour of others and collect data about manipulations in the wild (i.e., when players do not know other players' strategies). Manual inspection of the collected data reveals that 53 participants attempted manipulations in each round and the remaining 2 subjects manipulated in all but one round each, hence, we conclude that the data is collected under the alternative hypothesis of the presence of manipulations.

5.2 Exploratory data analysis

We now continue to the exploratory analysis of collected data and begin from analyzing the manipulating strategies employed by participants. In addition to rankings, in each round we asked participants to describe their reasoning in a textual form and we manually analyze these descriptions to identify the strategies people use. While these textual descriptions sometimes do not allow to unequivocally understand the general strategy of the player due to ambiguity, we are able to identify 6 broad clusters of strategies employed by participants. We underscore that each of these clusters may comprise several strategies that are similar in spirit but may slightly disagree in some situations. We now introduce these clusters by describing the most popular representative strategy that will be used in the subsequent analysis.

- `Reverse` This naive strategy prescribes to return the reversed ground truth ordering of players under comparison. Note that in contrast to other strategies we explain below, the ranking returned by a player who use this strategy is independent of their own value.
- `Distance` The idea behind this family of strategies is to identify the direct competitors and put them at the bottom of the ranking, while out of reach players and those with considerably smaller values are put at the top. The most popular incarnation of this strategy is to rank the other players in order of decreasing distance from *the player's* value: the furthest player gets the first place and the closest gets the last place.
- `See-Saw` This strategy follows `Reverse` if the value assigned to a player is in top 50% of all values (i.e., greater than 10) and follows the truthful strategy otherwise. None of the participants directly reported this strategy in the experiment, but we include it in the analysis as this strategy agrees with behaviour of several players.
- `Better-to-Bottom` This strategy is another simplification of the `Distance` strategy. Let v^* be the player's value. Then this strategy prescribes to put submissions with values smaller than v^* at the top (in order of increasing values) and submissions with values larger than v^* at the bottom (in order of decreasing values) of the ranking. For example, if the player's value is 10 and they are asked to rank other players whose values are $(16, 12, 7, 2)$, then this strategy would return $\pi = 2 \succ 7 \succ 16 \succ 12$.
- `Worse-to-Bottom` Submissions with values lower than v^* are placed at the bottom (in order of decreasing values) and submissions with values larger than v^* are placed at the top (in order of increasing values) of the ranking. In the earlier example with $v^* = 10$ and values $(16, 12, 7, 2)$ to be ranked, this strategy would return $\pi = 12 \succ 16 \succ 7 \succ 2$.

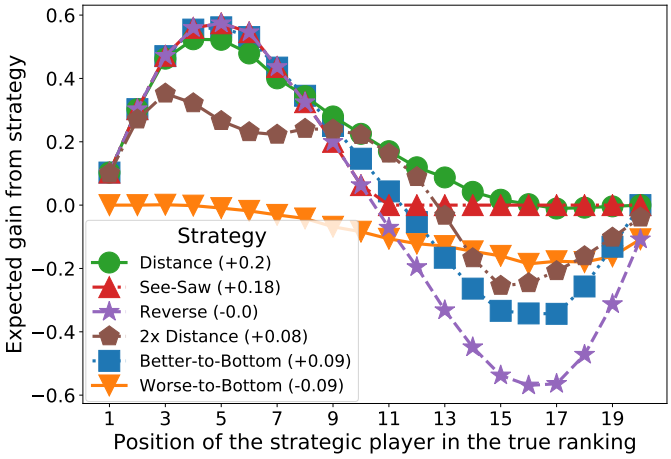


Figure 2: Expected gain from manipulation strategy when all but one player are truthful as a function of position of the strategic player in in the ground-truth ranking. The expectation is taken over randomness of the assignment procedure and values in brackets in the legend correspond to the mean gain over all positions. Borda count aggregation rule is used. A positive value of the expected gain indicates that the manipulation strategy in expectation delivers better position in the final ordering than the truthful strategy. Error bars are too small to be visible.

- `2x-Distance` This strategy was reported only in round 5, that is, when participants were competing against each other, and is targeted to respond to the `Distance` strategy. This strategy suggests redefining all values (including the value of the player) by the following rule:

$$v' = \min\{20 - v, v - 1\},$$

  and apply the `Distance` strategy over the updated values.

Figure 2 juxtaposes the identified strategies by comparing them to the truthful one in case when all but one player are truthful. For each position of the strategic reviewer i^* in the ground-truth total ordering, we compute the expected gain (measured in positions in the aggregated ordering) from using each of the 6 strategies. To this end, we first compute the expected position (expectation is taken over randomness in the assignment) of reviewer i^* if they use the truthful strategy. We then compute the same expectations for each of the 6 manipulation strategies and plot the differences as a function of the position of the strategic player in the true underlying ranking.

We make several observations from Figure 2. First, strategies `Distance` and `See-Saw` benefit the manipulating player irrespective of her/his position in the underlying ranking. In contrast, `Better-to-Bottom` and `2x-Distance` can both help and hurt the player depending on the position in the ground-truth ordering and different effects average out to the positive total gain. The `Reverse` strategy delivers zero gain in expectation over positions, being not better nor worse than the truthful strategy. Finally, the `Worse-to-Bottom` strategy is uniformly dominated by the truthful strategy, implying that the strategic player can only hurt their expected position by relying on this strategy.

To conclude the preliminary analysis of collected data, for each of the 5 rounds we manually allocate each player to one of the aforementioned manipulation strategies based on the ranking and textual description they provided. As we mentioned above, this information is sometimes not sufficient to unequivocally identify the strategy. To overcome this ambiguity, we use fractional allocation in case several strategies match the response and leave some players unclassified in hard cases (for example, when textual response contradicts the actual ranking). Note that players who employed the truthful strategy are also included in the unclassified category as the goal of the categorization is to understand the behaviour of strategic players.

Table 1 displays the resulting allocation of players to strategies informed by the data collected in the

experiment. First, in round 1 of the game half of strategic players employed the `Reverse` strategy which is not better than the truthful strategy and hence does not lead to a successful manipulation. Second, as the game proceeds and players understand the mechanics of the game better, they converge to the `Distance` strategy which in expectation delivers a positive gain irrespective of the position of the player in the underlying ranking. Third, note that most of the players continued with the `Distance` strategy even in Round 5, despite in this round they were no longer playing against truthful bots. However, a non-trivial fraction of students managed to predict this behaviour and employed the `2x-Distance` strategy to counteract the `Distance` strategy. Finally, many players were clueless about what strategy to employ in round 5, contributing to the increased number of unclassified participants.

	Round 1	Round 2	Round 3	Round 4	Round 5
Reverse	.50	.33	.05	.03	.06
Distance	.37	.53	.93	.96	.78
See-Saw	.09	.08	.02	.01	–
Better-to-Bottom	.02	.04	–	–	–
Worse-to-Bottom	.02	.02	–	–	–
2x-Distance	–	–	–	–	.16
Unclassified	5	7	4	4	18

Table 1: Manually encoded characterization of strategies used by manipulating participants. In the first 4 rounds of the game participants played against truthful bots and in the last round they played against each other.

6 Evaluation of the test

We now investigate the detection power of our test (Test 1). We begin from analysis of real data collected in the previous section and execute the following procedure. For each of the 1,000 iterations, we uniformly at random subset 20 out of the 55 participants such that together they impersonate all 20 game players. We then apply our test (with and without supervision) to rankings output by these participants in each of the 5 rounds, setting significance level at $\alpha = 0.05$ and sampling $k = 100$ authorship matrices in Step 3 of the test. The impartial rankings for testing with supervision comprise ground truth rankings.

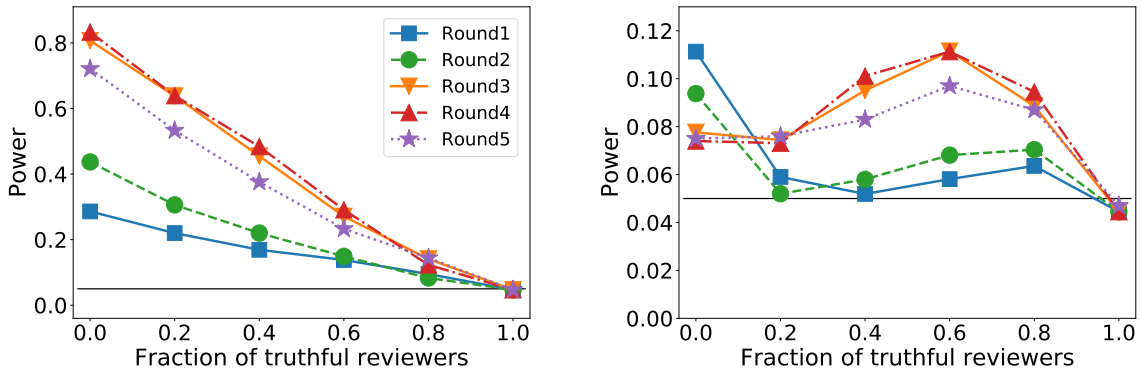
	Round 1	Round 2	Round 3	Round 4	Round 5
With supervision	0.61	0.57	0.87	1.00	0.09
Without supervision	0.17	0.02	0.16	0.01	0.08

Table 2: Detection rates of our test.

After performing all iterations, for each round we compute the mean detection rate and represent these values in Table 2. The results suggest that our test provided with the impartial set of rankings has a strong detection power, reliably detecting manipulations in the first 4 rounds. On the other hand, performance of our test without supervision is modest. The reason behind the difference in performance is that our test aims at detecting *successful* manipulations (i.e., those that improve the outcome of a player). In the first 4 rounds of the game, subjects were playing against truthful competitors and hence the test provided with the additional set of impartial rankings (which is targeted at detecting responses to the truthful strategy) has a good performance. However, the test without supervision is not able to detect such manipulations, because it evaluates success using rankings of other participants who also engage in manipulations and the response to

the truthful strategy is not necessarily successful in this case. As for the round 5, we will show in a moment that poor performance of our test appears to be due to random chance (i.e., the choice of the assignment which is hard for detection) and not due to any systematic issue.

Note that performance of our test depends not only on the strategies employed by players, but also on the assignment M realized in a particular round. Some realizations of random assignment make successful manipulations (and their detection) easier while under other realizations most of the players cannot improve their position even if they use the best strategy (and therefore our test cannot detect manipulations). To remove the impact of the specific assignments we used in the experiment, we now proceed to semi-synthetic trials. Specifically, we use the manual allocation of participants to manipulation strategies represented in Table 1 and create artificial agents who follow these strategies, replicating proportions learned from the real data. We then repeat our experiment with $m = 20$ artificial agents, simulating 1,000 assignments for each round of the game and computing the expectation of the power of our test over randomness of the assignment. Additionally, we enhance the set of synthetic agents with truthful agents and study how the detection power of our test changes with the fraction of truthful agents. Figure 3 displays the expected power of our test for proportions of strategies used by strategic agents informed by each round of the real game and for various fractions of truthful players. Note that when all players are truthful (rightmost points of both plots), the data is generated under the null hypothesis of absence of strategic behaviour, and the plots empirically verify the guarantee of Theorem 1 that our test indeed caps the false alarm rate at $\alpha = 0.05$.



(a) With supervision

(b) Without supervision

Figure 3: Expected power of our test for different allocations of strategic agents to strategies and different fractions of truthful agents. The black horizontal line is a baseline power achieved by a test that rejects the null with probability α=0.05 irrespective of the data. Error bars are too small to show.

Figure 3a shows that our test provided with optional rankings has a non-trivial power in every round, including the last round in which participants were playing against each other. Note that as game proceeds and participants understand the rules better (and find ways to manipulate efficiently), the power of the test increases. A surprising success of the test with supervision in round 5 is explained by the combination of two factors: (i) the majority of participants resorted to the response to the truthful strategy even in round 5 and (ii) a strategy that constitutes a response to the response to the truthful strategy is still a good response to the truthful strategy. Hence, our test provided with impartial rankings can detect manipulations even in case when participants play against each other.

Figure 3b shows that the test without supervision has considerably lower (but still non-trivial) power. We note, however, that the main feature of the test without supervision is that it can be readily applied to purely observational data and the power can be accumulated over multiple datasets (e.g., it can be applied to multiple iterations of a university course). An interesting feature of the test without supervision is the non-monotonicity of power with respect to the fraction of truthful reviewers, caused by a complex interplay between the fraction of truthful agents and the strategies employed by manipulating agents that determines success of manipulations.

Overall, the evaluations we conducted in this section suggests that the test we propose in this work has a non-trivial detection power, especially when an additional set of impartial rankings is available. In Appendix A1 we provide additional empirical evaluations of the test, including the case when reviewers are noisy and Assumption A2 (formulated in Section 3.3) is violated.

7 Discussion

In this work, we design a test for detection of strategic behaviour in the peer-assessment setup with rankings. We prove that it has a reliable control over the false alarm probability and demonstrate its non-trivial detection power on data we collected in a novel experiment. Thus, our work offers a tool for system designers to measure the presence of strategic behavior in the peer-assessment system (peer-grading of homeworks and exams, evaluation of grant proposals, and hiring at scale) and informs the trade off between the loss of accuracy due to manipulations and the loss of accuracy due to restrictions put by impartial aggregation mechanisms. Therefore, organizers can employ our test to make an informed decision on whether they need to switch to the impartial mechanism or not.

An important feature of our test is that it aims at detecting the manipulation on the aggregate level of all agents. As a result, our test does not allow for personal accusations and hence does not increase any pressure on individual agents. As a note of caution, we caveat, however, that selective application of our test (as well as of *any* statistical test) to a specific sub-population of agents may lead to discriminatory statements; to avoid this, experimenters need to follow pre-specified experimental routines and consider ethical issues when applying our test.

Our approach is conceptually different from the past literature which considers ratings (Balietti et al., 2016; Huang et al., 2019) as it does not assume any specific parametric model of manipulations and instead aims at detecting any *successful* manipulation of rankings, thereby giving flexibility of non-parametric tests. This flexibility, however, does not extend to the case when agents try to manipulate but do it *unsuccessfully* (see Appendix A1 for demonstration). Therefore, an interesting problem for future work is to design a test that possesses flexibility of our approach but is also able to detect any (and not only successful) manipulations.

Appendix

We provide supplementary materials and additional discussion. Appendix A1 is dedicated to additional evaluations of our test (Test 1). We show how to slightly relax Assumption A1 of random assignment in Appendix A2 and prove Theorem 1 in Appendix A3.

A1 Additional evaluations of the test

We now provide additional evaluations of our test and conduct simulations in the following settings:

- **Detecting pure strategies** First, we evaluate the detection power of the test against each of the strategies we learned from the experimental data (described in Section 5.2).
- **Noisy supervision** Next, we evaluate robustness of our test to the noise in the optional impartial rankings $\{\pi_i^*, i \in \mathcal{R}\}$.
- **Noise in reviewers' evaluations** We also study the detection power of our test when reviewers perceive quality scores of submissions assigned to them with noise.
- **Violation of Assumption A2** We then analyze the behavior of our test when Assumption A2 formulated in Section 3.3 is violated. To this end, we use a model that connects the quality of reviewer's submission to

the level of noise in their evaluations suggested by empirical research on peer grading and compute the false alarm rate of our test under this model.

- **Runtime of the test** In this work, we perform simulations in the small sample size setting ($n = m = 20$) to be able to run thousands of iterations and average out the impact of the specific assignment on the performance of our test. In practice, organizers will need to run the test only once (or several times if the test is applied over multiple datasets) and we provide runtimes of our naive implementation of the test for larger values of the problem parameters.

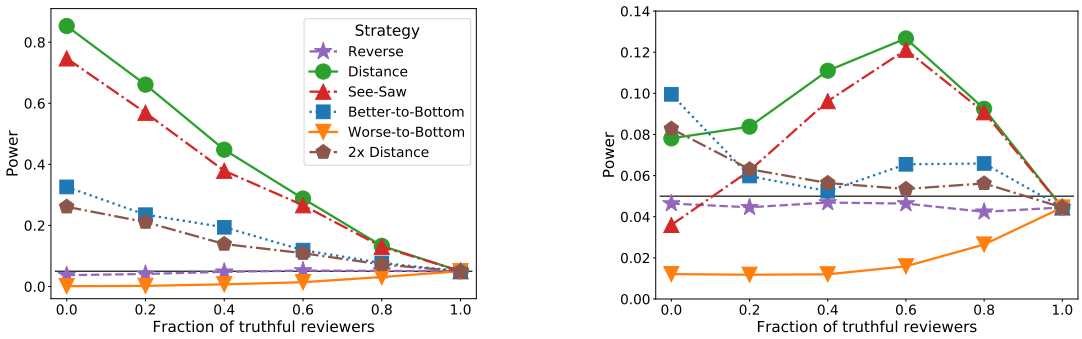
A1.1 Detecting pure strategies

To compute the power against specific strategies we identified in Section 5.2, we follow the same approach we used to evaluate the expected detection power of our test in each round of the game (Figure 3). Specifically, for each fraction of truthful agents and for each of the strategies we learned from data, we compute the detection power of the test with and without supervision over 1,000 assignments sampled uniformly at random from the set of all assignments valid for parameters:

$$A = C = I,\ n = m = 20,\ \lambda = \mu = 4. \tag{8.3}$$

Figure 4 compares the detection power of our test against each strategy used by participants of the experiment. Recall that our test aims at detecting manipulations that improve the final standing of the strategic reviewer. As shown in Figure 2, the `Reverse` and `Worse-to-Bottom` strategies do not improve the final standing of the manipulating agent and hence our test cannot detect strategic behaviour when these strategies are employed.

In contrast, the `See-Saw`, `Distance`, `Better-to-Bottom` and `2x-Distance` strategies in expectation improve the position of the strategic reviewer when all other players are truthful and hence our test with supervision can detect these manipulations with a non-trivial power, with power being greater for more successful strategies. The behaviour of the test without supervision against these 4 successful strategies involves a complex interplay (depicted in Figure 4b) between the fraction of non-strategic agents and the particular strategy employed by strategic players.



(a) With supervision

(b) Without supervision

Figure 4: Detection power of our test against different strategies employed by participants in the experiment. The black horizontal line is a baseline power achieved by a test that rejects the null with probability α=0.05 irrespective of the data. Error bars are too small to show.

A1.2 Noisy supervision

The key component of testing with supervision is the set of impartial rankings provided to the test. We now investigate the impact of the noise in the impartial rankings on the power of the test. To this end, we

continue with the simulation schema used in the previous section, with the exception that (i) we only consider the `Distance` strategy for the strategic agents and (ii) instead of varying strategies, we vary the level of noise in the impartial rankings. Specifically, we sample impartial rankings from the random utility model using values of the players as quality parameters and adding zero-centered Gaussian noise with standard deviation σ. We then vary parameter σ to obtain the power for different noise levels.

Figure 5 represents the results of simulations and demonstrates that our test is robust to a significant amount of noise in the impartial rankings. Note that under Gaussian noise with $\sigma = 3$ two players with values differing by 3 points are swapped in the impartial ranking with probability $p \approx 0.24$. Hence, our test with supervision is able to detect manipulations even under significant level of noise. Of course, as the level of noise increases and impartial rankings become random, the power of our test becomes trivial.

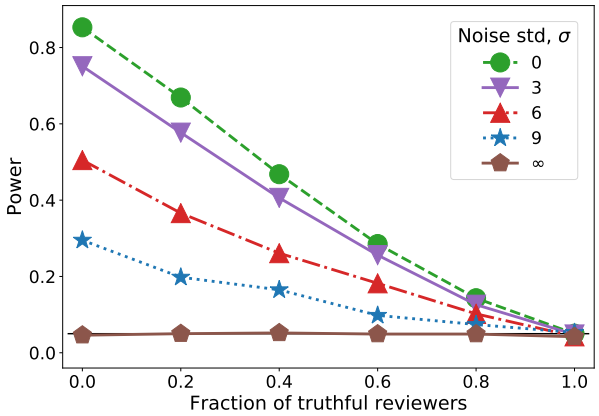


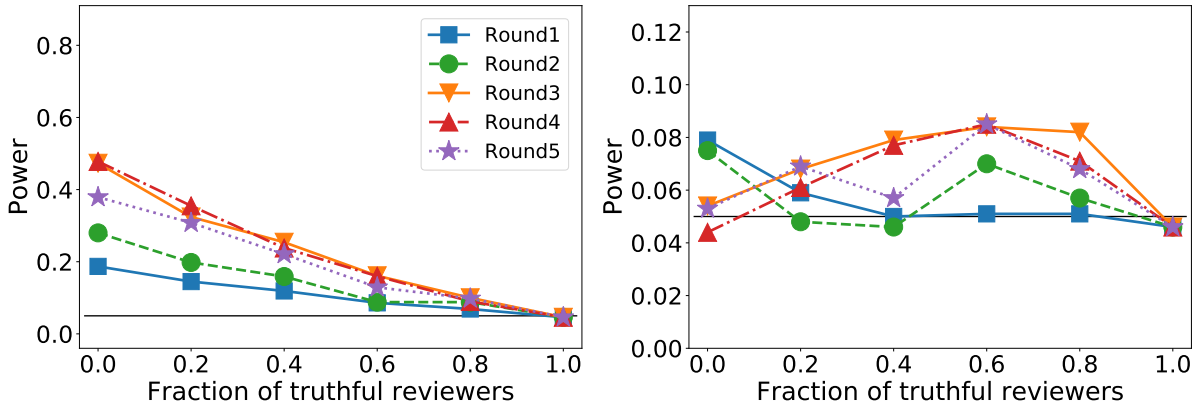
Figure 5: Detection power of our test with supervision for various levels of noise in the impartial rankings. The black horizontal line (hidden behind the line for $\sigma = \infty$) is a baseline power achieved by a test that rejects the null with probability α=0.05 irrespective of the data. Error bars are too small to show.

A1.3 Noise in reviewers' evaluations

In the experiment we conducted in Section 5 players were communicated exact values of the ground truth quality of submissions, that is, we assumed that reviewers can estimate the quality of submissions without noise. We now study the performance of our test when reviewers are noisy. To this end, we replicate the semi-synthetic setting described in Section 6 with the exception that we add a zero-mean Gaussian noise ($\sigma = 3$) to scores communicated to artificial agents and compute the detection power of our test under this noisy setup. Figure 6 summarizes the results of simulations and shows that out test with supervision continues to have a strong detection power even under significant amount of noise in reviewers' evaluations. Similarly, the test without supervision, while loosing some power as compared to the noiseless case, also manages to maintain a non-trivial power in the noisy case. For additional evaluations of our test when reviewers are noisy we refer the reader to the next section in which some application-specific noise model is considered.

A1.4 Violation of Assumption A2

The focus of our work is on the peer assessment process and student peer grading is one of the most prominent applications. The literature on peer grading (Piech et al., 2013; Shah et al., 2013) suggests that Assumption A2 that we formulated in Section 3.3 may be violated in this application: the models proposed in these works (which are used in Coursera and Wharton's peer-grading system) suggest that authors of stronger submissions are also more reliable graders. While our theoretical analysis does not guarantee control over the false alarm probability in this case, we note that in practice our test does not break its respective guarantees under such relationship.



(a) With noiseless supervision

(b) Without supervision

Figure 6: Detection power of our test when reviewers perceive the ground truth quality of submissions assigned to them with a zero mean Gaussian noise ($\sigma = 3$). The black horizontal line is a baseline power achieved by a test that rejects the null with probability α=0.05 irrespective of the data. Error bars are too small to show.

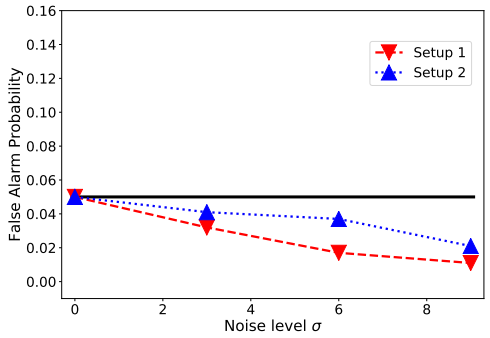
Intuitively and informally, Figure 2 suggests that authors of stronger works benefit from reversing the ranking of submissions assigned to them whereas authors of weaker submissions should play truthfully to maximize the outcome of their submission. In contrast, the aforementioned relationship between the quality of a submission and the grading ability of its author claims the converse: authors of top submissions return rankings that are closer to the ground truth than noisy rankings returned by authors of weaker submissions. Hence, truthful reviewers do not benefit from the difference in noise levels, suggesting that our test may be more conservative, but will not break its false alarm guarantees.

To validate this intuition, we consider the problem parameters given in (8.3) and assume that each reviewer $i \in \mathcal{R}$ samples the ranking of the works assigned to them from the random utility model with reviewer-specific noise level σ_i and quality parameters determined by the true values of the works. We then simulate the false alarm probability of our test under two setups with different definitions of noise:

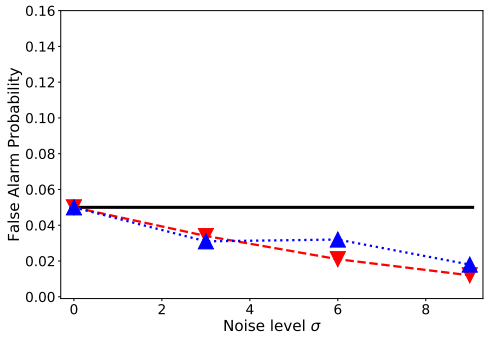
- **Setup 1** If reviewer i is the author of one of the top 10 works, they are noiselesee, that is, $\sigma_i = 0$. In contrast, if reviewer i is the author of one of the bottom 10 submissions, their noise level is non-zero: $\sigma_i = \sigma$.
- **Setup 2** Each reviewer i samples the ranking from the random utility model with noise level $\sigma \times {}^{k_i}/_{20}$, where k_i is the position of the work authored by reviewer i in the underlying ground-truth ordering.

