Catch Me if I Can: Detecting Strategic Behaviour in Peer Assessment

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

Abstract

We consider the issue of strategic behaviour in various peer-assessment tasks, including peer grading of exams or homeworks and peer review in hiring or promotions. When a peer-assessment task is competitive (e.g., when students are graded on a curve), agents 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 agents evaluate a subset of their peers and output rankings that are later aggregated to form a final ordering. In this paper, we investigate a statistical framework for this problem and design a principled test for detecting strategic behaviour. We prove that our test has strong false alarm guarantees and evaluate its 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 this data to run a series of real and semi-synthetic evaluations that reveal a strong detection power of our test.

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 academic 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 [4] and grading students' homeworks [31], and is now expanding to more novel applications of massive open online courses [14, 20] and hiring at freelancing platforms [13].

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 [10, 3, 8]. 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 [1, 2, 15, 12, 33] in 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 from 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 [12, 33] and reduction of efforts put by reviewers [13]. Implementation of such methods also introduces some additional logistical burden on the system designers; as a result, in many critical applications (e.g., NeurIPS 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

Preprint. Under review.

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 work we *focus on detecting strategic manipulations in peer-assessment processes.*

Specifically, in this work we consider a setup in which each reviewer is asked to evaluate a subset of works submitted by their counterparts. Several past studies [3, 10] demonstrated that standard statistical tools for two-sample testing and parametric inference can be used to detect strategic behaviour when evaluations are collected in the form of *ratings*. However, these methods are not applicable to the case when evaluations take the form of *rankings* (e.g., as in the NSF proposal review pilot [9]) due to an important difference between rankings and ratings that we highlight in the sequel. Therefore, in this work we aim to design tools to detect manipulations when evaluations are collected in the form of rankings.

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 attached in the supplementary materials and will be released publicly. 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.

Related work In this paper, we formulate the test for strategic behaviour as a test for independence of rankings returned by reviewers from their own items. Classical statistical works [16] 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 [13]. Second, existing works [18, 7, 11, 22] on two-sample testing with rankings ignore the authorship information that is crucial in our case as we show in the sequel (Section 2.2, Appendix A).

This paper falls in the line of several recent works in computer science on the peer-evaluation process that includes both empirical [30, 23, 13] and theoretical [32, 28, 19] studies. Particularly relevant works are recent papers [30, 27] 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., authored by females), making the methods designed in these works not applicable 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.

2 Problem Formulation

In this section we present our formulation of the manipulation-testing problem.

2.1 Preliminaries

In this paper, 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 $\mathcal{R} = \{1, 2, \ldots, m\}$ and $\mathcal{W} = \{1, 2, \ldots, n\}$ 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 reviewer i 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. $\Pi\left[M(i)\right]$ denotes a set of all $|M(i)|!$ rankings of these works and reviewer i returns a total ranking $\pi_i \in \Pi\left[M(i)\right]$. 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* work. 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 submissions) is employed.

2.2 Problem Setting

Let us now highlight an important difference between rankings and ratings in context of strategic manipulations. The most straightforward approach to aggregate the evaluations in the form of ratings, used explicitly or implicitly in the past works [3, 10], is to order submissions by mean scores they receive. It is not hard to see that in this setting the dominant strategy for each strategic agent is to give the lowest possible score to all submissions assigned to them. In contrast, in Appendix A we show that in a ranking-based setting, *a strategic reviewer must necessarily choose a strategy depending on how their own works compare to the works they review*. Therefore, dealing with rankings is more challenging both for strategic reviewers who no longer have a fixed optimal strategy, and for us, as we do not have a simple pattern of manipulations to look for.

With this intuition, 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 hypothesis (H_0):	$\forall i \in \mathcal{R}$ s.t. $A(i) \neq \emptyset$	$\pi_i \perp A(i)$.
Alternative hypothesis (H_1):	$\exists i \in \mathcal{R}$ s.t. $A(i) \neq \emptyset$	$\pi_i \not\perp A(i)$.

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 which 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[1] 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 in-class peer grading [6, 14] and NSF review of proposals [9].

