Proceedings of the Tenth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2022)

Near-Optimal Reviewer Splitting in Two-Phase Paper Reviewing and Conference Experiment Design

Steven Jecmen,[1] **Hanrui Zhang,**[1] **Ryan Liu,**[1]
Fei Fang,[1] **Vincent Conitzer,**[2] **Nihar B. Shah**[1]
[1] Carnegie Mellon University
[2] Duke University
sjecmen@cs.cmu.edu, hanruiz1@cs.cmu.edu, ryanliu@andrew.cmu.edu,
feif@cs.cmu.edu, conitzer@cs.duke.edu, nihars@cs.cmu.edu

Abstract

Many scientific conferences employ a two-phase paper review process, where some papers are assigned additional reviewers after the initial reviews are submitted. Many conferences also design and run experiments on their paper review process, where some papers are assigned reviewers who provide reviews under an experimental condition. In this paper, we consider the question: how should reviewers be divided between phases or conditions in order to maximize total assignment similarity? We make several contributions towards answering this question. First, we prove that when the set of papers requiring additional review is unknown, a simplified variant of this problem is NP-hard. Second, we empirically show that across several datasets pertaining to real conference data, dividing reviewers between phases/conditions uniformly at random allows an assignment that is nearly as good as the oracle optimal assignment. This uniformly random choice is practical for both the two-phase and conference experiment design settings. Third, we provide explanations of this phenomenon by providing theoretical bounds on the suboptimality of this random strategy under certain natural conditions. From these easily-interpretable conditions, we provide actionable insights to conference program chairs about whether a random reviewer split is suitable for their conference.

1 Introduction

Peer review is a widely-adopted method for evaluating scientific research and a highly useful application of human computation. Careful assignment of reviewers to papers is critically important in order to ensure that reviewers have the requisite expertise and that the resulting reviews are of high quality. At large scientific conferences, the paper assignment is usually chosen by solving an optimization problem. Given a set of papers, a set of reviewers, and a similarity matrix consisting of scores representing the level of expertise each reviewer has for each paper, the standard paper assignment problem is to find an assignment of reviewers to papers that maximizes total similarity, subject to constraints on the reviewer and paper loads. This standard paper assignment problem is a simple matching problem and so can be efficiently solved (for example, through linear programming). Our work is motivated by two scenarios that arise in the context of paper assignment in conference peer review.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Motivation 1: Two-phase paper assignment. Many conferences (e.g., AAAI 2021-2022, IJCAI 2022) have adopted a two-phase review process. After the initial reviews are submitted, a subset of papers proceed to a second phase of reviews with additional reviewers assigned. There are a variety of reasons that a two-phase reviewing process can be helpful. For example, the process can be used to triage papers based on reviews in the first phase. This can allow the conference to solicit additional reviews only on papers that obtained sufficiently high ratings in the first phase and have any chance of getting accepted (as done at AAAI 2021). The second phase can also help focus on evaluation of the papers in the "messy middle"—the papers at the borderline between acceptance and rejection. This messy middle model (Price 2014), which hypothesizes that the acceptance decisions for some percentage of submitted papers are effectively random, was proposed after the NeurIPS 2014 experiment (Lawrence 2014) in order to explain the observed inconsistency in acceptance decisions. Additional reviewers could improve the evaluation of these papers to more accurately discern which should be accepted. Later analysis of the NeurIPS 2015 and 2016 review process found that the size of the messy middle in these conferences was 45% and 30% of submissions respectively (Shah et al. 2018). A second phase of reviews can also be used to help compensate for reviewers who were unresponsive or minimally responsive in the first phase, who can no longer review due to problems in their personal lives, who discovered conflicts they had with a paper assigned to them and recused themselves from it, etc.

In all of these cases, the set of papers that will require additional review is unknown beforehand. While some venues choose to recruit new reviewers after knowing which papers proceed to phase two, the tight timeline of many conferences makes it hard to recruit new reviewers after phase one. In SIGMOD 2019 (Ailamaki et al. 2019): "*Additional reviews were solicited manually by the chairs and this was a huge time sink, especially when some reviewers refused to take on the additional assignment. The additional review solicitation needs to be automated and reviewer expectations need to be set appropriately beforehand.*" For this reason, it is best if all the reviewers are recruited at the beginning, and a key question is then how to assign reviewers to papers in the first phase such that enough review capacity is saved for the second phase.

Motivation 2: Conference experiment design. Reviewers also need to be split into two groups when conferences run controlled experiments on the paper review process. Conferences often run such experiments to test changes to the review process. For example, the WSDM 2017 conference conducted an experiment to test the effects of single-blind versus double-blind reviewing (Tomkins, Zhang, and Heavlin 2017). As another example, the NeurIPS 2014 and 2021 conferences ran experiments testing the consistency of acceptance decisions by providing some papers with a second set of reviews from a separate group of reviewers (Lawrence 2014; Price 2014; Beygelzimer et al. 2021). In these experiments, all papers receive reviews conducted in the usual manner (the control condition), but a random subset of papers are additionally assigned reviewers who provide reviews under an experimental condition. In the NeurIPS 2014/2021 experiments, a random 10% of papers were put in the experimental condition and received a second set of reviews. In the WSDM 2017 experiment, the subset of papers was the full paper set; that is, all papers were reviewed under both single-blind and double-blind conditions. The key question is then how to divide the reviewers between the control and experimental conditions. As in the NeurIPS 2014/2021 and WSDM 2017 experiments, this is often done randomly for statistical purposes. However, conferences still want to ensure that the resulting assignment of papers to reviewers is of high similarity.

As our results apply to both the two-phase and experiment design settings, we will use the generic terminology of "stages" to refer to both phases and conditions simultaneously across the two settings.

Problem outline. In this paper, we formally analyze the two-stage paper assignment problem, which encompasses both above motivations. As stated earlier, the standard paper assignment problem is to maximize the total similarity of the assignment subject to load constraints and is efficiently solvable. However, in the two-stage paper assignment problem, we must additionally decide how much of each reviewer's capacity should be saved to review papers in the second stage. We assume that the *fraction* of papers that will need additional reviews is known and that the set of second-stage papers is chosen uniformly at random.

Because of constraints present in each setting, the maximum-similarity paper assignment across the two stages cannot be achieved. In the two-phase setting, the set of papers that will need to be reviewed in the second stage is unobserved when the first-stage assignment is made, making the problem one of stochastic optimization. In the experiment design setting, reviewers are often randomized between stages for statistical purposes. We show that a simple strategy for choosing reviewers to save for the second stage performs near-optimally in terms of assignment similarity and can be used in either setting.

Contributions. Our contributions are as follows.

First, we identify and formulate the two-stage paper assignment problem, an issue of practical importance to modern conferences, with applications to two-phase paper assignment and conference experiment design (Section 3).

Second, we prove that a simplified version of the problem is NP-hard, suggesting that the problem may not be efficiently solvable (Section 4).

Third, we empirically show that a very simple "random split" strategy, which chooses a subset of reviewers uniformly at random to save for the second stage, gives near-optimal assignments on real conference similarity scores (Section 5.1). This result is summarized in Figure 1, which shows the assignment similarity achieved using random split as compared to the oracle optimal assignment (which views the set of second-stage papers before optimally assigning reviewers across both stages) for several datasets. We find that all random reviewer splits achieve at least 90% of the oracle optimal solution's similarity on all datasets and at least 94% on all but two experiments. These results hold across similarities constructed via a variety of methods used in practice (including text-matching, bidding, and subject areas), indicating that random split is robust across methods of similarity construction. They also hold both when the second-stage papers are drawn uniformly at random (as in Figure 1a) and when they are selected based on the review scores of the papers (as in Figure 1b). In practice, this means that program chairs planning a two-phase review process or a conference experiment can simply split reviewers across the two phases/conditions at random without concerning themselves with the potential reduction in assignment quality.

We also show that this good performance is not achieved in general: there exist similarity matrices on which random split performs very poorly (Section 5.2).

Fourth, we theoretically *explain* why random split performs well on our real conference similarity matrices by deriving theoretical bounds on the suboptimality of this random strategy under certain natural conditions (Sections 6 and 7). We consider two such sufficient conditions here, which are met by our datasets: if the reviewer-paper similarity matrix is low-rank, and if the similarity matrix allows for a high-value assignment (in terms of total similarity) with a large number of reviewers assigned to each paper. From these results, we give key actionable insights to conference program chairs to help them decide–well before the reviewers and/or papers are known–if random split is likely to perform well in their conference.

