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 (Edwards and Ewen 1996) and grading students' homeworks (Topping 1998), and is now expanding to more novel applications of massive open online courses (Kulkarni et al. 2013; Piech et al. 2013) and hiring at freelancing platforms (Kotturi et al. 2020).

The downside of such methods, however, is that reviewers are incentivized to evaluate their counterparts strategically to ensure a better outcome of their own item (Huang et al. 2019; Balietti, Goldstone, and Helbing 2016; Hassidim, Romm, and Shorrer 2018). Deviations from the truthful behaviour decrease the overall quality of the resulted ranking and undermine fairness of the process. This issue has led to a long line of work (Alon et al. 2009; Aziz et al. 2016; Kurokawa et al. 2015; Kahng et al. 2018; Xu et al. 2019) on designing "impartial" aggregation rules that can eliminate the impact of the ranking returned by a reviewer on the final position of their item.

While impartial methods remove the benefits of manipulations, such robustness may come at the cost of some accuracy loss when reviewers do not engage in strategic behaviour. This loss is caused by less efficient data usage (Kahng et al. 2018; Xu et al. 2019) and reduction of efforts put by reviewers (Kotturi et al. 2020). Implementation of such methods also introduces some additional logistical burden on the system designers; as a result, in many critical applications (e.g., conference peer review) the non-impartial mechanisms are employed. An important barrier that prevents stakeholders from making an informed choice to implement an impartial mechanism is a lack of tools to detect strategic behaviour. Indeed, to evaluate the trade off between the loss of accuracy due to manipulations and the loss of accuracy due to impartiality, one needs to be able to evaluate the extent of strategic behaviour in the system. With this motivation, in this work we *focus on detecting strategic manipulations in peer-assessment processes.*

Specifically, in this work we consider a setting where each reviewer is asked to evaluate a subset of works submitted by their counterparts. In a carefully designed randomized study of strategic behaviour when evaluations take the form of *ratings*, Balietti, Goldstone, and Helbing (2016) were able to detect manipulations by comparing the distribution of scores given by target reviewers to some truthful reference. However, other works (Huang et al. 2019; Barroga 2014) suggest that in more practical settings reviewers may strategically decrease some scores and increase others in attempt to mask their manipulations or intentionally promote weaker submissions, thereby keeping the distribution of output scores unchanged and making the distribution-based detection inapplicable. Inspired by this observation, we aim to design tools to detect manipulations when the distributions of scores output by reviewers are fixed, that is, we assume that evaluations are collected in the form of *rankings*. Ranking-based

evaluation is used in practice (Hazelrigg 2013) and has some theoretical properties that make it appealing for peer grading (Shah et al. 2013; Caragiannis, Krimpas, and Voudouris 2014) which provides additional motivation for our work.

Contributions In this work we present two sets of results.

- **Theoretical.** First, we propose a non-parametric test for detection of strategic manipulations in peer-assessment setup with rankings. Second, we prove that our test has a reliable control over the false alarm probability (probability of claiming existence of the effect when there is none). Conceptually, we avoid difficulties associated to dealing with rankings as covariates by carefully accounting for the fact that each reviewer is "connected" to their submission(s); therefore, the manipulation they employ is naturally not an arbitrary deviation from the truthful strategy, but instead the deviation that potentially improves the outcome of their works.
- **Empirical.** On the empirical front, we first design and conduct an experiment that incentivizes strategic behaviour of participants. This experiment yields a novel dataset of patterns of strategic behaviour that can be useful for other researchers (the dataset is attached in supplementary materials)[1]. 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 test has a non-trivial detection power, while not making strong assumptions on the manipulating strategies.

Related work Despite motivation for this work comes from studies of Balietti, Goldstone, and Helbing (2016) and Huang et al. (2019), an important difference between rankings and ratings that we highlight in Section 2.2 makes the models considered in these works inapplicable to our setup. Several other papers (Thurner and Hanel 2011; Cabotà, Grimaldo, and Squazzoni 2013; Paolucci and Grimaldo 2014) specialize on the problem of strategic behaviour in peer review and perform simulations to explore its detrimental impact on the quality of published works. These works are orthogonal to the present paper because they do not aim to detect the manipulations.