(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. Note that this assumption is satisfied by various popular models for generation of rankings (under absence of manipulations), including Plackett-Luce model [17, 21] and more general location family random utility models [26]. It also captures models that include agent-dependent parameters such as Thurstone's model [29] provided that parameters of agents are generated independently of their submissions.

Of course, the aforementioned assumptions may be violated in some practical applications. For example, in conference peer review, the reviewer assignment is performed in a manner that maximizes the similarity between papers and reviewers, and hence is not independent of the content of submissions. Moreover, one may expect reviewers who write stronger papers to provide reviews of higher quality than authors of weaker submissions. While the test we design subsequently does not control the false alarm probability in this case, we note below that the output of our test is still meaningful even when these assumptions are violated.

3 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 $\Pi\left[M(i)\right]$ 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 has no work 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 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 1). To understand 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, thereby biasing the test statistic to the negative side.

[1]This assumption can be relaxed (see Appendix C) to allow for assignments of any fixed topology.

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

$$\tau = \sum_{i \in \mathcal{R}} \sum_{j \in A(i)} \left(\Lambda_j(\pi'_1, \pi'_2, \ldots, \pi_i, \ldots, \pi'_m) - \mathbb{E}_{\tilde{\pi} \sim \mathcal{U}_i} \left[\Lambda_j(\pi'_1, \pi'_2, \ldots, \tilde{\pi}, \ldots, \pi'_m) \right] \right), \quad (1)$$

   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 (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

$$\Phi = \left\{ \varphi(A'), A' \in \mathcal{P}(M) \right\} \quad (2)$$

   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.

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, observe 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 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 4) 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 $\{\pi_1^*, \ldots, \pi_{i-1}^*, \pi_{i+1}^* \ldots, \pi_m^*\}$. As we show in Section 4, availability of supervision can significantly aid the detection power of the test.

The following theorem puts together the intuitions we described above and ensures a reliable control over the false alarm probability for our test (a proof is given in Appendix D).

Theorem 1. *Suppose that assumptions (A1) and (A2) specified in Section 2.2 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 4 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 statement of 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 Fisher-Yates shuffling algorithm [5] 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 (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. Importantly, the value of the effect size is meaningful even when the assumptions (A1) and (A2) are violated. Indeed, while in this case we cannot distinguish whether the observed effect is caused by manipulations or is due to some spurious correlations, the large absolute value of the effect size still suggests that some authors *benefit*, while perhaps not engaging in manipulations, from simultaneously being reviewers which potentially indicates unfairness in the system towards the authors who have their work in submission, but do not review.

4 Experimental evaluation

In this section, we empirically evaluate the detection power of our test. We first design a game that incentivizes players to behave strategically and collect a dataset of strategies employed by $N = 55$ students of a large university (name suppressed for double-blind peer review) participated in our experiment. We then evaluate our test in a series of runs on real and semi-synthetic data.

4.1 Data collection

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 strategic behavior and hence many subjects of the hypothetical experiment would behave truthfully, decreasing the efficiency of the process. To overcome this issue, we employ gamification and organize the experiment as follows (game interface is attached in supplementary materials).

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 counterparts (the truthful strategy is to rank by values in decreasing order) and the rankings are aggregated into the final ordering using the Borda count aggregation rule (tied submissions share the position in the final ordering).

	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 1: Detection rates of our test.

For the experiment, we create 5 rounds of the game, sampling a separate authorship matrix A_k and assignment $M_k, k \in \{1, 2, \ldots, 5\}$, for each of the rounds. Each of the $N = 55$ subjects then participates in all 5 rounds, impersonating one (the same for all rounds) of the 20 game players.[2] 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 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 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. Appendix B contains a thorough exploratory analysis of collected data and documents strategies employed by subjects in each of the rounds.

4.2 Evaluation of the test

We now investigate the detection power of our test (Test 1). We begin from analysis of real data 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 (both 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 set of rankings for testing with supervision consists of ground truth rankings.