The full version of the paper can be found online,[1] as can all of the code for our empirical results.[2]

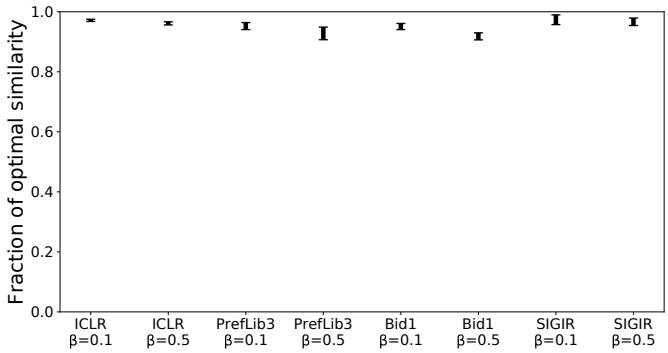
2 Related Work

Our work assumes that the "similarities" between reviewers and authors are given. In practice, there are several ways in which these similarities are computed, and different program chairs often make different decisions on how this computation is done. The similarities are generally computed using one or more of the following three sources of data:

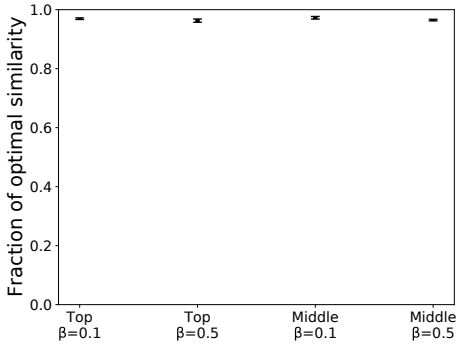
- **Text-matching of papers:** Natural language processing techniques (Mimno and McCallum 2007; Liu, Suel, and Memon 2014; Rodriguez and Bollen 2008; Tran, Cabanac, and Hubert 2017; Charlin and Zemel 2013) are used to

[1] https://arxiv.org/abs/2108.06371

[2] https://github.com/sjecmen/multistage_reviewing_bounds



(a) Second-stage papers drawn uniformly at random



(b) Second-stage papers chosen as the top- or middle-scoring papers from ICLR

Figure 1: Range of assignment similarities found over 10 random reviewer splits on real conference data, as a fraction of the oracle optimal assignment's similarity (computed after observing the second-stage papers). β indicates the fraction of papers in the second stage. The ICLR similarities (Xu et al. 2019) [911 papers, 2435 reviewers] are constructed from text-matching between papers and reviewers' past work, PrefLib3 (Mattei and Walsh 2013) [176 papers, 146 reviewers] and Bid1 (Meir et al. 2020) [600 papers, 400 reviewers] similarities are constructed from bidding data, and SIGIR (Karimzadehgan, Zhai, and Belford 2008) [73 papers, 189 reviewers] similarities are constructed from reviewer and paper subject areas.

match the text of the submitted paper with the text of the reviewers' past papers.

- **Subject areas:** The program chairs create a list of subject areas relevant to the conference. Each reviewer selects a subset of these subject areas that are representative of their expertise, and each submitted paper is accompanied by the authors selecting the subject areas relevant to the paper.
- **Bids:** A number of conferences adopt a bidding system, where reviewers are shown a list of (some of) the papers that are submitted to the conference (and which do not conflict with them) and asked to indicate the papers which they are willing to review (Cabanac and Preuss 2013; Fiez, Shah, and Ratliff 2020; Meir et al. 2020).

If more than one such source of data is used by the conference, they are combined in a manner deemed suitable by the program chairs (Shah et al. 2018). These computed similarities are then used to assign reviewers to papers. By far the most popular method of doing this assignment is to solve an optimization problem that maximizes the sum of the similarities of the assigned reviewer-paper pairs, subject to constraints on the reviewer and paper loads (Charlin and Zemel 2013; Long et al. 2013; Goldsmith and Sloan 2007; Tang, Tang, and Tan 2010; Flach et al. 2010; Taylor 2008). Given its widespread popularity, we analyze this sum-similarity objective in our paper.

That being said, there are other objectives that are also proposed for automated assignment using the similarities, such as the max-min fairness objective (Garg et al. 2010; Stelmakh, Shah, and Singh 2019a; Kobren, Saha, and McCallum 2019). A recent work (Jecmen et al. 2020) proposes assignments via optimizing the sum similarity but with some randomness in order to prevent fraud in peer review. Another line of work (Alon et al. 2011; Xu et al. 2019; Aziz et al. 2019) proposes assigning reviewers to papers in a manner that a reviewer cannot influence the outcome of their own paper by manipulating the reviews they provide. Finally, in practice, the conference organizers may also additionally apply manual tweaks to the outputs of any such automated procedure.

At a high level, the problem we study in the two-phase setting shares several common characteristics with problems in online (stochastic) matching (Karp, Vazirani, and Vazirani 1990; Feldman et al. 2009; Dickerson, Procaccia, and Sandholm 2012; Dickerson et al. 2018; Brubach et al. 2016), often considered in the context of ride-sharing, kidney exchange, or internet advertising. Particularly related to our results is the line of research on two-stage stochastic matching (Kong and Schaefer 2006; Katriel, Kenyon-Mathieu, and Upfal 2008; Escoffier et al. 2010; Lee and Singla 2020; Feng, Niazadeh, and Saberi 2021), which generally focuses on providing algorithms with tight approximation ratios that hold for any (i.e., worst-case) problem instances. To the best of our knowledge, the specific stochastic matching problem we consider (which arises in the context of paper assignment for peer review) has not previously been studied. Additionally, in contrast to this line of work, we aim to provide and justify simple and practical solutions (such as choosing reviewers uniformly at random) based on data-dependent conditions likely to hold in real-world paper assignment instances.

The simplified version of our problem considered in Section 4 can be seen as an instance of maximizing a submodular function subject to a cardinality constraint (see Appendix C). The paper (Buchbinder et al. 2014) gives an approximation algorithm achieving an approximation ratio of no greater than 0.5. However, this guarantee is very weak in the paper assignment setting since it can be trivially achieved by maximizing similarity in the first stage alone.

One motivation for our work is that of running controlled experiments in peer review. Controlled experiments pertaining to peer review are conducted in many different scientific fields (Armstrong 1980; Pier et al. 2017; Teplitskiy et al. 2019; Ceci and Peters 1982; Patat et al. 2019), including several controlled experiments recently conducted in computer science (Lawrence 2014; Tomkins, Zhang, and Heavlin 2017; Stelmakh et al. 2021, 2020). These experiments have also led to a relatively nascent line of work on careful design of experimental methods for peer review (Stelmakh, Shah, and Singh 2019b, 2021), and our work sheds some light in this direction in terms of trading off assignment quality with randomization in the assignment. Some other experiments in conferences (Madden and DeWitt 2006; Tung 2006; Manzoor and Shah 2021) do not operate under controlled settings, but exploit certain changes in the conference policy such as a switch from single blind to double blind reviewing (i.e., natural experiments). Overall, experiments offer important insights into the peer review process; see (Shah 2022) for more discussion on challenges in peer review and some solutions.

3 Problem Formulation

In this section, we formally define the two-stage paper assignment problem. Given a set of n papers $\mathcal{P} = [n]$ and a set of m reviewers $\mathcal{R} = [m]$, define $S \in [0,1]^{m \times n}$ as the similarity scores between each reviewer and paper. An assignment of papers to reviewers is represented as a matrix $A \in \{0,1\}^{m \times n}$, where $A_{rp} = 1$ if reviewer r is assigned to paper p and $A_{rp} = 0$ otherwise. In the standard paper assignment problem, the objective is to find an assignment A of reviewers to papers such that the total similarity $\sum_{r \in \mathcal{R}, p \in \mathcal{P}} A_{rp} S_{rp}$ is maximized, subject to constraints that each paper is assigned exactly a certain load of reviewers, each reviewer is assigned to at most a certain load of papers, and any reviewer-paper pairs with a conflict of interest are not assigned (Charlin and Zemel 2013; Charlin, Zemel, and Boutilier 2011; Goldsmith and Sloan 2007; Flach et al. 2010; Kobren, Saha, and McCallum 2019). In this work, we accommodate conflicts of interest by assuming the corresponding similarities are set to 0. This problem can be formulated as a min-cost flow problem or as a linear program, and can be efficiently solved.

