

# TRADEOFFS IN PREVENTING MANIPULATION IN PAPER BIDDING FOR REVIEWER ASSIGNMENT

**Steven Jecmen**

Carnegie Mellon University  
sjecmen@cs.cmu.edu

**Nihar B. Shah**

Carnegie Mellon University  
nihars@cs.cmu.edu

**Fei Fang**

Carnegie Mellon University  
feifang@cmu.edu

**Vincent Conitzer**

Duke University  
conitzer@cs.duke.edu

## ABSTRACT

Many conferences rely on paper bidding as a key component of their reviewer assignment procedure. These bids are then taken into account when assigning reviewers to help ensure that each reviewer is assigned to suitable papers. However, despite the benefits of using bids, reliance on paper bidding can allow malicious reviewers to manipulate the paper assignment for unethical purposes (e.g., getting assigned to a friend’s paper). Several different approaches to preventing this manipulation have been proposed and deployed. In this paper, we enumerate certain desirable properties that algorithms for addressing bid manipulation should satisfy. We then offer a high-level analysis of various approaches along with directions for future investigation.

## 1 INTRODUCTION

In peer review in computer science, paper submissions must be assigned reviewers with the expertise required to provide a high-quality review. The standard approach to this problem involves computing a similarity score for each reviewer-paper pair representing the estimated quality of review by that reviewer for that paper, incorporating both the reviewer’s expertise and preferences. These similarities are computed from various components (Shah, 2021), including text-matching with the reviewer’s past work (Mimno & McCallum, 2007; Liu et al., 2014; Rodriguez & Bollen, 2008; Tran et al., 2017; Charlin & Zemel, 2013), the paper and reviewer subject areas, and reviewer-provided “bids.” Typically, a reviewer assignment is then found that maximizes total similarity (Charlin & Zemel, 2013; Long et al., 2013; Goldsmith & Sloan, 2007; Tang et al., 2010; Flach et al., 2010; Taylor, 2008).

One major part of the similarity computation is the paper bidding process. During paper bidding, each reviewer has the option of indicating how interested they are in reviewing each of the submitted papers by choosing a “bid” from a list of options (e.g., “Not willing”, “In a pinch”, “Willing”, “Eager”). Reviewers make these decisions based on the paper title, subject areas, and abstract. Paper bidding is near-universally used in practice, and tends to have a major impact on the resulting reviewer assignment. At AAAI 2021 (Leyton-Brown et al., 2022): *“Reviewers were assigned papers for which they bid positively (willing or eager) 77.4% of the time. A back-of-the-envelope calculation leads us to estimate that 79.3% of these matches may not have happened had the reviewer not bid positively.”*

However, this reliance on paper bidding opens the door for malicious reviewers to take advantage of the paper assignment process. These malicious reviewers manipulate the assignment by providing dishonest bids in order to get assigned to a target paper. This target paper may be a friend’s work which the malicious reviewer wishes to provide a positive review for, or a rival’s work which the malicious reviewer wants to “torpedo” (Barroga, 2014; Paolucci & Grimaldo, 2014; Akst, 2010). Rings of colluding reviewers have been recently uncovered at a few computer science conferences, including this instance in an ACM conference (Vijaykumar, 2020; Littman, 2021): *“Another SIG community has had a collusion problem where the investigators found that a group of PC members*

and authors colluded to bid and push for each other’s papers violating the usual conflict-of-interest rules.” Beyond bidding, malicious reviewers can also potentially modify their subject areas or their record of past work in order to achieve a desired paper assignment. However, we focus primarily on bid manipulation in this work as the easiest and most obvious avenue through which the paper assignment can be manipulated.

Possible manipulation of the paper assignment is taken seriously by major conferences (e.g., AAAI 2021 (Leyton-Brown et al., 2022) and AAAI 2022 (Shah, 2021)), which have used a variety of approaches to prevent this sort of malicious behavior in recent years. Several techniques are described in recent research papers (Jecmen et al., 2020; Wu et al., 2021; Leyton-Brown et al., 2022; Shah, 2021). In this paper, we take a high-level look at several of these approaches and consider: to what extent do they satisfy properties that we would want paper assignment algorithms to satisfy? We enumerate a list of desiderata for assignment algorithms and present a preliminary evaluation of the strengths and weaknesses of various proposed approaches on these desiderata.

## 2 DESIDERATA

The simplest approach to handling the problem of bid manipulation is simply to not use paper bidding at all, relying solely on text-matching scores (*text similarities*) and subject areas for the assignment. However, bids are near-universally used in practice and some venues even assign reviewers based only on bids. This is because there are several significant benefits to considering bids when assigning papers.