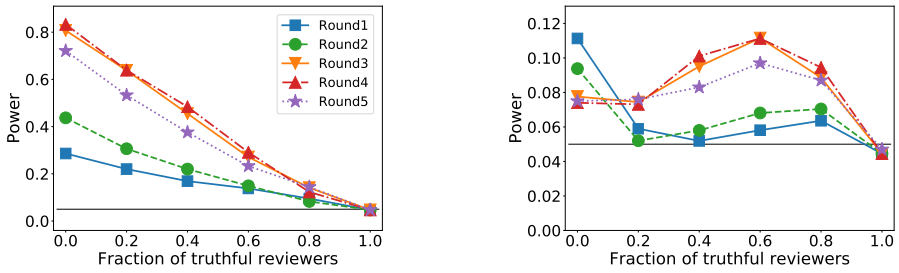
After performing all iterations, for each round we compute the mean detection rate and represent these values in Table 1. The results suggest that our test provided with the impartial set of rankings has a strong detection power, reliably detecting manipulations in the first 4 rounds. On the other hand, 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 last round, 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

[2]We sample a separate authorship matrix for each round so participants get different values between rounds.

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 under such assignments). To remove the impact of the specific assignments we used in the experiment, we now proceed to semi-synthetic trials.

Specifically, we manually annotate the strategies used by participants in each round (see Appendix B and Table 2 therein for the summary of strategies) and create artificial agents who follow these strategies, replicating proportions learned from real data. We then repeat our experiment with 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 1 displays the expected power of our test 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 1: 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 1a 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 1b 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 observation historical 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.

5 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. Our approach is conceptually different from the past literature (which considers ratings) [3, 10] 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 B 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.

Acknowledgments

This work was supported in part by NSF CAREER award 1942124 and in part by NSF CIF 1763734.

Broader Impact

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). It 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 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. Another important note is that one needs to carefully analyze Assumptions (A1) and (A2) 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.

References

[1] Noga Alon, Felix A. Fischer, Ariel D. Procaccia, and Moshe Tennenholtz. Sum of us: Strategyproof selection from the selectors. *CoRR*, abs/0910.4699, 2009.

[2] Haris Aziz, Omer Lev, Nicholas Mattei, Jeffrey S. Rosenschein, and Toby Walsh. Strategyproof peer selection: Mechanisms, analyses, and experiments. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 390–396. AAAI Press, 2016.

[3] S. Balietti, R. Goldstone, and D. Helbing. Peer review and competition in the art exhibition game. *Proceedings of the National Academy of Sciences*, 113(30):8414–8419, 2016.

[4] M.R. Edwards and A.J. Ewen. *360 Degree Feedback: The Powerful New Model for Employee Assessment & Performance Improvement*. AMACOM, 1996.

[5] R. A. Fisher and F. Yates. Statistical tables for biological, agricultural and medical research. *Biometrische Zeitschrift*, 7(2):124–125, 1965.

[6] Scott Freeman and John W. Parks. How accurate is peer grading? *CBE—Life Sciences Education*, 9:482–488, 2010.

[7] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[8] Avinatan Hassidim, Assaf Romm, and Ran I. Shorrer. 'Strategic' behavior in a strategy-proof environment. *SSRN*, pages 686–688, 02 2018.

[9] George A. Hazelrigg. Dear colleague letter: Information to principal investigators (PIs) planning to submit proposals to the Sensors and Sensing Systems (SSS) program October 1, 2013, deadline. 2013. `https://www.semanticscholar.org/paper/Dear-Colleague-Letter%3A-Information-to-Principal-to-Hazelrigg/2a560a95c872164a6316b3200504146ac977a2e6` [Last Retrieved on May 27, 2020.].

[10] Yifei Huang, Matt Shum, Xi Wu, and Jason Zezhong Xiao. Discovery of bias and strategic behavior in crowdsourced performance assessment, 2019.

[11] Y. Jiao and J. Vert. The kendall and mallows kernels for permutations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(7):1755–1769, 2018.

[12] Anson Kahng, Yasmine Kotturi, Chinmay Kulkarni, David Kurokawa, and Ariel Procaccia. Ranking wily people who rank each other, 2018.