In a two-stage assignment, all papers $\mathcal{P}$ require a certain number of reviewers in the first stage and a subset of papers $\mathcal{P}_2 \subseteq \mathcal{P}$ require additional review in the second stage. We assume that $\mathcal{P}_2$ consists of a fixed fraction β of papers and is drawn uniformly at random from $\mathcal{P}$. Specifically, for some $\beta \in \{\frac{1}{n}, \dots, \frac{n}{n}\}$, we assume that $\mathcal{P}_2 \sim \mathcal{U}_{\beta n}(\mathcal{P})$, the uniform distribution over all subsets of size βn of $\mathcal{P}$. In the two-phase setting, the fraction β itself can be viewed as a parameter that the program chairs set based on available reviewer resources, or it can be estimated from past editions of the conference. Our empirical results detailed in Section 5.1 also cover the case where papers are chosen for the second phase based on their first-phase review scores. In the conference experiment design setting, the value of β and the uniform distribution of $\mathcal{P}_2$ are both experiment design choices. The choice of a uniform distribution for $\mathcal{P}_2$ is common, as in the NeurIPS 2014/2021 and WSDM 2017 experiments. The question we analyze is: how should reviewers be assigned to papers across the two stages?

Before continuing further, we introduce some notation. For subsets $\mathcal{R}' \subseteq \mathcal{R}$ and $\mathcal{P}' \subseteq \mathcal{P}$, desired paper load $\ell_{pap} \in \mathbb{Z}_+$, and maximum reviewer load $\ell_{rev} \in \mathbb{Z}_+$, define $\mathcal{M}(\mathcal{R}', \mathcal{P}'; \ell_{rev}, \ell_{pap}) \subseteq \{0,1\}^{m \times n}$ as the set of valid assignment matrices on $\mathcal{R}'$ and $\mathcal{P}'$. Formally, $A \in \mathcal{M}(\mathcal{R}', \mathcal{P}'; \ell_{rev}, \ell_{pap})$ if and only if $\sum_{r \in \mathcal{R}'} A_{rp} = \ell_{pap}$ for all $p \in \mathcal{P}'$, $\sum_{p \in \mathcal{P}'} A_{rp} \le \ell_{rev}$ for all $r \in \mathcal{R}'$, and $A_{rp} = 0$ for all $(r,p) \notin \mathcal{R}' \times \mathcal{P}'$.

The two-stage paper assignment problem is to maximize the total similarity of the paper assignment across both stages. Without loss of generality, we instead consider the mean similarity so that later results will be easier to interpret. Fix a stage one paper load $\ell_{pap}^{(1)}$, a stage two paper load $\ell_{pap}^{(2)}$, and an overall reviewer load ℓ_{rev} such that $\ell_{pap}^{(1)} n + \ell_{pap}^{(2)} \beta n \le \ell_{rev} m$ (i.e., the number of reviews required by papers is no greater than the number of reviews that can be supplied by reviewers). Given $\mathcal{P}_2$, the oracle optimal assignment has mean similarity

$$
\begin{aligned}
Q^*(\mathcal{P}_2) = \max_{\substack{A \in \mathcal{M}(\mathcal{R}, \mathcal{P}; \ell_{rev}, \ell_{pap}^{(1)}), \\ B \in \mathcal{M}(\mathcal{R}, \mathcal{P}_2; \ell_{rev}, \ell_{pap}^{(2)})}} & \frac{1}{\ell_{pap}^{(1)} n + \ell_{pap}^{(2)} \beta n} \left[\sum_{r \in \mathcal{R}, p \in \mathcal{P}} A_{rp} S_{rp} \right. \\
& \left. + \sum_{r \in \mathcal{R}, p \in \mathcal{P}_2} B_{rp} S_{rp} \right] \\
\text{subject to } & \sum_{p \in \mathcal{P}} A_{rp} + B_{rp} \le \ell_{rev} \qquad \forall r \in \mathcal{R}.
\end{aligned}
$$

The last constraint ensures that each reviewer's assignment across both stages does not exceed the maximum reviewer load. Just like the standard paper assignment problem, the oracle optimal assignment for a given $\mathcal{P}_2$ can be found efficiently. However, in both the two-phase and experiment design settings, this oracle optimal assignment is either unachievable or undesirable. In the two-phase setting, the set of papers $\mathcal{P}_2$ requiring additional review is unknown until after the stage one assignment is chosen. Thus, the oracle optimal assignment cannot be computed beforehand. In the experiment design setting, the assignment of reviewers to conditions is commonly randomized in order to gain statistical power, as was done in the WSDM 2017 and NeurIPS 2014/2021 experiments. Thus, a deterministic choice of assignment may not be desirable. Additionally, depending on the experiment setup, it may not be possible for a reviewer to review papers in both conditions. In what follows, we use this oracle optimal assignment value as an unachievable baseline for comparison.

We instead consider simple strategies for the two-stage assignment problem that choose a subset $\mathcal{R}_2 \subseteq \mathcal{R}$ of reviewers to save for the second stage without observing $\mathcal{P}_2$, leaving reviewers $\mathcal{R}_1 = \mathcal{R} \setminus \mathcal{R}_2$ to be assigned to papers in the first stage. Unlike the oracle optimal assignment, such strategies are feasible to implement in both settings since they do not require knowledge of $\mathcal{P}_2$, do not split reviewer loads across

conditions, and allow for a random choice of $\mathcal{R}_2$. The mean similarity of the optimal assignment when reviewers $\mathcal{R}_2$ and papers $\mathcal{P}_2$ are in the second stage is

$$Q(\mathcal{R}_2, \mathcal{P}_2) = \frac{1}{\ell_{pap}^{(1)} n + \ell_{pap}^{(2)} \beta n} \left[\max_{\substack{A \in \mathcal{M}(\mathcal{R} \setminus \mathcal{R}_2, \mathcal{P}; \\ \ell_{rev}, \ell_{pap}^{(1)})}} \sum_{r \in \mathcal{R} \setminus \mathcal{R}_2, p \in \mathcal{P}} A_{rp} S_{rp} + \max_{\substack{B \in \mathcal{M}(\mathcal{R}_2, \mathcal{P}_2; \\ \ell_{rev}, \ell_{pap}^{(2)})}} \sum_{r \in \mathcal{R}_2, p \in \mathcal{P}_2} B_{rp} S_{rp} \right].$$

We require that $\ell_{rev}|\mathcal{R}_2| \geq \ell_{pap}^{(2)} \beta n$ and $\ell_{rev}(m - |\mathcal{R}_2|) \geq \ell_{pap}^{(1)} n$ for feasibility in both stages. Given $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{P}_2$, the optimal paper assignment in each stage can be efficiently computed using standard methods. Thus, the the difficulty of the problem lies entirely in choosing $\mathcal{R}_2$.

The expected mean similarity of the optimal assignment when saving reviewers $\mathcal{R}_2$ for the second stage is

$$f(\mathcal{R}_2) = \mathbb{E}_{\mathcal{P}_2 \sim \mathcal{U}_{\beta n}(\mathcal{P})} \left[Q(\mathcal{R}_2, \mathcal{P}_2) \right].$$

We can also evaluate the *suboptimality* of $\mathcal{R}_2$ as compared to the oracle optimal assignment as

$$Q^*(\mathcal{P}_2) - Q(\mathcal{R}_2, \mathcal{P}_2), \qquad \text{where } \mathcal{P}_2 \sim \mathcal{U}_{\beta n}(\mathcal{P}).$$

Note that Q^* and Q are bounded in $[0, 1]$, so that both f and the suboptimality are also bounded in $[0, 1]$.

In our theoretical analysis, for simplicity, we assume that $\ell_{rev} = \ell_{pap}^{(1)} = \ell_{pap}^{(2)} = 1$, leaving this implicit in f, Q, and Q^* throughout the paper. We also assume $m = (1 + \beta)n$ in our analysis unless specified otherwise, so that $|\mathcal{R}_2| = \beta n$. The intuition behind our results carries over to the cases of general loads and excess reviewers, which are covered by our empirical results in Section 5.1. Although we do not extend our theoretical results to formally handle these cases, we do not believe that doing so would provide any additional practical insights for program chairs. All asymptotic bounds are given as n grows.