Observe that in both setups data is generated under the null hypothesis of absence of manipulations. In simulations, we vary the noise level σ and sample impartial rankings for the test with supervision from the random utility model as described in Section A1.2 with noise level ${}^{\sigma}/_{2}$. Figure 7 depicts the false alarm probability of our test both with and without supervision and confirms the above intuition: our test indeed controls the false alarm probability when noise in evaluations decreases as the quality of submission authored by the reviewer increases.

Finally, we note that control over the false alarm probability under violation of the Assumption A2 does not come at the cost of trivial power. Figure 8 depicts the power of our test under the aforementioned setups when all reviewers manipulate using the `Distance` strategy on top of the noisy values of submissions they sample from the corresponding random utility models and confirms that our tests continue to have non-trivial power in this setup.

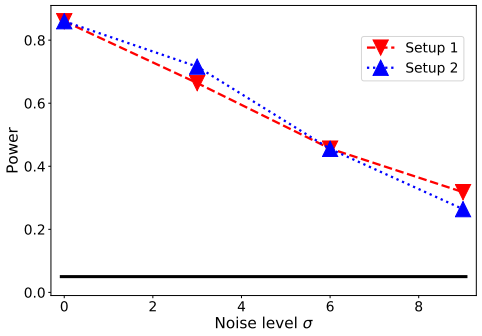


(a) With supervision

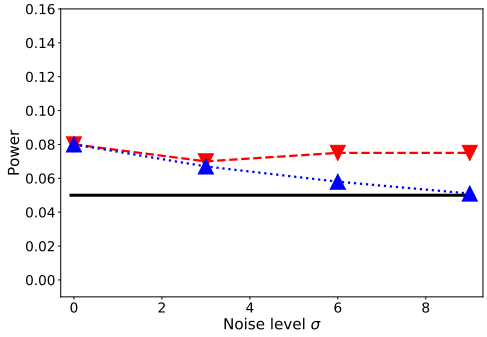


(b) Without supervision

Figure 7: False alarm probability of our test at the level $\alpha = 0.05$ when reviewer's noise depends on the quality of their submission. The black horizontal line represents the maximum false alarm probability that can be incurred by a valid test. Error bars are too small to show.



(a) With supervision



(b) Without supervision

Figure 8: Detection power of our test when reviewer's noise depends on the quality of their submission and all reviewers use the `Distance` strategy. The black horizontal line is a baseline power achieved by a test that rejects the null with probability α=0.05 irrespective of the data. Error bars are too small to show.

A1.5 Runtime of the test

In this section we continue working with the identity authorship and conflict matrices ($C = A = I$), considering student peer grading setup in which most of reviewers are conflicted only with their own work. Setting $\lambda = \mu = 4$, we estimate the runtime of our test for a wide range of sample sizes $n = m$. Specifically, we use a modification of the test that samples 100 valid authorship matrices in Step 3 of Test 1 and display the running time in Figure 9. We conclude that the running time of naive implementation of our test is feasible even for instances with thousands of reviewers and submissions.

A2 Random assignment

When evaluations are collected in the form of rankings, different structures of the assignment graph have different properties that may impact the quality of the final ordering (Shah et al., 2016). Therefore, one may want to choose a structure of the assignment graph instead of sampling it uniformly at random and we now show how to achieve any desirable structure without breaking the guarantees of our test.

Recall that m is the number of reviewers and n is the number of submissions. Let T be the desired structure of the assignment, that is, T is a bipartite graph with m nodes in the left part and n nodes in the

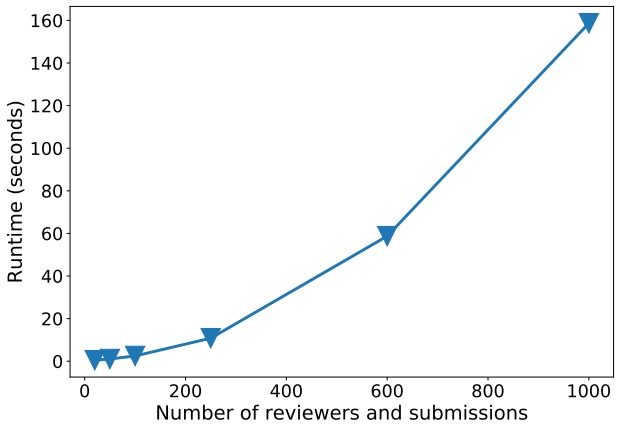


Figure 9: Running time of our test.

right part, such that each node in the left part has degree μ and each node in the right part has degree λ. Given a set of reviewers $\mathcal{R}$, a set of works $\mathcal{W}$ and a conflict matrix C, assignment M can be constructed by allocating reviewers and works to the nodes of the graph T uniformly at random, subject to the constraint that the assignment M does not violate the conflict matrix C. Note that the resulting assignment will by construction have the desired structure T.

The above procedure assumes that conflict matrix C admits structure T, that is, there exists an allocation of reviewers to nodes that does not violate T. In practice, it is always the case as reviewers are typically conflicted with only a handful of works (e.g., students in class are only conflicted with their own homework).

Observe that conditioned on topology T, the key component of the proof of Theorem 1 captured by equation (8.5) holds by construction of the assignment M, thereby ensuring that the assignment constructed in the described way does not brake the guarantees of our test. Overall, the requirement of random assignment procedure can be relaxed to the requirement of random assignment that respects a given topology T, thereby enabling deterministic selection of the assignment structure.

A3 Proof of Theorem 1

Recall that $\mathcal{W}$ is a set of works submitted for review and $\mathcal{R}$ is a set of reviewers. Let us also use A^* and C^* to denote authorship and conflict matrices up to permutations of rows and columns, that is, actual matrices A and C satisfy:

$$\begin{aligned} C &= LC^*R \\ A &= LA^*R \end{aligned} \tag{8.4}$$

where L and R are matrices of row- and column-permutations, respectively.

Let Π_m and Π_n be sets of all permutation matrices of m and n items, respectively. Conditioned on $\mathcal{W}, \mathcal{R}, C^*$ and A^*, we note that Assumptions A2 (exchangeability of reviewers and works) ensures that the actual pair of conflict and authorship matrices (C, A) follows a uniform distribution over the multiset

$$\mathcal{D} = \left\{ (LC^*R, LA^*R) \,\middle|\, (L, R) \in \Pi_m \times \Pi_n \right\}.$$

We will now show the statement of the theorem for any tuple $(\mathcal{W}, \mathcal{R}, C^*, A^*)$, thus yielding the general result. Let $(\widetilde{C}, \widetilde{A})$ be a random variable following a uniform distribution $\mathcal{U}(\mathcal{D})$ over the set $\mathcal{D}$. The proof of the theorem relies on the following fact: if assignment matrix M is selected uniformly at random from the set of all assignments valid for the conflict matrix $\widetilde{C}$, then for any pairs $(C_1, A_1) \in \mathcal{D}$ and $(C_2, A_2) \in \mathcal{D}$ that do

not violate the assignment M, it holds that:

$$\mathbb{P}\left[(\widetilde{C},\widetilde{A})=(C_1,A_1)\middle|M\right]=\mathbb{P}\left[(\widetilde{C},\widetilde{A})=(C_2,A_2)\middle|M\right], \tag{8.5}$$

where probability is taken over the randomness in the assignment procedure and uniform prior over the pair of conflict and authorship matrices. Indeed, it is not hard to see that:

$$\begin{aligned}\mathbb{P}\left[(\widetilde{C},\widetilde{A})=(C_1,A_1)\middle|M\right]&=\frac{\mathbb{P}\left[M\middle|(\widetilde{C},\widetilde{A})=(C_1,A_1)\right]\mathbb{P}\left[(\widetilde{C},\widetilde{A})=(C_1,A_1)\right]}{\mathbb{P}\left[M\right]}\\&=\frac{\mathbb{P}\left[M\middle|(\widetilde{C},\widetilde{A})=(C_2,A_2)\right]\mathbb{P}\left[(\widetilde{C},\widetilde{A})=(C_2,A_2)\right]}{\mathbb{P}\left[M\right]}\\&=\mathbb{P}\left[(\widetilde{C},\widetilde{A})=(C_2,A_2)\middle|M\right],\end{aligned}$$

where the second equality follows from the fact that

$$\mathbb{P}\left[M\middle|(\widetilde{C},\widetilde{A})=(C_1,A_1)\right]=\left|\left\{M'\middle|M' \text{ is a valid assignment for } C_1\right\}\right|^{-1}$$

and a simple observation that for any conflict matrix C' that satisfies (8.4) the number of valid assignments is the same by symmetry.

Equation (8.5) ensures that given the assignment matrix M, the randomness of the assignment procedure induces a uniform posterior distribution over the multiset $\mathcal{P}(M)$ of authorship matrices that do not violate the assignment M which we construct in Step 2 of the test. Therefore, the actual authorship matrix A is a sample from this distribution: $A\sim\mathcal{U}(\mathcal{P}(M))$.

Conditioned on the assignment M, under the null hypothesis, for each reviewer $i\in\mathcal{R}$ the ranking π_i is independent of works $A(i)$ and is therefore independent of A. Additionally, the optional impartial rankings $\{\pi^*_i, i\in\mathcal{R}\}$ are independent of the authorship matrix A by definition. Combining these observations, we deduce that conditioned on the assignment M, the test statistic τ has a uniform distribution over the multiset:

$$\Phi=\{\varphi(A')|\ A'\in\mathcal{P}(M)\},$$

where φ is defined as

$$\varphi(A')=\sum_{i\in\mathcal{R}}\sum_{j\in A'(i)}\left(\Lambda_j(\pi'_1,\pi'_2,\dots,\pi_i,\dots,\pi'_m)-\mathbb{E}_{\tilde{\pi}\sim\mathcal{U}_i}\left[\Lambda_j(\pi'_1,\pi'_2,\dots,\tilde{\pi},\dots,\pi'_m)\right]\right).$$

We finally observe that in Steps 3 and 4 of the test we reconstruct the set Φ and make decision based on the position of the actual value of the test statistic in the ordering of this set. It remains to note that probability of the event that observed value τ (sampled uniformly at random from the multiset Φ) is smaller than the value of k^{th} order statistic of the multiset Φ is upper bounded by $\frac{k-1}{|\Phi|}$. Substituting k with $(\lfloor\alpha|\Phi|\rfloor+1)$, we conclude the proof.

Chapter 9

A Study on Citation Bias in Peer Review

1 Introduction

The key decision-makers in peer review are fellow researchers with expertise in the research areas of the submissions they review. Exploiting this feature, many anecdotes suggest that adding citations to the works of potential reviewers is an effective (albeit unethical) way of increasing the chances that a submission will be accepted:

> *We all know of cases where including citations to journal editors or potential reviewers [. . .] will help a paper's chances of being accepted for publication in a specific journal.* Kostoff, 1998

The rationale behind this advice is that citations are one of the key success metrics of a researcher. A Google Scholar profile, for example, summarizes a researcher's output in the total number of citations to their work and several other citation-based metrics (h-index, i10-index). Citations are also a key factor in hiring and promotion decisions (Hirsch, 2005; Fuller, 2018). Thus, reviewers may consciously or subconsciously, be more lenient towards submissions that cite their work.

Existing research documents that the suggestion to pad reference lists with unnecessary citations is taken seriously by some authors. For example, a survey conducted by Fong and Wilhite (2017) indicates that over 40% of authors across several disciplines would preemptively add non-critical citations to their journal submission when the journal has a reputation of asking for such citations. The same observation applies to grant proposals, with 15% of authors willing to add citations even when "*those citations are of marginal import to their proposal*".

In this chapter, we investigate whether such a citation bias actually exists. We study whether a citation to a reviewer's past work induces a bias in the reviewer's evaluation. Note that citation of a reviewer's past work may impact the reviewer's evaluation of a submission in two ways: first, it can impact the scientific merit of the submission, thereby causing a *genuine change* in evaluation; second, it can induce an *undesirable bias* in evaluation that goes beyond the genuine change. We use the term "*citation bias*" to refer to the second mechanism. Formally, the research question we investigate in this work is as follows:

> **Research Question:** Does the citation of a reviewer's work in a submission *cause* the reviewer to be positively *biased* towards the submission, that is, *cause* a shift in reviewer's evaluation that goes beyond the genuine change in the submission's scientific merit?

Citation bias, if present, contributes to the unfairness of academia by making peer-review decisions dependent on factors irrelevant to the submission quality. It is therefore important for stakeholders to understand if citation bias is present, and whether it has a strong impact on the peer-review process.

Two studies have previously investigated citation bias in peer review (Sugimoto and Cronin, 2013; Beverly and Allman, 2013). These studies analyze journal and conference review data and report mixed evidence of citation bias in reviewers' recommendations. However, their analysis does not account for confounding factors such as paper quality (stronger papers may have longer bibliographies) or reviewer expertise (cited reviewers may have higher expertise). Thus, past works do not decisively answer the question of the presence of citation bias. A more detailed discussion of these and other relevant works is provided in Section 2.

Our contributions In this work, we investigate the research question in a large-scale study conducted in conjunction with the review process of two flagship publication venues: 2020 International Conference on Machine Learning (ICML 2020) and 2021 ACM Conference on Economics and Computation (EC 2021). We execute a carefully designed observational analysis that accounts for various confounding factors such as paper quality and reviewer expertise. Overall, our analysis identifies citation bias in both venues we consider: by adding a citation of a reviewer, a submission can increase the expectation of the score given by the reviewer by 0.23 (on a 5-point scale) in EC 2021 and by up to 0.42 (on a 6-point scale) in ICML 2020. For better interpretation of the effect size, we note that on average, a one-point increase in a score given by a single reviewer improves the position of a submission by 11%.

Finally, it is important to note that the bias we investigate is not necessarily an indicator of unethical behavior on the part of authors or reviewers. Citation bias may be present even when authors do not try to deliberately cite potential reviewers, and when reviewers do not consciously attempt to champion papers that cite their past work. Crucially, even subconscious citation bias is problematic for fairness reasons. Thus, understanding whether the bias is present is important for improving peer-review practices and policies.

2 Related literature

In this section, we discuss relevant past studies. We begin with an overview of cases, anecdotes, and surveys that document practices of coercive citations. We then discuss two works that perform statistical testing for citation bias in peer review. Finally, we conclude with a list of works that test for other biases in the peer-review process. We refer the reader to Shah (2022) for a broader overview of literature on peer review.

Coercive citations Fong and Wilhite (2017) study the practice of coercion by journal editors who, in order to increase the prestige of the journal, request authors to cite works previously published in the journal. They conduct a survey which reveals that 14.1% of approximately 12,000 respondents from different research areas have experienced coercion by journal editors. Resnik et al. (2008) notes that coercion happens not only at the journal level, but also at the level of individual reviewers. Specifically, 22.7% of 220 researchers from the National Institute of Environmental Health Sciences who participated in the survey reported that they have received reviews requesting them to include unnecessary references to publications authored by the reviewer.

In addition to the surveys, several works document examples of extreme cases of coercion. COPE (2018) reports that a handling editor of an unnamed journal asked authors to add citations to their work more than 50 times, three times more often than they asked authors to add citations of papers they did not co-author. The editorial team of the journal did not find a convincing scientific justification of such requests and the handling editor resigned from their duties. A similar case (Van Noorden, 2020) was uncovered in the Journal of Theoretical Biology where an editor was asking authors to add 35 citations on average to each submitted paper, and 90% of these requests were to cite papers authored by that editor. This behavior of the editor traced back to decades before being uncovered, and furthermore, authors had complied to such requests with an "apparently amazing frequency".

Given such evidence of coercion, it is not surprising that authors are willing to preemptively inflate bibliographies of their submissions either because journals they submit to have a reputation for coercion (Fong and Wilhite, 2017) or because they hope to bias reviewers and increase the chances of the submission (Meyer et al., 2009). That said, observe that evidence we discussed above is based on either case studies or surveys of authors' perceptions. We note, however, that (i) authors in peer review usually do not know identities

of reviewers, and hence may incorrectly perceive a reviewer's request to cite someone else's work as that of coercion to cite the reviewer's own work; and (ii) case studies describe only the most extreme cases and are not necessarily representative of the average practice. Thus, the aforementioned findings could overestimate the prevalence of coercion and do not necessarily imply that a submission can significantly boost its acceptance chances by strategically citing potential reviewers.

Citation bias We now describe several other works that investigate the presence of citation bias in peer review. First, Sugimoto and Cronin (2013) analyze the editorial data of the Journal of the American Society of Information Science and Technology and study the relationship between the reviewers' recommendations and the presence of references to reviewers' works in submissions. They find mixed evidence of citation bias: a statistically significant difference between *accept* and *reject* recommendations (cited reviewers are more likely to recommend acceptance than reviewers who are not cited) becomes insignificant if they additionally consider *minor/major revision* decisions. We note, however, that the analysis of Sugimoto and Cronin (2013) computes correlations and does not control for confounding factors associated with paper quality and reviewer identity (see discussion of potential confounding factors in Section 3.2.1). Thus, that analysis does not allow to test for the causal effect.

Another work (Beverly and Allman, 2013) performs data analysis of the 2010 edition of ACM Internet Measurement Conference and reports findings that suggest the presence of citation bias. As a first step of the analysis, they compute a correlation between acceptance decisions and the number of references to papers authored by 2010 TPC (technical program committee) members. For long papers, the correlation is 0.21 ($n = 109,\ p < 0.03$) and for short papers the correlation is 0.15 ($n = 102,\ p = 0.12$). Similar to the analysis of Sugimoto and Cronin (2013), these correlations do not establish causal relationship due to unaccounted confounding factors such as paper quality (papers relevant to the venue may be more likely to cite members of TPC than out-of-scope papers).

To mitigate confounding factors, Beverly and Allman (2013) perform a second step of the analysis. They recompute correlations but now use members of the *2009* TPC who are not in 2010 TPC as a target set of reviewers. Reviewers from this target set did not impact the decisions of the 2010 submissions and hence this second set of correlations can serve as an unbiased contrast. For long papers, the contrast correlation is 0.13 ($n = 109,\ p = 0.19$) and for short papers, the contrast correlation is -0.04 ($n = 102,\ p = 0.66$). While the *difference* between actual and contrast correlations hints at the presence of citation bias, we note that (i) the sample size of the study may not be sufficient to draw statistically significant conclusions (the paper does not formally test for significance of the difference); (ii) the overlap between 2010 and 2009 committees is itself a confounding factor — members in the overlap may be statistically different (e.g., more senior) from those present in only one of the two committees.

Testing for other biases in peer review A long line of literature (Mahoney, 1977; Blank, 1991; Lee, 2015; Tomkins et al., 2017; Stelmakh et al., 2020, 2021d; Manzoor and Shah, 2020, and many others) scrutinizes the peer-review process for various biases. These works investigate gender, fame, positive-outcome, and many other biases that can hurt the quality of the peer-review process. Our work continues this line by investigating citation bias.

3 Methods

In this section, we outline the design of the experiment we conduct to investigate the research question of this chapter. Section 3.1 introduces the venues in which our experiment was executed and discusses details of the experimental procedure. Section 3.2 describes our approach to the data analysis. In what follows, for a given pair of submission $\mathcal{S}$ and reviewer $\mathcal{R}$, we say that reviewer $\mathcal{R}$ is CITED in $\mathcal{S}$ if one or more of their past papers are cited in the submission. Otherwise, reviewer $\mathcal{R}$ is UNCITED.

	ICML 2020	EC 2021
# Reviewers	3,064	154
# Submissions	4,991	496
Number of submissions with at least one cited reviewer	1,513	287
Fraction of submissions with at least one cited reviewer	30%	58%

Table 1: Statistics on the venues where the experiment is executed. The number of reviewers includes all regular reviewers. The number of submissions includes all submissions that were not withdrawn from the conference by the end of the initial review period.

3.1 Experimental procedure

We begin with a discussion the details of the experiment we conduct in this work.

Experimental setting The experiment was conducted in the peer-review process of two conferences:[1]

- **ICML 2020** International Conference on Machine Learning is a flagship machine learning conference that receives thousands of paper submissions and manages a pool of thousands of reviewers.
- **EC 2021** ACM Conference on Economics and Computation is the top conference at the intersection of computer science and economics. The conference is smaller than ICML and handles several hundred submissions and reviewers.

Rows 1 and 2 of Table 1 display information about the size of the conferences used in the experiment.

The peer-review process in both venues is organized in a double-blind manner (neither authors nor reviewers know the identity of each other) and follows the conventional pipeline that we now outline. After the submission deadline, reviewers indicate their preference in reviewing the submissions. Additionally, program chairs compute measures of similarity between submissions and reviewers which are based on (i) overlap of research topics of submissions/reviewers (both conferences) and (ii) semantic overlap (Charlin and Zemel, 2013) between texts of submissions' and reviewers' past papers (ICML). All this information is then used to assign submissions to reviewers who have several weeks to independently write initial reviews. The initial reviews are then released to authors who have several days to respond to these reviews. Finally, reviewers together with more senior members of the program committee engage in the discussions and make final decisions, accepting about 20% of submissions to the conference.

Intervention As we do not have control over bibliographies of submissions, we cannot intervene on the citation relationship between submissions and reviewers. We rely instead on the analysis of observational data. As we explain in Section 3.2, for our analysis to have a strong detection power, it is important to assign a large number of submissions to both cited and uncited reviewers. In ICML, this requirement is naturally satisfied due to its large sample size, and we assign submissions to reviewers using the PR4A assignment algorithm (Stelmakh et al., 2021a) that does not specifically account for the citation relationship in the assignment.

The number of papers submitted to the EC 2021 conference is much smaller. Thus, we tweak the assignment process in a manner that gets us a larger sample size while retaining the conventional measures of the assignment quality. To explain our intervention, we note that, conventionally, the quality of the assignment in the EC conference is defined in terms of satisfaction of reviewers' preferences in reviewing the submissions, and research topic similarity. However, in addition to being useful for the sample size of our analysis, citation relationship has also been found (Beygelzimer et al., 2019) to be a good indicator for the review quality and was used in other studies to measure similarity between submissions and reviewers (Li, 2017). With

[1] In computer science, conferences are considered to be a final publication venue for research and are typically ranked higher than journals. Full papers are reviewed in CS conferences, and their publication has archival value

this motivation, in EC, we use an adaptation of the popular TMPS assignment algorithm (Charlin and Zemel, 2013) with the objective consisting of two parts: (i) conventional measure of the assignment quality and (ii) the number of CITED reviewers in the assignment. We then introduce a parameter that can be tuned to balance the two parts of the objective and find an assignment that has a large number of CITED reviewers while not compromising the conventional metrics of assignment quality. Additionally, the results of the automated assignment are validated by senior members of the program committee who can alter the assignment if some (submission, reviewer) pairs are found unsuitable. As a result, Table 1 demonstrates that in the final assignment more than half of the EC 2021 submissions were assigned to at least one CITED reviewer.

3.2 Analysis

As we mentioned in the previous section, in this work we rely on analysis of observational data. Specifically, our analysis operates with *initial reviews* that are written independently before author feedback and discussion stages (see description of the review process in Section 3.1). As is always the case for observational studies, our data can be affected by various confounding factors. Thus, we design our analysis procedure to alleviate the impact of several plausible confounders. In Section 3.2.1 we provide a list of relevant confounding factors that we identify and in Section 3.2.2 we explain how our analysis procedure accounts for them.

3.2.1 Confounding factors

We begin by listing the confounding factors that we account for in our analysis. For ease of exposition, we provide our description in the context of a naïve approach to the analysis and illustrate how each of the confounding factors can lead to false conclusions of this naïve analysis. The naïve analysis we consider compares the mean of numeric evaluations given by all CITED reviewers to the mean of numeric evaluations given by all UNCITED reviewers and declares bias if these means are found to be unequal for a given significance level. With these preliminaries, we now introduce the confounding factors.

C1 **Genuinely Missing Citations** Each reviewer is an expert in their own work. Hence, it is easy for reviewers to spot a genuinely missing citation to their own work, such as missing comparison to their own work that has a significant overlap with the submission. At the same time, reviewers may not be as familiar with the papers of other researchers and their evaluations may not reflect the presence of genuinely missing citations to these papers. Therefore, the scores given by UNCITED reviewers could be lower than scores of CITED reviewers even in absence of citation bias, which would result in the naïve test declaring the effect when the effect is absent.

C2 **Paper Quality** As shown in Table 1, not all papers submitted to the EC and ICML conferences were assigned to CITED reviewers. Thus, reviews by CITED and UNCITED reviewers were written for intersecting, but not identical, sets of papers. Among papers that were not assigned to CITED reviewers there could be papers which are clearly out of the conference's scope. Thus, even in absence of citation bias, there could be a difference in evaluations of CITED and UNCITED reviewers caused by the difference in relevance between two groups of papers the corresponding reviews were written for. The naïve test, however, will raise a false alarm and declare the bias even though the bias is absent.

C3 **Reviewer Expertise** The reviewer and submission pools of the ICML and EC conferences are diverse and submissions are assigned to reviewers of different expertise in reviewing them. The expertise of a reviewer can be simultaneously related to the citation relationship (expert reviewers may be more likely to be CITED) and to the stringency of evaluations (expert reviewers may be more lenient or strict). Thus, the naïve analysis that ignores this confounding factor is in danger of raising a false alarm or missing the effect when it is present.

C4 **Reviewer Preference** As we mentioned in Section 3.2.2, the assignment of submissions to reviewers is, in part, based on reviewers' preferences. Thus, (dis-)satisfaction of the preference may impact reviewers' evaluations — for example, reviewers may be more lenient towards their top choice submissions than to

submissions they do not want to review. Since citation relationships are not guaranteed to be independent of the reviewers' preferences, the naïve analysis can be impacted by this confounding factor.

