Eliciting Honest Information from Authors Using Sequential Review*

Yichi Zhang[1], **Grant Schoenebeck**[1], **Weijie Su**[2]

[1]School of Information, University of Michigan
[2]The Department of Computer and Information Science, University of Pennsylvania
yichiz@umich.edu, schoeneb@umich.edu, suw@wharton.upenn.edu

Abstract

In the setting of conference peer review, the conference aims to accept high-quality papers and reject low-quality papers based on noisy review scores. A recent work proposes the isotonic mechanism, which can elicit the ranking of paper qualities from an author with multiple submissions to help improve the conference's decisions. However, the isotonic mechanism relies on the assumption that the author's utility is both an increasing and a convex function with respect to the review score, which is often violated in realistic settings (e.g. when authors aim to maximize the number of accepted papers). In this paper, we propose a sequential review mechanism that can truthfully elicit the ranking information from authors while only assuming the agent's utility is increasing with respect to the true quality of her accepted papers. The key idea is to review the papers of an author in a sequence based on the provided ranking and conditioning the review of the next paper on the review scores of the previous papers. Advantages of the sequential review mechanism include: 1) eliciting truthful ranking information in a more realistic setting than prior work; 2) reducing the reviewing workload and increasing the average quality of papers being reviewed; 3) incentivizing authors to write fewer papers of higher quality.

1 Introduction

Peer review, the process of evaluating scientific research by volunteered experts, undergirds the success of a conference by ensuring the accepted papers are of high quality. However, the reliability of peer review (especially for large computer science conferences) has raised significant concerns. In a NeurIPS experiment conducted in 2014 (Lawrence and Cortes 2014), it was shown that within the set of papers recommended for acceptance by two independent committees, the disagreement rate was as high as 50%. This result was further confirmed in the repeated experiment conducted in 2021 (Cortes and Lawrence 2021). Even worse, the rapid growth of the reviewing workload and the shortage of qualified reviewers, have posed unprecedented challenges to our review system (Sculley, Snoek, and Wiltschko 2018).

This leads to the dilemma of conference peer review: the conference's objective of accepting only high-quality papers (from a large set of submissions) clashes with the shortage of reliable peer reviews upon which the conference must base its decisions. To mitigate this issue, we introduce a novel review mechanism called the *sequential review mechanism* that can 1) solicit high-quality information from authors to assist the acceptance/rejection decisions, 2) reduce the reviewing workload and 3) incentivize authors to write high-quality papers.

The main challenge is how to elicit useful information from the authors who have conflicting interests with the conference. For instance, authors may desire to have more publications, leading them to seek acceptance for more papers, regardless of the papers' quality. In this case, although authors possess the best signals of their own papers' quality compared with any reviewer, they may prefer not to disclose this information truthfully to the conference. For example, while being asked to report the true quality of their papers, authors may be inclined to inflate scores in order to increase the chance of acceptance.

Fortunately, positive results exist. Su (2021) shows that it is possible to elicit truthful rankings of paper quality from an author with multiple submissions. The main idea of the proposed *isotonic mechanism* is to shift the noisy review scores by running an isotonic regression based on the author's reported ranking. It is shown that reporting the ranking of papers truthfully is the best response for an author. Unfortunately, this result only holds when the agent's utility for each paper is an **increasing and convex** function of the review score and additive across papers. However, the assumption of convex utility is strong and likely violated in the setting of conference peer review. For example, this assumption is violated when an author aims to maximize the number of her accepted papers. Moreover, in this case, the isotonic mechanism can be gamed in a rather straightforward manner. Suppose such an author has several borderline papers and one outstanding paper to submit. Under the isotonic mechanism, her best response is nonetheless to rank the outstanding paper at the bottom such that the review scores of all borderline papers will be shifted up after the isotonic regression, which will almost certainly lead to the acceptance of all papers.[1]

In this work, we build on the idea of eliciting the author's

*The full version of this paper is available on Arxiv (Zhang, Schoenebeck, and Su 2023).

[1]For an introduction of the isotonic mechanism and a careful illustration of this example, please refer the full version.

ranking information from the previous work and primarily focus on addressing the incentive issue discussed above. We propose the sequential review mechanism in a natural conference review model. Our method can elicit the true ranking information as long as the author's utility is additive in terms of the rewards of all **accepted** papers, where the reward is a **non-decreasing** function of the accepted papers' qualities. The sequential review mechanism works by reviewing an author's submissions in sequence. In particular, papers with higher reported rankings are reviewed with priority, while papers with lower reported rankings will be conditionally reviewed depending on the review scores of the higher-ranked papers from the same author. If the review process terminates, e.g. due to a notably low review score of a paper, any remaining unreviewed papers will be rejected without further assessment. Intuitively, under the sequential review mechanism, any misreporting of the true ranking of the papers will result in an earlier termination of the review process, which penalizes dishonest behaviors.[2]

1.1 Contributions and Results

Our main contribution is a framework for designing theoretically robust mechanisms, with the potential to improve conference peer review in practice. The proposed sequential review mechanism not only addresses a key incentive issue that plagues prior work but also exhibits many additional appealing properties. Due to space limitations, please refer to our full version (Zhang, Schoenebeck, and Su 2023) for proofs, details of experiments, and additional results.