4 Hardness

In the two-phase setting, the oracle optimal assignment is unachievable because $\mathcal{R}_2$ must be chosen before observing $\mathcal{P}_2$. Therefore, conferences must choose $\mathcal{R}_2$ to maximize f, the expected mean similarity of the assignment across both stages. In this section, we demonstrate that maximizing a variant of f is NP-hard, indicating that it is unlikely that f can be optimized efficiently.

First, note that evaluating $f(\mathcal{R}_2)$ requires computing an expectation over the draw of $\mathcal{P}_2$, which naively requires evaluating a sum over the optimal assignment value for $\binom{n}{\beta n}$ possible choices of $\mathcal{P}_2$. This number is exponential in the input size, so an efficient algorithm for this problem would have to either optimize f without evaluating it or compute this expectation without computing the optimal assignment for each possible $\mathcal{P}_2$.

Instead of attempting to optimize f exactly, a standard approach from two-stage stochastic optimization is to simplify the problem by sampling as follows (Dai, Chen, and Birge 2000; King and Rockafellar 1993). First, take some fixed number of samples $\mathcal{P}_2^{(1)}, \ldots, \mathcal{P}_2^{(K)}$ from $\mathcal{U}_{\beta n}(\mathcal{P})$. Then, rather than optimizing an average over all $\mathcal{P}_2$ in the support of $\mathcal{U}_{\beta n}(\mathcal{P})$, choose $\mathcal{R}_2$ to optimize an average over only all sampled sets:

$$\overline{f}(\mathcal{R}_2) = \frac{1}{K} \sum_{i=1}^{K} Q(\mathcal{R}_2, \mathcal{P}_2^{(i)}).$$

This is a natural simplification of the two-stage paper assignment problem, because the sum in the objective is now taken over only a constant K subsets rather than an exponential number. However, this problem is still not efficiently solvable, as the following theorem shows.

Theorem 1. *It is NP-hard to find $\mathcal{R}_2 \subseteq \mathcal{R}$ such that $\overline{f}(\mathcal{R}_2)$ is maximized, even when $K = 3$.*

Proof sketch. We reduce from 3-Dimensional Matching (Karp 1972), which asks if there exists a way to select k tuples from a set $T \subseteq X \times Y \times Z$ where $|X| = |Y| = |Z| = k$ such that all elements of X, Y, and Z are selected exactly once. We construct 3 samples of second-stage papers corresponding to X, Y, and Z respectively, and construct reviewers corresponding to elements of T. These reviewers have 1 similarity with the papers in their tuple, and 0 similarity with all other papers. Thus, checking if there exists a choice of $\mathcal{R}_2$ which gives full expected similarity in the second stage would require solving 3-Dimensional Matching. We add additional reviewers and papers to ensure that this choice of $\mathcal{R}_2$ is optimal over both stages. □

The full proof is presented in Appendix D.1.

Since it is NP-hard to find the optimal $\mathcal{R}_2$ even when estimating the objective by sampling three random choices of $\mathcal{P}_2$, this suggests that the original objective f may be hard to optimize efficiently. Therefore, in the two-phase assignment setting, we instead look for efficient approximation algorithms.

5 Our Approach: Random Split

Our proposed approach for finding a two-stage assignment is extremely simple: choose $\mathcal{R}_2$ uniformly at random (i.e., $\mathcal{R}_2 \sim \mathcal{U}_{\beta n}(\mathcal{R})$). We refer to this as a "random split" of reviewers into the two review stages.

In the two-phase setting, random split is an efficient approximation algorithm for the problem of optimizing f, which is likely difficult (as shown in Section 4). Because random split does not execute f, it produces a two-stage paper assignment without needing to estimate f by sampling.

In the conference experiment design setting, our proposed random-split strategy corresponds to a uniform random choice of reviewers for the experimental condition. Recall that in this setting, assigning reviewers to conditions uniformly at random is already a common experimental setup. The performance of random split on f therefore indicates how

well this common setup performs in terms of the expected assignment similarity.

In our theoretical results, we often refer to the *suboptimality of random split*, defined as the suboptimality of $\mathcal{R}_2$ chosen via random split when $\mathcal{P}_2$ is chosen uniformly at random:

$$Q^*(\mathcal{P}_2) - Q(\mathcal{R}_2, \mathcal{P}_2), \quad \text{where } \mathcal{P}_2 \sim \mathcal{U}_{\beta n}(\mathcal{P}), \mathcal{R}_2 \sim \mathcal{U}_{\beta n}(\mathcal{R}). \tag{1}$$

Recall from Section 3 that $Q^*(\mathcal{P}_2)$ is the mean similarity of the oracle optimal assignment given second-stage papers $\mathcal{P}_2$ and that $Q(\mathcal{R}_2, \mathcal{P}_2)$ is the mean similarity of the optimal assignment given second-stage reviewers and papers $\mathcal{R}_2, \mathcal{P}_2$. Additionally, many of our results evaluate the expected mean similarity under random split:

$$\mathbb{E}_{\mathcal{R}_2 \sim \mathcal{U}_{\beta n}(\mathcal{R})}\left[f(\mathcal{R}_2)\right] = \mathbb{E}_{\mathcal{R}_2 \sim \mathcal{U}_{\beta n}(\mathcal{R}), \mathcal{P}_2 \sim \mathcal{U}_{\beta n}(\mathcal{P})}\left[Q(\mathcal{R}_2, \mathcal{P}_2)\right].$$

In the following subsections, we first elaborate on the good performance random split displays empirically before showing that there exist cases where random split performs very poorly.

5.1 Empirical Performance

As introduced earlier in Figure 1, random split performs very well in practice on four real conference similarity matrices. The first is a similarity matrix recreated using text-matching on ICLR 2018 data (Xu et al. 2019). The second is constructed using reviewer bid data for an AI conference (conference 3) from PrefLib dataset MD-00002 (Mattei and Walsh 2013). The third (denoted Bid1) is a sample of the bidding data from a major computer science conference (Meir et al. 2020). In both of these bidding datasets, we transformed "yes," "maybe," and "no response" bids into similarities of 1, 0.5, and 0.25 respectively, as is often done in practice (Shah et al. 2018). The fourth similarity matrix is constructed from the subject areas of ACM SIGIR 2007 papers and the subject areas of the past work of their authors (assumed to be the reviewers) (Karimzadehgan, Zhai, and Belford 2008); we set the similarity between each reviewer and paper to be equal to the number of matching subject areas out of the 25 total, normalized so that each entry of the matrix is in $[0, 1]$. In Appendix A, we present further empirical results including additional datasets. In Appendix B, we present additional empirical results particularly relevant to the conference experiment design setting.

We run several experiments, each corresponding to a choice of dataset and β. Each experiment consists of 10 trials, where in each trial we sample a random reviewer split and a set of second-stage papers. We then present the range of assignment values achieved across the trials as percentages of the oracle optimal assignments for each trial. The oracle optimal assignment for a trial is found by choosing the optimal assignment of reviewers across both stages after observing $\mathcal{P}_2$. We set paper loads of 2 in each stage (as done in AAAI 2021), and limit reviewer loads to be at most 6 (a realistic reviewer load (Shah et al. 2018)). Since these datasets have excess reviewers, we choose $\mathcal{R}_2$ to have size $\frac{\beta}{1+\beta}m$ so that the proportions of reviewers and papers in the second stage are equal.

In Figure 1a, $\mathcal{P}_2$ is drawn uniformly at random in each trial (as in the problem formulation). We see that all trials of random split achieve at least 90% of the oracle optimal solution's similarity on all datasets, with all trials on all but two experiments achieving at least 94%. We see additionally that the randomness of the reviewer choice does not cause much variance in the value of the assignment, as there is at most a 5% difference between the minimum and maximum similarity (as a percentage of oracle optimal) for each experiment. Note that this is true despite the fact that the similarity matrices of the different datasets are constructed in several different ways, indicating that random split is robust across methods of similarity construction.