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 (Lehmann and Romano 2005) for independence testing are not directly applicable to this problem due to the absence of low-dimensional representations of items. To avoid dealing with unstructured items, one could alternatively formulate the problem as a two-sample test and obtain a control sample of rankings from non-strategic reviewers. This approach, however, has two limitations. First, past work suggests that the test and control rankings may have different distributions even under the absence of manipulations due to misalignment of incentives (Kotturi et al. 2020). Second, existing works (Mania et al. 2018; Gretton et al. 2012; Jiao and Vert 2018; Rastogi et al. 2020) on two-sample testing with rankings ignore the authorship information that is crucial in our case as we show in the sequel (Section 2.2).

This paper also falls in the line of several recent works in computer science on the peer-evaluation process that includes both empirical (Tomkins, Zhang, and Heavlin 2017; Sajjadi, Alamgir, and von Luxburg 2016; Kotturi et al. 2020) and theoretical (Wang and Shah 2018; Stelmakh, Shah, and Singh 2018; Noothigattu, Shah, and Procaccia 2018) studies. Particularly relevant works are recent papers (Tomkins, Zhang, and Heavlin 2017; Stelmakh, Shah, and Singh 2019) that consider the problem of detecting biases (e.g., gender bias) in single-blind peer review. Biases studied therein manifest in reviewers being harsher to some subset of submissions (e.g., 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.

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 $\mathcal{R} = \{1, 2, \dots, m\}$ and $\mathcal{W} = \{1, 2, \dots, n\}$ denote the set of reviewers and works submitted for review, respectively. We let matrix $C \in \{0,1\}^{m \times n}$ represent conflicts of interests between reviewers and submissions, that is, $(i,j)^{\text{th}}$ entry of C equals 1 if reviewer i is in conflict with work j and 0 otherwise. Matrix C captures all kinds of conflicts of interest, including authorship, affiliation and others, and many of them can be irrelevant from the manipulation standpoint (e.g., affiliation may put a reviewer at conflict with dozens of submissions they are not even aware of). We use $A \in \{0,1\}^{m \times n}$ to denote a subset of "relevant" conflicts — those that reviewers may be incentivized to manipulate for — identified by stakeholders. For the ease of presentation, we assume that A represents the authorship conflicts, as reviewers are naturally interested in improving the final standing of their own works, but in general it can capture any subset of conflicts. For each reviewer $i \in \mathcal{R}$, non-zero entries of the corresponding row of matrix A indicate submissions that are (co-)authored by reviewer i. We let $C(i)$ and $A(i) \subseteq C(i)$ denote possibly empty sets of works conflicted with and authored by reviewer i, respectively.

Each work submitted for review is assigned to λ non-conflicting reviewers subject to a constraint that each reviewer gets assigned μ works. For brevity, we assume that parameters n, m, μ, λ are such that $n\lambda = m\mu$ so we can assign exactly μ works to each reviewer. The assignment is represented by a binary matrix $M \in \{0,1\}^{m \times n}$ whose $(i,j)^{\text{th}}$ entry equals 1 if reviewer i is assigned to work j and 0 otherwise. We call an assignment valid if it respects the (submission, reviewer)-loads and does not assign a reviewer to a conflicting work. Given a valid assignment M

[1] Supplementary materials and appendices are on the first author's website.

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 $|M(i)|!$ total rankings of these works and reviewer i returns a 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* works. The focus of this work is on designing tools to detect strategic behaviour of reviewers when a non-impartial aggregation rule Λ (e.g., a rule that theoretically allows reviewers to impact the final standing of their own submissions) is employed.