Truthful sequential review mechanisms. Under the sequential review mechanism framework, we first identify a sufficient constraint that ensures a sequential review mechanism to be truthful (i.e. reporting the true ranking of the paper quality is the best response for any author). While not necessary, this constraint provides a large space of truthful sequential review mechanisms. To show the effectiveness of our framework, we introduce two practical mechanisms as examples: the *memoryless coin-flip mechanism* that reviews the $i+1$th ranked paper with a probability determined by the review score of the ith ranked paper; and the *credit pool mechanism*, which counts the cumulative review scores (positive or negative) of the reviewed papers and terminates the review process when the "credit pool" is empty.

Conference utility and review burden. The sequential review mechanism utilizes the authors' information to prioritize the review of high-quality papers. Therefore, it can improve the conference's utility (as low-quality papers are less likely to be accepted), while reducing the review burden by reallocating more review resources to papers deemed likely to be of higher quality. To evaluate the performance of the sequential review mechanism, we use the *parallel review mechanism* as the baseline which unconditionally reviews all papers. We further use the isotonic mechanism with oracle access to the true ranking information as an upper bound.[3]

Our simulation results suggest that compared with the baseline, the sequential review mechanism can improve the conference utility towards the upper bound by over 40% when the author submits more than three papers. This effect is even more significant when 1) each author has more papers, 2) papers are more likely to be of low quality and 3) reviewers are more noisy. Moreover, we empirically investigate the number of reviews that a sequential review mechanism can save while achieving the same conference utility as the parallel review mechanism. We employ the ICLR OpenReview datasets spanning recent years and develop a more realistic review model. Our results indicate that about 20% of the review burden can be saved when utilizing the sequential review mechanism. Furthermore, this number will increase over time if the trend of a growing number of submissions per author continues.

Endogenous paper quality. In the setting where authors can choose the effort they exert on each of their papers, we show that compared with the parallel review mechanism, the sequential review mechanism always provides a stronger incentive for writing (fewer) papers of higher quality instead of (more) papers of low quality. This is because the sequential review mechanism decreases the marginal return of producing lower-quality papers by penalizing bottom-ranked papers with lower probabilities of being reviewed. We view this property particularly valuable, especially in light of the prevailing trend where authors submit an increasingly large number of papers to conferences, sometimes disregarding their inherent quality.

2 Related Works

Other than the isotonic mechanism (Su 2021), several attempts exist aiming to improve the peer review system with a focus on dealing with strategic interactions between conferences and authors. In a setting where authors can strategically decide the venues to submit their papers, Zhang et al. (2022) show how to design the review mechanism to achieve the Pareto optimal tradeoff between the conference quality and the review burden. In a recent work, Srinivasan and Morgenstern (2021) propose the idea of using the VCG mechanism to elicit bids from authors and using peer prediction mechanisms to evaluate reviews and reward the reviewers (with virtual money). In dealing with the malicious bidding problem, a stream of literature focuses on designing and optimizing the paper-reviewer assignment mechanism to guarantee strategyproofness (Aziz et al. 2019; Jecmen et al. 2020; Dhull et al. 2022; Xu et al. 2018).

Our work is also related to the impartial peer selection problem, where self-interested agents assess one another in such a way that none of them has an incentive to misrepresent their evaluation. A famous example is the NSF experiment, where each PI was asked to rank seven proposals from other PIs (Naghizadeh and Liu 2013). The primary

[2] One may be concerned that the sequential review mechanism will result in a significant delay in the review process. However, note that the mechanism works exactly the same if all papers are simultaneously reviewed or reviewed in batches, as long as the acceptance/rejection decisions are made in sequence. See more in Section 7.

[3] The isotonic mechanism is not truthful in our setting but we nonetheless provide oracle access to the true ranking to it.

goal of the peer selection literature is to improve the accuracy of assessments while guaranteeing strategyproofness (Naghizadeh and Liu 2013; Aziz et al. 2019; De Clippel, Moulin, and Tideman 2008). However, these investigations differ from our problem since we focus on eliciting evaluations of multiple items held by a single agent directly from the agent itself, rather than from the peers.

Additionally, there exists a considerable body of literature that aims to improve peer review from the reviewer's perspective. This includes investigations into single versus double blind reviewing (Blank 1991; Snodgrass 2006; Bazi 2020), assigning versus bidding papers (Cabanac and Preuss 2013; Meir et al. 2021), review scale and miscalibration (Siegelman 1991; Wang and Shah 2018; Spalvieri et al. 2014), and dishonest behaviors (Cohen et al. 2016; Fanelli 2009; Littman 2021). A recent survey by Shah (2022) provides additional contexts and perspectives on the problems of peer review.

3 Model

We view the peer review mechanism as an individual contract. That is, each paper is reviewed independently based on its review scores. Therefore, while reasoning about an agent's best response (Section 4 and 6), it is sufficient to assume that there is only one agent with n papers. When we investigate the optimization of a review mechanism (Section 5.3), we assume authors are drawn from a distribution.

Throughout the paper, we will use $[n]$ to denote the set $\{1, 2, \ldots, n\}$ and use $[n]_0$ to denote $\{0, 1, \ldots, n\}$. Suppose an author has n submissions indexed by $i \in [n]$, each with a quality of $q_i \in \mathbb{R}$. Suppose without loss of generality that $q_1 \geq q_2 \geq \cdots q_n$. We name each paper by its true ranking, e.g. paper 1 is the paper with the highest quality. To better present our results, unless otherwise specified, we assume that the author knows the true qualities of all her papers. Nonetheless, we note that our theoretical results can easily be generalized to the setting where the author observes a noisy signal $s_i = q_i + \xi_i$ for each of her papers where ξ_i are i.i.d. sampled from some distribution (see Section 4.2).