In Figure 1b, $\mathcal{P}_2$ is chosen as a fixed set for all trials based on the actual review scores received by the papers at ICLR 2018 (He 2020) (as review scores were not available for other datasets). We run trials where either the top-scoring papers or the messy-middle papers are given additional reviews. Since about 37% of papers were accepted, we define the messy middle as the range of $\frac{\beta}{1+\beta}m$ papers centered on the 63rd-percentile paper when ordered by score. These are sets of papers that a conference may potentially want to assign additional reviewers to. In all cases, random split shows consistently good performance, similar to when $\mathcal{P}_2$ was drawn uniformly at random. All trials achieve at least 95% of the oracle optimal similarity, with at most a 2% difference between the minimum and maximum for each experiment. This suggests that the good performance of random reviewer split naturally holds in these practical cases.

5.2 A Counterexample

The good results random split shows in practice may be somewhat surprising because random split does not perform well in all settings. The following theorem shows that for any β, there exist instances of the two-stage paper assignment problem where the suboptimality of random split (1) is $\Omega(1)$ in expectation.

Theorem 2. *For any constant $\beta \in [0, 1]$, there exists n_0 such that for all $n \geq n_0$ where $\beta n \in \mathbb{Z}_+$, there exist instances of the two-stage paper assignment problem where the suboptimality of random split is at least $\frac{\beta^4}{(1+\beta)^3}$ in expectation.*

Proof sketch. Consider $\beta = 1$. We construct a similarity matrix where every reviewer has similarity 1 with only 1 paper, and all papers have similarity 1 with only 2 reviewers. The optimal reviewer split puts the two good reviewers for each paper in separate stages and always achieves a mean similarity of 1. Random split puts both good reviewers in the same stage with at least constant probability for each paper, giving a constant mean similarity < 1. □

The full proof is presented in Appendix D.2.

Note that the above lower bound on the objective value of random split holds even in the easy case of $\beta = 1$, where the problem could be solved simply through standard paper assignment methods. This case is particularly relevant

in the conference experiment setting, where all papers are commonly reviewed under both conditions (as in the WSDM 2017 experiment).

Although the above lower bound demonstrates that random split cannot hope to do well in general, the constructed example is unrealistic for real conferences. However, program chairs may understandably want some guarantee that a random reviewer split will work well for their conference before deciding to use it. Ideally, this guarantee should be given before the precise similarity matrix for the conference is known, since the similarities may not be known in full until late in the planning process.

In the following sections, we provide such guarantees, thereby showing that the good performance of random split is not just an artifact of our specific datasets. We focus our attention on two sufficient conditions on the similarity matrix under which we show random split performs well. These conditions are natural for real similarity matrices, implying that random split will perform well for many real conferences, whether in the context of a two-phase review process or a conference experiment. Using these conditions, we provide actionable insights to program chairs based on simple properties of their conference's similarities that they may have intuition about. These insights are designed to be useful well before the full paper and reviewer sets are known.

6 Condition 1: Low-Rank Similarity Matrix

The first condition we consider is that the similarity matrix S has low rank k. This condition naturally arises in practice when reviewer-paper similarities are calculated from the number of subject area agreements between reviewers and papers; in such cases, the rank is no greater than the number of subject areas. For example, the SIGIR similarity matrix used in Figure 1 is constructed in this way and thus has rank no greater than 25 (the number of subject areas). In this section, we provide asymptotic upper and lower bounds on the suboptimality of random split for constant-rank similarity matrices.

6.1 Theoretical Bounds

We first provide an upper bound on the suboptimality of random split (1). This shows that random reviewer splits perform well on constant-rank similarity matrices, including the SIGIR similarity matrix examined earlier. More precisely, the following theorem shows that if the similarity matrix S has constant rank k, the suboptimality of random split is at most $\widetilde{O}(n^{-\frac{1}{2}})$ when $k = 1$, $\widetilde{O}(n^{-\frac{1}{2}+o(1)})$ when $k = 2$, and $\widetilde{O}(n^{-\frac{1}{k}+o(1)})$ when $k \geq 3$ with high probability.

Theorem 3. *Consider any constants $\beta \in [0, 1]$ and $k \in \mathbb{Z}_+$. There exists n_0 and constants C, η such that, for any $n \geq n_0$ where $\beta n \in \mathbb{Z}_+$ and for any similarity matrix $S \in [0, 1]^{(1+\beta)n \times n}$ of rank k, the suboptimality of random split is at most:*

- $C(\log n)^{\eta} n^{-\frac{1}{2}}$ *if* $k = 1$
- $C(\log n)^{\eta} n^{-\frac{1}{k}+\frac{1}{\log\log n}}$ *if* $k \geq 2$

with probability at least $1 - \frac{1}{n}$ (where log *indicates the base-*2 *logarithm).*

Proof sketch. By Lemma 4 of (Rothvoss 2014), a rank k similarity matrix $S \in [0, 1]^{m \times n}$ can be factored into vectors $u_r \in \mathbb{R}^k$ for each $r \in \mathcal{R}$ and $v_p \in \mathbb{R}^k$ for each $p \in \mathcal{P}$ such that $S_{rp} = \langle u_r, v_p \rangle$, $||u_r||_2 \leq k^{1/4}$, and $||v_p||_2 \leq k^{1/4}$. We cover the k-dimensional ball containing all paper vectors with smaller cells, and consider a reviewer to be in one of these cells if the oracle optimal assignment (given $\mathcal{P}_2$) assigns it to a paper in that cell. Using a concentration inequality on the number of reviewers and papers in each cell in each stage, we can upper bound the number of reviewers that we cannot match to papers within the same cell. We then increase the size of the cells and attempt to match the remaining reviewers in this way, continuing until all reviewers are matched. We upper bound the suboptimality of the resulting assignment by the L2 distance between a reviewer's assigned paper and the paper they are assigned by the oracle optimal assignment. □

The constants C and η may depend on k, which is itself assumed to be constant. The full proof is presented in Appendix D.3.

For constant-rank similarity matrices, the suboptimality diminishes as n grows, unlike when the rank of the similarity matrix is unrestricted. Conceptually, our proof technique of finding a minimum-distance matching between two samples of points resembles the optimal transport problem solved when finding the Wasserstein distance between a probability distribution and its empirical measure. Thus, our upper bounds nearly match those found in the literature on the expected empirical 1-Wasserstein distance for continuous measures (see (Panaretos and Zemel 2019) and references therein).

We now complement the above upper bound with lower bounds on the suboptimality of random split (1) for constant rank similarity matrices. The following theorem shows that, for similarity matrices of constant rank k, the suboptimality of random split is $\Omega(n^{-1/2})$ in expectation and $\Omega(n^{-2/k})$ with high probability.

Theorem 4. *Suppose $\beta = 1$. For any constant $k \in \mathbb{Z}_+$, there exists n_0 and constants C, ζ such that for all $n \geq n_0$:*

(a) There exist instances of the two-stage paper assignment problem with similarity matrices $S \in [0, 1]^{2n \times n}$ of rank k such that the suboptimality of random split is at least $Cn^{-1/2}$ in expectation.

(b) There exist instances of the two-stage paper assignment problem with similarity matrices $S \in [0, 1]^{2n \times n}$ of rank k such that the suboptimality of random split is at least $Cn^{-2/k}$ with probability $1 - \zeta e^{-n/10}$.

Proof sketch. (a) We construct k groups of reviewers and papers, where all reviewers and papers in the same group have similarity 1 with each other and similarity 0 with all other reviewers/papers. The first group contains $\frac{n}{2}$ papers and n reviewers. The optimal reviewer split puts half of each group's reviewers in each stage and assigns all reviewers to papers with similarity 1. By an anti-concentration inequality, with constant probability, at least $\Omega(\sqrt{n})$ reviewers in the first group cannot be assigned to a paper in their group under random split.

(b) We construct a vector in $\mathbb{R}^k$ for each reviewer and each paper and set the similarity between that reviewer and that paper to be the inner product of their corresponding vectors. We place one paper vector and two reviewer vectors at each point in an evenly-spaced grid throughout the cube $[0, 1/\sqrt{k}]^k$. The resulting similarity matrix has rank k. The optimal assignment assigns the two reviewers at each point to the paper at that point. With high probability, random split places $\Omega(n)$ pairs of reviewer vectors into the same stage. One of each of these reviewer pairs must be assigned to a paper at a different point, which is at least $\Omega(n^{-1/k})$ away in L2 distance. The suboptimality of the resulting assignment can be written in terms of the total squared L2 distance between each reviewer and their assigned paper, giving the stated bound. □

The constants C and ζ may depend on k, which is itself assumed to be constant. The full proof is presented in Appendix D.4.