C5 **Reviewer Seniority** Some past work has observed that junior reviewers may sometime be stricter than their senior colleagues (Toor, 2009; Tomiyama, 2007, note that some other works such as Shah et al. 2018; Stelmakh et al. 2021c do not observe this effect). If senior reviewers are more likely to be CITED (e.g., because they have more papers published) and simultaneously are more lenient, the seniority-related confounding factor can bias the naïve analysis.

3.2.2 Analysis procedure

Having introduced the confounding factors, we now discuss the analysis procedure that alleviates the impact of these confounding factors and enables us to investigate the research question. Specifically, our analysis consists of two steps: data filtering and inference. For ease of exposition, we first describe the inference step and then the filtering step.

Inference The key quantities of our inference procedure are overall scores (`score`) given in initial reviews and binary indicators of the citation relationship (`citation`). Overall scores represent recommendations given by reviewers and play a key role in the decision-making process. Thus, a causal connection between `citation` and `score` is a strong indicator of citation bias in peer review.

To test for causality, our inference procedure accounts for confounders C2–C5 (confounder C1 is accounted for in the filtering step). To account for these confounders, for each (submission, reviewer) pair we introduce several characteristics which we now describe, ignoring non-critical differences between EC and ICML. Appendix A1 provides more details on how these characteristics are defined in the two individual venues.

- `quality` Quality of the submission. The value of this quantity is, of course, unknown and below we explain how we accommodate this variable in our analysis to account for confounder C2.
- `expertise` Measure of expertise of the reviewer in reviewing the submission. In both ICML and EC, reviewers were asked to self-evaluate their ex post expertise in reviewing the assigned submissions. In ICML, two additional expertise-related measures were obtained: (i) ex post self-evaluation of the reviewer's confidence; (ii) an overlap between the text of each submitted paper and each reviewer's past papers (Charlin and Zemel, 2013). We use all these variables to control for confounding factor C3.
- `preference` Preference of the reviewer in reviewing the submission. As we mentioned in Section 3.1, both ICML and EC conferences elicited reviewers' preferences in reviewing the submissions. We use these quantities to alleviate confounder C4.
- `seniority` An indicator of reviewers' seniority. For the purpose of decision-making, both conferences categorized reviewers into two groups. While specific categorization criteria were different across conferences, conceptually, groups were chosen such that one contained more senior reviewers than the other. We use this categorization to account for the seniority confounding factor C5.

Having introduced the characteristics we use to control for confounding factors C2–C5, we now discuss the two approaches we take in our analysis.

Parametric approach (EC and ICML) First, following past observational studies of the peer-review procedure (Tomkins et al., 2017; Teplitskiy et al., 2019) we assume a linear approximation of the `score` given by a reviewer to a submission:[2]

$$\texttt{score} \sim \alpha_0 + \alpha_1 \cdot \texttt{quality} + \alpha_2 \cdot \texttt{expertise} + \alpha_3 \cdot \texttt{preference} + \alpha_4 \cdot \texttt{seniority} + \alpha^* \cdot \texttt{citation}. \quad (9.1)$$

Under this assumption, the test for citation bias as formulated in our research question reduces to the test for significance of α^* coefficient. However, we cannot directly fit the data we have into the model as the

[2]The notation $y \sim \alpha_0 + \sum_i^n \alpha_i x_i$ means that given values of $\{x_i\}_{i=1}^n$, dependent variable y is distributed as a Gaussian random variable with mean $\alpha_0 + \sum_i^n \alpha_i x_i$ and variance σ^2. The values of $\{\alpha_i\}_{i=0}^n$ and σ are unknown and need to be estimated from data. Variance σ^2 is independent of $\{x_i\}_{i=1}^n$.

values of `quality` are not readily available. Past work (Tomkins et al., 2017) uses a heuristic to estimate the values of paper quality, however, this approach was demonstrated (Stelmakh et al., 2019) to be unable to reliably control the false alarm probability.

To avoid the necessity to estimate `quality`, we restrict the set of papers used in the analysis to papers that were assigned to at least one CITED reviewer and at least one UNCITED reviewer. At the cost of the reduction of the sample size, we are now able to take a difference between scores given by CITED and UNCITED reviewers *to the same submission* and eliminate `quality` from the model (9.1). As a result, we apply a standard tools for the linear regression inference to test for the significance of the target coefficient α^*. We refer the reader to Appendix A2 for more details on the parametric approach.

Non-parametric approach (ICML) While the parametric approach we introduced above is conventionally used in observational studies of peer review and offers strong detection power even for small sample sizes, it relies on strong modeling assumptions that are not guaranteed to hold in the peer-review setting (Stelmakh et al., 2019). To overcome these limitations, we also execute an alternative non-parametric analysis that we now introduce.

The idea of the non-parametric analysis is to match (submission, reviewer) pairs on the values of all four characteristics (`quality`, `expertise`, `preference`, and `seniority`) while requiring that matched pairs have different values of `citation`. As in the parametric analysis, we overcome the absence of access to the values of `quality` by matching (submission, reviewer) pairs *within* each submission. In this way, we ensure that matched (submission, reviewer) pairs have the same values of confounding factors C2–C5. We then compare mean scores given by CITED and UNCITED reviewers, focusing on the restricted set of matched (submission, reviewer) pairs, and declare the presence of citation bias if the difference is statistically significant. Again, more details on the non-parametric analysis are given in Appendix A3.

Data filtering The purpose of the data-filtering procedure is twofold: first, we deal with missing values; second, we take steps to alleviate the genuinely missing citations confounding factor C1.

Missing Values As mentioned above, for a submission to qualify for our analysis, it should be assigned to at least one CITED reviewer and at least one UNCITED reviewer. In ICML data, 578 out of 3,335 (submission, reviewer) pairs that qualify for the analysis have values of certain variables corresponding to `expertise` and `preference` missing. The missingness of these values is due to various technicalities: reviewers not having profiles in the system used to compute textual overlap or not reporting preferences in reviewing submissions. Thus, given a large size of the ICML data, we remove such (submission, reviewer) pairs from the analysis.

In the EC conference, the only source of missing data is reviewers not entering their `preference` in reviewing some submissions. Out of 849 (submission, reviewer) pairs that qualify for the analysis, 154 have reviewer's `preference` missing. Due to a limited sample size, we do not remove such (submission, reviewer) pairs from the analysis and instead accommodate missing preferences in our parametric model (9.1) (see Appendix A1 and Appendix A2.1 for details).

Genuinely Missing Citation Another purpose of the filtering procedure is to account for the genuinely missing citations confounder C1. The idea of this confounder is that even in absence of citation bias, reviewers may legitimately decrease the score of a submission because citations to some of their own past papers are missing. The frequency of such legitimate decreases in scores may be different between CITED and UNCITED reviewers, resulting in a confounding factor. To alleviate this issue, we aim at identifying submissions with genuinely missing citations of reviewers' past papers and removing them from the analysis. More formally, to account for confounder C1, we introduce the following exclusion criteria:

> **Exclusion Criteria:** The reviewer flags a missing citation of *their own* work and this complaint is valid for reducing the score of the submission

The specific implementation of a procedure to identify submissions satisfying this criteria is different between ICML and EC conferences and we introduce it separately.

EC In the EC conference, we added a question to the reviewer form that asked reviewers to report if a submission has some important relevant work missing from the bibliography. Among 849 (submission,

reviewer) pairs that qualify for inclusion to our inference procedure, 110 had a corresponding flag raised in the review. For these 110 pairs, members of the research team (Charvi Rastogi, Federico Echenique) manually analyzed the submissions and the reviews, identifying submissions that satisfy the exclusion criteria.[3]

Overall, among the 110 target pairs, only three requests to add citations were found to satisfy the exclusion criteria. All (submission, reviewer) pairs for these three submissions were removed from the analysis, ensuring that reviews written in the remaining (submission, reviewer) pairs are not susceptible to confounding factor C1.

ICML In ICML, the reviewer form did not have a flag for missing citations. Hence, to fully alleviate the genuinely missing citations confounding factor, we would need to analyze all the 1,617 (submission, UNCITED reviewer)[4] pairs qualifying for the inference step to identify those satisfying the aforementioned exclusion criteria.

We begin from the analysis of (submission, UNCITED reviewer) pairs that qualify for our non-parametric analysis. There are 63 such pairs and analysis conducted by the author of the present thesis (who was a workflow chair of ICML 2020) found that three of them satisfy the exclusion criteria. The corresponding three submissions were removed from our non-parametric analysis.

The fraction of (submission, UNCITED reviewer) pairs with a genuinely missing citation of the reviewer's past paper in ICML is estimated to be 5% ($3/63$). As this number is relatively small, the impact of this confounding factor is limited. In absence of the missing citation flag in the reviewer form, we decided not to account for this confounding factor in the parametric analysis of the ICML data. Thus, we urge the reader to be aware of this confounding factor when interpreting the results of the parametric inference.

4 Results

As described in Section 3, we study our research question using data from two venues (ICML 2020 and EC 2021) and applying two types of analysis (parametric for both venues and non-parametric for ICML). While the analysis is conducted on observational data, we intervene in the assignment stage of the EC conference in order to increase the sample size of our study. Table 2 displays the key details of our analysis (first group of rows) and numbers of unique submissions, reviewers, and (submission, reviewer) pairs involved in our analysis (second group of rows).

The dependent variable in our analysis is the score given by a reviewer to a submission in the initial independent review. Therefore, the key quantity of our analysis (test statistic) is an expected increase in the reviewer's score due to citation bias. In EC, reviewers scored submissions on a 5-point Likert item while in ICML a 6-point Likert item was used. Thus, the test statistic can take values from -4 to 4 in EC and from -5 to 5 in ICML. Positive values of the test statistic indicate the positive direction of the bias and the absolute value of the test statistic captures the magnitude of the effect.

The third group of rows in Table 2 summarizes the key results of our study. Overall, we observe that after accounting for confounding factors, all three analyses detect statistically significant differences between the behavior of CITED and UNCITED reviewers (see the last row of the table for P values). Thus, we conclude that citation bias is present in both ICML 2020 and EC 2021 venues.

We note that conclusions of the parametric analysis are contingent upon satisfaction of the linear model assumptions and it is a priori unclear if these assumptions are satisfied to a reasonable extent. To investigate potential violation of these assumptions, in Appendix A4 we conduct analysis of model residuals. This analysis suggests that linear models provide a reasonable fit to both ICML and EC data, thereby supporting the conclusions we make in the main analysis. Additionally, we note that our non-parametric analysis makes less restrictive assumptions on reviewers' decision-making but still arrives at the same conclusion.

[3]Charvi Rastogi conducted an initial, basic screening and all cases that required a judgement were resolved by Federico Echenique – a program chair of the EC 2021 conference.

[4]Note that, in principle, CITED reviewers may also legitimately decrease the score because the submission misses some of their past papers. However, this reduction in score would lead us to an underestimation of the effect (or, under the absence of citation bias, to the counterintuitive direction of the effect) and hence we tolerate it.

		EC 2021	ICML 2020	ICML 2020
Analysis		Parametric	Parametric	Non-parametric
Intervention		Assignment Stage	No	No
Missing Values		Incorporated	Removed	Removed
Genuinely Missing Citations		Removed	Unaccounted (∼5%)	Removed
Sample Size	# Submissions (S)	283	1,031	60
	# Reviewers (R)	152	1,565	115
	# (S, R)-pairs	840	2,757	120
Test Statistic		0.23 on 5-point scale	0.16 on 6-point scale	0.42 on 6-point scale
Test Statistic (95% CI)		[0.06, 0.40]	[0.05, 0.27]	[0.10, 0.73]
P value		0.009	0.004	0.02

Table 2: Results of the analysis. The results suggest that citation bias is present in both EC 2021 and ICML 2020 conferences. P values and confidence intervals for parametric analysis are computed under the standard assumptions of linear regression. For non-parametric analysis, P value is computed using permutation test and the confidence interval is bootstrapped. All P values are two-sided.

Effect size To interpret the effect size, we note that the value of the test statistic captures the magnitude of the effect. In EC 2021, a citation of reviewer's paper would result in an expected increase of 0.23 in the score given by the reviewer. Similarly, in ICML 2020 the corresponding increase would be 0.16 according to the parametric analysis and 0.42 according to the non-parametric analysis. Confidence intervals for all three point estimates (rescaled to 5-point scale) overlap, suggesting that the magnitude of the effect is similar in both conferences. Overall, the values of the test statistic demonstrate that a citation of a reviewer results in a considerable improvement in the expected score given by the reviewer. In other words, there is a non-trivial probability of reviewer increasing their score by one or more points when cited. With this motivation, to provide another interpretation of the effect size, we now estimate the effect of a one-point increase in a score by a single reviewer on the outcome of the submission.

Specifically, we first rank all submissions by the mean score given in the initial reviews, breaking ties uniformly at random. For each submission, we then compute the improvement of its position in the ranking if one of the reviewers increases their score by one point. Finally, we compute the mean improvement over all submissions to arrive at the average improvement. As a result, on average, in both conferences a one-point increase in a score given by a single reviewer improves the position of a submission in a score-based ordering by 11%. Thus, having a reviewer who is cited in a submission can have a non-trivial implication on the acceptance chances of the submission.

As a note of caution, in actual conferences decisions are based not only on scores, but also on the textual content of reviews, author feedback, discussions between reviewers, and other factors. We use the readily available score-based measure to obtain a rough interpretation of the effect size, but we encourage the reader to keep these qualifications in mind when interpreting the result.

5 Discussion

We have reported the results of two observational studies of citation bias conducted in flagship machine learning (ICML 2020) and algorithmic economics (EC 2021) conferences. To test for the causal effect, we carefully account for various confounding factors and rely on two different analysis approaches. Overall, the results suggest that citation bias is present in peer-review processes of both venues. A considerable effect size of citation bias can (i) create a strong incentive for authors to add superfluous citations of potential reviewers, and (ii) result in unfairness of final decisions. Thus, the finding of this work may be informative for conference chairs and journal editors who may need to develop measures to counteract citation bias in peer review. In this section, we provide additional discussion of several aspects of our work.

Observational caveat First, we want to underscore that, while we try to carefully account for various confounding factors and our analysis employs different techniques, our study remains observational. Thus, the usual caveat of unaccounted confounding factors applies to our work. The main assumption that we implicitly make in this work is that the list of confounding factors C1–C5 is (i) exclusive and (ii) can be adequately modelled with the variables we have access to. As an example of a violation of these assumptions, consider that CITED reviewers could possess some characteristic that is not captured by `expertise`, `preference`, and `seniority` and makes them more lenient towards the submission they review. In this case, the effect we find in this work would not be a causation. That said, we note that to account for confounding factors, we used all the information that is routinely used in many publication venues to describe the competence of a reviewer in judging the quality of a submission.

Genuinely present citations In this work, we aim at decoupling citation bias from a genuine change in the scientific merit of a submission due to additional citation. For this, we account for the genuinely missing citations confounding factor C1 that manifests in reviewers *genuinely decreasing* their scores when their relevant past paper is not cited in the submission.

In principle, we could also consider a symmetric *genuinely present citations* confounding factor that manifests in reviewers *genuinely increasing* their scores when their relevant past work is adequately incorporated in the submission. However, while symmetric, these two confounding factors are different in an important aspect. When citation of a relevant work is missing from the submission, an author of that relevant work is in a better position to identify this issue than other reviewers and this asymmetry of information can bias the analysis. However, when citation of a relevant work is present in the paper, all reviewers observe this signal as they read the paper. The presence of the shared source of information reduces the aforementioned asymmetry across reviewers and alleviates the corresponding bias.

With this motivation, in this work we do not specifically account for the genuinely present citations confounding factor, but we urge the reader to be aware of our choice when interpreting the results of our study.

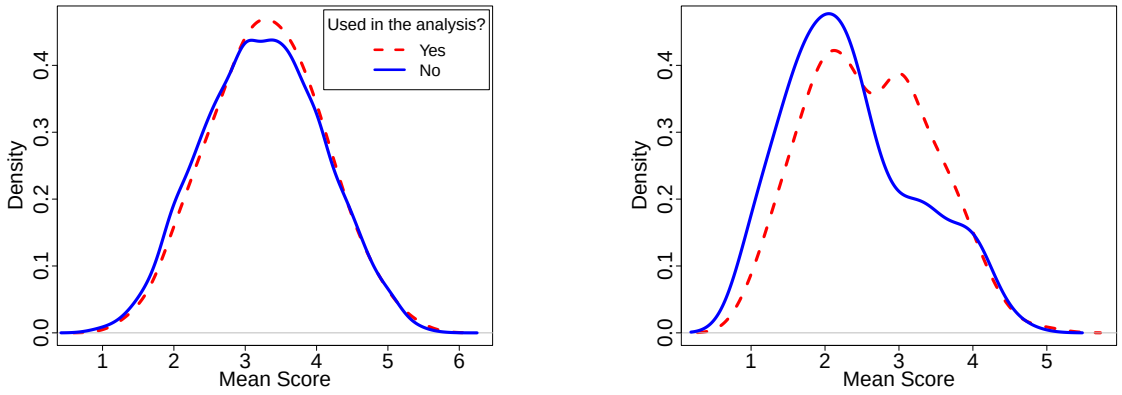
Fidelity of citation relationship Our analysis pertains to citation relationships between the submitted papers and the reviewers. In order to ensure that reviewers who are cited in the submissions are identified correctly, we developed a custom parsing tool. Our tool uses PDF text mining to (i) extract authors of papers cited in a submission (all common citation formats are accommodated) and (ii) match these authors against members of the reviewer pool. We note that there are several potential caveats associated with this procedure which we now discuss:

- **False positives** First, reviewers' names are not unique identifiers. Hence, if the name of a reviewer is present in the reference list of a submission, we cannot guarantee that it is the specific ICML or EC reviewer cited in the submission. To reduce the number of false positives, we took the following approach. First, for each reviewer we defined a *key*:

  {LAST NAME}_{FIRST LETTER OF FIRST NAME}

  Second, we considered all reviewers whose *key* is not unique in the conference they review for. For these reviewers, we manually verified all assigned (submission, reviewer) pairs in which reviewers were found to be CITED by our automated mechanism. We found that about 50% of more than 250 such cases were false positives and corrected these mistakes, ensuring that the analysis data did not have false positives among reviewers with non-unique values of their *key*.

  Third, for the remaining reviewers (those whose *key* was unique in the reviewer pool), we sampled 50 (submission, CITED reviewer) pairs from the actual assignment and manually verified the citation relationship. Among 50 target pairs, we identified only 1 false positive case and arrived at the estimate of 2% of false positives in our analysis.
- **False negatives** In addition to false positives, we could fail to identify some of the CITED reviewers. To estimate the fraction of false negatives, we sampled 50 (submission, UNCITED reviewer) pairs from the



(a) ICML 2020

(b) EC 2021

Figure 1: Distribution of mean overall scores given in initial reviews with a breakdown by whether a submission is used in our analysis or not.

actual assignment and manually verified the citation relationship. Among these 50 pairs we did not find any false negative case, which suggests that the number of false negatives is very small.

Finally, we note that both false positives and false negatives affect the power, but not the false alarm probability of our analysis. Thus, the conclusions of our analysis are stable with respect to imperfections of the procedure used to establish the citation relationship.

Generalizability of the results As discussed in Section 3, in this experiment we used submissions that were assigned to at least one CITED and one UNCITED reviewers and satisfied other inclusion criteria (see Data Filtering in Section 3.2.2). We now perform some additional analysis to juxtapose the population of submissions involved in our analysis to the general population of submissions.

Figure 1 compares distributions of mean overall scores given in initial reviews between submissions that satisfied the inclusion criteria of our analysis and submissions that were excluded from consideration. First, observe that Figure 1a suggests that in terms of the overall scores, ICML submissions used in the analysis are representative of the general ICML submission pool. However, in EC (Figure 1b), the submissions that were used in the analysis received on average higher scores than those that were excluded. Thus, we urge the reader to keep in mind that our analysis of the EC data may not be applicable to submissions that received lower scores.

One potential reason of the difference in generalizability of our EC and ICML analyses is the intervention we took in EC to increase the sample size. Indeed, by maximizing the number of submissions that are assigned to at least one CITED reviewer we could include most of the submissions that are *relevant* to the venue in the analysis, which results in the observed difference in Figure 1b.

Spurious correlations induced by reviewer identity In peer review, each reviewer is assigned to several papers. Our analysis implicitly assumes that conditioned on `quality`, `expertise`, `preference`, `seniority` characteristics, and on the value of the `citation` indicator, evaluations of different submissions made by the same reviewer are independent. Strictly speaking, this assumption may be violated by correlations introduced by various characteristics of the reviewer identity (e.g., some reviewers may be lenient while others are harsh). To fully alleviate this concern, we would need to significantly reduce the sample size by requiring that each reviewer contributes to at most one (submission, reviewer) pair used in the analysis. Given otherwise limited sample size, this requirement would put a significant strain on our testing procedure. Thus, in this work we follow previous empirical studies of the peer-review procedure (Lawrence and Cortes, 2014; Tomkins et al., 2017; Shah et al., 2018) and tolerate such potential spurious correlations. We note that simulations performed by Stelmakh et al. (2019) demonstrate that unless reviewers contribute to dozens

of data points, the impact of such spurious correlations is limited. In our analysis, reviewers on average contributed to 1.8 (submission, reviewer) pairs in ICML, and to 5.5 (submission, reviewer) pairs in EC, thereby limiting the impact of this caveat.

Counteracting the effect Our analysis raises an open question of counteracting the effect of citation bias in peer review. For example, one way to account for the bias is to increase the awareness about the bias among members of the program committee and add citation indicators to the list of information available to decision-makers. Another option is to try to equalize the number of CITED reviewers assigned to submissions. Given that Beygelzimer et al. (2019) found citation indicator to be a good proxy towards the quality of the review, enforcing the balance across submissions could be beneficial for the overall fairness of the process. More work may be needed to find more principled solutions against citation bias in peer review.

Appendix

In this section we provide additional details on our analysis procedure.

A1 Controlling for confounding factors

As described in Section 3.2.2, our analysis relies on a number of characteristics (`quality`, `expertise`, `preference`, `seniority`) to account for confounding factors C2–C5. The value of `quality` is, of course, unknown and we exclude it from the analysis by focusing on differences in reviewers' evaluations made for the same submission (details in Appendix A2.2). For the remaining characteristics, we use a number of auxiliary variables available to conference organizers to quantify these characteristics. These variables differ between conferences and Table 3 summarizes the details for both venues.

A2 Details of the parametric inference

Conceptually, the parametric analysis of both EC 2021 and ICML 2020 data is similar up to the specific implementation of our model (9.1) in these venues. In this section, we specify this parametric model for both venues using variables introduced in Table 3 (Section A2.1), and also introduce the procedure used to eliminate the `quality` variable whose values are unobserved (Section A2.2).

A2.1 Specification of parametric model

We begin by specifying the model (9.1) to each of the venues we consider in the analysis.

ICML With auxiliary variables introduced in Table 3, our model (9.1) for ICML reduces to the following specification:

$$\begin{aligned}\texttt{score} \sim \alpha_0 + \alpha_1 \cdot \texttt{quality} + \alpha_2^{(1)} \cdot \texttt{expertiseSRExp} + \alpha_2^{(2)} \cdot \texttt{expertiseSRConf} + \alpha_2^{(3)} \cdot \texttt{expertiseText} \\ + \alpha_3 \cdot \texttt{prefBid} + \alpha_4 \cdot \texttt{seniority} + \alpha^* \cdot \texttt{citation}.\end{aligned}$$

Characteristic	Auxiliary variable	EC 2021	ICML 2020
`expertise`	Self-reported expertise	In both venues, reviewers were asked to self-evaluate their ex post expertise in reviewing submissions using a 4-point Likert item. The evaluations were submitted together with initial reviews and higher values represent higher expertise. We encode these evaluations in a continuous variable `expertiseSRExp`.	
	Self-reported confidence	Not used in the conference.	Similar to expertise, reviewers were asked to evaluate their ex post confidence in their evaluation on a 4-point Likert item. We encode these evaluations in a continuous variable `expertiseSRConf`.
	Textual overlap	Not used in the conference.	TPMS measure of textual overlap (Charlin and Zemel, 2013) between a submission and a reviewer's past papers (real value between 0 and 1; higher values represent higher overlap). We denote this quantity `expertiseText`. Out of 3,335 (submission, reviewer) pairs that qualify for the analysis (before data filtering is executed), 439 pairs have the value of `expertiseText` missing due to reviewers not creating their TPMS accounts. Entries with missing values were removed from the analysis.
`preference`	Self-reported preference	Reviewers reported partial rankings of submissions in terms of their preference in reviewing them by assigning each submission a non-zero value from -100 to 100 (the higher the value the higher the preference; non-reported preferences are encoded as 0). In the automated assignment, reviewers were not assigned to papers with negative preferences. Assignment of submissions to reviewers who did not enter a preference was discouraged, but not forbidden. For analysis, we transform non-negative preferences into percentiles `prefPerc` (0 means top preference, 100 – bottom).	Reviewers bid on submissions by reporting a value from 2 (Not willing to review) to 5 (Eager to review). In the automated assignment, reviewers were not assigned to papers with bids of value 2. Assignment of submissions to reviewers who did not enter a bid was discouraged, but not forbidden. As a result, out of 3,335 (submission, reviewer) pairs that qualify for the analysis (before data filtering is executed), 159 pairs had the value of bid missing. Entries with missing values were removed from the analysis. Positive bids (3, 4, 5) are captured in the continuous variable `prefBid`.
	Missing preference	Out of 849 (submission, reviewer) pairs that qualify for the analysis (before data filtering is executed), 154 have the reviewer's preference missing. This missingness is captured in a binary indicator `missingPref`.	Not used in the analysis as data points with missing preferences are excluded from the analysis.
`seniority`	Manual classification	Program chairs split the reviewer pool in two groups: *curated* — reviewers with significant review experience or personally recommended by senior members of the program committee; *self-nominated* — reviewers who nominated themselves and satisfied mild qualification requirements.	Program chairs split the reviewer pool in two groups: *senior* and *junior*.
		In both venues, the split into groups was encoded in a binary variable `seniority` that equals 1 when a reviewer was assigned to *curated* or *senior* group and 0 otherwise.	