[13] Yasmine Kotturi, Anson Kahng, Ariel D. Procaccia, and Chinmay Kulkarni. 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, 2020.

[14] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. Peer and self assessment in massive online classes. *ACM Trans. Comput.-Hum. Interact.*, 20(6), 2013.

[15] David Kurokawa, Omer Lev, Jamie Morgenstern, and Ariel D. Procaccia. Impartial peer review. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 582–588. AAAI Press, 2015.

[16] E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

[17] R. Duncan Luce. *Individual Choice Behavior: A Theoretical analysis*. Wiley, New York, NY, USA, 1959.

[18] Horia Mania, Aaditya Ramdas, Martin J. Wainwright, Michael I. Jordan, and Benjamin Recht. On kernel methods for covariates that are rankings. *Electron. J. Statist.*, 12(2):2537–2577, 2018.

[19] Ritesh Noothigattu, Nihar Shah, and Ariel Procaccia. Choosing how to choose papers. *arXiv preprint arxiv:1808.09057*, 2018.

[20] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. Tuned models of peer assessment in moocs. *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)*, 07 2013.

[21] R. L. Plackett. The Analysis of Permutations. *Journal of the Royal Statistical Society Series C*, 24(2):193–202, June 1975.

[22] Charvi Rastogi, Nihar Shah, Sivaraman Balakrishnan, and Aarti Singh. Two-sample testing with pairwise comparison data and the role of modeling assumptions. In *IEEE International Symposium on Information Theory*. 2020.

[23] Mehdi S.M. Sajjadi, Morteza Alamgir, and Ulrike von Luxburg. 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, 2016. ACM.

[24] Nihar B. Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramch, ran, and Martin J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58):1–47, 2016.

[25] Nihar B. Shah, Joseph K. Bradley, Abhay Parekh, and Kannan Ramchandran. A case for ordinal peer-evaluation in moocs. 2013.

[26] Hossein Azari Soufiani, David C. Parkes, and Lirong Xia. 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, 2012. Curran Associates Inc.

[27] Ivan Stelmakh, Nihar Shah, and Aarti Singh. On testing for biases in peer review. In *Advances in Neural Information Processing Systems 32*, pages 5286–5296. 2019.

[28] Ivan Stelmakh, Nihar B. Shah, and Aarti Singh. PeerReview4All: Fair and accurate reviewer assignment in peer review. *arXiv preprint arXiv:1806.06237*, 2018.

[29] Louis Leon Thurstone. A law of comparative judgement. *Psychological Review*, 34:278–286, 1927.

[30] Andrew Tomkins, Min Zhang, and William D. Heavlin. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713, 2017.

[31] Keith Topping. Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3):249–276, 1998.

[32] Jingyan Wang and Nihar B. Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. *CoRR*, abs/1806.05085, 2018.

[33] Yichong Xu, Han Zhao, Xiaofei Shi, Jeremy Zhang, and Nihar Shah. On strategyproof conference review. *Arxiv preprint*, 2019.

Appendix

We provide supplementary materials and additional discussion. In Appendix A we highlight the difference between rankings and ratings in context of strategic manipulations. Appendix B is dedicated to details of the experiment and additional simulations. We show how to slightly relax Assumption (A1) of random assignment in Appendix C and prove Theorem 1 in Appendix D.

A Motivating Example

Let us consider the experiment conducted by Balietti et al. [3] 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.

In contrast, when reviewers are asked to output *rankings* of submissions, the situation is different. 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 also 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,[3] (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.

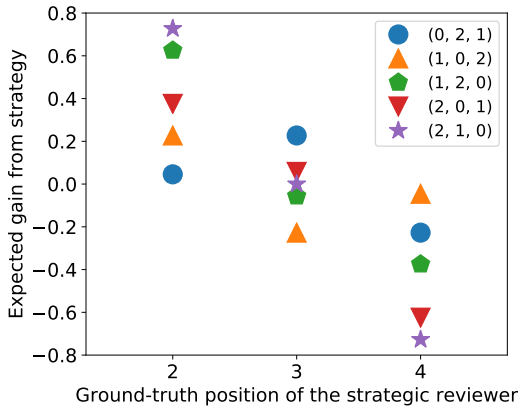


Figure 2: Comparison of fixed deterministic strategies available to a single strategic reviewer depending on position of their work in the true underlying ranking.