- Bids can capture aspects of a reviewer’s preferences or expertise not captured by text similarities, either because the text modeling failed to accurately represent the relationship between the submission and the reviewer’s past work or because relevant factors were not represented in the reviewer’s past work.
- Bidding allows reviewers to correct erroneous text similarities by expressing interest in papers that are truly a good match but with which the reviewer has a low text similarity.
- Reviewers may be more likely to provide high-quality reviews for papers that they explicitly expressed interest in reviewing during bidding. This is supported by Cabanac & Preuss (2013), who found that reviewers reported higher confidence in their reviews for papers that they bid on.

Thus, the assignment algorithms we consider here attempt to carefully use bids in order to achieve the above benefits while remaining robust against manipulation from malicious reviewers.

Based on these objectives, we present several desirable and potentially conflicting properties that an ideal assignment algorithm should satisfy.

- (A) **Assignment quality:** The algorithm should produce assignments with a high level of expertise, as represented by text similarities, subject areas, and bids.
- (B) **Preference expressiveness:** The algorithm should allow reviewers to express their true preferences in a flexible manner. In particular, this means that it should produce good assignments for reviewers with idiosyncratic preferences not captured by text similarities and for reviewers with erroneous text similarities.
- (C) **Incentives to bid:** The algorithm should incentivize reviewers to provide accurate bids by assigning reviewers to papers that match their own bids to some extent.
- (D) **Low attack success rate:** The algorithm should not allow a malicious reviewer to significantly increase their probability of assignment with a specific target paper through manipulating their bids. We call this assignment probability the “probability of successful manipulation” and call this manipulation of bids an “attack.”
- (E) **High attack cost:** A malicious reviewer should require extra information (e.g., other reviewers’ bids/text similarities) or resources (e.g., additional colluding reviewers) in order to effectively manipulate the paper assignment.
- (F) **Adjustability:** Conference program chairs should be able to easily adjust the algorithm in order to achieve a desired tradeoff between the other desiderata.

Algorithm	Strengths	Weaknesses
BID LIMIT	(A), (B), (C), (G)	(D), (E)
RANDOM DISPLAY	(B), (C), (F), (G)	(A), (E)
CYCLE PREVENTION (Guo et al., 2018)	(B), (C)	(D), (F), (G)
GEOGRAPHIC DIVERSITY	(A), (B), (C), (E)	(D), (F)
BID MODELING (Wu et al., 2021)	(A), (D), (E)	(B), (C),(F)
REVIEWER CLUSTERING	(D), (F)	(B), (C),(E), (G)
PROBABILITY-LIMITED	(A), (B), (C), (F), (G)	(E)
RANDOMIZED ASSIGNMENT (Jecmen et al., 2020)		

Table 1: Key strengths and weaknesses of algorithms.

(G) **Computational scalability:** The algorithm should be feasible to run at the large scale of modern conferences (with thousands of reviewers and papers), in terms of computational resources such as runtime and memory.

These objectives are often contradictory and cannot all be satisfied simultaneously. We instead hope for assignment algorithms that can effectively achieve a balance between them.

### 3 ALGORITHMS

Several different approaches have been proposed for paper assignment in the presence of malicious behavior, both in practice and in the literature. Although these approaches take a wide variety of forms, we view each of them as an end-to-end algorithm for the paper assignment process, encompassing the solicitation of bids and other features from reviewers and ending by outputting the final paper assignment. In this section, we present a brief description of some of these algorithms, along with what we see as their strengths and weaknesses on the various desiderata from Section 2. These strengths and weaknesses are summarized in Table 1.

#### 3.1 ALGORITHM: BID LIMIT

**Description:** This simple approach requires each reviewer to enter at least some number of positive bids, and may also limit the number of negative bids that can be placed. If a reviewer does not meet these bidding criteria, the assignment algorithm may down-weight their bids or ignore them entirely when computing similarities. Intuitively, if a reviewer must bid positively on several papers (and these bids are weighted heavily when computing similarities), a malicious reviewer will have high similarity with some papers other than their target paper and may be assigned to those papers instead of their target. This idea has been used at numerous conferences, including AAAI 2022.

**Evaluation:** On the strong side, this approach is minimally disruptive to the standard assignment process, since honest reviewers need only make additional positive bids or remove negative bids in order to meet the requirements. Thus, the approach maintains the benefits of using bids in the standard way: it finds a high-quality assignment (A), and works well for reviewers with inaccurate text similarities as they can bid positively on any papers they think are truly the best fit (B). This approach has benefits even in the absence of malicious behavior as it encourages honest reviewers to provide information (C). It also makes it more likely that each paper gets several positive bids, as Shah et al. (2018) observe that the standard bidding process leaves many papers with very few positive bids. The algorithm requires negligible additional computation (G).