6.2 Interpretation of Results

As discussed earlier in this section, certain methods of constructing similarities (such as counting subject area agreements) may inherently lead to low-rank similarity matrices. If a conference is using such a method, the results in this section provide guarantees to the program chairs that random split will perform well, particularly if the rank of the matrix is low compared to the number of papers and reviewers. Alternatively, program chairs may be able to estimate that their reviewers and papers can be grouped into a small number of communities with little variation within them, in which case the similarity matrix may also be low rank.

7 Condition 2: High-Value, Large-Load Assignment

A natural condition on the similarity matrix to consider is that each paper has a large number μ of reviewers with high similarity for that paper. It turns out that this condition is insufficient for guaranteeing good performance of random split, since the same group of μ reviewers could have high similarity with all papers, thus satisfying this condition without changing the assignment value by much (since we can only assign these reviewers to a few papers). In this section, we consider a condition on the similarity matrix that is similar in spirit: the existence of a high-value assignment (in terms of total similarity) on the full reviewer and paper sets where each paper is assigned a large number $(1 + \beta)\mu$ of reviewers. Our proposed condition handles the issue with the naive "large number of reviewers" condition by requiring that the high-value reviewers for each paper can all be simultaneously assigned.

In the following subsections, we first provide theoretical guarantees about the performance of random split under this condition. We then demonstrate that this condition helps to explain the good performance of random split on the real similarity matrices presented earlier.

7.1 Theoretical Bounds

The first result of this section gives a lower bound on the expected value of random split in terms of the value of a single, large-load assignment. All results in this section still hold if there are excess reviewers (i.e., if $m \geq (1 + \beta)n$ and $\mathcal{R}_2 \sim \mathcal{U}_{\frac{\beta}{1+\beta}m}(\mathcal{R})$).

Theorem 5. *Consider any $\mu \in [10, 000]$ and $\beta \in \left\{\frac{1}{100}, \ldots, \frac{100}{100}\right\}$ such that $\beta\mu \in \mathbb{Z}_+$. If there exists an assignment $A^{(\mu)} \in \mathcal{M}(\mathcal{R}, \mathcal{P}; \mu, (1 + \beta)\mu)$ with mean similarity $s^{(\mu)}$, choosing $\mathcal{R}_2$ via random split gives that*

$$\mathbb{E}_{\mathcal{R}_2}\left[f(\mathcal{R}_2)\right] \geq s^{(\mu)}\left[1 - \sqrt{\frac{\beta}{2\pi(1+\beta)^2\mu}}\left(2\sqrt{\frac{1}{1+\beta}} + \sqrt{1-\beta}\right)\right].$$

A similar bound holds when $\beta\mu$ is not integral, with some additional small terms due to rounding.

Proof sketch. We construct assignments with paper and reviewer loads of at most μ in stage one and at most $\beta\mu$ in stage two using the reviewer-paper pairs assigned by $A^{(\mu)}$. We drop any extra assignments at random so that no reviewers and papers are overloaded, and assume any pairs that must be assigned from outside of $A^{(\mu)}$ have similarity 0. From within each of these larger assignments, we can find an assignment with paper and reviewer loads of 1 with at least the same mean similarity. The expected mean similarity of these assignments can be written as the expectation of a function of binomial random variables. Approximating these by normal random variables and checking via simulation that this is in fact a lower bound for the stated values of β and μ, we get the stated bound. □

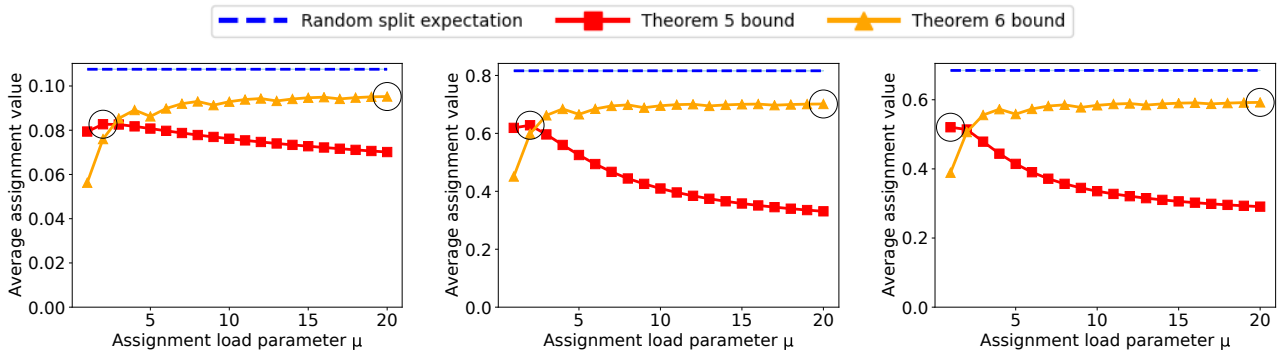
The more general version of the bound and the full proof are stated in Appendix D.5.

The above bound works well when the reviewer-paper pairs in the large-load assignment are all nearly equally valuable. However, it cannot take advantage of the fact that certain reviewers may be extremely valuable for a certain paper and can be prioritized for assignment to that paper when possible. The next result uses additional information about the value of an assignment with smaller loads, along with a large-load assignment disjoint from the small assignment, to make use of these highly valuable reviewer-paper pairs in the case where $\beta = 1$. Recall from Section 5.2 that $\beta = 1$ is still not an easy case for random split in general and is particularly relevant for the conference experiment setting.

Theorem 6. *Suppose $\beta = 1$, and consider any $\mu \in [10, 000]$ such that $\frac{\mu}{4} \in \mathbb{Z}_+$. Suppose there exists an assignment $A^{(1)} \in \mathcal{M}(\mathcal{R}, \mathcal{P}; 1, 2)$ with mean similarity $s^{(1)}$. Suppose there also exists an assignment $A^{(\mu)} \in \mathcal{M}(\mathcal{R}, \mathcal{P}; \mu, 2\mu)$ with mean similarity $s^{(\mu)}$ that does not contain any of the pairs assigned in $A^{(1)}$. Then, choosing $\mathcal{R}_2$ via random split gives that*

$$\mathbb{E}_{\mathcal{R}_2}\left[f(\mathcal{R}_2)\right] \geq \frac{3}{4}s^{(1)} + \left(1 - \frac{1.44}{\sqrt{\mu}}\right)\frac{1}{4}s^{(\mu)}.$$

A similar bound holds when $\frac{\mu}{4}$ is not integral, with some additional small terms due to rounding.



(a) ICLR (b) PrefLib3 (c) Bid1

Figure 2: Performance of the "high-value large-load" bounds on real conference datasets, $\beta = 1$. On the x-axis we vary the parameter μ, which determines the loads of the assignment $A^{(\mu)}$ used in the bound. The best setting of μ for each bound is circled.

Proof sketch. We first attempt to assign as many pairs as possible from within $A^{(1)}$; in expectation we can assign $\frac{3}{4}$ of them. Among the remaining reviewers and papers, we attempt to construct assignments with paper and reviewer loads of $\frac{\mu}{4}$ in both stages from within the reviewer-paper pairs assigned by $A^{(\mu)}$. This is done in a similar way as in Theorem 5. □

The more general version of the bound and the full proof are stated in Appendix D.6.

If we consider $A^{(1)}$ as the optimal assignment and assume that μ is divisible by 4, we get an approximation ratio (between the random split assignment and oracle optimal assignment's similarities) of $\frac{3}{4} + \frac{\gamma_\mu}{4}\left(1 - \frac{1.44}{\sqrt{\mu}}\right)$ where $\gamma_\mu = \frac{s^{(\mu)}}{s^{(1)}}$. With $\mu = 8$, we achieve an approximation ratio of at least $\frac{3}{4} + \frac{\gamma_8}{8}$. Additionally, if $\gamma_\mu \to 1$ as n grows for any $\mu = \omega(1)$, the suboptimality of random split (1) approaches 0. For example, this means that the suboptimality of random split approaches 0 as n grows if the mean similarity of an assignment with paper loads of $\log n$ improves faster than the mean similarity of the optimal assignment.