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 2 represents an expected gain of each strategy as compared to the truthful strategy for positions 2–4 of the work authored by reviewer i^* in the ground-truth ranking, where the expectation is taken over randomness in the assignment. 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 work compares to the works they rank in order to improve the outcome of their submission.

[3] We use the variant without tie-breaking — tied submissions share the same position in the final ordering.

B Details of the Experimental Evaluation

In this section we provide exploratory analysis of collected data (Section B.1) and evaluate our test in additional simulations (Section B.2).

B.1 Exploratory Data Analysis

We begin from summarizing some general analysis of the collected data. 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.
- `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 $\pi = 12 \succ 16 \succ 7 \succ 2$.
- `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 3 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 3. 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.

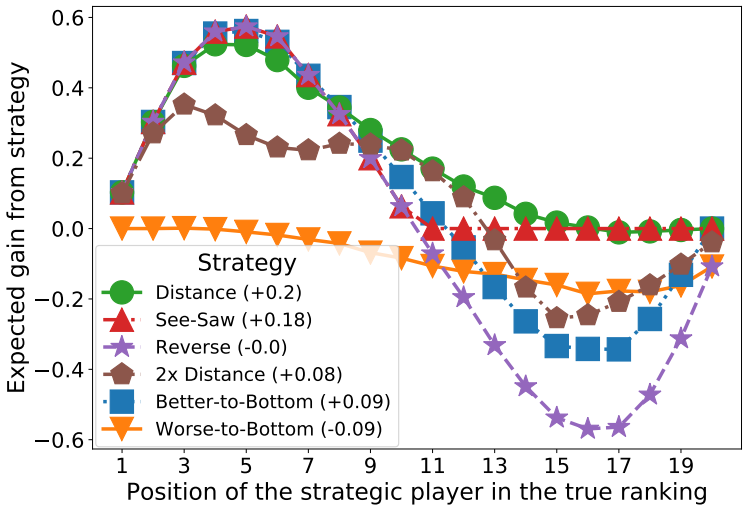


Figure 3: Expected gain from manipulation strategy when all but one player are truthful as a function of position of the strategic player in in the ground-truth ranking. The expectation is taken over randomness of the assignment procedure and values in brackets in the legend correspond to the mean gain over all positions. Borda count aggregation rule is used. Positive value of the gain indicates that the manipulation strategy in expectation delivers better position in the aggregated ordering than the truthful strategy. Error bars are too small to be visible.

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 2 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 one 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.

B.2 Additional Evaluations

We now provide additional evaluation of our test and conduct simulations in the following settings:

- **Detecting pure strategies.** In Section B.2.1 we evaluate the detection power of the test against each of the strategies we learned from data.
- **Noisy supervision.** Next, in Section B.2.2 we evaluate robustness of our test to the noise in the optional impartial rankings $\{\pi_i^*, i \in \mathcal{R}\}$.
- **Violation of Assumption (A2).** We then study the behavior of our test when Assumption (A2) is violated. To this end, in Section B.2.3 we use the model that connects the quality of reviewer's submission to noise in their evaluations suggested by the empirical research on peer grading and compute the false alarm rate of our test under this model.

	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 2: 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.

- **Runtime of the test.** In this paper, 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 in Section B.2.4 we provide runtimes of our naive implementation of the test for larger values of parameters.

