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

Extended Abstract

Steven Jecmen
Carnegie Mellon University
Pittsburgh, PA
sjecmen@cs.cmu.edu

Hanrui Zhang
Carnegie Mellon University
Pittsburgh, PA
hanruiz1@cs.cmu.edu

Ryan Liu
Carnegie Mellon University
Pittsburgh, PA
ryanliu@andrew.cmu.edu

Fei Fang
Carnegie Mellon University
Pittsburgh, PA
feifang@cmu.edu

Vincent Conitzer
Duke University
Durham, NC
conitzer@cs.duke.edu

Nihar B. Shah
Carnegie Mellon University
Pittsburgh, PA
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 show both empirically (on real conference data) and theoretically (under certain natural conditions) that dividing reviewers uniformly at random is near-optimal. The full paper is available at https://arxiv.org/abs/2108.06371.

KEYWORDS

Peer Review; Matching; Paper Assignment

ACM Reference Format:

Steven Jecmen, Hanrui Zhang, Ryan Liu, Fei Fang, Vincent Conitzer, and Nihar B. Shah. 2022. Near-Optimal Reviewer Splitting in Two-Phase Paper Reviewing and Conference Experiment Design: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

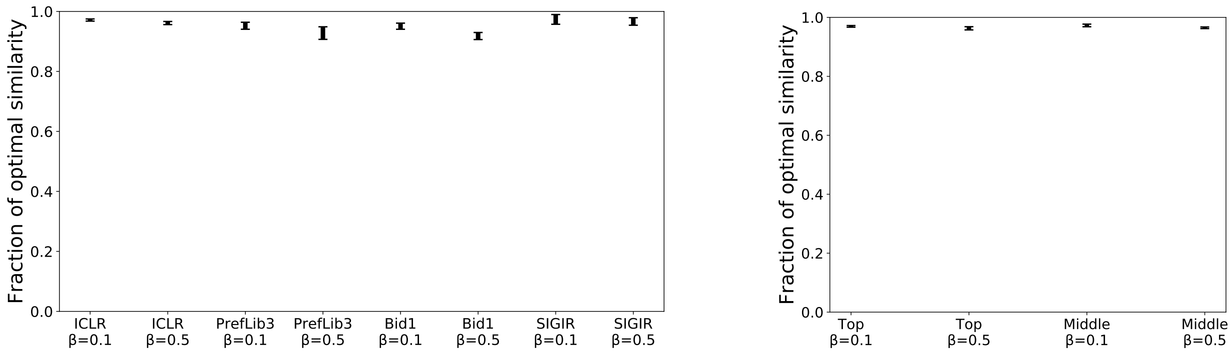
1 INTRODUCTION

Peer review is a widely-adopted method for evaluating scientific research [17]. Careful assignment of reviewers to papers is critically important in order to ensure 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 similarity scores representing the level of expertise each reviewer has for each paper [3–5, 10, 13, 14, 16, 22], 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 [4, 6, 7, 11, 19, 20]. 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.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

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 allow the conference to solicit additional reviews only on papers that obtained sufficiently high ratings in the first phase (as done at AAAI 2021/2022). 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 [15, 18]. In addition, a second phase of reviews can compensate for reviewers who were unresponsive in the first phase, who can no longer review due to personal problems, who discovered conflicts they had with an assigned paper, 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 [1]. 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 [21]. 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 [2, 9, 15]. 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. 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.



(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 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 is the fraction of papers in the second stage. The ICLR similarities [23] (911 papers, 2435 reviewers) are constructed from text-matching between papers and reviewers' past work, PrefLib3 [12] (176 papers, 146 reviewers) and Bid1 [13] (600 papers, 400 reviewers) similarities are constructed from bidding data, and SIGIR [8] similarities (73 papers, 189 reviewers) are constructed from reviewer and paper subject areas.

2 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 (i.e., the second phase/condition). 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 second-stage papers 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.

3 CONTRIBUTIONS

(1) 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.
(2) We prove that a simplified version of the problem is NP-hard, suggesting that the problem may not be efficiently solvable.
(3) 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. 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, 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.
(4) 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. 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.

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

[1] Anastasia Ailamaki, Periklis Chrysogelos, Amol Deshpande, and Tim Kraska. 2019. The SIGMOD 2019 Research Track Reviewing System. *ACM SIGMOD Record* 48, 2 (2019), 47–54.

[2] Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2021 (accessed January 25, 2022). The NeurIPS 2021 Consistency Experiment. https://blog.neurips.cc/2021/12/08/the-neurips-2021-consistency-experiment/.

[3] Guillaume Cabanac and Thomas Preuss. 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 (2013), 405–415.