Table 3: Description of variables used in the analysis.

EC An important difference between our ICML and EC analyses is that in the latter we do not remove entries with missing values of auxiliary variables but instead incorporate the data missingness in the model. For this, recall that in EC, the only source of missingness is reviewers not reporting their `preference` in reviewing submissions. To incorporate this missingness, we enhance the model by an auxiliary binary variable `missingPref` that equals one when the `preference` is missing and enables the model to accommodate associated dynamics:

$$\begin{aligned}\texttt{score} \sim \alpha_0 + \alpha_1 \cdot \texttt{quality} + \alpha_2 \cdot \texttt{expertiseSRExp} + \alpha_3^{(1)} \cdot \texttt{prefPerc} + \alpha_3^{(2)} \cdot \texttt{missingPref} \\ + \alpha_4 \cdot \texttt{seniority} + \alpha^* \cdot \texttt{citation}.\end{aligned}$$

A2.2 Elimination of submission quality from the model

Having the conference-specific models defined, we now execute the following procedure to exclude the unobserved variable `quality` from the analysis. For ease of exposition, we illustrate the procedure on the model (9.1) as details of this procedure do not differ between conferences.

Step 1. Averaging scores of CITED and UNCITED reviewers Each submission used in the analysis is assigned to at least one CITED and at least one UNCITED reviewer. Given that there may be more than one reviewer in each category, we begin by averaging the scores given by CITED and UNCITED reviewers to each submission. The linear model assumptions behind our model (9.1) ensure that for each submission, averaged scores $\texttt{score}_{\text{ctd}}$ and $\texttt{score}_{\text{unctd}}$ also adhere to the following linear models:

$$\texttt{score}_{\text{ctd}} \sim \alpha_0 + \alpha_1 \cdot \texttt{quality} + \alpha_2 \cdot \texttt{expertise}_{\text{ctd}} + \alpha_3 \cdot \texttt{preference}_{\text{ctd}} + \alpha_4 \cdot \texttt{seniority}_{\text{ctd}} + \alpha^*, \quad (9.2a)$$

$$\texttt{score}_{\text{unctd}} \sim \alpha_0 + \alpha_1 \cdot \texttt{quality} + \alpha_2 \cdot \texttt{expertise}_{\text{unctd}} + \alpha_3 \cdot \texttt{preference}_{\text{unctd}} + \alpha_4 \cdot \texttt{seniority}_{\text{unctd}}. \quad (9.2b)$$

In these equations, subscripts "ctd" and "unctd" represent means of the corresponding values taken over CITED and UNCITED reviewers, respectively. Variances of the corresponding Gaussian noise in these models are inversely proportional to the number of CITED reviewers (9.2a) and the number of UNCITED reviewers (9.2b).

Step 2. Taking difference between mean scores Next, for each submission, we take the difference between mean scores $\texttt{score}_{\text{ctd}}$ and $\texttt{score}_{\text{unctd}}$ and observe that the linear model assumptions again ensure that the difference ($\texttt{score}_\Delta$) also follows the linear model:

$$\texttt{score}_\Delta \sim \alpha_2 \cdot \texttt{expertise}_\Delta + \alpha_3 \cdot \texttt{preference}_\Delta + \alpha_4 \cdot \texttt{seniority}_\Delta + \alpha^*. \quad (9.3)$$

Subscript Δ in this equation denotes the difference between the mean values of the corresponding quantity across CITED and UNCITED conditions: $X_\Delta = X_{\text{ctd}} - X_{\text{unctd}}$. Observe that by taking a difference we exclude the original intercept α_0 and the unobserved `quality` variable from the model. Thus, all the variables in the resulting model (9.3) are known and we can fit the data we have into the model. Each submission used in the analysis contributes one data point that follows the model (9.3) with a submission-specific level of noise:

$$\sigma^2 = \sigma_0^2 \left({}^1\!/_{\#\text{CITED}} + {}^1\!/_{\#\text{UNCITED}} \right),$$

where σ_0^2 is the level of noise in the model (9.1) that defines individual behavior of each reviewer.

Step 3. Fitting the data Having removed the unobserved variable `quality` from the model, we use the weighted linear regression algorithm implemented in the R `stats` package (R Core Team, 2013) to test for significance of the target coefficient α^*.

A3 Details of the non-parametric inference

Non-parametric analysis conducted in ICML 2020 consists of two steps that we now discuss.

Step 1. Matching First, we conduct matching of (submission, reviewer) pairs by executing the following procedure separately for each submission. Working with a given submission $\mathcal{S}$, we consider two groups of reviewers assigned to $\mathcal{S}$: CITED and UNCITED. Next, we attempt to find CITED reviewer $\mathcal{R}_{\text{ctd}}$ and UNCITED reviewer $\mathcal{R}_{\text{unctd}}$ that are similar in terms of `expertise`, `preference`, and `seniority` characteristics. More formally, in terms of variables we introduced in Table 3, reviewers $\mathcal{R}_{\text{ctd}}$ and $\mathcal{R}_{\text{unctd}}$ should satisfy *all of the following criteria* with respect to $\mathcal{S}$:

- Self-reported expertise of reviewers in reviewing submission $\mathcal{S}$ is the same:

$$\texttt{expertiseSRExp}_{\text{ctd}} = \texttt{expertiseSRExp}_{\text{unctd}}$$

- Self-reported confidence of reviewers in their evaluation of submission $\mathcal{S}$ is the same:

$$\texttt{expertiseSRConf}_{\text{ctd}} = \texttt{expertiseSRConf}_{\text{unctd}}$$

- Textual overlap between submission $\mathcal{S}$ and papers of each of the reviewers differ by at most 0.1:

$$|\texttt{expertiseText}_{\text{ctd}} - \texttt{expertiseText}_{\text{unctd}}| \leq 0.1$$

- Reviewers' bids on sibmission $\mathcal{S}$ satisfy one of the two conditions:

  1. Both bids have value 3 ("In a pinch"):

$$\texttt{prefBid}_{\text{ctd}} = \texttt{prefBid}_{\text{unctd}} = 3$$

  2. Both bids have values greater than 3 (4-"Willing" or 5-"Eager"):

$$\texttt{prefBid}_{\text{ctd}} \in \{4, 5\} \quad \text{and} \quad \texttt{prefBid}_{\text{unctd}} \in \{4, 5\}$$

- Reviewers belong to the same seniority group:

$$\texttt{seniority}_{\text{ctd}} = \texttt{seniority}_{\text{unctd}}$$

We run this procedure for all submissions in the pool. If for submission $\mathcal{S}$ there are no reviewers $\mathcal{R}_{\text{ctd}}$ and $\mathcal{R}_{\text{unctd}}$ that satisfy these criteria, we remove submission $\mathcal{S}$ from the non-parametric analysis. Overall, we let K denote the number of such 1-1 matched pairs obtained and introduce the set of triples that the remaining analysis operates with:

$$\left\{ \left[(\mathcal{S}^{(i)}, \mathcal{R}_{\text{ctd}}^{(i)}, \mathcal{R}_{\text{unctd}}^{(i)}) \right] \right\}_{i=1}^{K}. \tag{9.4}$$

Each triple in this set consists of submission $\mathcal{S}$ and two reviewers $\mathcal{R}_{\text{ctd}}$ and $\mathcal{R}_{\text{unctd}}$ that (i) are assigned to $\mathcal{S}$ and (ii) satisfy the aforementioned conditions with respect to $\mathcal{S}$. Within each submission, each reviewer can be a part of only one triple.

Let us now consider two (submission, reviewer) pairs associated with a given triple. Observe that these pairs share the submission, thereby sharing the value of unobserved characteristic `quality`. Additionally, the criteria used to select reviewers $\mathcal{R}_{\text{ctd}}$ and $\mathcal{R}_{\text{unctd}}$ ensures that characteristics `expertise`, `preference`, and `seniority` are also similar across these pairs. Crucially, while being equal on all four characteristics, these pairs have different values of the `citation` indicator.

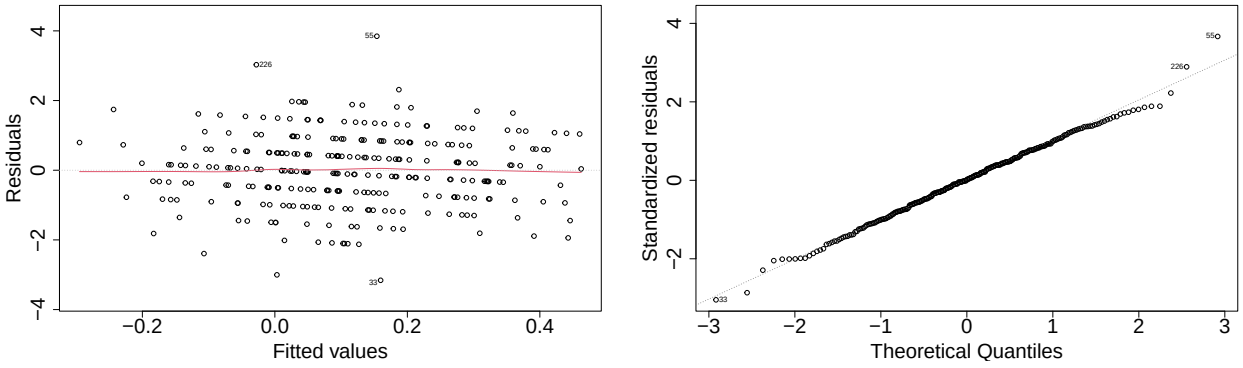
Step 2. Permutation test Having constructed the set of triples (9.4), we now compare scores given by CITED and UNCITED reviewers within these triples. Specifically, consider triple $i \in \{1, \dots, K\}$ and let $Y^{(i)}_{\text{ctd}}$ (respectively, $Y^{(i)}_{\text{unctd}}$) be the score given by CITED reviewer $\mathcal{R}^{(i)}_{\text{ctd}}$ (respectively, UNCITED reviewer $\mathcal{R}^{(i)}_{\text{unctd}}$) to submission $\mathcal{S}^{(i)}$. Then the test statistic τ of our analysis is defined as follows:

$$\tau = \frac{1}{K} \sum_{i=1}^{K} \left(Y^{(i)}_{\text{ctd}} - Y^{(i)}_{\text{unctd}} \right). \tag{9.5}$$

To quantify the significance of the difference between scores given by CITED and UNCITED reviewers, we execute the permutation test (Fisher, 1935). Specifically, at each of the 10,000 iterations, we independently permute the `citation` indicator within each triple $i \in \{1, \dots, K\}$. For each permuted sample, we recompute the value of the test statistic (9.5) and finally check whether the actual value of the test statistic τ appears to bee "too extreme" for the significance level 0.05.

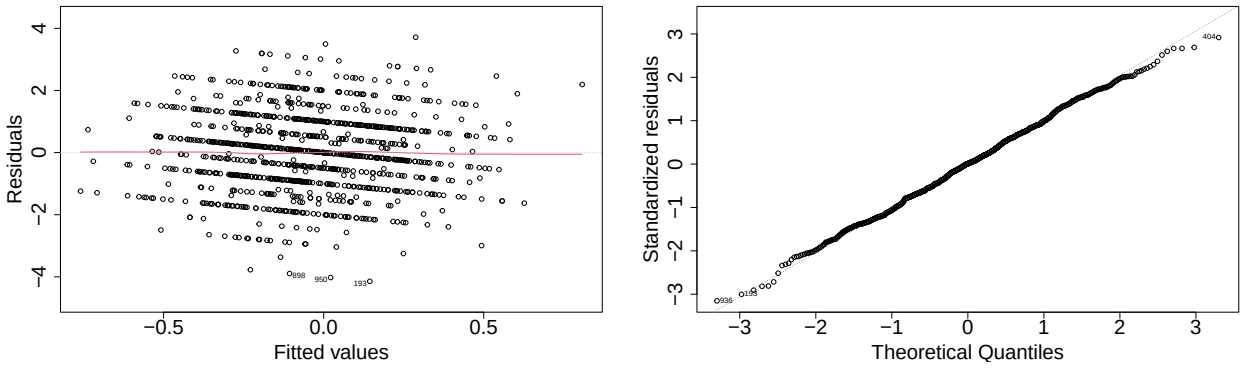
A4 Model diagnostics

Conclusions of our parametric analysis depend on the linear regression assumptions that we cannot a priori verify. To get some insight on whether these assumptions are satisfied, we conduct basic model diagnostics. Visualizations of these diagnostics are given in Figure 2 (EC 2021) and Figure 3 (ICML 2020). Overall, the diagnostics we conduct do not reveal any critical violations of the underlying modeling assumptions and suggest that our linear model (9.1) provides a reasonable fit to the data.



(a) Residuals vs Fitted (b) Normal Q-Q

Figure 2: Model diagnostics for the EC 2021 parametric analysis. Residuals do not suggest any critical violation of model assumptions.



(a) Residuals vs Fitted

(b) Normal Q-Q

Figure 3: Model diagnostics for the ICML 2020 parametric analysis. Residuals do not suggest any critical violation of model assumptions.

Chapter 10

A Novice-Reviewer Experiment to Address Scarcity of Qualified Reviewers in Large Conferences

1 Introduction

Over the last few years, Machine Learning (ML) and Artificial Intelligence (AI) conferences have been experiencing rapid growth in the number of submissions: for example, the number of submissions to AAAI and NeurIPS — popular AI and ML conferences — more than quadrupled in the last five years. The explosion in the number of submissions has challenged the sustainability of the peer-review process as the number of *qualified* reviewers is growing at a much slower rate (Sculley et al., 2019; Shah, 2022). While especially prominent in ML and AI, the problem is present in many other fields where "*submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint*" (McCook, 2006).

The disparity between growth rates of the submission and reviewer pools increases the burden on the reviewers, thereby putting a severe strain on the review process. According to the president of the International Conference on Machine Learning (ICML) board John Langford (Langford, 2018), *"There is significant evidence that the process of reviewing papers in machine learning is creaking under several years of exponentiating growth."* Hence, it is important to increase the number of qualified reviewers in the system to keep up with the growing number of submissions.

When the size of a conference is small, program chairs can extend the pool of reviewers by manually selecting new reviewers among researchers who have enough expertise in the area. The selection can be guided by the program chairs' understanding of who might be a good reviewer or by personal recommendations made by other senior members of the program committee. In what follows, we refer to the pool of reviewers manually constructed by the program chairs as the curated pool. However, with a massive increase in the scale of the conference, such a manual addition to the curated pool does not allow bringing in enough reviewers to cover the demand of the conference. The program chairs must then rely on alternative ways of reviewer recruiting.

With this motivation in mind, in the present chapter we aim to design and evaluate modifications to the reviewer recruiting process that simultaneously address two challenges:

- **Challenge 1.** To avoid overloading reviewers, conferences need to find new sources of reviewers as there are not enough curated reviewers to review all papers.
- **Challenge 2.** Conferences need to ensure that newly added reviewers do not compromise the quality of the process, that is, are able to write reviews of quality at least comparable to the curated reviewer pool.

In the past, conference organizers have been trying to expand the reviewer pool by relaxing the qualification bar, that is, by allowing researchers who meet some minimal requirements such as having one or two relevant

publications to join the pool of reviewers without further screening. For example, 1176 out of 3242 (that is, 36%) of the reviewers in the NeurIPS 2016 conference were recruited by requesting authors of each submission to name at least one author who is willing to become a reviewer, and 70% of these reviewers were PhD students: researchers at very early stages of their careers (Shah et al., 2018). Such practices have now become conventional and are adopted by many other conferences, including a flagship conference in artificial intelligence AAAI that in 2020 invited self-nominated individuals with publication history in top venues, and in 2021 requires authors of submissions to be willing to become reviewers on request. Similarly, ICML 2020 — a flagship ML conference — distributed a public call for reviewers and accepted self-nominated individuals with publication and reviewing history in top venues.

While the aforementioned innovations allow to enlarge the reviewer pool, little scientific evidence exists on the quality of reviews written by reviewers recruited through these novel procedures. NeurIPS 2016 compared the reviews written by curated reviewers with reviews sourced from authors of submissions in terms of numeric scores (overall score and several criteria scores) and inter-reviewer agreement (Shah et al., 2018). The analysis did not reveal a significant difference between populations, only showing that author-sourced reviews were slightly harsher in scoring the clarity of submissions. However, we note that this analysis only operates with scores given by reviewers and does not address the *quality* of reviews — perhaps the most important metric for success of the conference peer-review process — which is largely determined by the textual part of the review. Other works provide anecdotal and empirical evidence that junior reviewers are more critical than their senior counterparts (Mogul, 2013; Toor, 2009; Tomiyama, 2007) and that "*graduate students seem to be unable to provide very useful comments*" (Patat et al., 2019). Thus, while the methods employed by leading conferences address the first challenge, it remains unclear if and how they address the second challenge of high quality reviewing.[1]

In this work, in conjunction with the review process of ICML 2020 we conduct a threefold experiment:

- First, we recruit reviewers from the population not typically covered by the reviewer-selection process of major conferences. In that, we target the population of very junior researchers with limited or no publication/reviewing history most of whom do not pass the recruiting filters of ICML. Conceptually, in contrast to the standard approach of selecting reviewers based on some proxy towards reviewing ability (e.g., prior publication and reviewing history), we evaluate candidates' abilities to review in an auxiliary peer-review process organized for the experiment.
- Second, we add a select set of reviewers recruited through our experiment to the reviewer pool of the ICML conference and guide them through the peer-review process by offering mentoring.
- Finally, we evaluate the performance of these novice reviewers by comparing them with the general population of the ICML reviewer pool on multiple aspects. In doing so, we augment the past analysis of Shah et al. (2018) by using an explicit measure of review quality (evaluated by meta-reviewers) in addition to indirect proxies.

An important aspect of our experiment is that most of the reviewers brought to the reviewer pool through our experiment would not have been considered in standard ways of recruiting. Hence, our experiment offers a principled way to enlarge the reviewer pool. As a by-product, the new pool of reviewers contributes to diversity of peer review resonating with the virtues such as increased scrutiny and variety of opinions outlined by Garisto (2019). Moreover, we offer the new reviewers a more guided introduction to the reviewing process which is known to help novice reviewers to write better reviews (Patat et al., 2019) and improve their own writing skills (Kerzendorf et al., 2020). From the perspective of training reviewers, our experiment is conceptually similar to the initiative of *Journal of Neuroscience* (Picciotto, 2018) and SIGCOMM conference (Feldmann, 2005) that attempt to help novices in becoming reviewers.

This work falls in the line of empirical works that study various behavioral aspects of human computation, including motivational aspect (Kaufmann et al., 2011) and the impact of the task framing on performance (Kotturi et al., 2020; Levy and Sarne, 2018; Chandler and Kapelner, 2013). The findings we report in this chapter can be combined with insights from the aforementioned works to improve the design of

[1] After the experiment described in this chapter was completed, the NeurIPS 2020 conference released an analysis of their review process (Lin et al., 2020b) which compared quality of reviews sourced from authors of submissions and reviews written by the curated pool of reviewers. In Section 4 we compare the results presented therein with those of the present study.

the review process with a goal of achieving better efficiency and engagement of reviewers. Additionally, this work is complementary to a direction of research that aims at improving computational and statistical aspects of peer review (Kurokawa et al., 2015; Wang and Shah, 2019; Xu et al., 2019; Lian et al., 2018; Stelmakh et al., 2019; Fiez et al., 2020; Jecmen et al., 2020; Noothigattu et al., 2020) and results of the present study can be used to motivate future theoretical research.

The rest of the chapter is structured as follows. In Section 2 we discuss the methodology of each component of the novice-reviewer experiment. We then present the main results in Section 3. Finally, in Section 4 we conclude the chapter with a discussion of various aspects of the experiment.

2 Methodology

In this section we discuss the setup of our experiment. Specifically, we introduce the selection and mentoring mechanisms and explain the methodology of evaluation of reviewers recruited through our experiment in the ICML 2020 conference — a large venue that receives thousands of paper submissions and has more than three thousand reviewers.

2.1 Selection mechanism

The high-level idea of our selection mechanism is to pretest abilities of candidates to write high-quality reviews. To this end, we frame the experiment as an auxiliary peer-review process that mimics the pipeline of the real ML conferences as explained below and ask participants to serve as reviewers in this process. Let us now describe the experiment in detail by discussing the pools of participants and papers, the organization of the auxiliary review process, and the selection criteria we used to identify the best reviews whose authors were invited to join the ICML reviewer pool.

Papers We solicited 19 anonymized preprints in various sub-areas of ML from colleagues at various research labs, ensuring that authors of these manuscripts do not participate in the experiment as subjects. Some ML and AI conferences publicly release reviews for accepted/submitted papers, making these papers inappropriate for our experiment as our goal is to elicit independent reviews from participants. Thus, we used only those papers that did not have reviews publicly available. The final pool of papers consisted of working papers, papers under review at other conferences, workshop publications and unpublished manuscripts. The papers were 6–12 pages long excluding references and appendices (a standard range for many ML conferences) and were formatted in various popular journals' and conferences' templates with all explicit venue identifiers removed.

Participants Since we had a small quota of approximately 50 reviewers allocated for the experiment in the reviewer pool of the ICML 2020 conference, in this positional experiment we limited the target study population to graduate students or recent graduates of five large, top US universities (CMU, MIT, UMD, UC Berkeley and Stanford). To recruit participants, we messaged mailing lists of these universities and targeted master's and junior PhD students working in ML-related fields. The invitation also propagated to a small number of students outside of these schools through the word of mouth. The recruiting materials contained an invitation to participate in the ICML reviewer-selection experiment. Specifically, we notified participants that they will need to review one paper and that those who write strong reviews will be invited to join the the ICML reviewer pool. Being a reviewer in the top ML conference is a recognition of one's expertise and we envisaged that this potential benefit is a good motivation for junior researchers to join our experiment. As a result, we received responses from 200 candidates, more than 90% of whom were students/recent graduates from the aforementioned schools. All of these candidates were added to the pool of participants without further screening. We provide additional discussion of the demography of participants (including their research and reviewing experience) in Section 4.

Auxiliary peer-review process The selection procedure closely followed the initial stages of the standard double-blind ML conference peer-review pipeline and was hosted using Microsoft Conference Management Toolkit (`https://cmt3.research.microsoft.com`) which is also used in ICML. First, we asked participants

to indicate their preferences in what papers they would like to review by entering bids that take the following values: "Not Willing", "In a Pinch", "Willing" and "Eager". Thirteen participants did not enter any bids and were removed from the pool. The remaining 187 participants were active in bidding (mean number of "Willing" and "Eager" bids is 4.7) and we assigned all of them to 1 paper each, where we tried to satisfy reviewer bids, subject to a constraint that each paper is assigned to at least 8 reviewers.[2] As a result, 186 participants were assigned to a paper they bid either "Willing" or "Eager" and 1 participant was assigned to a paper they bid "In a Pinch" (this participant did not bid "Eager" or "Willing" on any paper).

Finally, we instructed participants that they should review the paper as if it was submitted to the real ICML conference with the exception that the relevance to ICML, formatting issues (e.g., page limit, margins) and potential anonymity issues should not be considered as criteria. To help participants in writing their reviews, we provided reviewer guidelines (included in supplementary materials on website of the author of this thesis) that discuss the best practices of reviewing. We gave participants 15 days to complete the review and then extended the deadline for 16 more days to accommodate late reviews as our original deadline interfered with the final exams at various US universities and the US holiday period.

Selection of participants Out of 187 participants who were assigned a paper for review, 134 handed in the reviews (response rate of 71.7%). Upon receipt of reviews, we removed numeric scores given by participants to the papers and relied on the combination of the following approaches to identify individuals to be invited to join the ICML reviewer pool:

- **Author evaluation** We asked authors of papers used in the experiment to read the reviews and rate/comment on their qualities. Authors of 14 of the 19 submissions responded to our request.
- **Internal evaluation** We analyzed reviews for 17 papers falling in the study team members' areas of expertise.
- **External evaluation** We called upon an independent domain expert to help with 2 papers that are outside of the study team members' areas of expertise.

It is natural to assume that authors are at the best position to evaluate the reviews written for their papers. Indeed, they know all the technical details of their papers, thereby being able to evaluate objective points made by reviewers. Additionally, we hypothesize that the non-competitive nature of the auxiliary review process may reduce potential biases related to a more negative perception of critical reviews which in the past were observed in some real conferences (Weber et al., 2002; Papagiannaki, 2007; Khosla et al., 2013). With this motivation, we requested authors to provide feedback on the received reviews and most of the authors fulfilled our request.

To validate our expectations regarding the quality of the author feedback, all the reviews together with the author feedback (when available) were additionally analyzed by the study team members and the aforementioned domain expert. We qualitatively observed that the author feedback is helpful to identify the strongest reviews and our selection decisions were well-aligned with the authors' evaluations. Overall, we invited 52 participants whose reviews received excellent feedback from all the evaluators who read the review to join the ICML reviewer pool; all 52 accepted the invitation. For the rest of the chapter, we will refer to these reviewers as EXPERIMENTAL reviewers.

2.2 Mentoring mechanism

Throughout the conference review process, the EXPERIMENTAL reviewers were offered additional mentorship:

- The reviewers were provided with a senior researcher as a point of contact, and were offered to ask any questions pertaining to the review process at any point in the process. There were several questions asked and answered as a part of the mentorship.
- The reviewers were provided with examples on various parts of the process, for instance, on how to lead a discussion among the reviewers.

[2]The constraint on the number of reviewers per paper was enforced to facilitate another experiment conducted in parallel with the present study and described in the companion paper (Stelmakh et al., 2021d).