B.2.1 Detecting Pure Strategies

To compute the power against specific strategies, we follow the approach we used to build Figure 1, that is, for each fraction of truthful agents and for each strategy used by manipulating agents, we compute the detection power of the test with and without supervision over 1000 assignments sampled uniformly at random from the set of all assignments valid for parameters:

$$A = C = I,\ n = m = 20,\ \lambda = \mu = 4. \tag{3}$$

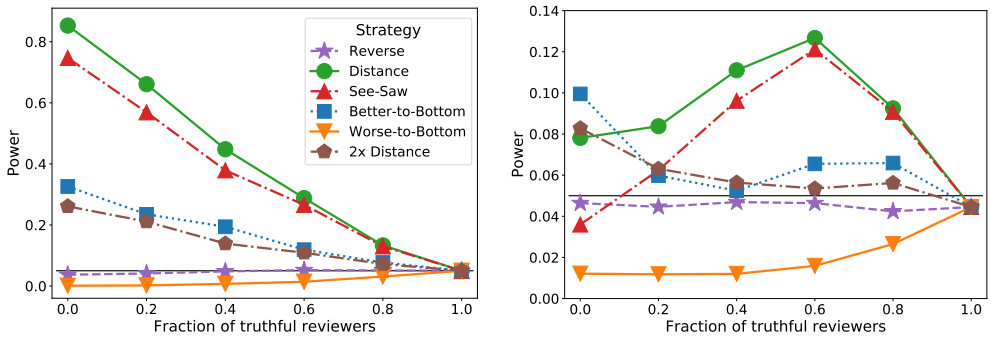
Figure 4 compares detection power of the test against each strategy used by participants of the experiment. Recall that our test aims at detecting manipulations that improve the final standing of reviewer. As shown in Figure 3, the `Reverse` and `Worse-to-Bottom` strategies do not improve the final standing 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 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.

B.2.2 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 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 $\sigma = 6$ two players with values differing by 4 points are swapped in the impartial ranking with probability $p \approx 0.32$. 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.



(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.

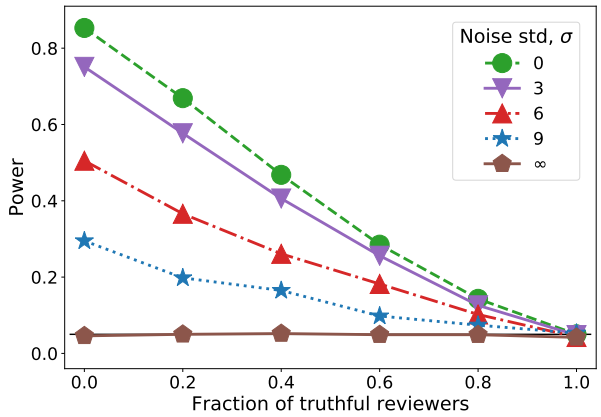


Figure 5: Detection power of our test with supervision for various levels of noise in the impartial rankings.

B.2.3 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 [20, 25] suggests that Assumption (A2) 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.

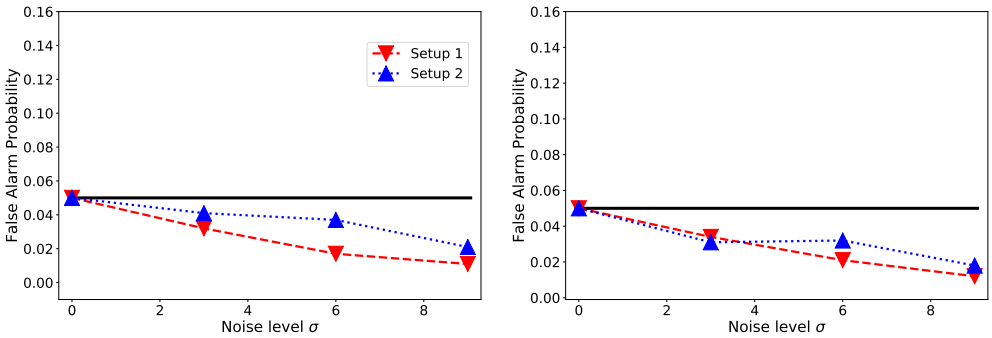
Intuitively and informally, Figure 3 suggests that authors of stronger works benefit from reversing the ranking of submissions assigned to them whereas authors of weaker submissions should play truthfully to maximize the outcome of their submission. In contrast, the aforementioned relationship between the quality of submission and 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 students. Hence, under this model the noise in evaluations of both strong and weak students hurts the final standing of their submission as compared to the constant-noise case, making our test more conservative as it aims at detecting successful manipulations.