As for weaknesses, this approach is not robust against malicious behavior if malicious reviewers are behaving strategically (D), since they can choose to bid positively only on papers with which they have very low text similarity and thus are unlikely to be assigned to. Furthermore, this attack is simple to execute (E). While the parameter denoting the number of required bids is easily adjustable, the connection between this parameter and the algorithm’s performance on other desiderata (e.g., the probability of successful manipulation) is unclear (F).

### 3.2 ALGORITHM: RANDOM DISPLAY

**Description:** Under this algorithm, each reviewer is shown a randomly-chosen subset of papers during the bidding process and can only bid on these papers. A similar procedure was used for bidding at AAAI 2020, where only a limited number of papers were shown to each reviewer. Since a malicious reviewer only has a limited probability of being able to bid on their target paper, this can lower the likelihood that they succeed at getting assigned. If desired, a conference can provide a hard limit on the probability of successful manipulation by disallowing the assignment of any reviewer to a paper not shown to them for bidding; we refer to this as the hard-constraint variant of RANDOM DISPLAY. In other words, if half of the papers are displayed to each reviewer under the hard-constraint variant, the probability of successful manipulation would be limited at 0.5 since the target paper is not displayed to the malicious reviewer half of the time.

**Evaluation:** One strength is that the subset of papers shown to each reviewer should be representative of the conference as a whole, so an honest reviewer should not have difficulty finding good matches to bid on (B). An honest reviewer also has a strong incentive to bid since bids are used in the same way as under the standard assignment algorithm (C). Under the hard-constraint variant, the program chairs can easily achieve a desired maximum probability of successful manipulation by appropriately choosing the proportion of displayed papers (F). The algorithm requires negligible additional computation (G).

On the weak side, the optimal strategy for a malicious reviewer is simple (E): bid positively on the target paper if it is displayed and bid negatively on all others. Further, one can show that the hard-constraint variant of RANDOM DISPLAY is dominated by PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT (another algorithm described later in Section 3.7), in terms of expected similarity (A) when they control the probability of successful manipulation at the same level. See Appendix A for the formal result. Note that the RANDOM DISPLAY and the PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT algorithms are directly comparable because they both use the same similarity objective and provide a guarantee on the probability of successful manipulation.

Overall, the algorithm’s ability to effectively limit the probability of successful manipulation (D) is unclear. Regardless of whether the hard-constraint variant is used, sufficiently limiting the probability of successful manipulation may require imposing impractical restrictions on the bidding options for honest reviewers. Furthermore, if the hard-constraint variant is not used, then a malicious reviewer may still be able to succeed even if their target paper is not displayed for bidding. By bidding negatively on all displayed papers, they may be able to lower their similarity with enough papers so that their target paper is one of the highest-similarity papers remaining (even though it was not displayed). This issue can be solved by using the hard-constraint variant, but this comes at the cost of severely restricting the assignments for honest reviewers.

### 3.3 ALGORITHM: CYCLE PREVENTION

**Description:** In some cases, malicious reviewers who have authored a paper may collude with other reviewers who have also authored a paper at the same conference. These reviewers will attempt to get assigned to each others’ papers through bidding as part of a deal to benefit each other. This algorithm (Guo et al., 2018; Boehmer et al., 2021) attempts to prevent this collusion by restricting the assignment so that it cannot contain any 2-cycles of reviewers: that is, if Alice is assigned to review Bob’s paper, then Bob cannot be assigned to review Alice’s paper. 3-cycles and larger may also be restricted if computational resources allow. This approach has been taken by AAAI 2021 (Leyton-Brown et al., 2022).

**Evaluation:** We first consider strengths. Note that unlike most of the other algorithms we discuss, this algorithm assumes that the malicious reviewers are part of a colluding group. As mentioned in Section 1, there is reason to believe that collusion rings are a common form of manipulation. If so, this algorithm can provide some robustness without impacting the expressiveness of bids (B) or the incentives to bid (C).

As for weaknesses, this algorithm does not do anything to stop a malicious reviewer who is not colluding with others (D). For example, this may be a reviewer aiming to torpedo-review a rival’s paper. Furthermore, this algorithm can be circumvented by groups of reviewers who decide to

collude across multiple different conferences or otherwise compensate each other outside the scope of a single conference’s peer review process. Program chairs cannot effectively adjust the algorithm to their needs, as even increasing the size of the removed cycles is computationally difficult (F). This computational difficulty poses a challenge for scalability (G), as finding a maximum-similarity assignment subject to cycle constraints requires solving an integer program.