2.2 Motivating Example

To set the stage, we start from highlighting an important difference between rankings and ratings in the peer-assessment setup. To this end, let us consider the experiment conducted by Balietti, Goldstone, and Helbing (2016) in which reviewers are asked to give a score to each work assigned to them for review and the final ranking is computed based on the mean score received by each submission. It is not hard to see that in their setting, the dominant strategy for each rational reviewer who wants to maximize the positions of their own works in the final ranking is to give the lowest possible score to all submissions assigned to them. Observe that this strategy is fixed, that is, it does not depend on the quality of reviewer's work — irrespective of position of their work in the underlying ordering, each reviewer benefits from assigning the lowest score to all submissions they review.

Similarly, Huang et al. (2019) in their work also operate with ratings and consider a fixed model of manipulations in which strategic agents increase the scores of low-quality submissions and decrease the scores of high-quality submissions, irrespective of the quality of reviewers' works.

In contrast, when reviewers are asked to output *rankings* of submissions, the situation is different and reviewers can no longer rely on fixed strategies to gain the most for their own submission. To highlight this difference, let us consider a toy example of the problem with 5 reviewers and 5 submissions ($m = n = 5$), authorship and conflict matrix given by an identity matrix ($C = A = I$), and three works (reviewers) assigned to each reviewer (work), that is, $\lambda = \mu = 3$. In this example, we additionally assume that: (i) assignment of reviewers to works is selected uniformly at random from the set of all valid assignments, (ii) aggregation rule Λ is the Borda count, that is, the positional scoring rule with weights equal to positions in the ranking,[2] (iii) reviewers are able to reconstruct the ground-truth ranking of submissions assigned to them without noise, and (iv) all but one reviewers are truthful.

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

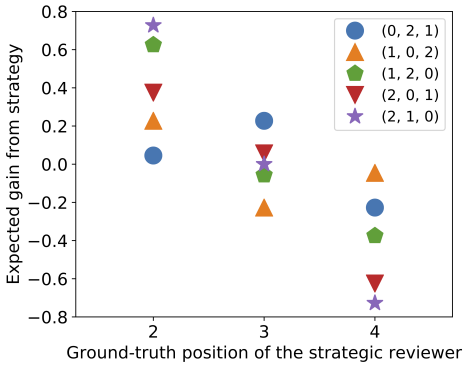


Figure 1: Comparison of fixed deterministic strategies available to a single strategic reviewer depending on position of their work in the true underlying ranking.

Under this simple formulation, we qualitatively analyze the strategies available to the strategic reviewer, say reviewer i^*. Specifically, following the rating setup, we consider the fixed deterministic strategies that do not depend on the work created by reviewer i^*. Such strategies are limited to permutations of the ground-truth ranking of submissions in $M(i^*)$. Figure 1 represents an expected gain of each strategy 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.

2.3 Problem Setting

With the motivation given in Section 2.2, we are ready to present the formal hypothesis-testing problem we consider in this work. When deciding on how to rank the works, the information available to reviewers is the content of the works they review and the content of their own works. Observe that while a truthful reviewer does not take into account their own submissions when ranking works of others, the aforementioned intuition suggests that the ranking output by a strategic agent should depend on their own works. Our formulation of the test for manipulations as an independence test captures this motivation.

Problem 1 (Testing for strategic behaviour)**.** Given a non-impartial aggregation rule Λ, assignment of works to reviewers M, rankings returned by reviewers $\{\pi_i, i \in \mathcal{R}\}$, conflict matrix C, authorship matrix A and set of works submitted for review $\mathcal{W}$, the goal is to test the following hypotheses:

Null (H_0) : $\quad \forall i \in \mathcal{R}$ s.t. $A(i) \neq \emptyset \qquad \pi_i \perp A(i)$.

Alternative (H_1) : $\quad \exists i \in \mathcal{R}$ s.t. $A(i) \neq \emptyset \qquad \pi_i \not\perp A(i)$.

In words, under the null hypothesis reviewers who have their submissions under review do not take into account their own works when evaluating works of others and hence are

not engaged in manipulations that can improve the outcome of their own submissions. In contrast, under the alternative hypothesis some reviewers choose the ranking depending on how their own works compare to works they rank, suggesting that they are engaged in manipulations.