The conference decides whether to accept or to reject each of the n submissions based on its review score. Given the true quality q_i, the paper's review score is observed by adding an error term, i.e. $r_i = q_i + \epsilon_i$, where ϵ_i is i.i.d. sampled from some distribution. The conference commits to an acceptance policy such that a paper with review score r (if it is reviewed) is accepted with probability $P_{\text{acc}}(r)$. For example, for a threshold acceptance policy, $P_{\text{acc}}(r) = 1$ if $r \geq \tau_{acc}$ and 0 otherwise, where $\tau_{acc} \in \mathcal{R}$ is a threshold. We assume that the utility of the conference is the sum of the accepted papers' quality, i.e. $U_c(\mathcal{M}) = \sum_{i \in [n]} q_i \cdot \mathbb{1}[\text{paper } i \text{ is accepted under mechanism } \mathcal{M}]$.

In addition to soliciting review scores, the conference can solicit a ranking of the author's submissions. That is, the author reports a permutation π of her papers, where $\pi(i)$ is the rank of paper i after the permutation. The truthful report is the original ranking, i.e. $\pi^*(i) = i$. We assume that the author's utility is the sum of the rewards of her accepted papers: each paper's reward is zero if rejected and $u_a(q_i)$ if accepted, where u_a is a non-negative and non-decreasing reward function. For example, if $u_a(q) = 1$ for any q, the author's goal is to maximize the expected number of accepted papers. The author can strategically report a ranking π so as to maximize its expected utility. Let $U_a(\pi)$ be the expected author utility under the permutation π, where the randomness is with respect to the review noise and the mechanism.

4 Truthful Sequential Review Mechanisms

This section presents a framework for designing truthful review mechanisms. In particular, we introduce the *sequential review mechanism framework* and show a sufficient condition for a sequential review mechanism to be truthful. We further provide two concrete and practical truthful sequential review mechanisms under this framework as examples.

4.1 The Sequential Review Mechanism Framework

We first introduce the *naive sequential review mechanism* as an illustrative example of the more general sequential review mechanism framework.

Definition 4.1. Given an author with n papers and a ranking of these papers, the naive sequential review mechanism reviews one paper at a time based on the order of the reported ranking. The first paper is always reviewed. For i ranging from 2 to n, the paper ranked in the ith place is reviewed if and only if the paper ranked in $i-1$st place is accepted.

Intuitively, the naive sequential review mechanism incentivizes truth-telling because any manipulation of the true ranking will more likely result in an early stop of the review process which harms the author. However, the naive mechanism can be too stringent in reality, especially when authors are likely to produce good papers. To address this, we generalize this idea and present the sequential review mechanism framework. This framework offers a large set of mechanisms that can be fine-tuned to optimize performance in various settings.

At a high level, the idea is to condition the review of lower-ranked papers on the acceptance (and thus the review score) of the higher-ranked papers. If the mechanism decides not to review the paper in round i, any paper in round $j > i$ will be rejected without review. We then say that the mechanism terminates in round i. Now, we formally introduce the sequential review mechanism framework.

Definition 4.2. A sequential review mechanism $\mathcal{M}_s = (P_{\text{acc}}, P_{\text{rev}}, \boldsymbol{\mu})$ has three components:

- An *acceptance policy* P_{acc} that maps from a review score to a probability of accepting the corresponding paper.
- A *review policy* P_{rev} that maps from a review state in round i to a probability of reviewing the paper in round $i+1$, for $i \in [n-1]_0$.[4]
- A *state transition mapping* μ_i that maps from a review state in round i and the review score of the paper in round $i+1$ to a distribution of states in round $i+1$, for $i \in [n-1]_0$.

[4]Here, we assume $i < n$ because there is no need to discuss the review policy in round n.

In the above definition, a *review state* in round i, denoted as $\phi_i \in \Phi_i$, is sufficient to determine the probability that the paper in round $i+1$ will be reviewed, where Φ_i is the space of review states in round i. Specifically, Φ_0 is the space of initial states before the review of the first paper. The review states are weakly ordered such that the author always prefers to be in a higher-ordered state. That is, for each pair of states in round i, one must have a (weakly) higher order than the other, denoted as $\phi_i' \succeq \phi_i$ for every $\phi_i', \phi_i \in \Phi_i$ and every $i \in [n-1]_0$. We use $\phi_i' \succ \phi_i$ to represent the strict ordering. Furthermore, if the author is indifferent between two states, we say $\phi_i' \sim \phi_i$. When comparing two vectors of states, we use the same notations to indicate term-wise preference. For example, if $\boldsymbol{\phi}, \boldsymbol{\phi}' \in \Phi^m$, $\boldsymbol{\phi}' \succeq \boldsymbol{\phi}$ implies that $\phi_i' \succeq \phi_i$ for any $1 \le i \le m$.