[4] Laurent Charlin and Richard S. Zemel. 2013. The Toronto Paper Matching System: An Automated Paper-Reviewer Assignment system. In *ICML Workshop on Peer Reviewing and Publishing Models*.

[5] Tanner Fiez, Nihar Shah, and Lillian Ratliff. 2020. A SUPER* Algorithm to Optimize Paper Bidding in Peer Review. In *36th Conference on Uncertainty in Artificial Intelligence*.

[6] Peter A. Flach, Sebastian Spiegler, Bruno Golénia, Simon Price, John Guiver, Ralf Herbrich, Thore Graepel, and Mohammed J. Zaki. 2010. Novel Tools to Streamline the Conference Review Process: Experiences from SIGKDD'09. *SIGKDD Explorations Newsletter* 11, 2 (May 2010), 63–67. https://doi.org/10.1145/1809400.1809413

[7] Judy Goldsmith and Robert H. Sloan. 2007. The AI Conference Paper Assignment Problem. *AAAI Workshop* WS-07-10 (12 2007), 53–57.

[8] Maryam Karimzadehgan, ChengXiang Zhai, and Geneva Belford. 2008. Multi-Aspect Expertise Matching for Review Assignment. In *17th ACM Conference on Information and Knowledge Management*. 1113–1122.

[9] Neil D. Lawrence. 2014 (accessed May 17, 2021). The NIPS Experiment. https://inverseprobability.com/2014/12/16/the-nips-experiment.

[10] Xiang Liu, Torsten Suel, and Nasir Memon. 2014. A Robust Model for Paper Reviewer Assignment. In *8th ACM Conference on Recommender Systems (RecSys '14)*. ACM, New York, NY, USA, 25–32. https://doi.org/10.1145/2645710.2645749

[11] Cheng Long, Raymond Wong, Yu Peng, and Liangliang Ye. 2013. On Good and Fair Paper-Reviewer Assignment. In *IEEE International Conference on Data Mining*. 1145–1150.

[12] Nicholas Mattei and Toby Walsh. 2013. PrefLib: A Library of Preference Data HTTP://PREFLIB.ORG. In *3rd International Conference on Algorithmic Decision Theory (Lecture Notes in Artificial Intelligence)*. Springer.

[13] Reshef Meir, Jérôme Lang, Julien Lesca, Natan Kaminsky, and Nicholas Mattei. 2020. A Market-Inspired Bidding Scheme for Peer Review Paper Assignment. In *Games, Agents, and Incentives Workshop at AAMAS*.

[14] David Mimno and Andrew McCallum. 2007. Expertise Modeling for Matching Papers with Reviewers. In *13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07)*. ACM, New York, NY, USA, 500–509. https://doi.org/10.1145/1281192.1281247

[15] Eric Price. 2014 (accessed May 17, 2021). The NIPS Experiment. http://blog.mrtz.org/2014/12/15/the-nips-experiment.html.

[16] Marko A. Rodriguez and Johan Bollen. 2008. An Algorithm to Determine Peer-Reviewers. In *17th ACM Conference on Information and Knowledge Management (CIKM '08)*. ACM, New York, NY, USA, 319–328. https://doi.org/10.1145/1458082.1458127

[17] Nihar B. Shah. 2021. Systemic Challenges and Solutions on Bias and Unfairness in Peer Review. Preprint http://www.cs.cmu.edu/~nihars/preprints/Shah_Survey_PeerReview.pdf.

[18] Nihar B. Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. 2018. Design and Analysis of the NIPS 2016 Review Process. *The Journal of Machine Learning Research* 19, 1 (2018), 1913–1946.

[19] Wenbin Tang, Jie Tang, and Chenhao Tan. 2010. Expertise Matching via Constraint-Based Optimization. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT '10)*. IEEE Computer Society, Washington, DC, USA, 34–41. https://doi.org/10.1109/WI-IAT.2010.133

[20] Camillo J. Taylor. 2008. On the Optimal Assignment of Conference Papers to Reviewers. *University of Pennsylvania Department of Computer and Information Science Technical Report* (2008).

[21] Andrew Tomkins, Min Zhang, and William D. Heavlin. 2017. Reviewer Bias in Single- Versus Double-Blind Peer Review. *Proceedings of the National Academy of Sciences* 114, 48 (2017), 12708–12713. https://doi.org/10.1073/pnas.1707323114 arXiv:https://www.pnas.org/content/114/48/12708.full.pdf

[22] Hong Diep Tran, Guillaume Cabanac, and Gilles Hubert. 2017. Expert Suggestion for Conference Program Committees. In *11th International Conference on Research Challenges in Information Science*. 221–232. https://doi.org/10.1109/RCIS.2017.7956540

[23] Yichong Xu, Han Zhao, Xiaofei Shi, and Nihar Shah. 2019. On Strategyproof Conference Review. In *IJCAI*.