The impact of this algorithm on the quality of the assignment is unclear (A). With enough expert reviewers for each topic, it’s possible that most honest reviewers involved in a high-similarity cycle can be replaced with a similarly-qualified reviewer; at AAI 2021, preventing 2-cycles lowered the total assignment similarity by only 0.01% Leyton-Brown et al. (2022). However, the conference in question may not have a deep enough reviewer pool and this claim may not hold even if it does. Additionally, the difficulty of attacking this algorithm is dependent on the type of attacker (E). For colluding pairs of reviewers, the algorithm is not trivial to circumvent, since either an additional collaborator must be recruited or the submission venue of one of the papers must be changed; however, large colluding groups can easily set up cycles of higher length to avoid detection.

### 3.4 ALGORITHM: GEOGRAPHIC DIVERSITY

**Description:** Like the CYCLE PREVENTION algorithm, this approach focuses on defending against malicious reviewers who collude in groups. It specifically defends against groups of colluding reviewers that are based in a single geographic region by adding some form of geographic diversity constraint on the reviewer assignment. For example, AAI 2021 used a constraint that no two reviewers assigned to the same paper belonged to the same region (Leyton-Brown et al., 2022), and AAI 2022 used a constraint that at least one assigned reviewer must be from a different region as the paper’s authors. This approach is motivated by the idea that colluding groups are more likely to be from a single region, since reviewers from different areas are less likely to know each other or be able to communicate easily.

**Evaluation:** Large conferences include reviewers from a wide range of geographic regions, and experts in any particular topic exist in many regions. Thus, a strength is that this algorithm should not impose significant limitations on the assignments for honest reviewers. The overall assignment quality (A), expressiveness of bids (B), and incentive to bid (C) should all remain quite high, even if some expert reviewers are blocked from their optimal assignment. For example, the geographic diversity constraint imposed by AAI 2021 lowered the assignment similarity by only 0.85% Leyton-Brown et al. (2022). Malicious reviewers who would be stopped by this algorithm can attempt to avoid detection by recruiting colluders from a different region or by changing their location and affiliation in the conference system. However, recruiting colluders from other regions may be difficult and falsified locations can be detected by careful program chairs, making effectively circumventing this algorithm difficult (E).

One weakness is that, like CYCLE PREVENTION, this algorithm does not defend against a malicious reviewer who is not colluding with others or against colluding reviewers who compensate each other outside of the conference’s peer review process (D). It further does not defend well against colluding groups containing reviewers from several different regions, which could have formed because the reviewers previously met in some professional setting or because reviewers have moved institutions to a different region. The program chairs can choose the specific form of geographic constraint that is desired, but cannot easily see how effectively this will prevent collusion (F) since the geographical distribution of colluding groups is unknown. The computational cost of the algorithm depends on the exact form of geographic constraint posed (G).

### 3.5 ALGORITHM: BID MODELING

**Description:** This algorithm, proposed in Wu et al. (2021), uses the submitted bids from all reviewers to train a linear regression model. This model aims to predict the bid value for each reviewer-paper pair as a function of various features of that reviewer-paper pair, including the text similarity and the subject area intersection. The paper assignment is then chosen to maximize the total *predicted* bid value of the assigned reviewers. The authors propose further techniques to defend against groups of colluding reviewers.

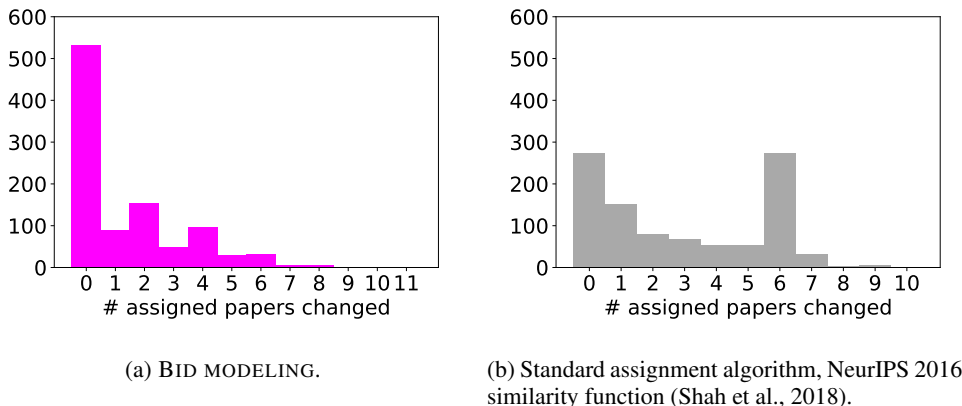


Figure 1: Symmetric differences between the sets of papers assigned to 1000 reviewers with honest bids and with no bids. Each reviewer is assigned at most 6 papers and each paper is assigned to 3 reviewers, within a dataset of around 2500 papers and reviewers (Wu et al., 2021).