Taking the naive sequential review mechanism as an example, the review state is $\phi_i = 1$ if the paper in round i is accepted and 0 otherwise. In round 0, $\phi_0 = 1$. Then, the review policy is $P_{\text{rev}}^{std}(1) = 1$ and $P_{\text{rev}}^{std}(0) = 0$ for any round. The state transition mapping of the naive sequential review mechanism is that $\mu_i^{std}(1, r_i) = 1$ with probability $P_{\text{acc}}(r_i)$, $\mu_i^{std}(1, r_i) = 0$ with probability $1 - P_{\text{acc}}(r_i)$, and $\mu_i^{std}(0, \cdot) = 0$ with probability 1 for any i.

We further note that by our definition, the acceptance policy is assumed to be memoryless, where the acceptance of a paper only depends on its own review score. However, the review policy can have memory such that previous review scores may affect the distribution of the review state in round i which affects the probability of the paper in round $i+1$ being reviewed.

4.2 A Sufficient Condition For Truthfulness

Now, we investigate what conditions on $P_{\text{acc}}, P_{\text{rev}}$ and $\boldsymbol{\mu}$ are sufficient for a sequential review mechanism to be truthful. At a high level, we need both policies to be monotone which rewards higher review scores and punishes lower review scores.

Definition 4.3. An acceptance policy is monotone if P_{acc} is (weakly) increasing, i.e. $P_{\text{acc}}(r') \ge P_{\text{acc}}(r)$ for any $r' \ge r$.

Definition 4.4. A review policy is monotone if $P_{\text{rev}}(\phi_i') \ge P_{\text{rev}}(\phi_i)$ for every $\phi_i' \succeq \phi_i$.

A monotone acceptance policy rewards a paper with a higher review score by accepting it with a higher probability, while a monotone review policy rewards a higher-ordered review state with an increased probability of reviewing the paper in the next round. However, the requirements for the state transition mapping are more complicated, where we use the concept of stochastic dominance.

Definition 4.5. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two m-dimension random vectors of review states $\boldsymbol{X}, \boldsymbol{Y} \in \Phi^m$ for some review state space Φ. We say $\boldsymbol{X}$ first-order stochastic dominates $\boldsymbol{Y}$ if $\Pr(\boldsymbol{X} \succeq \boldsymbol{\phi}) \ge \Pr(\boldsymbol{Y} \succeq \boldsymbol{\phi})$ for any $\boldsymbol{\phi} \in \Phi^m$.

For simplicity, let $\tilde{\mu}(r_1, r_2|\phi_i) = \mu_{i+1}(\mu_i(\phi_i, r_1), r_2)$ be the state distribution in round $i+1$ conditioned on having review state ϕ_i in round $i-1$, and having review scores r_1 and r_2 in round i and round $i+1$ respectively.

Definition 4.6. We say the state transition mapping $\boldsymbol{\mu}$ is *monotone* if for any review round $i \in [n-1]_0$

1. it is monotone in score: For any state $\phi_i \in \Phi_i$, $\mu_i(\phi, r')$ first-order stochastic dominates $\mu_i(\phi, r)$ for any $r' \ge r$;
2. it is monotone in state: For any review score r, $\mu_i(\phi', r)$ first-order stochastic dominates $\mu_i(\phi, r)$ for any state $\phi' \succeq \phi$;
3. it is monotone in ordering: For any state $\phi_i \in \Phi_i$ and review scores $r' \ge r$, $\tilde{\mu}(r', r|\phi_i)$ first-order stochastic dominates $\tilde{\mu}(r, r'|\phi_i)$. Furthermore, for any $r_1 \le r_2 \le r_3 \le r_4$ such that $r_1 + r_4 = r_2 + r_3$, let $X \sim \tilde{\mu}(r_4, r_1|\phi_i)$, $Y \sim \tilde{\mu}(r_2, r_3|\phi_i)$, $X' \sim \tilde{\mu}(r_1, r_4|\phi_i)$ and $Y' \sim \tilde{\mu}(r_3, r_2|\phi_i)$. Let $\bar{Z} = \max(X, Y)$, $\underline{Z} = \min(X, Y)$, $\bar{Z}' = \max(X', Y')$, $\underline{Z}' = \min(X', Y')$.[5] Then, $(\bar{Z}, \underline{Z})$ first-order stochastic dominates $(\bar{Z}', \underline{Z}')$.

The monotonicities in score and state suggest that the state transition mapping results in a better state when the review score is higher and the review state is higher ordered, respectively. The monotonicity in ordering deals with the cases where the review scores in two rounds are swapped. First, it requires the review state distribution to be better if the higher review score is put earlier. Furthermore, supposing there are four ordered review scores, it compares the distribution of a pair of review states. In particular, putting the largest score r_4 earlier in round i with the lowest score r_1 in round $i+1$ and putting r_2 in round i with r_3 in round $i+1$ should lead to a better distribution of a pair of review states than swapping the review scores in round i and $i+1$.

We are ready to present the main theorem.

Theorem 4.7. *The sequential review mechanism $\mathcal{M}^s = (P_{acc}, P_{rev}, \boldsymbol{\mu})$ is truthful if P_{acc}, P_{rev} and $\boldsymbol{\mu}$ are monotone.*

At a high level, the proof follows by coupling the realizations of review noises. Then, due to the monotonicity of $P_{\text{acc}}, P_{\text{rev}}$, and $\boldsymbol{\mu}$, flipping the true order of any two papers will result in a review state that is always dominated by truthful reporting.