Assumptions Our formulation of the testing problem makes two assumptions about the data-generation process to ensure that association between works authored by reviewer i and ranking π_i may be caused only by strategic manipulations and not by some intermediate mediator variables.

(A1) **Random assignment.** We assume that the assignment of works to reviewers is selected uniformly at random from the set of all assignments that respect the conflict matrix C. This assumption ensures that the works authored by a reviewer do not impact the set of works assigned to them for review. The assumption of random assignment holds in many applications, including peer grading (Freeman and Parks 2010; Kulkarni et al. 2013) and NSF review of proposals (Hazelrigg 2013).

(A2) **Independence of ranking noise.** We assume that under the null hypothesis of absence of strategic behaviour, the reviewer identity is independent of the works they author, that is, the noise in reviewers' evaluations (e.g., the noise due to subjectivity of the truthful opinion) is not correlated with their submissions. This assumption is satisfied by various popular models for generation of rankings, including Plackett-Luce model (Luce 1959; Plackett 1975) and more general location family random utility models (Soufiani, Parkes, and Xia 2012).

Of course, the aforementioned assumptions may be violated in some practical applications.[3] 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. 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.

[3] Assumption (A1) can be relaxed (Appendix B) to allow for assignments of any fixed topology. We examine the behaviour of our test under realistic violation of Assumption (A2) in Appendix A.2

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 get the motivation behind this choice of the test statistic, note that if a reviewer i is truthful then the ranking they return may be either better or worse for *their own* submissions than a random ranking, depending on how their submissions compare to works they review. In contrast, a strategic reviewer may choose the ranking that delivers a better final standing for their submissions, biasing the test statistic to the negative side.

Having defined the test statistic, we now understand its behaviour under the null hypothesis to quantify when its value is too large to be observed under the absence of manipulations for a given significance level α. To this end, we note that for a given assignment matrix M, there are many pairs of conflict and authorship matrices (C', A') that (i) are equal to the actual matrices C and A up to permutations of rows and columns and (ii) do not violate the assignment M, that is, do not declare a conflict between any pair of reviewer i and submission j such that submission j is assigned to reviewer i in M. Next, 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

Test 1 Test for strategic behaviour

Input: Reviewers' rankings $\{\pi_i, i \in \mathcal{R}\}$
Assignment M of works to reviewers
Conflict and authorship matrices (C, A)
Significance level α, aggregation rule Λ
Optional Argument: Impartial rankings $\{\pi^*_i, i \in \mathcal{R}\}$

1. Compute the test statistic τ as

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

   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

$$\Phi = \big\{ \varphi(A'), A' \in \mathcal{P}(M) \big\} \tag{2}$$

   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.

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 combines the above intuitions and ensures a reliable control over the false alarm probability for our test (a proof is given in Appendix C).

Theorem 1. *Suppose that assumptions (A1) and (A2) specified in Section 2.3 hold. Then, under the null hypothesis of absence of manipulations, for any significance level $\alpha \in (0, 1)$ and for any aggregation rule Λ, Test 1 (both with and without supervision) is guaranteed to reject the null with probability at most α. Therefore, Test 1 controls the false alarm probability at the level α.*

Remark. 1. In Section 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 theorem holds if instead of using the full multiset $\mathcal{P}(M)$, when defining Φ, we only sample some k authorship matrices uniformly at random from the multiset $\mathcal{P}(M)$. The value of k should be chosen large enough to ensure that $(\lfloor \alpha |\Phi| \rfloor + 1)$ is greater than 1. The sampling can be performed by generating random permutations using the shuffling algorithm of Fisher and Yates (1965) 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 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 $N = 55$ attendees of a graduate-level AI course at Carnegie Mellon University who 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 cheating and hence many subjects of the hypothetical experiment would behave truthfully, reducing the efficiency of the process. To overcome this issue, we use gamification and organize the experiment as follows (game interface 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 $v \in \{1, 2, \dots, 20\}$ and this value is privately communicated to the respective player; therefore, players are associated to values and in the sequel we do not distinguish between a player's value and their "submission". We then communicate values of some $\mu = 4$ other contestants to each player subject to the constraint that a value of each player becomes known to $\lambda = 4$ counterparts. To do so, we sample an assignment M from the set of assignments respecting the conflict matrix $C = A$ uniformly at random. Note that players do not get to see the full assignment and only observe the values of players assigned to them. The rest of the game replicates the peer grading setup: participants are asked to rank their peers (the truthful strategy is to rank by values in decreasing order) and the rankings are aggregated using the Borda count aggregation rule (tied submissions share the position in the final ordering).