7.2 Empirical Evaluation

We now show the performance of these bounds on our real conference datasets in order to evaluate the extent to which they explain the good performance of random split. We use three of the conference datasets introduced earlier with $\beta = 1$. In Appendix A, we evaluate the bounds on additional datasets (including the SIGIR dataset). On PrefLib3 and Bid1, the problem is infeasible with paper and reviewer load constraints of 1 since $m < 2n$, so we modify the datasets by splitting each reviewer into 3 copies as follows. For each paper, we arbitrarily give one of the copies the same similarity as the original reviewer and give the other copies similarity 0. In this way, the similarity of the optimal assignment on this modified dataset is no greater than the similarity of the optimal assignment on the original dataset.

In Figure 2, we vary the value of the parameter μ (indicating the loads of the assignment $A^{(\mu)}$) and show the bounds of Theorem 5 and Theorem 6 as compared to the estimated expected value of random split. The estimated expected value is averaged over 10 trials with the standard error of the mean shaded, although it is sometimes not visible because it is small. We see that on these datasets, the bound of Theorem 5 performs best for low values of μ and not very well for higher values, likely due to the presence of a few "star" reviewers for each paper which hold a lot of the value. By making use of extra information about the values of these reviewers, the bound of Theorem 6 achieves a high fraction of the actual random split value. Although this bound is maximized at large values of μ on these datasets, it is close to its maximum even with reasonably low values of μ. For example, on ICLR, the lower bound achieves 86% of the estimated expected value of random split with $\mu = 8$. This indicates the good performance of random split is explained well by the presence of just a few good reviewers per paper that can be simultaneously assigned.

7.3 Interpretation of Results

Although our results in this section are stated in terms of the precise values of high-load assignments, they can be interpreted by program chairs in a simple and practical way. Roughly, our results indicate that if several good reviewers can be *simultaneously* assigned to each paper (as was the case for the three conference similarity matrices in Figure 2), random split will perform well. When considering the potential performance of randomly splitting reviewers, program chairs should consider the reviewer and paper pools they expect to have at their conference and make a judgement about how many good-quality reviewers they think could be assigned to each paper (if the reviewer loads are scaled up proportionately). For example, the program chairs of a large AI conference might be confident that the top several reviewers for each paper are about equally valuable (due to the depth of the reviewer pool) and could be assigned to each paper with only a modest loss in average review quality; this

would imply that random split would perform very well for this conference.

8 Conclusion

We showed that randomly splitting reviewers between two reviewing phases or two reviewing conditions produces near-optimal assignments on realistic conference similarity matrices. Our analysis of this phenomenon can help future program chairs make decisions about whether random split will work well for their conference's two-phase review process, based on their assessment of whether a few simple conditions are applicable to their case. In the setting of conference experiment design, our analysis allows program chairs to understand if running an experiment on their review process will significantly impact their assignment quality.

In addition, our results can potentially be further generalized to related reviewing models such as those of academic journals (which accept submissions on a rolling basis), or to other multi-stage resource allocation problems that involve matching resources based on similarities. For example, datacenters receiving a large batch of jobs may have to select some to run on various servers immediately and some to run later when additional servers have been freed, or hospitals may want to assign nurses to shifts based on expertise but without knowledge of which expertise will be most applicable in later shifts. Within the field of human computation, researchers may want to run experiments in real-world task-assignment settings where maintaining a high-quality assignment of tasks to participants is important.

One limitation of our work is that while our empirical results demonstrate the effectiveness of the random-split strategy with real conference data, our theoretical results make the simplifying assumption that paper and reviewer loads are 1, which is unrealistic for real conferences. However, we believe that incorporating this detail would not change our explanations for the good performance of random split. Another limitation is that we assume the set of papers requiring reviews in the second stage is drawn uniformly at random. Although this is a reasonable belief without further information in the two-phase setting, one direction for future work is to consider non-uniform distributions of second-stage papers and analyze if a form of random split still performs well there.

Our work could potentially produce negative outcomes in the form of worse paper assignments if program chairs decide to use random split on an incorrect belief that their conference will fit our conditions. However, program chairs are required to make such decisions about how to perform the paper assignment anyway, so this is not a significant increase in risk. The use of random reviewer splits, as opposed to some alternate strategy where reviewers can self-select their stage, could also negatively impact reviewers with strong preferences over which stage they review in (e.g., due to schedule constraints). These preferences should ideally be taken into account along with the similarity of the resulting assignment when choosing the reviewer split; we leave this as an interesting direction for future work.

Acknowledgments

This work was supported by NSF CAREER awards 1942124 and 2046640, NSF grant CIF-1763734, NSF grants IIS-1850477 and IIS-1814056, and a Google Research Scholar Award.

References

Ailamaki, A.; Chrysogelos, P.; Deshpande, A.; and Kraska, T. 2019. The SIGMOD 2019 Research Track Reviewing System. *ACM SIGMOD Record*, 48(2): 47–54.

Alon, N.; Fischer, F.; Procaccia, A.; and Tennenholtz, M. 2011. Sum of Us: Strategyproof Selection From the Selectors. In *Conference on Theoretical Aspects of Rationality and Knowledge*, 101–110.

Armstrong, J. S. 1980. Unintelligible Management Research and Academic Prestige. *Interfaces*, 10(2): 80–86.

Aziz, H.; Lev, O.; Mattei, N.; Rosenschein, J. S.; and Walsh, T. 2019. Strategyproof Peer Selection Using Randomization, Partitioning, and Apportionment. *Artificial Intelligence*, 275: 295–309.

Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W. 2021. The NeurIPS 2021 Consistency Experiment. https://blog.neurips.cc/2021/12/08/the-neurips-2021-consistency-experiment/. Accessed January 25, 2022.

Brubach, B.; Sankararaman, K. A.; Srinivasan, A.; and Xu, P. 2016. New Algorithms, Better Bounds, and a Novel Model for Online Stochastic Matching. In *24th Annual European Symposium on Algorithms*.

Buchbinder, N.; Feldman, M.; Naor, J.; and Schwartz, R. 2014. Submodular Maximization with Cardinality Constraints. In *ACM-SIAM Symposium on Discrete Algorithms*, 1433–1452. SIAM.

Cabanac, G.; and Preuss, T. 2013. Capitalizing on Order Effects in the Bids of Peer-Reviewed Conferences to Secure Reviews by Expert Referees. *Journal of the Association for Information Science and Technology*, 64(2): 405–415.

Ceci, S. J.; and Peters, D. P. 1982. Peer Review: A Study of Reliability. *Change: The Magazine of Higher Learning*, 14(6): 44–48.

Charlin, L.; Zemel, R.; and Boutilier, C. 2011. A Framework for Optimizing Paper Matching. In *27th Conference on Uncertainty in Artificial Intelligence*, 86–95.

Charlin, L.; and Zemel, R. S. 2013. The Toronto Paper Matching System: An Automated Paper-Reviewer Assignment system. In *ICML Workshop on Peer Reviewing and Publishing Models*.

Dai, L.; Chen, C.-H.; and Birge, J. R. 2000. Convergence Properties of Two-Stage Stochastic Programming. *Journal of Optimization Theory and Applications*, 106(3): 489–509.

Dickerson, J.; Sankararaman, K.; Srinivasan, A.; and Xu, P. 2018. Allocation Problems in Ride-Sharing Platforms: Online Matching With Offline Reusable Resources. In *AAAI Conference on Artificial Intelligence*, volume 32.

Dickerson, J. P.; Procaccia, A. D.; and Sandholm, T. 2012. Dynamic Matching via Weighted Myopia With Application

to Kidney Exchange. In *AAAI Conference on Artificial Intelligence*.

Escoffier, B.; Gourvès, L.; Monnot, J.; and Spanjaard, O. 2010. Two-Stage Stochastic Matching and Spanning Tree Problems: Polynomial Instances and Approximation. *European Journal of Operational Research*, 205(1): 19–30.

Feldman, J.; Mehta, A.; Mirrokni, V.; and Muthukrishnan, S. 2009. Online Stochastic Matching: Beating 1-1/e. In *50th Annual IEEE Symposium on Foundations of Computer Science*, 117–126. IEEE.

Feng, Y.; Niazadeh, R.; and Saberi, A. 2021. Two-Stage Stochastic Matching With Application to Ride Hailing. In *ACM-SIAM Symposium on Discrete Algorithms*, 2862–2877. SIAM.