Remark 4.8 (Noisy Authors)**.** Although the proof of Theorem 4.7 assumes that the author perfectly knows her paper qualities, we emphasize that it can be straightforwardly generalized to the setting where the author observes a signal $s_i = q_i + \xi_i$ with i.i.d. noise term ξ_i for every paper i. Intuitively, this works because the author's noise affects her reasoning about the ranking of papers in the exactly same way as the review noise. Let $\hat{\epsilon}_i = \epsilon_i - \xi_i$ such that $r_i = s_i + \hat{\epsilon}_i$. In this way, by coupling the new noise term $\hat{\epsilon}_i$, the same proof can be used to show that the author will truthfully rank her papers based on her signals.

4.3 The Memoryless Coin-Flip Mechanism

Here, we provide an example of how to use our framework to design a truthful sequential review mechanism. In this example, the acceptance of a paper in round i will guarantee the review of the paper in the next round; while if a paper is rejected, the mechanism will review the paper in round

[5] Here, the max and min function select the higher-ordered state and the lower-ordered state respectively.

$i+1$ with probability $\rho(r_i)$ determined by the review score of the paper in round i. Furthermore, the mechanism always reviews the first paper. We call this mechanism the *memoryless coin-flip mechanism*.

We show that the memoryless coin-flip mechanism is truthful by mapping it to the sufficient conditions of truthfulness as shown in Theorem 4.7. The following proposition states our result.

Proposition 4.9. *The memoryless coin-flip mechanism is truthful if P_{acc} is monotone and ρ is increasing.*

4.4 The Credit Pool Mechanism

Another example of the sequential reviewing framework implements the idea of a reputation system. Suppose the conference keeps a record of a credit pool, which is initialized at $B_0 \geq 0$. For every reviewed paper, the mechanism will increase (or decrease) the credit pool by a credit score which is determined by the review of that paper. Let $\beta : \mathbb{R} \to \mathbb{R}$ be a credit function which maps from a review score to a review credit. Note that β can be negative, indicating a punishment of papers with low review scores. The credit pool mechanism reviews the paper in round $i+1$ if and only if $B_i \geq 0$. Therefore, $B_{i+1} = B_i + \beta(r_i)$ if $B_i \geq 0$ and $B_{i+1} = B_i$ otherwise. We present the following result.

Proposition 4.10. *The credit pool mechanism is truthful if P_{acc} is monotone and β is increasing and convex.*

Intuitively, the credit pool mechanism is truthful because any untruthful permutation is more likely to result in an earlier termination of the review process.

5 Evaluating Sequential Review Mechanisms

We evaluate a mechanism from two dimensions: conference utility and review burden. The former is measured by the sum of accepted papers' quality, and the latter quantifies the number of reviewed papers, while both are normalized by the total number of submitted papers. For both computational and practical considerations, we focus on the threshold sequential review mechanism, a special case of the memoryless coin-flip mechanism. In comparison, we use the threshold parallel review mechanism as a baseline and the threshold isotonic mechanism with the underlying ranking information as an unreachable upper bound.

We conduct experiments on both a simple model with Gaussian review noise and a more complicated real-data estimated model where each paper has multiple integer-valued review scores. Our results suggest that the sequential review mechanism can achieve conference utility that is competitive compared to the upper bound and can significantly reduce the review burden. Moreover, in the real-data estimated model, we show that the sequential review mechanism can save more than 20% review burden compared with the baseline conditioned on a weakly better conference utility.

5.1 Mechanisms of Comparison

Here, we introduce how we implement and optimize the three types of mechanisms in our experiments. For both computational and practical considerations, we focus on the threshold acceptance policy, i.e. a paper is accepted if and only if its (modified) review score is larger than a threshold.

The threshold sequential review mechanism. The threshold sequential review mechanism is a memoryless coin-flip mechanism with P_{acc} and ρ being threshold functions: $P_{\text{acc}}(r) = 1$ if $r \geq \tau^s_{acc}$ and 0 otherwise; $\rho(r) = 1$ if $r \geq \tau^s_{rev}$ and 0 otherwise. Furthermore, $\tau^s_{rev} \leq \tau^s_{acc}$. In words, the threshold sequential review mechanism accepts a paper, conditioned on it being reviewed, if its review score is larger than a threshold τ^s_{acc}; and it reviews the next paper if the review score is larger than a lower threshold τ^s_{rev}.

The threshold parallel review mechanism. The parallel review mechanism, which operates without soliciting the ranking information from authors, guarantees to independently review all papers. The threshold parallel review mechanism is characterized by the acceptance threshold τ^p_{acc} where papers are accepted if and only if their review scores are no less than τ^p_{acc}. Note that the threshold parallel review mechanism is a special case of the threshold sequential review mechanism by setting $\tau^s_{rev} = -\infty$.

The threshold isotonic mechanism with true ranking information. In general, the isotonic mechanism cannot truthfully elicit the ranking information from authors in our setting where the author's utility is not convex with respect to the review score. However, we assume the author still reports the ranking of their papers' quality truthfully and uses the isotonic mechanism with this information as the upper bound. The isotonic mechanism modifies the original review scores by solving the isotonic regression conditioned on the author's ranking. Then, a threshold acceptance policy is applied to the modified review scores such that the paper with score r is accepted if and only if $r \geq \tau^i_{acc}$.

To emphasize the importance of truthfulness, we note that in the absence of truthful ranking information, the isotonic mechanism can yield a conference utility that is even worse than the baseline.