- When the initial reviews were submitted, certain issues were identified that were common across many reviews from the EXPERIMENTAL pool (e.g., many reviews were initially written about the authors rather than the paper). The EXPERIMENTAL reviewers were requested to address these issues.
- The EXPERIMENTAL reviewers were sent a few more reminders than the conventional reviewers.

The total amount of time and effort in the mentorship (across all 52 EXPERIMENTAL reviewers) was equal to about half the time and effort for a meta-reviewer's job.

2.3 Methodology of evaluation

The main pool of the ICML 2020 reviewers was recruited through a combination of conventional approaches and consisted of 3,012 reviewers[3] belonging to two disjoint groups. The first group, which we refer to as CURATED, made up about 68% of the main pool and included reviewers who were invited by program chairs based on satisfaction of at least one of the following criteria: (i) several years of reviewing and publishing experience for top ML venues, (ii) above-average performance in reviewing for NeurIPS 2019 or (iii) personal recommendation by a meta-reviewer. The remaining 32% of reviewers constituted the second group that we call SELF-NOMINATED: this group comprised individuals who self-nominated and satisfied the selection criteria of (i) having at least two papers published in some top ML venues, and (ii) being a reviewer for at least one top ML conference in the past. On average, the CURATED group consisted of more senior researchers while the SELF-NOMINATED pool mostly comprised researchers at early stages of their careers.

In the sequel, we compare the performance of 52 EXPERIMENTAL reviewers who joined the ICML reviewer pool through our experiment with the performance of the reviewers from the main pool. Let us now discuss some important details of the evaluation.

Affiliation caveat 51 out of 52 EXPERIMENTAL reviewers recruited through our selection procedure are current master's and PhD students or recent graduates of the aforementioned universities (one reviewer is a graduate of another US school), whereas reviewers from the main pool represent universities as well as private companies, government organizations, non-profits and more, from all over the world. Hence, the reviewers in the main pool have different backgrounds from the EXPERIMENTAL reviewers and this difference can serve as an undesirable confounder (orthogonal to the selection procedure and mentoring) in our analysis.

To counteract this confounding factor, we identify a subset of the main pool of reviewers, whom we call COLLEAGUE reviewers. The COLLEAGUE group comprises 305 reviewers from the main pool who share an affiliation (i.e., email domain or affiliation listed on the conference management system) with the 5 schools mentioned above. In our evaluations subsequently, we additionally juxtapose the EXPERIMENTAL reviewers to this group to evaluate how they compare to reviewers of similar background, thereby alleviating the affiliation confounder.

Metrics and tools of comparison We use a set of indirect indicators of review quality (e.g., review length and discussion participation) as well as direct evaluations of review quality made by meta-reviewers of the ICML conference — senior reviewers, each of whom is in charge of overseeing the review process for approximately 20 submissions. To quantify significance of the difference in these metrics, we use the permutation test (Fisher, 1935), treating each paper-reviewer pair as a unit of analysis. Error bars presented in figures below represent bootstrapped 90% confidence intervals unless stated otherwise.

Finally, throughout the review process, meta-reviewers were calling upon additional external reviewers to help with some submissions or asking reviewers from the main pool to review additional papers; these paper-reviewer pairs are not included into comparison because new reviewers typically had less time to complete the assigned reviews.

[3]Some reviewers who initially accepted the invitation dropped out in the early stages of the review process and are not included in this number and in the subsequent analysis.

Criteria (R = reviewer, P = paper)	Range	Experimental	Main Pool	P val.
1.* Mean number of positive bids per R	$[0, 5052]$	34.6	27.4	.043
2.* Frac. of Rs with > 0 reviews completed in time	$[0, 1]$	0.92	0.81	.041
3.* Mean review length (in symbols)	$[0, \infty)$	4759	2858	$< .001$
4. Mean initial overall score given by Rs	$[1, 6]$	3.34	3.25	.373
5. Mean self-reported confidence	$[1, 4]$	3.05	3.03	.841
6.* Mean self-reported expertise	$[1, 4]$	2.83	2.98	.026
7.* Frac. of (P, R) pairs with R active in P discussion	$[0, 1]$	0.68	0.58	.033
8.* Frac. of (P, R) pairs with post-rebuttal review update	$[0, 1]$	0.61	0.43	$< .001$
9.* Mean review quality evaluated by meta-R	$[1, 3]$	2.26	2.08	$< .001$

Table 1: Performance comparison of reviewers from the main pool and experimental reviewers on various criteria. Asterisks indicate criteria with significant difference at the level 0.05.

3 Evaluation

In the previous section we described our approach towards recruiting novice reviewers and mentoring them. In this section we move to the real ICML conference and evaluate the benefit of the proposal by juxtaposing the performance of experimental reviewers to the main reviewer pool which consists of self-nominated and curated reviewers, some of whom belong to the group of colleague reviewers. For this, we compare performance of reviewers at different stages of the review process: bidding, reviewing (in-time submission, review length, self-assessed confidence and others) and discussion (activity, attention to the author feedback). Finally, we complement the comparison by overall evaluation of the review quality made by meta-reviewers.

Table 1 summarizes the results of comparison of experimental reviewers with reviewers from the main pool; subsequently, we will present a more detailed analysis with breakdown by reviewer groups. The main message of Table 1 is that from various angles the reviews written by experimental reviewers are comparable to or sometimes even better than reviews written by reviewers from the main pool. With this general observation, we now provide details and background for each row of Table 1.

Bidding activity (Row 1 of Table 1) Algorithms for automated paper-reviewer matching significantly rely on reviewer bids (Fiez et al., 2020). Hence, activity of reviewers in the bidding stage is crucial to ensure that submissions are assigned to reviewers with appropriate expertise. To give matching algorithms enough flexibility, ICML program chairs requested reviewers to positively bid (i.e., indicate papers they are "Willing" or "Eager" to review) on at least 30-40 submissions (out of approximately 5,000 submitted for review).

Figure 1 compares mean numbers of positive and non-negative ("Willing", "Eager" and "In a Pinch") bids

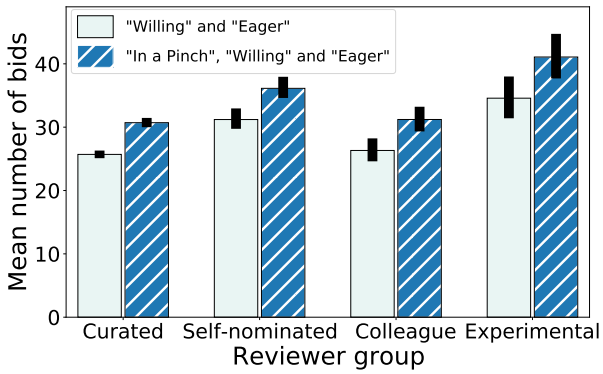


Figure 1: Mean number of positive/non-negative bids per reviewer. Experimental reviewers positively bid on more papers than reviewers from each of the comparison groups.

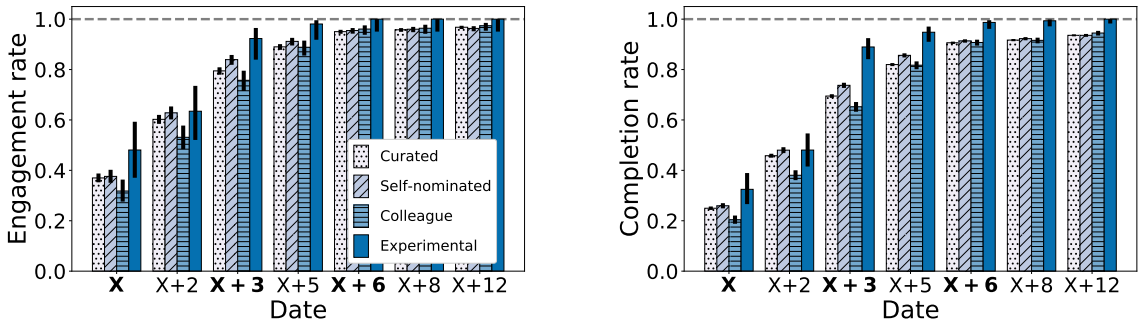
Bids		Experimental	Main Pool	Curated	Self-Nominated	Colleague
Positive	Sample Size	52	3012	2060	952	305
	Mean Value	34.6	27.4	25.7	31.2	26.3
	90% CI	[31.4; 38.1]	[26.8; 28.1]	[25.1; 26.3]	[29.8; 33.0]	[24.6; 28.3]
	P value	–	.043	.003	.221	.011
Non-Negative	Sample Size	52	3011	2059	952	305
	Mean Value	41.1	32.4	30.7	36.1	31.2
	90% CI	[37.6; 44.7]	[31.7; 33.2]	[30.0; 31.5]	[34.6; 37.9]	[29.3; 33.3]
	P value	–	.046	.007	.165	.005

Table 2: Comparison of bidding activity of the reviewers. P values are for the test of the difference of means between experimental and each of the other groups of reviewers.

made by reviewers from different groups. Note that to compare non-negative bids we remove one reviewer who bulk bid "In a Pinch" on all non-conflicting submissions. Several reviewers bid on hundreds of submissions (possibly by bulk bidding on some specific areas or keywords), but we do not exclude such reviewers from the analysis.

Overall, we observe that experimental reviewers are more active than other categories of reviewers with a qualification that the difference with self-nominated reviewers ($\Delta_{\text{positive}} = 3.4,\ \Delta_{\text{non-negative}} = 5.0$) is not statistically significant at the 0.05 significance level as demonstrated in Table 2 that summarizes the results of comparison.[4]

Timely review submission (Row 2 of Table 1) A typical conference timeline is very tight and it is crucial that reviewers complete their reviews in a timely manner. We now compare how different groups of reviewers respect the deadlines. For this, we use two metrics: first, Figure 2a juxtaposes engagement rates — fractions of reviewers who submitted at least one review by a given date — of different reviewer groups. Second, Figure 2b compares completion rates — the total number of submitted reviews divided by the total number of assigned papers. While the completion rate is perhaps a more intuitive choice of metric, it is artificially favourable to experimental reviewers due to a difference in the reviewer loads between experimental reviewers and reviewers from the main pool (see more discussion in Section 4). To counteract



(a) Engagement rates.

(b) Completion rates.

Figure 2: Timely review submission. Bold labels indicate dates at which deadlines were set with the original deadline on day X and two extensions. 90% confidence intervals are computed using the method of Wilson (1927). Experimental reviewers have higher engagement and completion rates than other reviewers.

[4]We also observe that self-nominated reviewers are more active than curated reviewers, suggesting that the bidding activity may be decreasing with seniority.

	Experimental	Main Pool	Curated	Self-Nominated	Colleague
Sample Size	52	3007	2055	952	305
Mean Value	0.92	0.81	0.79	0.84	0.76
90% Wilson CI	[0.84; 0.96]	[0.80; 0.82]	[0.78; 0.81]	[0.82; 0.86]	[0.71; 0.80]
P value	–	.041	.028	.140	.006

Table 3: Comparison of engagement rates of reviewers on the first extended deadline. P values are for the test of the difference of means between experimental and each of the other groups of reviewers.

the bias in objective, for formal comparisons presented in Tables 1 and 3 we use the engagement rate which is less impacted by the difference in loads.

Looking at Figure 2, we again observe the trend of junior self-nominated and experimental reviewers being consistently more active than their senior curated counterparts throughout the whole review-submission period, with experimental reviewers achieving the highest engagement and completion rates across all reviewer groups. Note that due to the impact of the COVID-19 pandemic, the initial deadline for review submission on day X was extended twice (deadline dates are highlighted in bold in Figure 2) with the first extension announced well in advance and hence only a small fraction of reviews was submitted by day X. Thus, in Table 1 we use data for the first extended deadline on day X+3.

Table 3 extends the comparison reported in Table 1 by displaying a breakdown by reviewer groups.[5] The results of the permutation test qualify the observations we made from Figure 2 by showing that the difference between experimental and self-nominated reviewers ($\Delta = 0.08$), who represent the more junior population of the main reviewer pool, is not significant at the requested level.

Before we proceed to other dimensions of comparison, we note that a small number of reviewers from the main pool never submitted reviews for some of the assigned papers (less than 5% of paper-reviewer pairs had no review submitted). Corresponding paper-reviewer pairs are excluded from the analysis of various aspects of review quality we perform below.

Review length (Row 3 of Table 1) We continue the analysis by juxtaposing the lengths of textual comments submitted by reviewers in Figure 3. We observe that different categories of reviewers from the main pool appear to write reviews of comparable length whereas experimental reviewers write considerably longer reviews. The distribution of lengths of reviews written by reviewers from the main pool is very similar to that of several major ML conferences (Beygelzimer et al., 2019), and thus we conclude that experimental reviewers produced longer reviews than standard in the field. Table 4 compares mean lengths of reviews written by reviewers from different groups and confirms the intuition represented in Figure 3.

	Experimental	Main Pool	Curated	Self-Nominated	Colleague
Sample Size	154	15206	10502	4704	1593
Mean Value	4759	2858	2953	2647	2985
90% CI	[4432; 5089]	[2836; 2880]	[2926; 2979]	[2609; 2685]	[2924; 3047]
P value	–	$< .001$	$< .001$	$< .001$	$< .001$

Table 4: Comparison of mean lengths (in symbols) of reviews. P values are for the test of the difference of means between experimental and each of the other groups of reviewers.

[5]The number of reviewers used in the comparison is smaller than the total number of reviewers because we only use paper-reviewer pairs that were in the assignment from the beginning of the review period and 5 reviewers from the curated group with small initial loads had a set of their papers fully changed throughout the process.

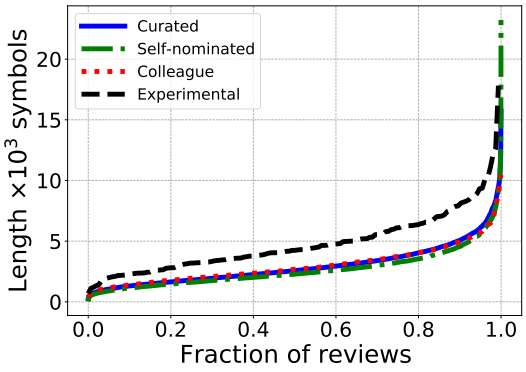


Figure 3: Distribution of review lengths. EXPERIMENTAL reviewers write longer reviews than other reviewers.

Hypercriticality (Row 4 of Table 1) Although junior reviewers are often perceived to be more critical (Mogul, 2013; Tomiyama, 2007; Toor, 2009), the analysis of the NeurIPS 2016 conference conducted by Shah et al. (2018) does not reveal a significant difference between overall scores given by junior and senior reviewers. We now perform such an analysis for the ICML conference: in ICML 2020 reviewers were asked to give the overall score on a 6-point Likert item and we encode the options with integers from 1 to 6 such that the larger number indicates the better score. The mean overall scores given in *initial* reviews (i.e., before reviewers got to see other reviews and the author rebuttal) are compared across different groups of reviewers in Figure 4.

Mean initial overall scores given by different groups of reviewers appear to be comparable; SELF-NOMINATED and EXPERIMENTAL reviewers seem to be slightly more lenient than CURATED reviewers, but the sample size of EXPERIMENTAL reviewers is not sufficient to draw definitive conclusions (Table 5 summarizes the comparison). However, we note that SELF-NOMINATED reviewers are indeed more lenient than CURATED reviewers ($\Delta = 0.14, P < .001$), contradicting the aforementioned observations of hypercriticality in junior reviewers. This misalignment may be specific to the field of computer science where hypercriticality is not limited to junior reviewers, but is prevalent in the whole area (Vardi, 2010), or, alternatively, it is possible that hypercriticality of junior reviewers manifests not in numeric scores but in textual reviews.

Expertise and confidence (Rows 5 and 6 of Table 1) We now continue with the analysis of self-assessed confidence and expertise of different reviewer groups (using values reported in initial reviews). The reviewer form of the ICML 2020 conference contained two questions in which reviewers were asked to evaluate their expertise and confidence in their review on 4-point Likert items. We encode the options of the Likert

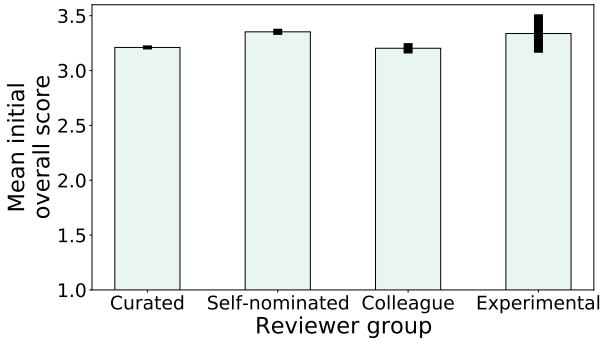
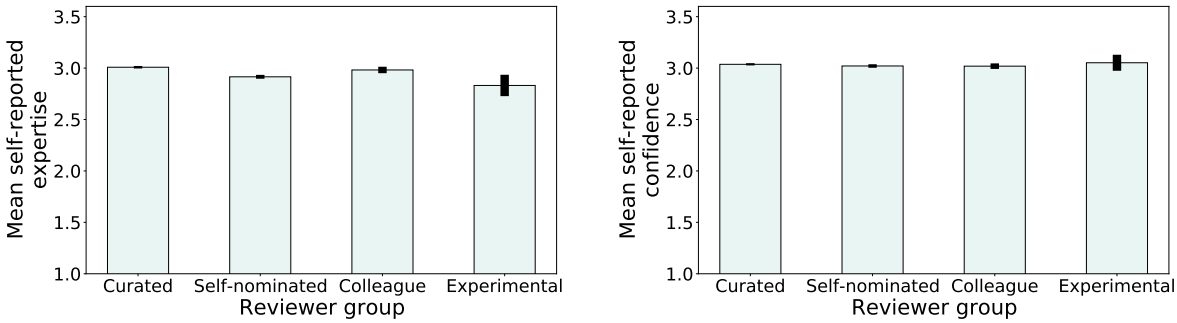


Figure 4: Mean initial overall scores. SELF-NOMINATED and EXPERIMENTAL reviewers appear to be more lenient than CURATED reviewers.

	Experimental	Main Pool	Curated	Self-Nominated	Colleague
Sample Size	154	15206	10502	4704	1593
Mean Value	3.34	3.25	3.21	3.35	3.20
90% CI	[3.17; 3.51]	[3.24; 3.27]	[3.19; 3.23]	[3.33; 3.38]	[3.16; 3.25]
P value	–	.373	.199	.895	.175

Table 5: Comparison of mean initial overall scores. P values are for the test of the difference of means between experimental and each of the other groups of reviewers.



(a) Self-assessed expertise.

(b) Self-assessed confidence.

Figure 5: Comparison of self-assessed expertise and confidence. Experimental reviewers report considerably lower expertise than other groups of reviewers, but there is no significant difference in self-assessed confidence.

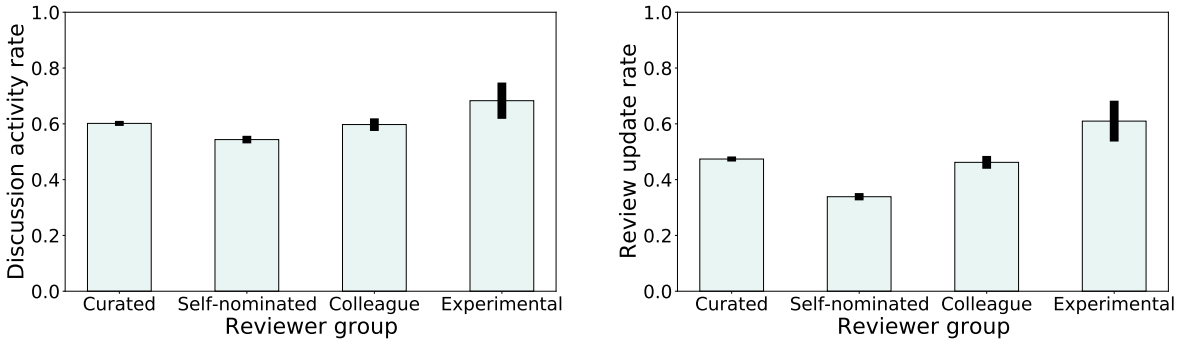
items with integer numbers from 1 to 4 such that larger numbers indicate higher expertise/confidence. We then compare mean scores across reviewer groups and report the results in Figure 5 and Table 6.

Not surprisingly, experimental reviewers reported lower self-assessed expertise (Figure 5a) with a caveat that the difference with self-nominated reviewers is not statistically significant. That said, the difference in expertise does not seem to result in the difference in self-assessed confidence (Figure 5b).

Rebuttals and discussion (Rows 7 and 8 of Table 1) The review process of ICML allows authors to respond to initial reviews written for their papers by submitting a short rebuttal that is followed by a private discussion between reviewers and the meta-reviewer. Past analysis (Shah et al., 2018; Gao et al., 2019; ACLCommittee, 2018) provide mixed evidence regarding the usefulness of rebuttals, and in this work we do

Criteria		Experimental	Main Pool	Curated	Self-Nominated	Colleague
Expertise	Sample Size	154	15206	10502	4704	1593
	Mean Value	2.83	2.98	3.01	2.92	2.98
	90% CI	[2.73; 2.94]	[2.97; 2.99]	[3.00; 3.02]	[2.90; 2.93]	[2.95; 3.01]
	P value	–	.026	.005	.204	.021
Confidence	Sample Size	154	15206	10502	4704	1593
	Mean Value	3.05	3.03	3.04	3.02	3.02
	90% CI	[2.97; 3.13]	[3.02; 3.04]	[3.03; 3.05]	[3.00; 3.04]	[2.99; 3.05]
	P value	–	.841	.829	.646	.579

Table 6: Comparison of self-assessed expertise (first 4 rows) and confidence (last 4 rows). P values are for the test of the difference of means between experimental and each of the other groups of reviewers.



(a) Participation in discussion.

(b) Post-rebuttal review update.

Figure 6: Activity in the last stage of the review process. EXPERIMENTAL reviewers participate in the discussion and update textual reviews more often than other reviewers.

not aim to judge the overall efficacy of the rebuttal process. However, in order for the rebuttal or discussion to change the reviewer's opinion, reviewers at the very least need to consider the rebuttal and be engaged in the discussion. We now investigate this aspect, conditioning on papers whose authors supplied a response to initial reviews (approximately 80% of submissions had the author response provided).

Figure 6 compares the fractions of paper-reviewer pairs such that the reviewer posted at least one message in the discussion thread (discussion activity rate, Figure 6a) / updated the textual review after the rebuttal (review update rate, Figure 6b), formal results of comparison are summarized in Table 7. We note that in both dimensions EXPERIMENTAL reviewers are more active than other categories of reviewers.[6]

Review quality (Row 9 of Table 1) So far we have observed that EXPERIMENTAL reviewers are more active in all stages of the review process than reviewers from the main pool. However, the comparisons above do not decisively answer the question of *quality* of reviews written by the new reviewers. To bridge this gap, we now report the evaluations of review quality made by meta-reviewers. At the end of the review process, meta-reviewers were asked to evaluate the quality of each review on a 3-point Likert item with the following options: "Failed to meet expectations" (Score 1), "Met expectations" (Score 2), "Exceeded expectations" (Score 3). Not all meta-reviewers completed the evaluation (approximately 25% of all paper-reviewer pairs

CRITERIA		EXPERIMENTAL	MAIN POOL	CURATED	SELF-NOMINATED	COLLEAGUE
DISCUSSION ACTIVITY	SAMPLE SIZE	123	11985	8298	3687	1262
	MEAN VALUE	0.68	0.58	0.60	0.54	0.60
	90% CI	[0.61; 0.75]	[0.58; 0.59]	[0.59; 0.61]	[0.53; 0.56]	[0.57; 0.62]
	P VALUE	–	.033	.083	.004	.077
REVIEW UPDATE	SAMPLE SIZE	123	11985	8298	3687	1262
	MEAN VALUE	0.61	0.43	0.47	0.34	0.46
	90% CI	[0.54; 0.68]	[0.42; 0.44]	[0.46; 0.48]	[0.33; 0.35]	[0.44; 0.48]
	P VALUE	–	$< .001$	.005	$< .001$	.003

Table 7: Comparison of post-rebuttal reviewer activity: participation in discussion (first 4 rows) and review update rate (last 4 rows). P values are for the test of the difference of means between EXPERIMENTAL and each of the other groups of reviewers.

[6]Interestingly, Figure 6 shows that SELF-NOMINATED reviewers are less engaged in the last stage of the review process than senior CURATED reviewers. This observation suggests that the relative engagement of junior SELF-NOMINATED reviewers decreases as the review process progresses. We do not see this in the EXPERIMENTAL reviewers and hypothesize that more tailored mentoring leads to a consistent engagement of EXPERIMENTAL reviewers.

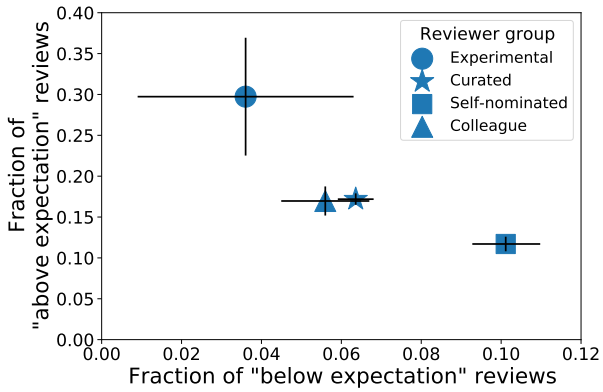


Figure 7: Evaluation of review quality by meta-reviewers. The closer the point to the upper-left corner, the better. EXPERIMENTAL reviewers dominate other groups of reviewers.

did not have a meta-reviewer evaluation) so for comparison of review quality we limit the attention to paper-reviewer pairs for which the corresponding meta-reviewer rated the quality of the review. Importantly, meta-reviewers were not aware of the group affiliation of reviewers. In the corresponding row of Table 1, we compare mean quality scores between different categories of reviewers and Table 8 complements the comparison by additionally presenting a breakdown by reviewer groups, confirming that EXPERIMENTAL reviewers receive higher scores than their counterparts.