Fiez, T.; Shah, N.; and Ratliff, L. 2020. A SUPER* Algorithm to Optimize Paper Bidding in Peer Review. In *36th Conference on Uncertainty in Artificial Intelligence*.

Flach, P. A.; Spiegler, S.; Golénia, B.; Price, S.; Guiver, J.; Herbrich, R.; Graepel, T.; and Zaki, M. J. 2010. Novel Tools to Streamline the Conference Review Process: Experiences from SIGKDD'09. *SIGKDD Explorations Newsletter*, 11(2): 63–67.

Garg, N.; Kavitha, T.; Kumar, A.; Mehlhorn, K.; and Mestre, J. 2010. Assigning Papers to Referees. *Algorithmica*, 58(1): 119–136.

Goldsmith, J.; and Sloan, R. 2007. The AI Conference Paper Assignment Problem. *AAAI Workshop*, WS-07-10: 53–57.

He, H. 2020. OpenReview Explorer. https://github.com/Chillee/OpenReviewExplorer. Accessed May 26, 2021.

Jecmen, S.; Zhang, H.; Liu, R.; Shah, N. B.; Conitzer, V.; and Fang, F. 2020. Mitigating Manipulation in Peer Review via Randomized Reviewer Assignments. In *Advances in Neural Information Processing Systems*.

Karimzadehgan, M.; Zhai, C.; and Belford, G. 2008. Multi-Aspect Expertise Matching for Review Assignment. In *17th ACM Conference on Information and Knowledge Management*, 1113–1122.

Karp, R. M. 1972. Reducibility Among Combinatorial Problems. In *Complexity of Computer Computations*, 85–103. Springer.

Karp, R. M.; Vazirani, U. V.; and Vazirani, V. V. 1990. An Optimal Algorithm for Online Bipartite Matching. In *22nd Annual ACM Symposium on Theory of Computing*, 352–358.

Katriel, I.; Kenyon-Mathieu, C.; and Upfal, E. 2008. Commitment Under Uncertainty: Two-Stage Stochastic Matching Problems. *Theoretical Computer Science*, 408(2-3): 213–223.

King, A. J.; and Rockafellar, R. T. 1993. Asymptotic Theory for Solutions in Statistical Estimation and Stochastic Programming. *Mathematics of Operations Research*, 18(1): 148–162.

Kobren, A.; Saha, B.; and McCallum, A. 2019. Paper Matching with Local Fairness Constraints. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Kong, N.; and Schaefer, A. J. 2006. A Factor 1/2 Approximation Algorithm for Two-Stage Stochastic Matching Problems. *European Journal of Operational Research*, 172(3): 740–746.

Lawrence, N. D. 2014. The NIPS Experiment. https://inverseprobability.com/2014/12/16/the-nips-experiment. Accessed May 17, 2021.

Lee, E.; and Singla, S. 2020. Maximum Matching in the Online Batch-Arrival Model. *ACM Transactions on Algorithms*, 16(4): 1–31.

Liu, X.; Suel, T.; and Memon, N. 2014. A Robust Model for Paper Reviewer Assignment. In *8th ACM Conference on Recommender Systems*, RecSys '14, 25–32. New York, NY, USA: ACM. ISBN 978-1-4503-2668-1.

Long, C.; Wong, R.; Peng, Y.; and Ye, L. 2013. On Good and Fair Paper-Reviewer Assignment. In *IEEE International Conference on Data Mining*, 1145–1150. ISBN 978-0-7695-5108-1.

Madden, S.; and DeWitt, D. 2006. Impact of Double-Blind Reviewing on SIGMOD Publication Rates. *ACM SIGMOD Record*, 35(2): 29–32.

Manzoor, E.; and Shah, N. B. 2021. Uncovering Latent Biases in Text: Method and Application to Peer Review. In *AAAI Conference on Artificial Intelligence*.

Mattei, N.; and Walsh, T. 2013. PrefLib: A Library of Preference Data HTTP://PREFLIB.ORG. In *3rd International Conference on Algorithmic Decision Theory*, Lecture Notes in Artificial Intelligence. Springer.

Meir, R.; Lang, J.; Lesca, J.; Kaminsky, N.; and Mattei, N. 2020. A Market-Inspired Bidding Scheme for Peer Review Paper Assignment. In *Games, Agents, and Incentives Workshop at AAMAS*.

Mimno, D.; and McCallum, A. 2007. Expertise Modeling for Matching Papers with Reviewers. In *13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, 500–509. New York, NY, USA: ACM. ISBN 978-1-59593-609-7.

Panaretos, V. M.; and Zemel, Y. 2019. Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Application*, 6: 405–431.

Patat, F.; Kerzendorf, W.; Bordelon, D.; Van de Ven, G.; and Pritchard, T. 2019. The Distributed Peer Review Experiment. *The Messenger*, 177: 3–13.

Pier, E.; Raclaw, J.; Kaatz, A.; Brauer, M.; Carnes, M.; Nathan, M.; and Ford, C. 2017. Your Comments Are Meaner Than Your Score: Score Calibration Talk Influences Intra- and Inter-Panel Variability During Scientific Grant Peer Review. *Research Evaluation*, 26(1): 1–14.

Price, E. 2014. The NIPS Experiment. http://blog.mrtz.org/2014/12/15/the-nips-experiment.html. Accessed May 17, 2021.

Rodriguez, M. A.; and Bollen, J. 2008. An Algorithm to Determine Peer-Reviewers. In *17th ACM Conference on Information and Knowledge Management*, CIKM '08, 319–328. New York, NY, USA: ACM. ISBN 978-1-59593-991-3.

Rothvoss, T. 2014. A Direct Proof for Lovett's Bound on the Communication Complexity of Low Rank Matrices. *arXiv preprint arXiv:1409.6366*.

Shah, N. B. 2022. An Overview of Challenges, Experiments, and Computational Solutions in Peer Review. Forthcoming.

Shah, N. B.; Tabibian, B.; Muandet, K.; Guyon, I.; and Von Luxburg, U. 2018. Design and Analysis of the NIPS 2016 Review Process. *The Journal of Machine Learning Research*, 19(1): 1913–1946.

Stelmakh, I.; Rastogi, C.; Shah, N. B.; Singh, A.; and Daumé III, H. 2020. A Large Scale Randomized Controlled Trial on Herding in Peer-Review Discussions. *arXiv preprint arXiv:2011.15083*.

Stelmakh, I.; Shah, N.; and Singh, A. 2019a. PeerReview4All: Fair and Accurate Reviewer Assignment in Peer Review. In *Conference on Algorithmic Learning Theory*.

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

Stelmakh, I.; Shah, N. B.; and Singh, A. 2021. Catch Me if I Can: Detecting Strategic Behaviour in Peer Assessment. In *AAAI Conference on Artificial Intelligence*.

Stelmakh, I.; Shah, N. B.; Singh, A.; and Daumé III, H. 2021. Prior and Prejudice: The Novice Reviewers' Bias against Resubmissions in Conference Peer Review. In *ACM Conference on Computer-Supported Cooperative Work and Social Computing*.

Tang, W.; Tang, J.; and Tan, C. 2010. Expertise Matching via Constraint-Based Optimization. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT '10, 34–41. Washington, DC, USA: IEEE Computer Society. ISBN 978-0-7695-4191-4.

Taylor, C. J. 2008. On the Optimal Assignment of Conference Papers to Reviewers. *University of Pennsylvania Department of Computer and Information Science Technical Report*.

Teplitskiy, M.; Ranu, H.; Gray, G.; Menietti, M.; Guinan, E.; and Lakhani, K. R. 2019. Do Experts Listen to Other Experts? Field Experimental Evidence From Peer Review. Forthcoming.

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.

Tran, H. D.; Cabanac, G.; and Hubert, G. 2017. Expert Suggestion for Conference Program Committees. In *11th International Conference on Research Challenges in Information Science*, 221–232.

Tung, A. K. 2006. Impact of Double Blind Reviewing on SIGMOD Publication: A More Detailed Analysis. *ACM SIGMOD Record*, 35(3): 6–7.

Xu, Y.; Zhao, H.; Shi, X.; and Shah, N. 2019. On Strategyproof Conference Review. In *IJCAI*.