5.2 Gaussian Review Noise

We first introduce a simple yet intuitive model where each paper is associated with a single review score, which is the true quality plus an additive Gaussian review noise. The simplicity of this model offers computational convenience, enabling us to efficiently explore the parameter space. Note that under this simple model, we again assume that there is only one author. Furthermore, we note that in the rest of this section, when we mention a mechanism, we are referring to its threshold implementation.

Model and Experiment Setup Now, we introduce the parametric setting that is used to generate synthetic data for our experiments. First, suppose the author draws n i.i.d. paper qualities from $\mathcal{N}(\mu_q, \sigma_q)$. Then, the conference observes the true ranking of these samples. Let $\boldsymbol{q}$ be the ordered vector of paper qualities from high to low. Next, the conference draws an i.i.d. review noise $\epsilon_i \sim \mathcal{N}(0, \sigma_r)$ for each $i \in [n]$. Finally, review scores are observed: $r_i = q_i + \epsilon_i$. The parameters $(n, \mu_q, \sigma_q, \sigma_r)$ defines a *Gaussian review model* φ_g.

Given φ_g and a mechanism $\mathcal{M}$, we use the Monte-Carlo method with $10,000$ samples of $\boldsymbol{q}$ to estimate the expected conference utility. We refer the details of this estimation to the full version. For each parameter setting, we further optimize the threshold(s) of the three types of mechanisms using stochastic gradient descent. Again, the gradient function can be estimated using a Monte Carlo method.

Results We first define two dimensions of our evaluations.

Definition 5.1. Let U_c^p, U_c^s, and U_c^i be the expected conference utility under the parallel review mechanism, the sequential review mechanism, and the isotonic mechanism with the optimized thresholds, respectively. The *relative conference utility* is defined as $\hat{U}_c^s = (U_c^s - U_c^p)/(U_c^i - U_c^p)$.

That is, the relative conference utility is the conference utility of the threshold sequential review mechanism while normalizing the conference utility of the baseline to 0 and normalizing that of the upper bound to 1.

Definition 5.2. Let B^p and B^s be the review burden of the parallel review mechanism and the sequential review mechanism respectively. The *relative review burden* is defined as $\hat{B}^s = B^s/B^p$ conditioned on achieving the same conference utility.

Note that by the design of mechanisms, $B^p = n$ and $B^s \leq 1$. Thus, $0 < \hat{B}^s \leq 1$ and a smaller relative review burden imply that the sequential review mechanism can save more reviews compared with the parallel review mechanism without harming the conference utility.

In Fig. 1, we observe that the sequential review mechanism exhibits a significant improvement over the baseline on both dimensions: over 40% improvement on the relative conference utility and $10-30\%$ reduction on review burden even when $n = 2$. Furthermore, we observe that the sequential review mechanism can significantly improve the average quality of the reviewed papers, as bottom-ranked low-quality papers are more likely to be rejected without review.

In addition, we observe that the sequential review mechanism is particularly more effective when 1) the author is more likely to write low-quality papers (μ_q is smaller), 2) the author has more papers (n is larger) and 3) reviews are noisier (σ_r is larger). The intuition behind these observations is that the number of papers that are rejected by the sequential review mechanism but are mistakenly accepted by the parallel review mechanism increases in these three cases.

5.3 Softmax Review Noise With Real Data

The real review data has two features that are not adequately addressed by the Gaussian noise model. First, review scores are integers, not continuous real values. Second, each paper has multiple independent review scores, rather than a single score. To incorporate these distinctions, we present a more fine-grained model, wherein the review score is characterized by a softmax function. Furthermore, in this section, we optimize the mechanisms for the entire population, rather than an individual agent, based on the empirical quality distribution estimated from real data.

Another challenge of fitting the model with real data is the coauthorship problem. Our solution is to assign each paper to one of its authors. That is, we iteratively find the author with the largest number of papers, assign those papers to that author, and remove the author and the assigned papers until every paper is paired with one author.

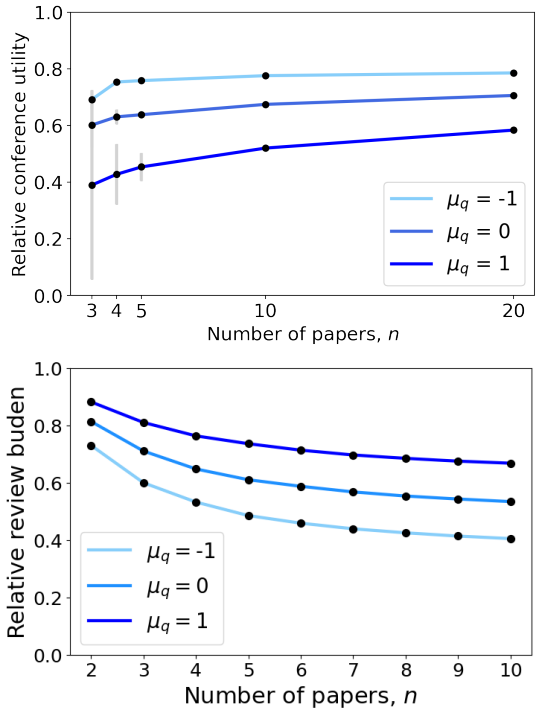

Figure 1: The relative conference utility and relative review burden under different parameter settings. The parameter setting is $\varphi_g = (n = *, \mu_q = *, \sigma_q = 2, \sigma_r = 1)$. The error bars are particularly large for small n and large μ_q because in these cases the difference between the performances of three mechanisms is small.