To validate this intuition, we consider the problem parameters given in (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 equal to the negated positions of the works in the underlying ground-truth ordering (the best work has quality -1, the second best has

quality -2 and so on). We then simulate the false alarm probability of our test under two setups with different definition 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.

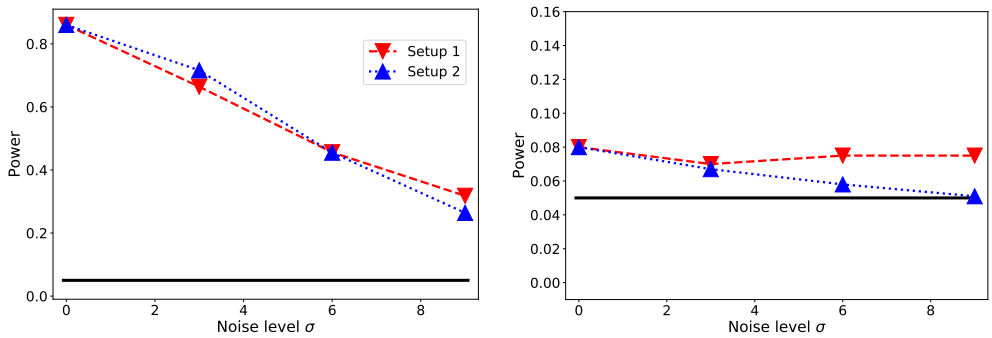
Observe that in both setups data is generated under the null hypothesis of absence of manipulations. In simulations, we vary the noise level σ and sample impartial rankings from the random utility model with noise level ${}^{\sigma}/_{2}$. Figure 6 depicts the false alarm probability of our test both 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.



(a) With supervision (b) Without supervision

Figure 6: 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.

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 7 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 the sample from the corresponding random utility models.



(a) With supervision (b) Without supervision

Figure 7: 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.

B.2.4 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 8. We conclude that the running time of naive implementation of our test is feasible even for instances with thousands of reviewers and submissions.

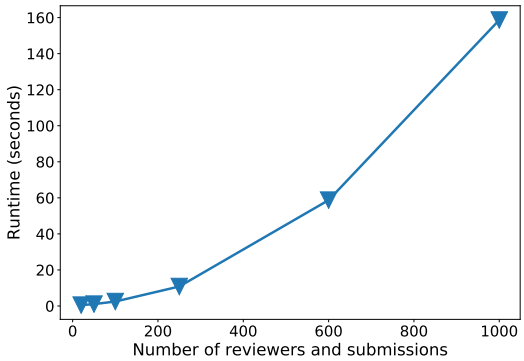


Figure 8: Running time of our test.

C 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 [24]. 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 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.

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 (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.

D 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:

$$\begin{aligned} C &= LC^*R \\ A &= LA^*R \end{aligned} \tag{4}$$

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:

$$\mathbb{P}\left[(\widetilde{C}, \widetilde{A}) = (C_1, A_1) \middle| M\right] = \mathbb{P}\left[(\widetilde{C}, \widetilde{A}) = (C_2, A_2) \middle| M\right], \tag{5}$$

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 (4) the number of valid assignments is the same by symmetry.

Equation (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:

$$\Phi = \left\{ \varphi(A') |\ A' \in \mathcal{P}(M) \right\},$$

where φ is defined as

$$\varphi(A') = \sum_{i \in \mathcal{R}} \sum_{j \in A'(i)} \left(\Lambda_j(\pi'_1, \pi'_2, \ldots, \pi_i, \ldots, \pi'_m) - \mathbb{E}_{\tilde{\pi} \sim \mathcal{U}_i} \left[\Lambda_j(\pi'_1, \pi'_2, \ldots, \tilde{\pi}, \ldots, \pi'_m) \right] \right).$$

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 $\left(\lfloor \alpha |\Phi| \rfloor + 1 \right)$, we conclude the proof.