In contrasts to Tables 1 and 8 that compare mean ratings, Figure 7 visualizes the fraction of reviews below and above the expectations of meta-reviewers within each group. Observe that reviews written by EXPERIMENTAL reviewers exceed expectations of meta-reviewers more often than reviews of CURATED and SELF-NOMINATED reviewers, and conditioning on the affiliation does not impact the comparison. Figure 7 also shows that EXPERIMENTAL reviewers produced substandard reviews less often than other reviewers.

4 Discussion

In this work we designed and executed an experimental procedure for novice reviewer recruiting and mentoring with a goal to address scarcity of qualified reviewers in large conferences. We evaluated the results of the experiment by juxtaposing the performance of the new reviewers to the traditional reviewer pool in the real ICML 2020 conference. We now provide additional discussion of the recruiting and evaluation procedures, comment on the scalability of the experiment, and suggest directions for future work.

4.1 Discussion of the recruiting experiment

We begin from some important aspects of the auxiliary peer-review process we conducted to recruit reviewers. First, we perform the analysis of the demography of participants. After that, we mention another potentially

	EXPERIMENTAL	MAIN POOL	CURATED	SELF-NOMINATED	COLLEAGUE
SAMPLE SIZE	111	11624	8035	3589	1179
MEAN VALUE	2.26	2.08	2.11	2.02	2.11
90% CI	$[2.18; 2.34]$	$[2.07; 2.09]$	$[2.10; 2.12]$	$[2.00; 2.03]$	$[2.09; 2.14]$
P VALUE	–	$< .001$	.002	$< .001$	.003

Table 8: Comparison of mean review qualities as evaluated by meta-reviewers. P values are for the test of the difference of means between EXPERIMENTAL and each of the other groups of reviewers.

	All	Invited
Total number	134	52
With prior review experience	36%	37%
With publications	72%	77%
Pass reviewing filter	21%	21%
Pass publication filter	24%	23%
Pass self-nominated filters	**13%**	**12%**

Table 9: Demography of subjects of our experiment.

useful aspect of the experiment related to reviews written in this auxiliary review process.

Demography of participants Recall that self-nominated individuals had to pass a publication and reviewing filters (see Section 2.3 for details) to join the self-nominated reviewer pool of ICML 2020. We now test whether the subjects of our experiment satisfy these criteria. Table 9 comprises relevant demographic information for 134 subjects of the selection experiment who completed the participation (that is, submitted the review of the assigned paper) and for 52 subjects who were eventually invited to join the ICML 2020 experimental reviewer pool. Importantly, this demographic information was hidden from evaluators who performed the selection. Observe that most of the participants of our experiment (including those that were selected to join the ICML reviewer pool) do not pass at least one of the filters mandatory for the self-nominated reviewers. Similarly, none of the participants was invited to join the curated reviewer pool. Therefore, we conclude that most of the participants of our experiment would not have been invited to ICML 2020 through conventional ways of reviewer recruiting.

Reviews As a byproduct of our experiment, authors of manuscripts we used in the auxiliary peer-review process received a set of reviews. They generally admitted a high quality of reviews: several authors mentioned that reviewers found some errors/important typos in their papers or suggested some ways to improve the presentation. An author of one paper comments:

> "*Very high quality reviews, in my opinion. Most of them [...] are clearly more detailed and more useful than reviews we received at [another top ML venue]*"

and an author of another manuscript says:

> "*Those [reviews] were super informative and helpful!*"

As a minor qualification, we underscore that while the overall feedback was positive, authors also helped us to identify a number of reviews of substandard quality for ICML (with serious factual errors or dismissive criticism) and we found the author feedback to be very useful for the selection. That said, the positive overall feedback from authors hints that a large scale version of our experiment can give researchers a set of useful reviews before they submit a paper to a real conference, potentially decreasing the load on the actual conferences.

4.2 Discussion of the evaluation

We now discuss some aspects important for interpretation of the results of the comparison between experimental reviewers and the main reviewer pool of ICML 2020.

Aspect 1. Assignment procedure Each paper submitted to the ICML conference was first automatically assigned to three reviewers from the main pool, two of whom were curated reviewers and one was a self-nominated reviewer. In that, we tried to satisfy reviewer bids and optimize for the notion of textual similarity (Charlin and Zemel, 2013) between submissions and assigned reviewers, subject to a requirement that each reviewer is assigned at most six papers (a small set of reviewers requested a lower quota) and under

	Main Pool	Experimental
# Reviewers	3012	52
Mean reviewer load	5.4	3.0
Fraction of positive bids	0.88	0.99
Mean paper↔reviewer similarity	0.77	0.88

Table 10: Assignment quality. "Fraction of positive bids" represents a fraction of paper-reviewer pairs in the assignment such that the reviewer has bid positively on the paper. Similarities between papers and reviewers take values in the interval $[0, 1]$ and are computed using the TPMS system (Charlin and Zemel, 2013). 706 reviewers in the main pool and 4 EXPERIMENTAL reviewers did not have similarities computed and are excluded from the computation of mean similarity.

a fairness constraint that aims at balancing assignment quality across submissions (Stelmakh et al., 2021a). After that, each EXPERIMENTAL reviewer was manually assigned to three submissions as a fourth reviewer. All assignments were finally adjusted by meta-reviewers before being released to the reviewers. Given the small number of EXPERIMENTAL reviewers and the large number of submissions, the constraints we had to satisfy in the manual assignment were mild as compared to the main assignment. Hence, EXPERIMENTAL reviewers could receive submissions that better fit their expertise than reviewers from the main pool.

Table 10 compares several metrics of assignment quality across reviewers from the main pool and EXPERIMENTAL reviewers. First, we note that EXPERIMENTAL reviewers were intentionally assigned fewer papers than reviewers from the main pool to ensure a gentle introduction to the review process. We underscore that in this study we aim to show that reviewers from the population not covered by current recruiting methods can usefully augment the reviewer pool given some special treatment (careful recruiting, reduced load and mentoring). Hence, we do not consider the difference in load to be a confounder in our comparisons.

Second, EXPERIMENTAL reviewers were assigned to papers they positively bid on and to papers with high textual similarity more often than reviewers from the main pool. Hence, we caveat that the quality of the assignment differs significantly between reviewers from the main pool and EXPERIMENTAL reviewers, introducing a confounding factor that can potentially impact the comparison. On the other hand, the difference in the assignment quality may in part be due to the difference in bidding activity that we outlined in Section 3: a large number of positive bids made by EXPERIMENTAL reviewers gives more flexibility in satisfying them, thereby increasing the quality of the assignment. It will be of interest to investigate, in any future larger scale studies, whether a larger number of EXPERIMENTAL reviewers also continue to have such a higher quality of assignment due to higher bidding activity, and if not, then it will be of interest to observe how that impacts the other metrics.

Aspect 2. Quality evaluation Following a standard approach in AI/ML conferences that conduct a survey of meta-reviewers on the review quality, when asking meta-reviewers to evaluate the quality of reviews, we left it for the meta-reviewers to decide on their expectations and did not precisely define the term "quality". As a result, different meta-reviewers could have different standards in mind, leading to some inconsistency in evaluations. To account for this issue, in the Appendix we complement the above analysis of review quality and some other metrics with additional comparisons performed on a restricted set of submissions that had at least one EXPERIMENTAL reviewer assigned (by doing so we equalize the sets of meta-reviewers who rate EXPERIMENTAL reviewers and other categories of reviewers as well as other characteristics of papers assigned to different groups of reviewers). Importantly, this analysis leads to the same conclusions as the analysis we described in Section 3.

Another related caveat is that the absence of a well-defined notion of quality could result in the substitution bias in meta-reviewers' judgments of review quality. For example, meta-reviewers' evaluations could be driven by the length of the review or some other computationally inexpensive, but suboptimal, proxy, resulting in a biased evaluation of quality. We urge the reader to be aware of this issue when interpreting the results of the quality evaluations.

Aspect 3. Insights from NeurIPS 2020 review process After the experiment we describe in this chapter was completed, the NeurIPS 2020 conference released the analysis of its review process (Lin et al., 2020b) in which they compared the quality of reviews written by curated and author-sourced reviewers (to avoid ambiguity, in this section we refer to curated NeurIPS reviewers as invited reviewers). In terms of the selection criteria, these groups of reviewers roughly correspond to CURATED and SELF-NOMINATED groups we consider in the present study, with an important exception that in our case SELF-NOMINATED reviewers were not required to have their paper submitted to ICML 2020. There may also be some subtle differences in how the pools of invited NeurIPS reviewers and CURATED ICML reviewers were constructed. With these caveats, the analysis of NeurIPS 2020 data presents two key insights relevant to our study that we now discuss.

First, NeurIPS data suggests that the quality (as measured by meta-reviewers) of reviews written by author-sourced reviewers is only marginally worse than that of invited reviewers. This observation agrees with what we report in Table 8 where CURATED reviewers have slightly higher mean review quality than their SELF-NOMINATED counterparts, but the difference is somewhat more pronounced in our case.

Second, and perhaps more importantly, it appears that in NeurIPS the quality of reviews was negatively correlated with experience of reviewers: reviewers for whom NeurIPS 2020 was the first big ML conference they serve for appear to produce reviews of higher quality than their counterparts who have reviewed for major conferences before. In ICML 2020, all but perhaps 5–10% of members of the main reviewer pool had past experience of being a reviewer for some top ML venues,[7] while most of EXPERIMENTAL reviewers did not have such experience. Hence, the result of NeurIPS analysis highlights another potential confounding factor in our study: past reviewing experience. Indeed, hypothetically the results of our experiment can be explained by the fact that reviewers put more efforts into their first reviews and then become less engaged in future conferences.

Let us now qualify the above caveat. First, in our experiment the difference between EXPERIMENTAL reviewers and reviewers from the main ICML reviewer pool appears to be considerably larger than the difference between the first-time reviewers and those who have reviewed before observed in NeurIPS. The larger effect size hints at the potential effect of our intervention. Second, in the present work we carefully account for the affiliation confounding factor and additionally juxtapose the EXPERIMENTAL reviewers with the COLLEAGUE reviewers who share the same affiliation. The NeurIPS analysis does not provide such comparison and hence we cannot remove this confounding factor. Finally, in the present work we compare the performance of reviewers on various metrics beyond the review quality and it would be interesting to see how novice NeurIPS reviewers perform on these metrics.

Aspect 4. Additional caveats In addition to the caveats mentioned above, we would like to make several other remarks:

- First, while we measured a number of metrics that were possible to measure and have been considered in the literature, we cannot exclude the possibility that EXPERIMENTAL reviewers may be worse than the main pool of reviewers in some other aspect not considered here.
- Second, it is possible that behavior of EXPERIMENTAL reviewers was affected by demand characteristics McCambridge et al. (2012), that is, EXPERIMENTAL reviewers could hypothesize that we want them to perform better than reviewers from the main pool and hence they could adjust their behaviour to meet these perceived expectations.
- Third, the review process of the ICML 2020 conference was impacted by the COVID-19 pandemic and the impact of the pandemic on different reviewer groups could be unequal. For example, senior reviewers from the CURATED pool could have more family-related duties (and hence could be more restricted in reviewing ability) than junior EXPERIMENTAL reviewers.
- Finally, in extrapolating any results to other conferences, one should carefully consider any idiosyncrasies of specific conferences.

[7] Some CURATED reviewers were recommended by the meta-reviewers and are not guaranteed to have the past review experience

All the aforementioned caveats coupled with sensitivity of the subject matter underscore the importance of a careful experimentation with the proposed procedure before its implementation in the routine review process.

Aspect 5. The role of reviewers With the above caveats, the experiment demonstrated that EXPERIMENTAL reviewers are comparable to and sometimes even better than reviewers recruited in conventional ways in terms of various metrics analyzed in Section 3. However, we qualify that this observation absolutely does not imply that EXPERIMENTAL reviewers can entirely substitute the pool of experienced reviewers. Instead, we conclude that if recruited and mentored appropriately, EXPERIMENTAL reviewers can form a useful augmentation to the traditional pool. The EXPERIMENTAL and experienced reviewers may have different strengths that can be combined to achieve an overall improvement of the peer-review quality. For instance (Shah, 2019), EXPERIMENTAL and, more generally, junior reviewers can be used to evaluate nuanced technical details of submissions (e.g., proofs) while senior researchers can focus on the broader picture and more subjective criteria such as impact where their expertise is extremely important.

4.3 Scalability of the experiment

In this section we provide some ideas on how the experiment we described in this work can be scaled to increase the number of recruited reviewers from 52 to several hundred. To this end, recall that the experiment is based on the two major components: the selection and mentoring mechanisms. We now comment on how to scale each of these components.

Selection Mechanism The current selection pipeline requires an amount of work equivalent to 2 to 4 days of the conference workflow chair's work and 2 to 4 hours of the conference program chairs' work to execute the experiment. Hence, it is important to design a version of the selection mechanism that can handle more reviewers while not resulting in a proportional increase of the load on the organizers.

First, we note that a large share of work in the selection stage was spent on finding papers to use in the auxiliary review. For this initial experiment, we had not made the call for papers public and instead personally reached to dozens of colleagues asking them to contribute their manuscripts which resulted in a large amount of communication-related work. However, this initial experiment demonstrated a lot of enthusiasm from authors of the papers used in the experiment who appreciated a set of useful reviews they received. Hence, we believe that we can easily extend the pool of papers by widely distributing the call for papers.

Similarly, participants of the selection mechanism executed in the present study were active in signing up for the experiment and appreciated an opportunity to join the ICML reviewer pool. In this initial experiment, the population of participants was limited to students of 5 large US universities. Hence, by making an open call for participants on behalf of a large ML conference, we expect to increase the pool of candidates to several hundred participants without much additional efforts.

The major part of the selection mechanism that requires a close attention of organizers is the review evaluation and final decision making. As noted in Section 2.1, we found author feedback to be very helpful for evaluating the quality of reviews and hence the selection part can be streamlined if the authors are required to evaluate all the reviews received in the experiment. Given that authors of papers used in the experiment generally found these reviews useful, we think that such a requirement is feasible as these reviews serve as a good incentive for authors to put some efforts in the experiment.

Mentoring As mentioned in Section 2.2, the total amount of time and effort in the mentorship of 52 EXPERIMENTAL reviewers was equal to about half the time and effort for a meta-reviewer's job. We note that the time demand for mentorship does not increase proportionally to the number of reviewers as some parts of the mentorship have a fixed cost (e.g., sending general guideline emails or multiple reminders to non-responsive reviewers). Thus, several additional committee members recruited specifically for mentoring will allow to handle several hundred novice reviewers in the real conference. Alternatively, mentoring can be distributed across many senior researchers who do not have time for meta-reviewing, but can contribute a smaller amount to mentoring.

Overall, we believe that the procedure outlined above can produce several hundred experimental reviewers without overburdening the organizers of the experiment.

4.4 Future work

An important direction for future work is to compare EXPERIMENTAL reviewers with the main reviewer pool in a larger scale study whose design we outlined above. A larger experiment would enable a deeper analysis of textual reviews written by different reviewer groups. It will also be of interest to design and execute experiments that address the caveats discussed above.

Another important direction is a principled design of a mentoring protocol to support novice reviewers. Since ML/AI conferences have hundreds of meta-reviewers, it may be prudent to assign a small number of meta-reviewers as mentors for junior reviewers and reduce their meta-reviewer workload accordingly. Future editions could also involve sharing more material on how to review with reviewers (e.g., Köhler et al., 2020) and holding webinars with Q&A sessions.

Finally, the feedback from the EXPERIMENTAL reviewers was that it was helpful for them to experience and gain insights into the review process, which will also help in their own research dissemination in the future. It would be interesting to measure the impact of the guided introduction to the review process in the early stages of career on the future trajectory of the individuals as researchers and reviewers.

Appendix

In this section we provide additional details for comparison of EXPERIMENTAL reviewers with reviewers from the main reviewer pool of the ICML 2020 conference. Specifically, where applicable we replicate comparisons described in the main paper, conditioning on a target set of papers with at least one EXPERIMENTAL reviewer assigned. Note that this conditioning significantly reduces the sample size and hence we may not always have enough data to establish statistically significant differences. Nevertheless, the results we present below allow to make some qualitative conclusions.

Before we proceed, recall that the main reviewer pool consists of complementary sets of CURATED and SELF-NOMINATED reviewers and we additionally consider a subset of COLLEAGUE reviewers who are affiliated with one of the 5 US schools we were recruiting EXPERIMENTAL reviewers from.

Review length (Row 3 of Table 1) Bidding activity as well as in-time review submission are independent of a set of papers assigned to reviewers so we begin our additional analysis from a comparison of mean review lengths. Figure 8 compares mean lengths of initial reviews across different reviewer groups, where mean values are computed using all reviews and also using reviews written for papers that have at least one EXPERIMENTAL reviewer assigned. Overall, we observe that EXPERIMENTAL reviewers write considerably longer reviews than other reviewers even after conditioning on the aforementioned subset of paper. Table 11 mimics Table 4 with the exception that only papers with EXPERIMENTAL reviewers assigned are used for the comparison.

	Experimental	Main Pool	Curated	Self-Nominated	Colleague
Sample Size	154	423	294	129	51
Mean Value	4759	2959	3073	2700	2917
90% CI	[4432; 5089]	[2827; 3098]	[2908; 3248]	[2487; 2928]	[2592; 3261]
P value	–	$< .001$	$< .001$	$< .001$	$< .001$

Table 11: Comparison of mean lengths (in symbols) of reviews using papers with at least one EXPERIMENTAL reviewer assigned. P values are for the test of the difference of means between EXPERIMENTAL and each of the other groups of reviewers.

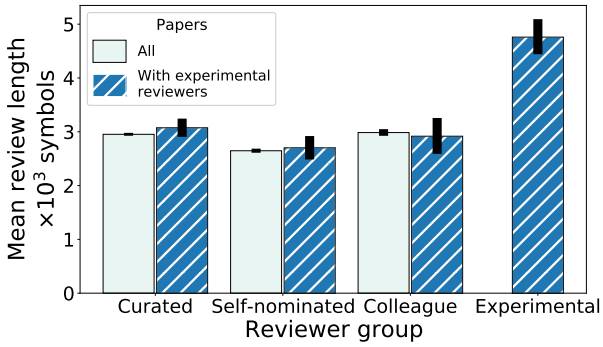


Figure 8: Mean review lengths (in symbols). EXPERIMENTAL reviewers write longer reviews than other reviewers.

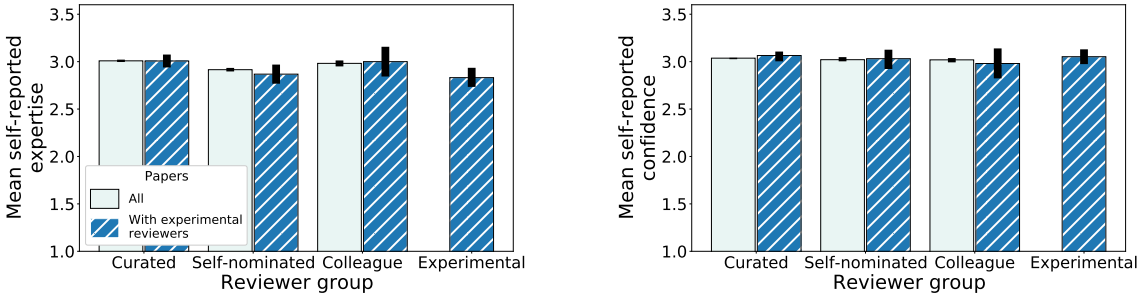
Hypercriticality (Row 4 of Table 1) Analysis presented in the main paper did not reveal hypercriticality in EXPERIMENTAL reviewers. Table 12 replicates the analysis displayed in Table 5 on the target subset of papers and also does not show any evidence of hypercriticality in EXPERIMENTAL reviewers.

	EXPERIMENTAL	MAIN POOL	CURATED	SELF-NOMINATED	COLLEAGUE
SAMPLE SIZE	154	423	294	129	51
MEAN VALUE	3.34	3.22	3.20	3.27	3.12
90% CI	[3.17; 3.51]	[3.13; 3.32]	[3.10; 3.32]	[3.09; 3.46]	[2.86; 3.37]
P VALUE	–	.361	.282	.691	.325

Table 12: Comparison of mean initial overall scores using papers with at least one EXPERIMENTAL reviewer assigned. P values are for the test of the difference of means between EXPERIMENTAL and each of the other groups of reviewers.

Expertise and confidence (Rows 5 and 6 of Table 1) Figure 9 juxtaposes the mean self-assessed confidence and expertise of reviewers computed over all papers and over papers with at least one EXPERIMENTAL reviewer assigned. Observe that conditioning on the target subset of papers does not change the conclusions we made in the main paper (see Table 13 for formal comparison).

Rebuttals and discussion (Rows 7 and 8 of Table 1) We now provide additional details on comparison of reviewers' activity in the post-rebuttal stage of the conference peer-review process. Recall that for this



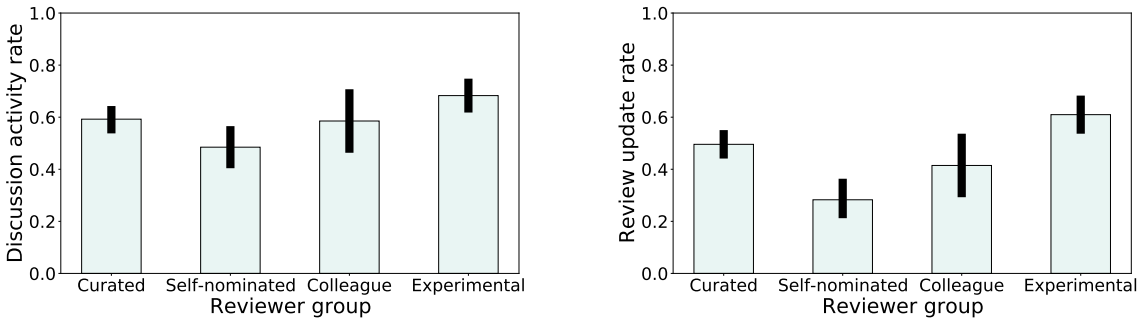
(a) Mean self-reported expertise. EXPERIMENTAL reviewers report considerably lower expertise than other groups of reviewers.

(b) Mean self-reported confidence. We do not observe significant difference in mean confidence of different reviewer groups.

Figure 9: Comparison of self-assessed expertise and confidence.

Criteria		Experimental	Main Pool	Curated	Self-Nominated	Colleague
Expertise	Sample Size	154	423	294	129	51
	Mean Value	2.83	2.96	3.01	2.87	3.00
	90% CI	[2.73; 2.94]	[2.91; 3.02]	[2.94; 3.07]	[2.77; 2.97]	[2.84; 3.16]
	P value	–	.064	.021	.737	.211
Confidence	Sample Size	154	423	294	129	51
	Mean Value	3.05	3.05	3.06	3.03	2.98
	90% CI	[2.97; 3.13]	[3.00; 3.11]	[3.00; 3.13]	[2.92; 3.14]	[2.82; 3.14]
	P value	–	.993	.878	.863	.556

Table 13: Comparison of self-assessed expertise (first 4 rows) and confidence (last 4 rows) using papers with at least one experimental reviewer assigned. P values are for the test of the difference of means between experimental and each of the other groups of reviewers.



(a) Participation in discussion.

(b) Post-rebuttal review update.

Figure 10: Activity of reviewers in the last stage of the review process conditioned on papers with at least one experimental reviewer assigned. Experimental reviewers participate in the discussion and update reviews more actively than other reviewers.

comparison we use only paper-reviewer pairs such that the authors of the paper supplied the response to the initial reviews. Figure 10 replicates Figure 6 with the exception that it is computed using papers that have at least one experimental reviewer assigned. Note that even after conditioning on this subset of papers, experimental reviewers remain to be more active in the post-rebuttal stage of the review process than other categories of reviewers. (see Table 14 for the formal comparison).

Review quality (Row 9 of Table 1) We conclude the analysis with comparison of review quality evaluated by meta-reviewers. First, Figure 11 replicates Figure 7 with the exception that we condition on papers that have at least one experimental reviewer assigned. As before, conditioning on this subset of papers does not change the qualitative relationship between different groups of reviewers with experimental pool dominating others. Getting back to the comparison of the mean quality scores that we reported in Tables 1 and 8, we now complement this analysis with results reported in Table 15. Again, we conclude that experimental reviewer remain to have higher mean quality of reviews even after we equalize the sets of meta-reviewers who rate the reviews written by different groups of reviewers.

Criteria		Experimental	Main Pool	Curated	Self-Nominated	Colleague
Discussion Activity	Sample Size	123	337	238	99	41
	Mean Value	0.68	0.56	0.59	0.48	0.59
	90% CI	[0.61; 0.75]	[0.52; 0.61]	[0.54; 0.64]	[0.40; 0.57]	[0.46; 0.71]
	P value	–	.026	.119	.003	.338
Review Update	Sample Size	123	337	238	99	41
	Mean Value	0.61	0.43	0.50	0.28	0.41
	90% CI	[0.54; 0.68]	[0.39; 0.48]	[0.44; 0.55]	[0.21; 0.35]	[0.29; 0.54]
	P value	–	.002	.050	$< .001$	.045

Table 14: Comparison of post-rebuttal reviewer activity using papers with at least one experimental reviewer assigned: participation in discussion (first 4 rows) and review update rate (last 4 rows). P values are for the test of the difference of means between experimental and each of the other groups of reviewers.