We defer the detailed discussion of the softmax review model and how we fit it with ICLR OpenReview datasets to the full version while highlighting some key takeaways.

First, we observe a consistent increase in both the total number of papers and the average number of papers per author, where the latter increases from 1.81 to 1.93 over three years. Second, we find that the relative review burden of the threshold sequential review mechanism is around 0.8 for all three tested years, indicating a potential reduction of 20% in review burden while using the sequential review mechanism. Furthermore, this effect has become more pronounced over the years as more and more authors submit multiple papers to the same conference.

6 Endogenous Paper Quality

This section considers the setting where authors have the choice of the quality of papers they write. Papers of higher quality have a higher probability of being accepted and bring higher rewards to the author if accepted. However, producing a high-quality paper usually requires greater effort and time from the author. The main result in this section suggests

that compared with the parallel review mechanism, the naive sequential review mechanism can better incentivize authors to improve the quality of papers instead of producing more papers with lower quality. In this section, we assume that the author reports the ranking truthfully.

We first consider a binary effort setting as a toy example.

Suppose the author can either exert a high effort to write a high-quality paper with an acceptance probability p_h or exert a low effort to write a low-quality paper with an acceptance probability $p_l < p_h$. Furthermore, the acceptance of a low-quality paper gives the author a reward of u_l while the acceptance of a high-quality paper brings a reward of $u_h \geq u_l$. Let $U_a^s(n_h, n_l)$ and $U_a^p(n_h, n_l)$ be the expected utility of writing n_h high-quality papers and n_l low-quality papers under naive sequential review mechanism and the parallel review mechanism respectively.

Theorem 6.1. *For any $n_h, n_l, n_h', n_l' \in \mathbb{N}_0$ such that $n_h' > n_h$ and $n_l' > n_l$, if $U_a^p(n_h', n_l) \geq U_a^p(n_h, n_l')$, $U_a^s(n_h', n_l) \geq U_a^s(n_h, n_l')$. Furthermore, there exist settings where $U_a^s(n_h', n_l) \geq U_a^s(n_h, n_l')$, but $U_a^p(n_h', n_l) < U_a^p(n_h, n_l')$.*

In words, Theorem 6.1 shows that whenever the author wants to write $\Delta n_h = n_h' - n_h$ more high-quality papers compared with writing $\Delta n_l = n_l' - n_l$ more low-quality papers under the parallel review mechanism, she is always willing to do so under the naive sequential review mechanism. However, the opposite is not true.

We defer the proof of Theorem 6.1 and the generalization to the finite effort setting to our full version.

7 Limitations, Discussions and Future Work

Here, we discuss the limitations of our analysis and how to possibly implement the proposed method in practice. However, the implementation details can likely be further improved by future work.

Coauthorship We assume that every paper has only one author. However, in practice, coauthorship is an inevitable issue for the implementation of our method. A straightforward solution is to assign each paper to one of its authors and only solicit the ranking information from that author. For example, in Section 5.3, we greedily assign each paper to the author with the largest number of submissions. In a recent work (Wu et al. 2023), it is shown that this greedy assignment is truthful and has appealing robust approximation guarantees for the isotonic mechanism. Alternatively, we can assign each paper only to its first author, driven by the notion that the first author may possess the most accurate insight into the paper's quality. However, this assignment weakens the sequential review mechanism, as many first authors have fewer than three submissions to their name.

The trade-off between review burden and delay One way to implement the sequential review mechanism is to divide the review process into n phases, where n is the maximum number of papers owned by any single author. In each phase i, the ith ranked paper (if any) for all authors is reviewed, and in phase $i+1$ the sequential review mechanism is applied to determine which papers (if any) necessitate review. This implementation minimizes the review burden but significantly delays the review process. Alternatively, we can have only one review phase where all the papers are simultaneously reviewed. Then, for each author, the acceptance/rejection decision is made in sequence based on the sequential review mechanism. This implementation benefits the conference utility and does not delay the review process, but it does not help with the review burden.

In reality, perhaps a two-phase implementation of the sequential review mechanism can achieve a desirable trade-off. For example, some conferences such as AAAI and EC are already implementing a two-phase review mechanism: all papers are assigned with two reviews in the first phase and only papers with at least one good review will enter the second phase where two more reviews are assigned. We can integrate the sequential review mechanism with this framework. In the first phase, the top $\min(3, \lfloor n_i/2 \rfloor)$ ranked papers of each author with n_i papers are assigned with two reviews, and the remaining papers are assigned with one review (so that no paper is rejected without reviewing). Then, any paper with two negative reviews and those papers that are ranked lower than them by authors are rejected with no further review. The surviving papers enter the second phase, wherein they are assigned additional reviews to reach a total of four reviews each and a sequential review mechanism is implemented. The advantage of this implementation is that the author's information can be leveraged to prioritize the reviewing of high-quality papers.

An interesting future work lies in optimizing the trade-off between the review burden, conference utility, and fairness in a two-phase review mechanism involving multiple reviews per paper while guaranteeing truthfulness.