**Evaluation:** The primary strength of the BID MODELING algorithm is its robustness against malicious behavior (D): assuming that malicious reviewers cannot manipulate the reviewer-paper features, Wu et al. (2021, Figure 1-2) demonstrate that a single malicious reviewer is unable to improve their probability of assignment to a target paper using a naive attack and has limited success with a more advanced heuristic attack. Furthermore, computing an effective attack against this model requires knowledge of the features and bids of other reviewers (E), which is unlikely to be available to malicious reviewers. The authors also find that the text similarity and bid values of the resulting assignment remain comparable to standard assignment methods (A): BID MODELING achieves a 16% increase in the average text-similarity score of the assignment over a standard assignment algorithm with the NeurIPS 2014 similarity function, and a 38% increase in the average bid value of the assignment over a standard assignment algorithm using only text similarities (Wu et al., 2021, Table 1).

As for weaknesses, the algorithm pays a price for this robustness in terms of its flexibility to reviewer preferences (B), as a reviewer with incorrect text similarities may find their predicted bid values to be incorrect. The algorithm also does not allow for easy tuning by the program chairs (F), since the hyperparameters are not clearly connected to any desiderata. Additionally, if reviewer and paper features such as the subject areas and text similarities can be strategically manipulated, this approach may not be effective. Computing appropriate reviewer-paper features and fitting the model will add some additional time to the assignment algorithm at scale (G), but the algorithm does run in polynomial time.

Additionally, we conducted experiments which indicate that honest reviewers may not be sufficiently incentivized to provide bids to the algorithm (C). We sample 1000 reviewers from the dataset provided in Wu et al. (2021) and for each compute the assignments that would result if they provide their honest bids and if they provide no bids. In Figure 1a, we plot the size of the symmetric difference between the set of papers assigned to this reviewer in these two cases under BID MODELING. We see that a majority of reviewers have identical assignments under BID MODELING, regardless of whether or not they provide bids; the mean number of papers changed is 1.394 and the median is 0. For comparison, we also plot in Figure 1b the same metric under the standard paper assignment algorithm using the NeurIPS 2016 similarity function (Shah et al., 2018); the mean change is 2.973 and the median is 2.

### 3.6 ALGORITHM: REVIEWER CLUSTERING

**Description:** Similar to BID MODELING, this algorithm takes as input various features for each reviewer, such as their subject areas and their text similarity scores with each paper. Based on these features, it clusters reviewers into groups of some fixed size  $m$ . Papers are then assigned to each group based on the averaged bids of that group and randomly distributed among reviewers within the

group. This algorithm is our attempt to capture some of the ideas behind BID MODELING in a simple manner while also providing a guarantee on the maximum probability of successful manipulation: at most  $1/m$ . The idea of clustering reviewers by their features and arbitrarily distributing papers within each cluster is already used in contexts where reviewer assignment is done entirely by subject area (Merrifield & Saari, 2009).

**Evaluation:** On the strong side, the algorithm appears to limit much of the control that a malicious reviewer has over their assignment in the same manner as BID MODELING (D), and it also provides a parameter that can easily be tuned to adjust the tradeoff between assignment quality and probability of successful manipulation (F).

However, weaknesses of the algorithm are that it would not work well for reviewers with inaccurate text similarities (B) and that a malicious reviewer does not require knowledge of other reviewers’ features in order to determine how to bid (E). Further, some honest reviewers may choose to not submit bids in the hopes that the bids of their cluster are suitable enough (C). It could also be computationally expensive to find good fixed-size clusters, since heuristic approaches may perform poorly (G). The quality of the resulting assignment depends strongly on how well the reviewer pool can be clustered into groups of similar expertise and interests, which may vary by conference (A).

### 3.7 ALGORITHM: PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT

**Description:** This algorithm, proposed in Jecmen et al. (2020), adds a randomized aspect to the standard assignment algorithm. Like the standard assignment algorithm, it takes bids and computes similarities as normal. Then, given a parameter  $q \in [0, 1]$ , it finds a randomized assignment with maximum expected similarity, subject to the constraint that the maximum probability of any reviewer-paper assignment is at most  $q$ .

**Evaluation:** We first consider strengths. By definition, PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT finds the assignment with highest similarity among all assignments that provide a guarantee on the maximum probability of successful manipulation (A). On data from ICLR 2018, PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT achieves 90.8% of the standard assignment algorithm’s similarity with  $q = 0.5$  (Jecmen et al., 2020, Figure 1). Program chairs can compute this percentage for various values of  $q$  before choosing one to use in deployment, allowing them to easily control the tradeoff between the assignment similarity and the maximum probability of successful manipulation (F). Additionally, the algorithm’s guarantees on the maximum probability of successful manipulation hold without any assumptions on the malicious reviewers’ capabilities, so it is still effective even if aspects like the subject areas and text similarities can be manipulated. Since reviewers’ bids are used without modification, the expressiveness of bids is fully preserved (B) and honest reviewers are still incentivized to bid (C). The randomized assignment can be found with the same computational resources as the standard assignment algorithm, and sampling the assignment adds little additional overhead (G).