For the experiment, we create 5 rounds of the game, sampling a separate authorship matrix A_k and assignment $M_k, k \in \{1, 2, \dots, 5\}$, 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.[4] 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 A contains a thorough exploratory analysis of collected data, documents strategies employed by subjects and has evaluations of the test in addition to those we perform in the next section.

4.2 Evaluation of the Test

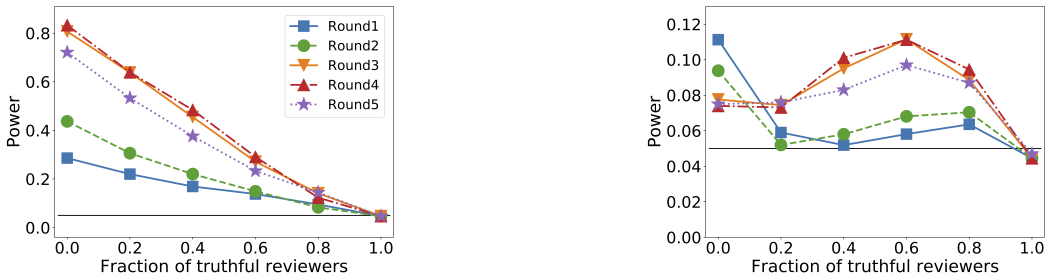
We now investigate the detection power of our test (Test 1). We begin from analysis of real data and execute the following procedure. For each of the 1,000 iterations, we uniformly at random subset 20 out of the 55 participants such that together they impersonate all 20 game players. We then apply our test (with and without supervision) to rankings output by these participants in each of the 5 rounds, setting significance level at $\alpha = 0.05$ and sampling $k = 100$ authorship matrices in Step 3 of the test. The impartial rankings for testing with supervision comprise ground truth rankings.

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 round 5, we will show in a moment that poor performance of our test appears to be due to random chance (i.e., the choice of the assignment which

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

	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.



(a) With supervision

(b) Without supervision

Figure 2: 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.

is hard for detection) and not due to any systematic issue.

Note that performance of our test depends not only on the strategies employed by players, but also on the assignment M realized in a particular round. Some realizations of random assignment make successful manipulations (and their detection) easier while under other realizations most of the players cannot improve their position even if they use the best strategy (and therefore our test cannot detect manipulations). To remove the impact of the specific assignments we used in the experiment, we now proceed to semi-synthetic trials. Specifically, we manually annotate the strategies used by participants in each round (see Appendix A 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 $m = 20$ artificial agents, simulating 1,000 assignments for each round of the game and computing the expectation of the power of our test over randomness of the assignment. Additionally, we enhance the set of synthetic agents with truthful agents and study how the detection power of our test changes with the fraction of truthful agents. Figure 2 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 $\alpha = 0.05$.

Figure 2a 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 2b shows that the test without supervision has considerably lower (but still non-trivial) power. We note, however, that the main feature of the test without supervision is that it can be readily applied to purely observational data and the power can be accumulated over multiple datasets (e.g., it can be applied to multiple iterations of a university course). An interesting feature of the test without supervision is the non-monotonicity of power with respect to the fraction of truthful reviewers, caused by an 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 (Balietti, Goldstone, and Helbing 2016; Huang et al. 2019) as it does not assume any specific parametric model of manipulations and instead aims at detecting any *successful* manipulation of rankings, thereby giving flexibility of non-parametric tests. This flexibility, however, does not extend to the case when agents try to manipulate but do it *unsuccessfully* (see Appendix A 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

The human subject experiment introduced in this paper was approved by Carnegie Mellon University Institutional Review Board.