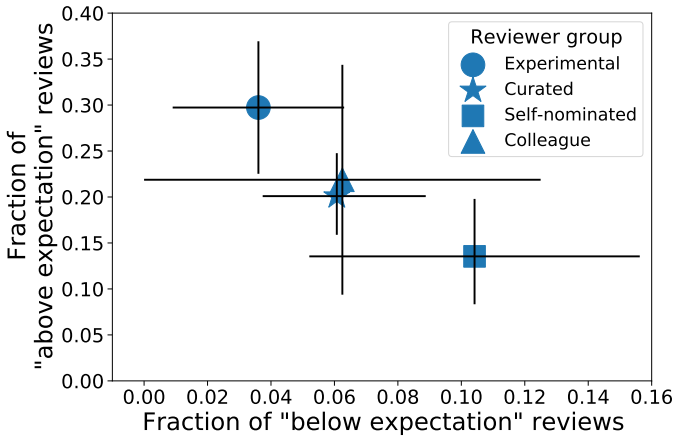


Figure 11: Evaluation of mean review quality by meta-reviewers conditioned on papers with at least one experimental reviewer assigned. The closer the point to the upper-left corner, the better. Experimental reviewers dominate other groups of reviewers.

	Experimental	Main Pool	Curated	Self-Nominated	Colleague
Sample Size	111	310	214	96	32
Mean Value	2.26	2.11	2.14	2.03	2.16
90% CI	[2.18; 2.34]	[2.06; 2.15]	[2.08; 2.20]	[1.95; 2.11]	[2.00; 2.31]
P value	–	.008	.058	.002	.431

Table 15: Comparison of mean review qualities as evaluated by meta-reviewers using papers with at least one experimental reviewer assigned. P values are for the test of the difference of means between experimental and each of the other groups of reviewers.

Part IV

Conclusions

Chapter 11

Conclusions

Peer review is the backbone of academia and an important prerequisite for fair, equitable, and efficient progression of science. Many pieces of empirical and anecdotal evidence demonstrate various shortcomings of the peer-review system and the broad scientific community agrees on the importance of making scientific peer review scientific. In this thesis, we investigate three broad challenges on the way to this goal: noise, bias, and strategic behavior.

In the first part, we study the impact of the assignment stage on the noise in the final decisions. Specifically, we (i) design a novel assignment algorithm that balances noise across submissions with the dual goal of fairness and accuracy in mind, and (ii) collect a novel dataset that helps researchers to develop better algorithms for predicting expertise (i.e., the expected level of noise) of reviewers in reviewing submissions.

In the second part, we investigate identity-related and policy-related biases. In that, we conduct real-world experiments to test for the presence of biases in the review process of flagship computer science conferences. Our experiments help stakeholders in making policy decisions in an evidence-based manner. Additionally, we design tools that enable organizers of any single-blind conference to test for identity-related biases in their specific venue.

Finally, in the third part, we focus on the problem of incentives. To investigate the impact of incentives on the behavior of reviewers, we design statistical tools and execute real-world experiments that identify deviations from the expected behavior. Additionally, we propose and evaluate in practice a novel procedure for recruiting highly-motivated reviewers.

All in all, this thesis contributes to the general goal of making scientific peer review scientific from both theoretical and empirical sides. It has also had a real-world impact with some tools being employed in a top-tier machine learning conference and some experiments leading to changes in venues' policies.

More broadly, the principles and insights we learned while working on this thesis are applicable to the general problem of distributed human decision-making. Hiring, university admissions, and crowdsourcing—these are examples of applications where similar problems arise and where our tools and insights might be useful. More concretely, let us mention several works that the author of this thesis contributed to during his PhD, but that were excluded from this thesis for the purpose of brevity. For illustration, we associate each work with the primary application area:

- **Evaluation of teaching quality** In our work with Jingyan Wang, Yuting Wei, and Nihar Shah (Wang et al., 2021) we study the problem of biases in evaluations induced by external "outcomes". For example, when students evaluate teaching quality, they may be biased by the course outcome—the grade received in the course. Such biases may lead to wrong conclusions and create wrong incentives (e.g., instructors may be incentivized to inflate course grades). Thus, we propose a debiasing method that helps to decouple the useful signal from the outcome-induced bias.
- **Applied crowdsourcing** A paper with Nikita Pavlichenko, and Dmitry Ustalov (Pavlichenko et al., 2021) contributes to the long line of work that designs aggregation methods for crowdsourced data. Specifically, we collect a high-quality dataset that can be used to develop novel methods for aggregation

of crowdsourced text sequences—data format that often arises in speech recognition. Our dataset has already enabled strong progress on the problem with novel methods built on our data outperforming existing baselines.

- **Theoretical crowdsourcing** Finally, a work with Charvi Rastogi, Nihar Shah and Sivaraman Balakrishnan (Rastogi et al., 2022a) studies the problem of aggregation of crowdsourced data from a theoretical standpoint. In that, we demonstrate that in a natural setting where annotators have different expertise levels, the standard maximum likelihood estimator (MLE) is asymptotically inadmissible. We show this by constructing an alternative estimator that we prove to be significantly better than the MLE in certain parameter regimes and at least as good elsewhere.

Ongoing and Future work

This thesis inspires several open problems and directions for future research. Many of the previous chapters conclude with a list of open problems. In addition, let us discuss several broader research directions that emanate from this thesis.

Evaluation of review quality A big challenge in peer review is a lack of tools to measure the effects of interventions. Did a new assignment algorithm improve the process? Is one reviewer form better than the other? To answer these questions, it is imperative to have a reliable measure of the quality of the reviews written in the process.

However, if we ask area chairs, reviewers, or authors to review the reviews, then the problems faced in evaluations of papers may arise again. This possibility is especially concerning, given that many conferences have already been using evaluations of reviews made by area chairs to distribute reviewer awards. Although these evaluations are supposed to create the right incentives for reviewers, noise, bias, and strategic behavior may actually hurt the motivation. Indeed, strong reviewers who are not recognized due to erroneous evaluations may lose their motivation. Simultaneously, reviewers who need improvement may not receive the right feedback and have no motivation to improve.

Overall, building a reliable procedure for measuring the quality of reviews is instrumental for improving peer review. One interesting direction is to combine evaluations made by area chairs, reviewers, and authors who have different points of view on the process into a single reliable evaluation. Our ongoing NeurIPS 2022 experiment aims at measuring the reliability of evaluations made by these parties and exploring various extraneous factors that may impact these evaluations.

Another promising direction is to employ automated tools to identify issues such as impoliteness, unjustified subjectivity, and bias in reviews. Such automated tools can be used to assist area chairs in identifying problematic reviews or to provide immediate feedback to reviewers.

Self-selection of authors Another big challenge faced by the peer-review process is the growth in the number of submissions. One way to address this challenge is to design tools and techniques that help to handle this growth. In this thesis, we presented a number of works along this direction. However, an appealing complementary direction is to transfer a part of the selection duties from reviewers to authors. Indeed, authors know the content of their paper best and, hypothetically, should be able to estimate whether their submission is strong and relevant enough to secure a place at a given venue.

However, the current mechanism of the review process does not encourage self-selection. Even if rejected, submission of a half-baked work to most of the venues does not hurt authors. At the same time, the noise in the review process results in non-zero acceptance chances for a large fraction of submitted papers. Overall, authors are incentivised to "*buy more lottery tickets*".

A promising direction to reduce the burden on the system is to develop interventions that incentivize self-selection among authors. An important caveat, though, is that any intervention that incentivizes self-selection may contribute to a disparity between different populations of authors and put otherwise disadvantaged authors under an additional strain. As a preventive step, our ongoing NeurIPS 2021 experiment investigates whether there is any difference in calibration among authors. For this, we asked authors to predict acceptance

chances of their submissions and will evaluate these predictions through the lens of actual outcomes of the review process. The findings of this experiment may be useful for any subsequent work on inducing self-selection.

Argument-based decision-making The third direction is more general and applies to the whole class of human decision-making problems. In many applications, including peer review, university admissions, hiring, and crowdsourcing, final decisions for submissions are made by aggregating decisions of independent individual evaluators. To make decisions across different submissions consistent, system designers often rely on score-based aggregation methods: scores given by committee members who handle the submission are averaged to arrive at a final ranking of submissions. While this approach may be optimal under some strong statistical assumptions, in practice, people are known to be subjective, miscalibrated, and noisy and all these properties hurt the quality of final decisions.

A long line of work in computer science (Paul, 1981; Roos et al., 2011; Baba and Kashima, 2013; Mackay et al., 2017; Wang and Shah, 2019) has been attempting to address this problem by analyzing scores given by committee members and adjusting them based on hypothesized behavioral patterns. However, most of such *score-based methods* did not end up being used in practice for various reasons, including negative feedback from stakeholders.

We believe that to address the problems of subjectivity, miscalibration, and noise in a principled manner, it is instrumental to build algorithms that make decisions not only based on scores, but also on textual comments made by evaluators. Indeed, research in psychology indicates that when being faced with a challenging problem, people often rely on *arguments* rather than numeric evaluations of alternatives (Shafir et al., 1993). Thus, reasons indicated in textual comments may enable algorithms to better understand the rationale of the given decision-maker, improving the quality of algorithmic recommendations as compared to the score-based methods. With this motivation, *a promising direction of future work is to investigate the use of textual information for an algorithmic approach towards subjectivity, miscalibration, and noise in human decision-making.*

Bibliography

ACLCommittee (2018). A report on the review process of ACL 2018. `https://acl2018.org/2018/05/19/how-decisions-made/` [Accessed: 9/7/2020].

Akoglu, L., Chandy, R., and Faloutsos, C. (2013). Opinion fraud detection in online reviews by network effects. *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, pages 2–11.

Akst, J. (2010). I hate your paper. Many say the peer review system is broken. Here's how some journals are trying to fix it. *The Scientist*, 24(8):36.

Alberts, B., Kirschner, M. W., Tilghman, S., and Varmus, H. (2014). Rescuing us biomedical research from its systemic flaws. *Proceedings of the National Academy of Sciences*, 111(16):5773–5777.

Allison, S. T., Mackie, D. M., and Messick, D. M. (1996). Outcome biases in social perception: Implications for dispositional inference, attitude change, stereotyping, and social behavior. In Zanna, M. P., editor, *Advances in Experimental Social Psychology*, volume 28, pages 53 – 93. Academic Press.

Alon, N., Fischer, F. A., Procaccia, A. D., and Tennenholtz, M. (2009). Sum of us: Strategyproof selection from the selectors. *CoRR*, abs/0910.4699.

Aman, V. (2014). Is there any measurable benefit in publishing preprints in the arXiv section quantitative biology? *arXiv preprint arXiv:1411.1955*.

Anderson, M. S., Ronning, E. A., De Vries, R., and Martinson, B. C. (2007). The perverse effects of competition on scientists' work and relationships. *Science and engineering ethics*, 13(4):437–461.

Anjum, O., Gong, H., Bhat, S., Hwu, W.-M., and Xiong, J. (2019). PaRe: A paper-reviewer matching approach using a common topic space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 518–528, Hong Kong, China. Association for Computational Linguistics.

Arnold, D., Dobbie, W., and Yang, C. S. (2018). Racial bias in bail decisions*. *The Quarterly Journal of Economics*, 133(4):1885–1932.

Asadpour, A. and Saberi, A. (2010). An approximation algorithm for max-min fair allocation of indivisible goods. *SIAM Journal on Computing*, 39(7):2970–2989.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In Guetzkow, H., editor, *Groups, leadership and men; research in human relations*, page 177–190. Carnegie Press.

Aziz, H., Lev, O., Mattei, N., Rosenschein, J. S., and Walsh, T. (2016). Strategyproof peer selection: Mechanisms, analyses, and experiments. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 390–396. AAAI Press.

Aziz, H., Lev, O., Mattei, N., Rosenschein, J. S., and Walsh, T. (2019). Strategyproof peer selection using randomization, partitioning, and apportionment. *Artificial Intelligence*, 275:295–309.

Baba, Y. and Kashima, H. (2013). Statistical quality estimation for general crowdsourcing tasks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 554–562, New York, NY, USA. Association for Computing Machinery.

Babbage, C. (1864). *Passages from the Life of a Philosopher*. Cambridge Library Collection - Technology. Cambridge University Press.

Baguley, T. (2008). Standardized or simple effect size: What should be reported? *British journal of psychology (London, England : 1953)*, 100:603–17.

Bakanic, V., McPhail, C., and Simon, R. J. (1987). The manuscript review and decision-making process. *American Sociological Review*, pages 631–642.

Balietti, S., Goldstone, R. L., and Helbing, D. (2016). Peer review and competition in the art exhibition game. *Proceedings of the National Academy of Sciences*, 113(30):8414–8419.

Banerjee, A. V. (1992). A simple model of herd behavior. *QUART. J. ECONOM*, 107(3):797–818.

Bansal, N. and Sviridenko, M. (2006). The Santa Claus problem. In *Proceedings of the Thirty-eighth Annual ACM Symposium on Theory of Computing*, STOC '06, pages 31–40, New York, NY, USA. ACM.

Baron, J. and Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, 54(4):569–579.

Baron, R. S., Vandello, J. A., and Brunsman, B. (1996). The forgotten variable in conformity research: Impact of task importance on social influence. *Journal of Personality and Social Psychology*, 71(5):915–927.

Barroga, E. F. (2014). Safeguarding the integrity of science communication by restraining 'rational cheating'in peer review. *Journal of Korean medical science*, 29(11):1450–1452.

Bendick, M., Jackson, C., and Romero, J. (1996). Employment discrimination against older workers: An experimental study of hiring practices. *Journal of aging & social policy*, 8:25–46.

Bendick, M. and Nunes, A. (2011). Developing the research basis for controlling bias in hiring. *Journal of Social Issues*, 68:238–262.

Benferhat, S. and Lang, J. (2001). Conference paper assignment. *International Journal of Intelligent Systems*, 16(10):1183–1192.

Bernstein, R. (2015). PLOS ONE ousts reviewer, editor after sexist peer-review storm. *Science*.

Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, 94(4):991–1013.

Beverly, R. and Allman, M. (2013). Findings and implications from data mining the imc review process. *SIGCOMM 2013*, 43(1):22–29.

Beygelzimer, A., Dauphin, Y., Liang, P., and Wortman Vaughan, J. (2021). The NeurIPS 2021 consistency experiment. `https://blog.neurips.cc/2021/12/08/the-neurips-2021-consistency-experiment/`.

Beygelzimer, A., Fox, E., dÁlche Buc, F., and Larochelle, H. (2019). What we learned from NeurIPS 2019 data. `https://medium.com/@NeurIPSConf/what-we-learned-from-neurips-2019-data-111ab996462c` [Accessed: 9/7/2020].

Bharadhwaj, H., Turpin, D., Garg, A., and Anderson, A. (2020). De-anonymization of authors through arXiv submissions during double-blind review. *arXiv preprint arXiv:2007.00177*.

Bianchi, F. and Squazzoni, F. (2015). Is three better than one? Simulating the effect of reviewer selection and behavior on the quality and efficiency of peer review. In *Proceedings of the 2015 Winter Simulation Conference*, pages 4081–4089. IEEE Press.

Bikhchandani, S. and Sharma, S. (2000). Herd behavior in financial markets. *IMF Staff Papers*, 47:279–310.

Black, N., Van Rooyen, S., Godlee, F., Smith, R., and Evans, S. (1998). What makes a good reviewer and a good review for a general medical journal? *Jama*, 280(3):231–233.

Blank, R. M. (1991). The effects of double-blind versus single-blind reviewing: Experimental evidence from the american economic review. *American Economic Review*, 81(5):1041–1067.

Bonald, T., Massoulié, L., Proutiere, A., and Virtamo, J. (2006). A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queueing systems*, 53(1-2):65–84.

Bornmann, L. (2011). Scientific peer review. *Annual review of information science and technology*, 45(1):197–245.

Brown, T. (2004). *Peer Review and the Acceptance of New Scientific Ideas: Discussion Paper from a Working Party on Equipping the Public with an Understanding of Peer Review: November 2002-May 2004*. Sense About Science.

Brunner, J. and Austin, P. C. (2009). Inflation of Type I error rate in multiple regression when independent variables are measured with error. *Canadian Journal of Statistics*, 37(1):33–46.

Budden, A. E., Tregenza, T., Aarssen, L. W., Koricheva, J., Leimu, R., and Lortie, C. J. (2008). Double-blind review favours increased representation of female authors. *Trends in Ecology and Evolution*, 23(1):4 – 6.

Cabotà, J. B., Grimaldo, F., and Squazzoni, F. (2013). When competition is pushed too hard. an agent-based model of strategic behaviour of referees in peer review. In *ECMS*, pages 881–887.

Callaham, M. L., Wears, R. L., Weber, E. J., Barton, C., and Young, G. (1998). Positive-outcome bias and other limitations in the outcome of research abstracts submitted to a scientific meeting. *JAMA*, 280(3):254–257.

Cals, J. W., Mallen, C. D., Glynn, L. G., and Kotz, D. (2013). Should authors submit previous peer-review reports when submitting research papers? views of general medical journal editors. *Annals of family medicine*, 11(2):179–181.

Caragiannis, I., Krimpas, G. A., and Voudouris, A. A. (2014). Aggregating partial rankings with applications to peer grading in massive online open courses. *CoRR*, abs/1411.4619.

Carretta, T. R. and Moreland, R. L. (1983). The direct and indirect effects of inadmissible evidence1. *Journal of Applied Social Psychology*, 13(4):291–309.

Chandler, D. and Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90:123 – 133.

Charlin, L. and Zemel, R. S. (2013). The Toronto Paper Matching System: An automated paper-reviewer assignment system.

Charlin, L., Zemel, R. S., and Boutilier, C. (2012). A framework for optimizing paper matching. *CoRR*, abs/1202.3706.

Cialdini, R. and Goldstein, N. (2004). Social influence: Compliance and conformity. *Annual review of psychology*, 55:591–621.

Cleveland, W. S. and Loader, C. (1996). Smoothing by local regression: Principles and methods. In Härdle, W. and Schimek, M. G., editors, *Statistical Theory and Computational Aspects of Smoothing*, pages 10–49, Heidelberg. Physica-Verlag HD.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. S. (2020). Specter: Document-level representation learning using citation-informed transformers.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1):155–159.

Cole, S., Simon, G. A., et al. (1981). Chance and consensus in peer review. *Science*, 214(4523):881–886.

COPE (2018). Editor and reviewers requiring authors to cite their own work. `https://publicationethics.org/case/editor-and-reviewers-requiring-authors-cite-their-own-work` [Accessed: 1/21/2022].

Corey, D., Dunlap, W., and Burke, M. (1998). Averaging correlations: Expected values and bias in combined pearson rs and fisher's z transformations. *Journal of General Psychology - J GEN PSYCHOL*, 125:245–261.

Cover, T. M. and Thomas, J. A. (2005). *Entropy, Relative Entropy, and Mutual Information*, pages 13–55. John Wiley & Sons, Inc.

CSRankings (2021). Computer Science Rankings: Economics and Computation. `http://csrankings.org/#/fromyear/2011/toyear/2021/index?ecom&world` [Last Accessed: 10/15/2021].

Dai, W., Jin, G. Z., Lee, J., and Luca, M. (2012). Optimal aggregation of consumer ratings: An application to yelp.com. Working Paper 18567, National Bureau of Economic Research.

DeStefano, F. and Chen, R. T. (1999). Negative association between MMR and autism. *Lancet (London, England)*, 353(9169):1987–1988.

Dubrovsky, V. J., Kiesler, S., and Sethna, B. N. (1991). The equalization phenomenon: Status effects in computer-mediated and face-to-face decision-making groups. *Human–Computer Interaction*, 6(2):119–146.

Edwards, M. and Ewen, A. (1996). *360 Degree Feedback: The Powerful New Model for Employee Assessment & Performance Improvement*. AMACOM.

Emerson, G., Warme, W., Wolf, F., Heckman, J., Brand, R., and Leopold, S. (2010). Testing for the presence of positive-outcome bias in peer review: A randomized controlled trial. *Archives of Internal Medicine*, 170(21):1934–1939.

Ernst, E. and Resch, K.-L. (1994). Reviewer bias: a blinded experimental study. *The Journal of laboratory and clinical medicine*, 124(2):178–182.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS one*, 4(5):e5738.

Feldmann, A. (2005). Experiences from the sigcomm 2005 european shadow pc experiment. *ACM SIGCOMM Computer Communication Review*, 35(3):97–102.

Fiez, T., Shah, N., and Ratliff, L. (2020). A SUPER* algorithm to optimize paper bidding in peer review. In *Conference on Uncertainty in Artificial Intelligence*.

Fischhoff, B. and Beyth, R. (1975). I knew it would happen: Remembered probabilities of once—future things. *Organizational Behavior and Human Performance*, 13(1):1 – 16.

Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd, Oxford, England.

Flach, P. A., Spiegler, S., Golénia, B., Price, S., Guiver, J., Herbrich, R., Graepel, T., and Zaki, M. J. (2010). Novel tools to streamline the conference review process: Experiences from SIGKDD'09. *SIGKDD Explor. Newsl.*, 11(2):63–67.

Fogelholm, M., Leppinen, S., Auvinen, A., Raitanen, J., Nuutinen, A., and Väänänen, K. (2011). Panel discussion does not improve reliability of peer review for medical research grant proposals. *Journal of clinical epidemiology*, 65:47–52.

Fong, E. A. and Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *PLoS ONE 12*, 12.

Francis, P. (2008). Thoughts on improving review quality. In *Proceedings of the Conference on Organizing Workshops, Conferences, and Symposia for Computer Systems*, WOWCS'08, USA. USENIX Association.

Freeman, S. and Parks, J. W. (2010). How accurate is peer grading? *CBE—Life Sciences Education*, 9:482–488.

Friedman, M. and Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of Political Economy*, 56(4):279–304.

Fuller, S. (2018). Must academic evaluation be so citation data driven? `https://www.universityworldnews.com/post.php?story=20180925094651499` [Last Accessed: 3/15/2022].

Gans, J. S. and Shepherd, G. B. (1994). How are the mighty fallen: Rejected classic articles by leading economists. *Journal of Economic Perspectives*, 8(1):165–179.

Gao, Y., Eger, S., Kuznetsov, I., Gurevych, I., and Miyao, Y. (2019). Does my rebuttal matter? Insights from a major NLP conference. *CoRR*, abs/1903.11367.

Garcia, J. A., Rodriguez-Sánchez, R., and Fdez-Valdivia, J. (2020). Confirmatory bias in peer review. *Scientometrics*, 123:517 – 533.

Garfinkel, R. S. (1971). Technical note. An improved algorithm for the bottleneck assignment problem. *Operations Research*, 19(7):1747–1751.

Garg, N., Kavitha, T., Kumar, A., Mehlhorn, K., and Mestre, J. (2010). Assigning papers to referees. *Algorithmica*, 58(1):119–136.

Garisto, D. (2019). Diversifying peer review by adding junior scientists. `https://www.natureindex.com/news-blog/diversifying-peer-review-by-adding-junior-scientists` [Accessed: 9/7/2020].

Ge, H., Welling, M., and Ghahramani, Z. (2013). A Bayesian model for calibrating conference review scores.

Gilovich, T., Griffin, D., and Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press, Cambridge.

Gino, F., Moore, D., and Bazerman, M. (2008). No harm, no foul: The outcome bias in ethical judgments. *Harvard Business School, Harvard Business School Working Papers*.

Goldsmith, J. and Sloan, R. (2007). The AI conference paper assignment problem. *AAAI Workshop - Technical Report*, WS-07-10:53–57.

Goues, C. L., Brun, Y., Apel, S., Berger, E., Khurshid, S., and Smaragdakis, Y. (2018). Effectiveness of anonymization in double-blind review. *Communications of the ACM*, 61:30 – 33.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

Hahne, E. L. (1991). Round-robin scheduling for max-min fairness in data networks. *IEEE Journal on Selected Areas in communications*, 9(7):1024–1039.

Hartvigsen, D., Wei, J. C., and Czuchlewski, R. (1999). The conference paper-reviewer assignment problem. *Decision Sciences*, 30(3):865–876.

Hassidim, A., Romm, A., and Shorrer, R. I. (2018). 'Strategic' behavior in a strategy-proof environment. *SSRN*, pages 686–688.

Hawkins, S. A. and Hastie, R. (1990). Hindsight: Biased judgment of past events after the outcomes are known. *Psychological Bulletin*, 107(3):311–327.

Hazelrigg, G. A. (2013). Dear colleague letter: Information to principal investigators (PIs) planning to submit proposals to the Sensors and Sensing Systems (SSS) program October 1, 2013, deadline. `https://www.semanticscholar.org/paper/Dear-Colleague-Letter%3A-Information-to-Principal-to-Hazelrigg/2a560a95c872164a6316b3200504146ac977a2e6` [Accessed: 8/3/2022.].

Hill, S. and J. Provost, F. (2003). The myth of the double-blind review? Author identification using only citations. *SIGKDD Explorations*, 5:179–184.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572.

Hofer, T. P., Bernstein, S. J., DeMonner, S., and Hayward, R. A. (2000). Discussion between reviewers does not improve reliability of peer review of hospital quality. *Medical Care*, 38(2):152–161.

Hollander, E. (1978). *Leadership Dynamics: A Practical Guide to Effective Relationships*. Free Press/Macmillan.

Huang, Y., Shum, M., Wu, X., and Xiao, J. Z. (2019). Discovery of bias and strategic behavior in crowdsourced performance assessment. *arXiv preprint arXiv:1908.01718*.

Ivanov, S. (2020). ICML 2020. Comprehensive analysis of authors, organizations, and countries. `https://medium.com/criteo-engineering/icml-2020-comprehensive-analysis-of-authors-organizations-and-countries-c4d1bb847fde` [Last Accessed: 3/15/2022].

Janis, I. (1982). *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. Houghton Mifflin.