However, one weakness of the algorithm is that it’s easy for a malicious reviewer to determine their best strategy (E): bid the maximum value on their target paper and the minimum value on all others. In this manner, malicious reviewers may easily be able to achieve this theoretical maximum probability in practice, as demonstrated in simulations by Jecmen et al. (2020). Additionally, although PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT is optimal in terms of similarity (subject to the constraint on the probability of successful manipulation), it remains agnostic to the computation of similarities. If some similarity components (e.g., text similarity) are believed to be more trustworthy than the bids, this algorithm may not be able to control the probability of successful manipulation as efficiently as other algorithms that leverage this distinction (D). Although one can place greater weight on trustworthy components when computing similarities, this approach may not be the optimal way to accommodate such assumptions.

In Section 3.2, we mention that PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT dominates the hard-constraint variant of RANDOM DISPLAY in terms of expected similarity. However, one downside of PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT is that reviewers may waste time bidding on papers that they will not be assigned due to the subsequent randomization. In contrast, by doing the randomization before bidding, RANDOM DISPLAY ensures that reviewers only spend time bidding on papers for which they are eligible to be assigned.

## 4 DISCUSSION

Addressing bid manipulation in a manner that maintains the valuable properties of paper bidding is a pressing issue, given the scale and importance of modern conferences. The approaches we consider tackle the issue in a variety of ways, with different strengths and weaknesses. The least intrusive approaches (BID LIMIT, RANDOM DISPLAY, CYCLE PREVENTION, and GEOGRAPHIC DIVERSITY) keep the paper assignment process largely the same as under the standard assignment algorithm, which make them easier to deploy in practice. These algorithms preserve the essential benefits of bids but may not do enough to prevent manipulation effectively, as they have not been rigorously examined.

The other algorithms can be divided into two categories based on how they use the non-bid similarity features (e.g., text similarities). BID MODELING, along with the related REVIEWER CLUSTERING algorithm, gains significant power to stop manipulation under the assumption that these features are harder for an adversary to change. If the adversary can manipulate these features (e.g., via falsifying their TPMS profile or strategically providing subject areas (Shah, 2021, Section 4.2)), these algorithms may lose some effectiveness. In contrast, PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT entirely abstracts away the similarity computation, ignoring any differences in the cost of manipulating different features. This algorithm thus may be most appropriate for a worst-case setting where program chairs are not willing to make assumptions about the capabilities of malicious reviewers.

CYCLE PREVENTION and GEOGRAPHIC DIVERSITY specifically focus on defending against colluding reviewers, but other approaches also can be extended to handle collusion. The formulation of the BID MODELING algorithm as proposed by Wu et al. (2021) includes a component that effectively prevents colluding groups of a known size from manipulating the learned model. Jecmen et al. (2020, Section 5.2) provides an extension to their PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT algorithm that additionally enforces that each paper be assigned diverse reviewers, essentially combining the PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT and GEOGRAPHIC DIVERSITY approaches.

The algorithms we consider in this work sit at different positions on the tradeoffs between our proposed desiderata, but many other positions on these tradeoffs remain unfilled. We hope that our list of desiderata can help direct the development of additional algorithms to address bid manipulation. For example, we proposed the REVIEWER CLUSTERING algorithm as a simplified variant of the BID MODELING algorithm that improves on desideratum (F). Further study on the bid manipulation problem can improve on the balance between these various desired properties.

In addition, some past conferences have used multiple of these approaches at the same time. AAAI 2021 used both CYCLE PREVENTION and GEOGRAPHIC DIVERSITY, and AAAI 2022 used forms of BID LIMIT, GEOGRAPHIC DIVERSITY, and PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT. A useful direction of future work is to develop new algorithms that combine multiple previous approaches in order to simultaneously achieve their benefits.

Finally, our analysis indirectly compares algorithms based on whether they satisfy our desiderata. One might hope to additionally conduct some form of direct comparison between algorithms, e.g., by comparing the assignment quality of each algorithm at a given probability of successful manipulation. However, there are numerous challenges in making such a comparison. Different algorithms make different assumptions about adversary capabilities and may optimize different objectives, such that both “probability of successful manipulation” and “assignment quality” may be incomparable between algorithms. Furthermore, non-malicious reviewers may behave differently under different algorithms (e.g., by providing more bids under BID LIMIT than under another algorithm). Determining from past data how these reviewers might have behaved in a different environment is difficult, as seen in the literature on valuation estimation in auctions Jiang & Leyton-Brown (2007). We leave addressing these challenges for future work.