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

Ethics Statement

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

Alon, N.; Fischer, F. A.; Procaccia, A. D.; and Tennenholtz, M. 2009. Sum of Us: Strategyproof Selection from the Selectors. *CoRR* abs/0910.4699. URL http://arxiv.org/abs/0910.4699.

Aziz, H.; Lev, O.; Mattei, N.; Rosenschein, J. S.; and Walsh, T. 2016. Strategyproof Peer Selection: Mechanisms, Analyses, and Experiments. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, 390–396. AAAI Press.

Balietti, S.; Goldstone, R.; and Helbing, D. 2016. Peer review and competition in the Art Exhibition Game. *Proceedings of the National Academy of Sciences* 113(30): 8414–8419.

Barroga, E. F. 2014. Safeguarding the integrity of science communication by restraining 'rational cheating'in peer review. *Journal of Korean medical science* 29(11): 1450–1452.

Cabotà, J. B.; Grimaldo, F.; and Squazzoni, F. 2013. When Competition Is Pushed Too Hard. An Agent-Based Model Of Strategic Behaviour Of Referees In Peer Review. In *ECMS*, 881–887.

Caragiannis, I.; Krimpas, G. A.; and Voudouris, A. A. 2014. Aggregating partial rankings with applications to peer grading in massive online open courses. *CoRR* abs/1411.4619. URL http://arxiv.org/abs/1411.4619.

Edwards, M.; and Ewen, A. 1996. *360 Degree Feedback: The Powerful New Model for Employee Assessment & Performance Improvement*. AMACOM. ISBN 9780814403266.

Fisher, R. A.; and Yates, F. 1965. Statistical tables for biological, agricultural and medical research. *Biometrische Zeitschrift* 7(2): 124–125. doi:10.1002/bimj.19650070219.

Freeman, S.; and Parks, J. W. 2010. How Accurate Is Peer Grading? *CBE—Life Sciences Education* 9: 482–488.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A Kernel Two-Sample Test. *Journal of Machine Learning Research* 13(25): 723–773. URL http://jmlr.org/papers/v13/gretton12a.html.

Hassidim, A.; Romm, A.; and Shorrer, R. I. 2018. 'Strategic' Behavior in a Strategy-Proof Environment. *SSRN* 686–688. doi:10.2139/ssrn.2784659.

Hazelrigg, G. A. 2013. Dear Colleague Letter: Information to Principal Investigators (PIs) Planning to Submit Proposals to the Sensors and Sensing Systems (SSS) Program October 1, 2013, Deadline. https://www.semanticscholar.org/paper/Dear-Colleague-Letter\%3A-Information-to-Principal-to-Hazelrigg/2a560a95c872164a6316b3200504146ac977a2e6 [Last Retrieved on May 27, 2020.].

Huang, Y.; Shum, M.; Wu, X.; and Xiao, J. Z. 2019. Discovery of Bias and Strategic Behavior in Crowdsourced Performance Assessment. *arXiv preprint arXiv:1908.01718* .

Jiao, Y.; and Vert, J. 2018. The Kendall and Mallows Kernels for Permutations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(7): 1755–1769.

Kahng, A.; Kotturi, Y.; Kulkarni, C.; Kurokawa, D.; and Procaccia, A. 2018. Ranking Wily People Who Rank Each Other. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI'18. AAAI Press.

Kotturi, Y.; Kahng, A.; Procaccia, A. D.; and Kulkarni, C. 2020. HirePeer: Impartial Peer-Assessed Hiring at Scale in Expert Crowdsourcing Markets. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'20. AAAI Press.

Kulkarni, C.; Wei, K. P.; Le, H.; Chia, D.; Papadopoulos, K.; Cheng, J.; Koller, D.; and Klemmer, S. R. 2013. Peer and Self Assessment in Massive Online Classes. *ACM Trans. Comput.-Hum. Interact.* 20(6). ISSN 1073-0516. doi:10.1145/2505057. URL https://doi.org/10.1145/2505057.