Jecmen, S., Yoon, M., Conitzer, V., Shah, N. B., and Fang, F. (2022). A dataset on malicious paper bidding in peer review.

Jecmen, S., Zhang, H., Liu, R., Shah, N. B., Conitzer, V., and Fang, F. (2020). Mitigating manipulation in peer review via randomized reviewer assignments. In *NeurIPS*.

Jefferson, T., Alderson, P., Wager, E., and Davidoff, F. (2002). Effects of editorial peer review: a systematic review. *Jama*, 287(21):2784–2786.

Jiao, Y. and Vert, J. (2018). The kendall and mallows kernels for permutations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(7):1755–1769.

Jindal, N. and Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, page 219–230, New York, NY, USA. Association for Computing Machinery.

Kaghazgaran, P., Caverlee, J., and Squicciarini, A. (2018). Combating crowdsourced review manipulators: A neighborhood-based approach. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 306–314, New York, NY, USA. Association for Computing Machinery.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux, New York.

Kahneman, D. and Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, 49:49–81.

Kahneman, D., Sibony, O., and Sunstein, C. (2021). *Noise: A Flaw in Human Judgment*. Little, Brown.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.

Kahng, A., Kotturi, Y., Kulkarni, C., Kurokawa, D., and Procaccia, A. (2018). Ranking wily people who rank each other. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI'18. AAAI Press.

Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E. H., and Schwartz, R. (2018). A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. *CoRR*, abs/1804.09635.

Karimzadehgan, M., Zhai, C., and Belford, G. (2008). Multi-aspect expertise matching for review assignment. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 1113–1122, New York, NY, USA. ACM.

Kaufmann, N., Schulze, T., and Veit, D. (2011). More than fun and money. worker motivation in crowdsourcing–a study on mechanical turk. *Proceedings of the Seventeenth Americas Conference on Information Systems*.

Keeney, R. and Raiffa, H. (1976). *Decisions with multiple objectives–preferences and value tradeoffs*. Wiley, New York.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Kerr, S., Tolliver, J., and Petree, D. (1977). Manuscript characteristics which influence acceptance for management and social science journals. *Academy of Management Journal*, 20(1):132–141.

Kerzendorf, W. E., Patat, F., Bordelon, D., van de Ven, G., and Pritchard, T. A. (2020). Distributed peer review enhanced with natural language processing and machine learning.

Khosla, A., Hoiem, D., and Belongie, S. (2013). Analysis of reviews for CVPR 2012.

King, V., Rao, S., and Tarjan, R. (1992). A faster deterministic maximum flow algorithm. In *Proceedings of the Third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '92, pages 157–164, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Kobren, A., Saha, B., and McCallum, A. (2019). Paper matching with local fairness constraints. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Köhler, T., González-Morales, M. G., Banks, G. C., O'Boyle, E. H., Allen, J. A., Sinha, R., Woo, S. E., and Gulick, L. M. (2020). Supporting robust, rigorous, and reliable reviewing as the cornerstone of our profession: Introducing a competency framework for peer review. *Industrial and Organizational Psychology*, 13(1):1–27.

Kostoff, R. N. (1998). The use and misuse of citation analysis in research evaluation. *Scientometrics*, 43(1):27–43.

Kotturi, Y., Kahng, A., Procaccia, A. D., and Kulkarni, C. (2020). Hirepeer: Impartial peer-assessed hiring at scale in expert crowdsourcing markets. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'20. AAAI Press.

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., and Klemmer, S. R. (2013). Peer and self assessment in massive online classes. *ACM Trans. Comput.-Hum. Interact.*, 20(6).

Kurokawa, D., Lev, O., Morgenstern, J., and Procaccia, A. D. (2015). Impartial peer review. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 582–588. AAAI Press.

Lamont, M. (2009). *How professors think*. Harvard University Press.

Langford, J. (2008). Adversarial academia. `http://hunch.net/?p=499` [Accessed: 8/03/2022].

Langford, J. (2012a). Bidding problems. `https://hunch.net/?p=407` [Accessed: 8/03/2022].

Langford, J. (2012b). ICML acceptance statistics. `http://hunch.net/?p=2517` [Accessed: 8/03/2022].

Langford, J. (2018). When the bubble bursts.... `https://hunch.net/?p=9604328` [Accessed: 8/03/2022].

Largent, E. and Snodgrass, R. (2016). Blind peer review by academic journals. In Robertson, C. and Kesselheim, A., editors, *Blinding as a Solution to Bias: Strengthening Biomedical Science, Forensic Science, and Law*, pages 75–95. Cambridge.

Larson, H. J., Cooper, L. Z., Eskola, J., Katz, S. L., and Ratzan, S. (2011). Addressing the vaccine confidence gap. *The Lancet*, 378(9790):526–535.

Lavi, R., Mu'Alem, A., and Nisan, N. (2003). Towards a characterization of truthful combinatorial auctions. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pages 574–583. IEEE.

Lawrence, N. and Cortes, C. (2014). The NIPS experiment. `http://inverseprobability.com/2014/12/16/the-nips-experiment`. [Accessed: 8/03/2022].

Lee, C. J. (2015). Commensuration bias in peer review. *Philosophy of Science*, 82(5):1272–1283.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition.

Lenstra, J. K., Shmoys, D. B., and Tardos, É. (1990). Approximation algorithms for scheduling unrelated parallel machines. *Mathematical Programming*, 46(1):259–271.

Levenshtein, V. I. (1971). Upper-bound estimates for fixed-weight codes. *Problemy Peredachi Informatsii*, 7(4):3–12.

Levy, P. and Sarne, D. (2018). Understanding over participation in simple contests. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI'18. AAAI Press.

Li, D. (2017). Expertise versus bias in evaluation: Evidence from the NIH. *American Economic Journal: Applied Economics*, 9(2):60–92.

Li, L., Wang, Y., Liu, G., Wang, M., and Wu, X. (2015). Context-aware reviewer assignment for trust enhanced peer review. *PLOS ONE*, 10(6):1–28.

Lian, J. W., Mattei, N., Noble, R., and Walsh, T. (2018). The conference paper assignment problem: Using order weighted averages to assign indivisible goods. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI'18. AAAI Press.

Lin, H. and Bilmes, J. (2011). A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 510–520, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lin, H.-T., Balcan, M.-F., Hadsell, R., and Ranzato, M. (2020a). Getting started with NeurIPS 2020. `https://medium.com/@NeurIPSConf/getting-started-with-neurips-2020-e350f9b39c28`. [Accessed: 8/03/2022].

Lin, H.-T., Balcan, M.-F., Hadsell, R., and Ranzato, M. (2020b). What we learned from NeurIPS 2020 reviewing process. `https://medium.com/@NeurIPSConf/what-we-learned-from-neurips-2020-reviewing-process-e24549eea38f` [Accessed: 8/03/2022].

Liu, X., Suel, T., and Memon, N. (2014). A robust model for paper reviewer assignment. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 25–32, New York, NY, USA. ACM.

Long, C., Wong, R., Peng, Y., and Ye, L. (2013). On good and fair paper-reviewer assignment. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 1145–1150.

Lord, C. G., Ross, L. D., and Lepper, M. R. (1979). Biased assimilation and attitude polarization : The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098–2109.

Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences of the United States of America*, 108(22):9020–9025.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical analysis*. Wiley, New York, NY, USA.

Lynch, J., Cunningham, M., Warme, W., Schaad, D., Wolf, F., and Leopold, S. (2007). Commercially funded and united states-based research is more likely to be published; good-quality studies with negative outcomes are not. *The Journal of bone and joint surgery. American volume*, 89:1010–8.

Mackay, R., Kenna, R., Low, R., and Parker, S. (2017). Calibration with confidence: A principled method for panel assessment. *Royal Society Open Science*, 4:160760.

Madden, S. and DeWitt, D. (2006). Impact of double-blind reviewing on SIGMOD publication rates. *ACM SIGMOD Record*, 35(2):29–32.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research*, 1(2):161–175.

Mania, H., Ramdas, A., Wainwright, M. J., Jordan, M. I., and Recht, B. (2018). On kernel methods for covariates that are rankings. *Electron. J. Statist.*, 12(2):2537–2577.

Manzoor, E. and Shah, N. B. (2020). Uncovering latent biases in text: Method and application to peer review.

Marshall, G. W. and Mowen, J. C. (1993). An experimental investigation of the outcome bias in salesperson performance evaluations. *Journal of Personal Selling & Sales Management*, 13(3):31–47.

Martinson, B. C., Anderson, M. S., and De Vries, R. (2005). Scientists behaving badly. *Nature*, 435(7043):737–738.

McCambridge, J., De Bruin, M., and Witton, J. (2012). The effects of demand characteristics on research participant behaviours in non-laboratory settings: a systematic review. *PloS one*, 7(6):e39116.

McCook, A. (2006). Is peer review broken? *The Scientist*, 20(2):26–34.

McGlohon, M., Glance, N., and Reiter, Z. (2010). Star quality: Aggregating reviews to rank products and merchants. In *Proceedings of Fourth International Conference on Weblogs and Social Media (ICWSM)*.

McGuire, T. W., Kiesler, S., and Siegel, J. (1987). Group and computer-mediated discussion effects in risk decision making. *American Psychological Association*, 52:917–930.

Merton, R. K. (1968). The Matthew effect in science. *Science*, 159:56–63.

Meyer, B., Choppy, C., Staunstrup, J., and van Leeuwen, J. (2009). Research evaluation for computer science. *Communications of the ACM*, 52(4):31–34.

Mimno, D. and McCallum, A. (2007). Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 500–509, New York, NY, USA. ACM.

Mogul, J. C. (2013). Towards more constructive reviewing of SIGCOMM papers. *SIGCOMM Comput. Commun. Rev.*, 43(3):90–94.

Moscati, R., Jehle, D., Ellis, D., Fiorello, A., and Landi, M. (1994). Positive-outcome bias: Comparison of emergency medicine and general medicine literatures. *Academic Emergency Medicine*, 1(3):267–271.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., and Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474–16479.

Mulligan, A., Hall, L., and Raphael, E. (2013). Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the Association for Information Science and Technology*, 64(1):132–161.

Mussweiler, T. and Strack, F. (2001). Considering the impossible: Explaining the effects of implausible anchors. *Social Cognition - SOC COGNITION*, 19:145–160.

Myers, A. (2003). Condorcet internet voting service. `https://civs1.civs.us/`. [Accessed: 8/03/2022].

Nobarany, S., Booth, K. S., and Hsieh, G. (2016). What motivates people to review articles? The case of the human-computer interaction community. *Journal of the Association for Information Science and Technology*, 67(6):1358–1371.

Noothigattu, R., Shah, N. B., and Procaccia, A. D. (2020). Loss functions, axioms, and peer review. In *ICML workshop on Incentives in Machine Learning*.

Obrecht, M., Tibelius, K., and D'Aloisio, G. (2007). Examining the value added by committee discussion in the review of applications for research awards. *Research Evaluation*, 16(2):79–91.

Okike, K., Hug, K. T., Kocher, M. S., and Leopold, S. S. (2016). Single-blind vs double-blind peer review in the setting of author prestige. *JAMA*, 316(12):1315–1316.

OpenReview (2022). Paper-reviewer affinity modeling for openreview. `https://github.com/openreview/openreview-expertise`.

Orlin, J. B. (2013). Max flows in $\mathcal{O}(nm)$ time, or better. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 765–774, New York, NY, USA. ACM.

Paolucci, M. and Grimaldo, F. (2014). Mechanism change in a simulation of peer review: from junk support to elitism. *Scientometrics*, 99(3):663–688.

Papagiannaki, K. (2007). Author feedback experiment at pam 2007. *SIGCOMM Comput. Commun. Rev.*, 37(3):73–78.

Patat, F., Kerzendorf, W., Bordelon, D., Van de Ven, G., and Pritchard, T. (2019). The Distributed Peer Review Experiment. *The Messenger*, 177:3–13.

Paul, S. (1981). Bayesian methods for calibration of examiners. *British Journal of Mathematical and Statistical Psychology*, 34:213–223.

Pavlichenko, N., Stelmakh, I., and Ustalov, D. (2021). Crowdspeech and vox diy: Benchmark dataset for crowdsourced audio transcription. In Vanschoren, J. and Yeung, S., editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Payan, J. (2022). Fair allocation problems in reviewer assignment. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, page 1857–1859, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Perez-Diaz, A., Gerding, E., and McGroarty, F. (2018). Detecting strategic manipulation in distributed optimisation of electric vehicle aggregators. *CoRR*, abs/1810.07063.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Picciotto, M. (2018). New reviewer mentoring program. *Journal of Neuroscience*, 38(3):511–511.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. (2013). Tuned models of peer assessment in moocs. *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)*.

Pier, E., Raclaw, J., Kaatz, A., Brauer, M., Carnes, M., Nathan, M., and Ford, C. (2017). 'Your comments are meaner than your score': Score calibration talk influences intra- and inter-panel variability during scientific grant peer review. *Research Evaluation*, 26:1–14.

Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., Ford, C. E., and Carnes, M. (2018). Low agreement among reviewers evaluating the same nih grant applications. *Proceedings of the National Academy of Sciences*, 115(12):2952–2957.

Plackett, R. L. (1975). The Analysis of Permutations. *Journal of the Royal Statistical Society Series C*, 24(2):193–202.

Price, E. (2014). The NeurIPS experiment. `http://blog.mrtz.org/2014/12/15/the-nips-experiment.html`. [Accessed: 8/03/2022].

Price, S. and Flach, P. A. (2017). Computational support for academic peer review: a perspective from artificial intelligence. *Communications of the ACM*, 60(3):70–79.

QS (2021a). QS world university rankings by subject 2021: Computer Science and Information Systems. `https://www.topuniversities.com/university-rankings/university-subject-rankings/2021/computer-science-information-systems` [Last Accessed: 10/15/2021].

QS (2021b). QS world university rankings by subject 2021: Economics & Econometrics. `https://www.topuniversities.com/university-rankings/university-subject-rankings/2021/economics-econometrics` [Last Accessed: 10/15/2021].

Quillian, L., Pager, D., Hexel, O., and Midtbøen, A. H. (2017). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences*, 114(41):10870–10875.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rabin, M. and Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics*, 114(1):37–82.

Rao, T. S. S. and Andrade, C. (2011). The MMR vaccine and autism: Sensation, refutation, retraction, and fraud. *Indian journal of psychiatry*, 53(2):95—96.

Rastogi, C., Shah, N., Balakrishnan, S., and Singh, A. (2020). Two-sample testing with pairwise comparison data and the role of modeling assumptions. *IEEE International Symposium on Information Theory*.

Rastogi, C., Stelmakh, I., Shah, N. B., and Balakrishnan, S. (2022a). No rose for MLE: Indamissibility of MLE for evaluation aggregation under levels of expertise. In *Proceedings of IEEE ISIT 2022*.

Rastogi, C., Stelmakh, I., Shen, X., Meila, M., Echenique, F., Chawla, S., and Shah, N. B. (2022b). To ArXiv or not to ArXiv: A study quantifying pros and cons of posting preprints online. *arXiv:2203.17259*.

Rawls, J. (1971). *A theory of justice: Revised edition*. Harvard university press.

Rennie, D. (2016). Make peer review scientific. *Nature*, 535:31–33.

Resnik, D. B., Gutierrez-Ford, C., and Peddada, S. (2008). Perceptions of ethical problems with scientific journal peer review: an exploratory study. *Science and engineering ethics*, 14(3):305–310.

Resnik, D. B. and Smith, E. M. (2020). *Bias and Groupthink in Science's Peer-Review System*, pages 99–113. Springer International Publishing, Cham.

Rodriguez, M. A. and Bollen, J. (2008). An algorithm to determine peer-reviewers. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 319–328, New York, NY, USA. ACM.

Rodriguez, M. A., Bollen, J., and Van de Sompel, H. (2007). Mapping the bid behavior of conference referees. *Journal of Informetrics*, 1(1):68–82.

Roese, N. J. and Vohs, K. D. (2012). Hindsight bias. *Perspectives on Psychological Science*, 7(5):411–426.

Roos, M., Rothe, J., and Scheuermann, B. (2011). How to calibrate the scores of biased reviewers by quadratic programming. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, AAAI'11, page 255–260. AAAI Press.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641.

Ross, L., Lepper, M., and Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of personality and social psychology*, 32:880–92.

Sajjadi, M. S., Alamgir, M., and von Luxburg, U. (2016). Peer grading in a course on algorithms and data structures: Machine learning algorithms do not improve over simple baselines. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, L@S '16, pages 369–378, New York, NY, USA. ACM.

Sculley, D., Snoek, J., and Wiltschko, A. B. (2019). Avoiding a tragedy of the commons in the peer review process. *CoRR*, abs/1901.06246.

Seeber, M. and Bacchelli, A. (2017). Does single blind peer review hinder newcomers? *Scientometrics*, 113(1):567–585.

Shafir, E., Simonson, I., and Tversky, A. (1993). Reason-based choice. *Cognition*, 49(1):11 – 36.

Shah, N. (2019). Double decker peer review. Research on Research blog. `https://researchonresearch.blog/2019/02/23/double-decker-peer-review/`.

Shah, N. B. (2017). Learning from people. `https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-133.pdf` [Accessed: 7/21/2022].

Shah, N. B. (2022). An overview of challenges, experiments, and computational solutions in peer review. Communications of the ACM (to appear). Preprint available at `http://bit.ly/PeerReviewOverview`.

Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramch, K., ran, and Wainwright, M. J. (2016). Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58):1–47.

Shah, N. B., Bradley, J. K., Parekh, A., and Ramchandran, K. (2013). A case for ordinal peer-evaluation in moocs. `http://www.cs.cmu.edu/~jkbradle/papers/shahetal.pdf` [Accessed: 8/03/2022.].

Shah, N. B., Tabibian, B., Muandet, K., Guyon, I., and Von Luxburg, U. (2018). Design and analysis of the NIPS 2016 review process. *The Journal of Machine Learning Research*, 19(1):1913–1946.

Shah, N. B. and Wainwright, M. J. (2015). Simple, robust and optimal ranking from pairwise comparisons. *CoRR*, abs/1512.08949.

Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99(4):178–182.

Snodgrass, R. (2006). Single- versus double-blind reviewing: An analysis of the literature. *SIGMOD Record*, 35:8–21.

Soufiani, H. A., Parkes, D. C., and Xia, L. (2012). Random utility theory for social choice. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 126–134, Red Hook, NY, USA. Curran Associates Inc.

Squazzoni, F. and Gandelli, C. (2012). Saint Matthew strikes again: An agent-based model of peer review and the scientific community structure. *Journal of Informetrics*, 6(2):265–275.

Stanovich, K. (1999). *Who Is Rational? Studies of Individual Differences in Reasoning*. Mahwah, NJ: Erlbaum.

Stanovich, K. and West, R. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of personality and social psychology*, 94:672–95.

Stefanski, L. A. and Carroll, R. J. (1985). Covariate measurement error in logistic regression. *Ann. Statist.*, 13(4):1335–1351.

Stelmakh, I., Rastogi, C., Liu, R., Chawla, S., Echenique, F., and Shah, N. B. (2022). Cite-seeing and reviewing: A study on citation bias in peer review. *arXiv:2203.17239*.

Stelmakh, I., Rastogi, C., Shah, N. B., Singh, A., and Daumé III, H. (2020). A large scale randomized controlled trial on herding in peer-review discussions. *arXiv:2011.15083*.

Stelmakh, I., Shah, N., and Singh, A. (2019). On testing for biases in peer review. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Stelmakh, I., Shah, N., and Singh, A. (2021a). PeerReview4All: Fair and accurate reviewer assignment in peer review. *Journal of Machine Learning Research*, 22(163):1–66.

Stelmakh, I., Shah, N. B., and Singh, A. (2021b). Catch me if I can: Detecting strategic behaviour in peer assessment. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 4794–4802.

Stelmakh, I., Shah, N. B., Singh, A., and Daumé III, H. (2021c). A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 4785–4793.

Stelmakh, I., Shah, N. B., Singh, A., and Daumé III, H. (2021d). Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW1):1–17.

Strack, F. and Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, 73:437–446.

Sugimoto, C. R. and Cronin, B. (2013). Citation gamesmanship: Testing for evidence of ego bias in peer review. *Scientometrics*, 95(3):851–862.

Tan, M., Dai, Z., Ren, Y., Walsh, T., and Aleksandrov, M. (2021). Minimal-envy conference paper assignment: Formulation and a fast iterative algorithm. In *2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT)*, pages 667–674.

Tang, W., Tang, J., and Tan, C. (2010). Expertise matching via constraint-based optimization. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 34–41, Washington, DC, USA. IEEE Computer Society.

Taylor, C. J. (2008). On the optimal assignment of conference papers to reviewers. Technical report, Department of Computer and Information Science, University of Pennsylvania.

Taylor and Francis group (2015). Peer review in 2015 a global view. `https://authorservices.taylorandfrancis.com/publishing-your-research/peer-review/peer-review-global-view/`.

Teixeira da Silva, J. and Al-Khatib, A. (2017). The Clarivate Analytics acquisition of Publons–an evolution or commodification of peer review? *Research Ethics*.

Teplitskiy, M., Ranub, H., Grayb, G. S., Meniettid, M., Guinan, E. C., and Lakhani, K. R. (2019). Social influence among experts: Field experimental evidence from peer review.

Thorngate, W. and Chowdhury, W. (2014). By the numbers: Track record, flawed reviews, journal space, and the fate of talented authors. In *Advances in Social Simulation*, pages 177–188. Springer.

Thornhill, T. (2018). We want black students, just not you: How white admissions counselors screen black prospective students. *Sociology of Race and Ethnicity*, page 2332649218792579.

Thurner, S. and Hanel, R. (2011). Peer-review in a world with rational scientists: Toward selection of the average. *The European Physical Journal B*, 84(4):707–711.

Tite, L. and Schroter, S. (2007). Why do peer reviews decline to review? A survey. *Journal of epidemiology and community health*, 61:9–12.

Tomiyama, A. J. (2007). Getting involved in the peer review process. `https://www.apa.org/science/about/psa/2007/06/student-council` [Accessed: 9/7/2020].

Tomkins, A., Zhang, M., and Heavlin, W. D. (2017). Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713.

Toor, R. (2009). Reading like a graduate student. `https://www.chronicle.com/article/Reading-Like-a-Graduate/47922` [Accessed: 9/7/2020].

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3):249–276.

Tran, H. D., Cabanac, G., and Hubert, G. (2017). Expert suggestion for conference program committees. In *2017 11th International Conference on Research Challenges in Information Science (RCIS)*, pages 221–232.

Travis, G. D. L. and Collins, H. M. (1991). New light on old boys: Cognitive and institutional particularism in the peer review system. *Science, Technology, & Human Values*, 16(3):322–341.

Triggle, C. and Triggle, D. (2007). What is the future of peer review? why is there fraud in science? is plagiarism out of control? why do scientists do bad things? is it all a case of: " all that is necessary for the triumph of evil is that good men do nothing?". *Vascular health and risk management*, 3:39–53.

Tung, A. K. (2006). Impact of double blind reviewing on SIGMOD publication: A more detail analysis. *ACM SIGMOD Record*, 35(3):6–7.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Van Noorden, R. (2020). Highly cited researcher banned from journal board for citation abuse. *Nature*, 578:200–201.

Vardi, M. Y. (2010). Hypercriticality. *Communications of the ACM*, 53(7):5–5.

Vargha, A. and Delaney, H. D. (2000). A critique and improvement of the cl common language effect size statistics of mcgraw and wong. *Journal of Educational and Behavioral Statistics*, 25(2):101–132.

Vrettas, G. and Sanderson, M. (2015). Conferences versus journals in computer science. *Journal of the Association for Information Science and Technology*, 66(12):2674–2684.

Wade, A. D. (2022). The semantic scholar academic graph (S2AG). In *Companion Proceedings of the Web Conference 2022 (WWW'22 Companion)*.

Wakefield, A., Murch, S., Anthony, A., Linnell, J., Casson, D., Malik, M., Berelowitz, M., Dhillon, A., Thomson, M., Harvey, P., Valentine, A., Davies, S., and Walker-Smith, J. (1998). Retracted: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, 351(9103):637–641.

Wang, J. and Shah, N. B. (2019). Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *AAMAS*.

Wang, J., Stelmakh, I., Wei, Y., and Shah, N. B. (2021). Debiasing evaluations that are biased by evaluations. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 10120–10128.

Ware, M. (2016). Publishing research consortium peer review survey 2015. *Publishing Research Consortium*.

Webb, T. J., O'Hara, B., and Freckleton, R. P. (2008). Does double-blind review benefit female authors? *Trends in Ecology and Evolution*, 23(7):351 – 353.

Weber, E. J., Katz, P. P., Waeckerle, J. F., and Callaham, M. L. (2002). Author perception of peer review: impact of review quality and acceptance on satisfaction. *JAMA*, 287(21):2790–2793.

Weisband, S. P. (1992). Group discussion and first advocacy effects in computer-mediated and face-to-face decision making groups. *Organizational Behavior and Human Decision Processes*, 53(3):352 – 380.

Weisberg, S. (2005). *Applied Linear Regression*. Wiley, Hoboken NJ, third edition.

West, S., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D., Holtgrave, D., Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M., and Mullen, P. D. (2008). Alternatives to the randomised controlled trial. *American journal of public health*, 98:1359–66.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Wu, R., Guo, C., Wu, F., Kidambi, R., Van Der Maaten, L., and Weinberger, K. (2021). Making paper reviewing robust to bid manipulation attacks. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11240–11250. PMLR.

Xie, B., Shen, Z., and Wang, K. (2021). Is preprint the future of science? A thirty year journey of online preprint services. *ArXiv*, abs/2102.09066.

Xu, Y., Zhao, H., Shi, X., and Shah, N. (2019). On strategyproof conference review. In *IJCAI*.