### ACKNOWLEDGEMENTS

This work was supported by NSF CAREER award 1942124, NSF CAREER award 2046640, and NSF 1763734. We thank Hanrui Zhang for helpful discussions.



## REFERENCES

- Jef Akst. I Hate Your Paper. Many say the peer review system is broken. Here's how some journals are trying to fix it. *The Scientist*, 24(8):36, 2010.
- Edward F Barroga. Safeguarding the integrity of science communication by restraining 'rational cheating' in peer review. *Journal of Korean Medical Science*, 29(11):1450–1452, 2014.
- Niclas Boehmer, Robert Bredereck, and André Nichterlein. Combating collusion rings is hard but possible. *CoRR*, abs/2112.08444, 2021. URL <https://arxiv.org/abs/2112.08444>.
- Guillaume Cabanac and Thomas Preuss. Capitalizing on order effects in the bids of peer-reviewed conferences to secure reviews by expert referees. *Journal of the American Society for Information Science and Technology*, 64(2):405–415, 2013.
- Laurent Charlin and Richard S. Zemel. The Toronto Paper Matching System: An automated paper-reviewer assignment system. In *ICML Workshop on Peer Reviewing and Publishing Models*, 2013.
- Peter A. Flach, Sebastian Spiegler, Bruno Golénia, Simon Price, John Guiver, Ralf Herbrich, Thore Graepel, and Mohammed J. Zaki. Novel tools to streamline the conference review process: Experiences from SIGKDD'09. *SIGKDD Explor. Newsl.*, 11(2):63–67, May 2010. ISSN 1931-0145. doi: 10.1145/1809400.1809413. URL <http://doi.acm.org/10.1145/1809400.1809413>.
- Judy Goldsmith and Robert H. Sloan. The AI conference paper assignment problem. *WS-07-10: 53–57*, 12 2007.
- Longhua Guo, Jie Wu, Wei Chang, Jun Wu, and Jianhua Li. K-loop free assignment in conference review systems. In *2018 International Conference on Computing, Networking and Communications (ICNC)*, pp. 542–547. IEEE, 2018.
- Steven Jecmen, Hanrui Zhang, Ryan Liu, Nihar B. Shah, Vincent Conitzer, and Fei Fang. Mitigating manipulation in peer review via randomized reviewer assignments. In *NeurIPS*, 2020.
- Albert Xin Jiang and Kevin Leyton-Brown. Bidding agents for online auctions with hidden bids. *Mach. Learn.*, 67(1-2):117–143, 2007. doi: 10.1007/s10994-006-0477-8. URL <https://doi.org/10.1007/s10994-006-0477-8>.
- Kevin Leyton-Brown, Mausam, Yatin Nandwani, Hedayat Zarkoob, Chris Cameron, Neil Newman, and Dinesh Raghu. Matching papers and reviewers at large conferences, 2022.
- Michael L Littman. Collusion rings threaten the integrity of computer science research. *Communications of the ACM*, 64(6):43–44, 2021.
- Xiang Liu, Torsten Suel, and Nasir Memon. A robust model for paper reviewer assignment. In *8th ACM Conference on Recommender Systems, RecSys '14*, pp. 25–32, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2668-1. doi: 10.1145/2645710.2645749. URL <http://doi.acm.org/10.1145/2645710.2645749>.
- Cheng Long, Raymond Wong, Yu Peng, and Liangliang Ye. On good and fair paper-reviewer assignment. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 1145–1150, 12 2013. ISBN 978-0-7695-5108-1.
- Michael R Merrifield and Donald G Saari. Telescope time without tears: a distributed approach to peer review. *Astronomy & Geophysics*, 50(4):4–16, 2009.
- David Mimno and Andrew McCallum. Expertise modeling for matching papers with reviewers. In *13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07*, pp. 500–509, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-609-7. doi: 10.1145/1281192.1281247. URL <http://doi.acm.org/10.1145/1281192.1281247>.
- Mario Paolucci and Francisco Grimaldo. Mechanism change in a simulation of peer review: From junk support to elitism. *Scientometrics*, 99(3):663–688, 2014.