Kurokawa, D.; Lev, O.; Morgenstern, J.; and Procaccia, A. D. 2015. Impartial Peer Review. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, 582–588. AAAI Press. ISBN 9781577357384.

Lehmann, E. L.; and Romano, J. P. 2005. *Testing statistical hypotheses*. Springer Texts in Statistics. New York: Springer, third edition. ISBN 0-387-98864-5.

Luce, R. D. 1959. *Individual Choice Behavior: A Theoretical analysis*. New York, NY, USA: Wiley.

Mania, H.; Ramdas, A.; Wainwright, M. J.; Jordan, M. I.; and Recht, B. 2018. On kernel methods for covariates that are rankings. *Electron. J. Statist.* 12(2): 2537–2577. doi:10.1214/18-EJS1437. URL https://doi.org/10.1214/18-EJS1437.

Noothigattu, R.; Shah, N.; and Procaccia, A. 2018. Choosing how to choose papers. *arXiv preprint arxiv:1808.09057* .

Paolucci, M.; and Grimaldo, F. 2014. Mechanism change in a simulation of peer review: from junk support to elitism. *Scientometrics* 99(3): 663–688.

Piech, C.; Huang, J.; Chen, Z.; Do, C.; Ng, A.; and Koller, D. 2013. Tuned models of peer assessment in MOOCs. *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)* .

Plackett, R. L. 1975. The Analysis of Permutations. *Journal of the Royal Statistical Society Series C* 24(2): 193–202. doi:10.2307/2346567.

Rastogi, C.; Shah, N.; Balakrishnan, S.; and Singh, A. 2020. Two-Sample Testing with Pairwise Comparison Data and the Role of Modeling Assumptions. In *IEEE International Symposium on Information Theory*.

Sajjadi, M. S.; Alamgir, M.; and von Luxburg, U. 2016. Peer Grading in a Course on Algorithms and Data Structures: Machine Learning Algorithms Do Not Improve over Simple Baselines. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, L@S '16, 369–378. New York, NY, USA: ACM. ISBN 978-1-4503-3726-7. doi:10.1145/2876034.2876036. URL http://doi.acm.org/10.1145/2876034.2876036.

Shah, N. B.; Bradley, J. K.; Parekh, A.; and Ramchandran, K. 2013. A Case for Ordinal Peer-evaluation in MOOCs http://www.cs.cmu.edu/~jkbradle/papers/shahetal.pdf [Last Retrieved on May 27, 2020.].

Soufiani, H. A.; Parkes, D. C.; and Xia, L. 2012. Random Utility Theory for Social Choice. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, 126–134. Red Hook, NY, USA: Curran Associates Inc.

Stelmakh, I.; Shah, N.; and Singh, A. 2019. On Testing for Biases in Peer Review. In *Advances in Neural Information Processing Systems 32*, 5286–5296.

Stelmakh, I.; Shah, N. B.; and Singh, A. 2018. PeerReview4All: Fair and Accurate Reviewer Assignment in Peer Review. *arXiv preprint arXiv:1806.06237* .

Thurner, S.; and Hanel, R. 2011. Peer-review in a world with rational scientists: Toward selection of the average. *The European Physical Journal B* 84(4): 707–711.

Tomkins, A.; Zhang, M.; and Heavlin, W. D. 2017. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences* 114(48): 12708–12713. ISSN 0027-8424. doi:10.1073/pnas.1707323114. URL https://www.pnas.org/content/114/48/12708.

Topping, K. 1998. Peer Assessment between Students in Colleges and Universities. *Review of Educational Research* 68(3): 249–276. ISSN 00346543, 19351046. URL http://www.jstor.org/stable/1170598.

Wang, J.; and Shah, N. B. 2018. Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings. *CoRR* abs/1806.05085. URL http://arxiv.org/abs/1806.05085.

Xu, Y.; Zhao, H.; Shi, X.; Zhang, J.; and Shah, N. 2019. On Strategyproof Conference Review. *arXiv preprint arXiv:1806.06266* .