Broader Applications The insights in this paper can be potentially applied to address general principal-agent problems where decisions rely on noisy evaluations. For example, on content-recommendation platforms, the designer can solicit a ranking of the quality of content from producers and use this information to provide better recommendations. Additional applications suitable for our method include employee recruitment, Wikipedia article reviewing, and second-hand product trading markets. Nonetheless, adapting our method to various applications requires further in-depth modeling and analysis.

8 Conclusion

In the setting of (conference) peer review, we study the problem of how to elicit honest information from authors, who themselves are interested in the outcome. Our main contribution is a framework for designing mechanisms capable of eliciting quality rankings from authors with multiple submissions. Compared with the previous isotonic mechanism, our mechanism works within a more realistic utility model for peer review and addresses a key incentive issue that plagued the previous method. We further investigate the advantages of our mechanism from the aspects of reducing reviewing workload, improving the average quality of the reviewed papers, and incentivizing authors to focus more on the quality of papers rather than the quantity.

Acknowledgments

We would like to thank the Elicitation Mechanisms in Practice Workshop (2022) hosted by the Institute for Data, Econometrics, Algorithms, and Learning (IDEAL), where this work was initiated. This work is supported by the National Science Foundation under Grants #2007256 and #2313137.

References

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.

Bazi, T. 2020. Peer review: single-blind, double-blind, or all the way-blind? *International Urogynecology Journal*, 31(3): 481–483.

Blank, R. M. 1991. The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review. *The American Economic Review*, 1041–1067.

Cabanac, G.; and Preuss, T. 2013. Capitalizing on order effects in the bids of peer-reviewed conferences to secure reviews by expert referees. *Journal of the American Society for Information Science and Technology*, 64(2): 405–415.

Cohen, A.; Pattanaik, S.; Kumar, P.; Bies, R. R.; De Boer, A.; Ferro, A.; Gilchrist, A.; Isbister, G. K.; Ross, S.; and Webb, A. J. 2016. Organised crime against the academic peer review system. *British Journal of Clinical Pharmacology*, 81(6): 1012.

Cortes, C.; and Lawrence, N. D. 2021. Inconsistency in Conference Peer Review: Revisiting the 2014 NeurIPS Experiment. arXiv:2109.09774.

De Clippel, G.; Moulin, H.; and Tideman, N. 2008. Impartial division of a dollar. *Journal of Economic Theory*, 139(1): 176–191.

Dhull, K.; Jecmen, S.; Kothari, P.; and Shah, N. B. 2022. Strategyproofing peer assessment via partitioning: The price in terms of evaluators' expertise. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, 53–63.

Fanelli, D. 2009. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS one*, 4(5): e5738.

Jecmen, S.; Zhang, H.; Liu, R.; Shah, N.; Conitzer, V.; and Fang, F. 2020. Mitigating manipulation in peer review via randomized reviewer assignments. *Advances in Neural Information Processing Systems*, 33: 12533–12545.

Lawrence, N.; and Cortes, C. 2014. The NIPS experiment. *See http://inverseprobability. com/2014/12/16/the-nips-experiment (accessed 3 March 2021)*.

Littman, M. L. 2021. Collusion rings threaten the integrity of computer science research. *Communications of the ACM*, 64(6): 43–44.

Meir, R.; Lang, J.; Lesca, J.; Mattei, N.; and Kaminsky, N. 2021. A market-inspired bidding scheme for peer review paper assignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4776–4784.

Naghizadeh, P.; and Liu, M. 2013. Incentives, Quality, and Risks: A Look Into the NSF Proposal Review Pilot. arXiv:1307.6528.

Sculley, D.; Snoek, J.; and Wiltschko, A. 2018. Avoiding a Tragedy of the Commons in the Peer Review Process. arXiv:1901.06246.

Shah, N. B. 2022. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6): 76–87.

Siegelman, S. S. 1991. Assassins and zealots: variations in peer review. Special report. *Radiology*, 178(3): 637–642. PMID: 1994394.

Snodgrass, R. 2006. Single-versus double-blind reviewing: An analysis of the literature. *ACM Sigmod Record*, 35(3): 8–21.

Spalvieri, A.; Mandelli, S.; Magarini, M.; and Bianchi, G. 2014. Weighting peer reviewers. In *2014 Twelfth Annual International Conference on Privacy, Security and Trust*, 414–419. IEEE.

Srinivasan, S.; and Morgenstern, J. 2021. Auctions and prediction markets for scientific peer review. *arXiv preprint arXiv:2109.00923*.

Su, W. J. 2021. You Are the Best Reviewer of Your Own Papers: An Owner-Assisted Scoring Mechanism. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Wang, J.; and Shah, N. B. 2018. Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings. arXiv:1806.05085.

Wu, J.; Xu, H.; Guo, Y.; and Su, W. 2023. An Isotonic Mechanism for Overlapping Ownership. arXiv:2306.11154.

Xu, Y.; Zhao, H.; Shi, X.; Zhang, J.; and Shah, N. B. 2018. On strategyproof conference peer review. *arXiv preprint arXiv:1806.06266*.

Zhang, Y.; Schoenebeck, G.; and Su, W. 2023. Eliciting Honest Information From Authors Using Sequential Review. arXiv:2311.14619.

Zhang, Y.; Yu, F.-Y.; Schoenebeck, G.; and Kempe, D. 2022. A System-Level Analysis of Conference Peer Review. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, 1041–1080. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391504.