- Marko A. Rodriguez and Johan Bollen. An algorithm to determine peer-reviewers. In *17th ACM Conference on Information and Knowledge Management, CIKM '08*, pp. 319–328, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458127. URL <http://doi.acm.org/10.1145/1458082.1458127>.
- Nihar B Shah. An overview of challenges, experiments, and computational solutions in peer review. *Communications of the ACM* (to appear). Preprint available at <http://bit.ly/PeerReviewOverview>, July 2021.
- Nihar B Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. Design and analysis of the NIPS 2016 review process. *The Journal of Machine Learning Research*, 19(1):1913–1946, 2018.
- Wenbin Tang, Jie Tang, and Chenhao Tan. Expertise matching via constraint-based optimization. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010.
- Camillo J. Taylor. On the optimal assignment of conference papers to reviewers. Technical report, Department of Computer and Information Science, University of Pennsylvania, 2008.
- Hong Diep Tran, Guillaume Cabanac, and Gilles Hubert. Expert suggestion for conference program committees. In *11th International Conference on Research Challenges in Information Science*, pp. 221–232, May 2017. doi: 10.1109/RCIS.2017.7956540.
- T. N. Vijaykumar. Potential organized fraud in ACM/IEEE computer architecture conferences. <https://medium.com/@tnvijayk/potential-organized-fraud-in-acm-ieee-computer-architecture-conferences-ccd61169370d> (accessed August 17, 2020), 2020.
- Ruihan Wu, Chuan Guo, Felix Wu, Rahul Kidambi, Laurens Van Der Maaten, and Kilian Weinberger. Making paper reviewing robust to bid manipulation attacks. In *International Conference on Machine Learning*, pp. 11240–11250. PMLR, 2021.

## APPENDIX

### A COMPARISON BETWEEN THE EXPECTED SIMILARITY OF RANDOM DISPLAY AND PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT

We consider the hard-constraint variant of RANDOM DISPLAY, described in Section 3.2, which does not allow a reviewer to be assigned to papers that were not displayed to them during bidding. Define the “display fraction” of RANDOM DISPLAY as the proportion of papers in the subset displayed to each reviewer. In this section, we compare the hard-constraint variant of RANDOM DISPLAY with display fraction  $q$  to PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT (from Section 3.7) with probability limit  $q$ , in terms of expected similarity. These algorithms are directly comparable, since both limit the maximum probability of successful manipulation at  $q$ .

We first introduce some notation. Call  $n$  the number of reviewers and  $m$  the number of papers. Define  $S \in [0, 1]^{m \times n}$  as the matrix of similarities used by both algorithms, where  $S_{p,r}$  is the similarity of paper  $p$  with reviewer  $r$ .  $S$  can be computed from the bids along with other features using any method, since both algorithms are agnostic to the method of similarity computation. We assume that the bids of each reviewer are the same regardless of which algorithm is used or which papers are displayed to that reviewer.

The following result shows that PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT outperforms RANDOM DISPLAY in terms of expected similarity.

**Theorem 1.** *For any  $q \in [0, 1]$ , the expected similarity of the assignment produced by PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT with probability limit  $q$  is no less than the expected similarity of the assignment produced by the hard-constraint variant of RANDOM DISPLAY with display fraction  $q$ .*

*Proof.* Define the matrix  $Q \in \{0, 1\}^{m \times n}$  as the random variable representing the papers displayed to each reviewer by RANDOM DISPLAY;  $Q_{p,r} = 1$  if paper  $p$  is displayed to reviewer  $r$ . Since  $qm$  of the  $m$  papers are chosen uniformly at random for each reviewer,  $\mathbb{E}[Q_{p,r}] = q$ . Call  $Q^{(1)}, \dots, Q^{(N)}$  the possible realizations of  $Q$ , from which  $Q$  is chosen uniformly at random. For each  $i \in [N]$ , define  $A^{(i)} \in \{0, 1\}^{m \times n}$  as the matrix representing the assignment produced by RANDOM DISPLAY if  $Q^{(i)}$  was displayed;  $A_{p,r}^{(i)} = 1$  if paper  $p$  is assigned to reviewer  $r$ .

The expected similarity of the assignment produced by RANDOM DISPLAY is

$$\frac{1}{N} \sum_{i=1}^N \sum_{r=1}^n \sum_{p=1}^m A_{p,r}^{(i)} S_{p,r}.$$

The matrix  $F = \frac{1}{N} \sum_{i=1}^N A^{(i)}$  satisfies  $F_{p,r} \leq q$  for all entries  $(p, r)$ , since

$$\frac{1}{N} \sum_{i=1}^N A_{p,r}^{(i)} \leq \frac{1}{N} \sum_{i=1}^N Q_{p,r}^{(i)} = \mathbb{E}[Q_{p,r}] = q.$$

Consider the randomized assignment represented by  $F$ , where  $F_{p,r}$  represents the marginal probability of assigning paper  $p$  to reviewer  $r$ . This randomized assignment has the same expected similarity as the assignment from RANDOM DISPLAY. Further, this is a feasible randomized assignment for PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT with probability limit  $q$ , meaning that PROBABILITY-LIMITED RANDOMIZED ASSIGNMENT will return an assignment with at least this expected similarity.  $\square$