Group Fairness in Peer Review

Extended Abstract

Haris Aziz
UNSW Sydney
Sydney, Australia
haris.aziz@unsw.edu.au

Evi Micha
University of Toronto
Toronto, Canada
emicha@cs.toronto.edu

Nisarg Shah
University of Toronto
Toronto, Canada
nisarg@cs.toronto.edu

ABSTRACT

Conferences like AAMAS and NeurIPS have attracted submissions from a large number of communities. This has resulted in a poor reviewing experience for communities, whose submissions are assigned to less qualified reviewers outside of their communities. An often-advocated solution is to break up such large conferences into smaller conferences, but this can lead to the isolation of various communities. We tackle this challenge by introducing a notion of group fairness, called core, which requires every subset of researchers to be treated in such a manner such that they cannot benefit from organizing a smaller conference on their own.

We study a simple peer review model, prove that it always admits a reviewing assignment in the core, and design an efficient algorithm to find one such assignment. On the negative side, we show that the core is incompatible with achieving a good worst-case approximation of social welfare, an often-sought desideratum. We complement these results by conducting experiments with real data.

KEYWORDS

peer review, group fairness, core, stable

ACM Reference Format:
Haris Aziz, Evi Micha, and Nisarg Shah. 2023. Group Fairness in Peer Review: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

In conferences, such as AAMAS, AAAI, and NeurIPS, the assignment of the papers to reviewers is usually an automated procedure, due to their massive scale. Famous automated systems that are used are the Toronto Paper Matching System [1], Microsoft CMT[1], and OpenReview[2]. The authors of the submissions are usually very interested to receive useful feedback from their peers, regarding how they could improve their paper [6, 9, 13]. Thus, the overall experience of an author for a peer review procedure depends on the quality of the reviews that her manuscripts receive.

In many large conferences, the typical procedure of selecting the reviewers of each manuscript is the following one. First, for each paper-reviewer pair is calculated a similarity score based on various parameters such as the subject area of the paper and the reviewer, the bidding of the reviewer, etc. [1, 4, 5, 8, 12]. Then, an assignment is calculated through an optimization problem where the usual objectives are either to maximize the utilitarian social welfare, which is equal to the total similarity, or the egalitarian social welfare, which is equal to the minimum score of each submission, subject to constraints related to the total number of papers that each reviewer can review and the total number of reviewers that each paper should be assigned to.

Peng et al. [7] recently mentioned that a major problem with the prestigious mega conferences is that they constitute the main venues for several communities, and as a result, in some cases, people are asked to review submissions that are beyond their main areas of work. They claim that a reasonable solution is to move to a de-centralized publication process by creating more specialized conferences appropriate for different communities. However, this solution could cause the isolation of different communities which in its turn could cause various other problems such as the difficulty of emerging interdisciplinary ideas. So, a reasonable question is: *how can we treat each group of researchers in a fair way in the current review and publication processes?*

To answer this question, we use the concept of fairness, which, to the best of our knowledge, we are the first that introduce in a peer review setting, called *core* [2]. In this context, this notion requires that given an assignment there is no subset of authors– who can also serve as reviewers– that can deviate as following: They can find an assignment of their submissions among themselves such that (a) no author reviews her own submissions, (b) each paper is reviewed by as many reviewers as in the given assignment, (c) each reviewer reviews no more papers than in the given assignment, and (d) the submissions of each author are assigned to better reviewers than in the given assignment. Intuitively, this notion of fairness requires that any group of authors is treated in a way that it does not have any incentive to deviate from the given assignment and create its own assignment that meets the constraints of the peer review procedure. In other words, any sub-community in a big conference is treated in a way that it does not have any incentive to deviate from the conference and create its own smaller conference. Note that this definition provides fairness to *every* sub-community and not only to pre-defined ones, and hence it guarantees that even emerged interdisciplinary communities, are treated fairly.

2 THEORETICAL RESULTS

In this work, we consider the case that each submission is authored by one agent that also serves as reviewer. A reviewing assignment is valid if each paper is reviewed by k_p reviewers, each reviewer reviews up to k_a papers and no agent reviews her own submissions. To ensure that a valid assignment always exists, we assume that

[1] https://cmt3.research.microsoft.com/
[2] https://github.com/openreview/openreview-matcher

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

the maximum number of papers that each agent can submit is at most $\lfloor k_a/k_p \rfloor$.

There is a set of agents $N = [n]$ where each agent can serve as reviewer and may also author some papers. Let $P_i = \{p_{i,1}, \ldots, p_{i,m_i}\}$ be the set of submissions of agent i where $m_i \in \mathbb{N}$ and $P = (P_1, \ldots, P_n)$. We call $p_{i,\ell}$ as the ℓ-th submission of agent i. A *reviewing assignment* (sometimes simply called as assignment) $R \in \{0,1\}^{n \times m}$ is a binary matrix such that $R(i,j) = 1$, if agent i is assigned to review submission j. With a slight abuse of notation, we denote with $R_i^a = \{j \in [m] : R(i,j) = 1\}$, i.e. the submissions that agent i reviews and with $R_j^p = \{i \in [n] : R(i,j) = 1\}$, i.e. the agents that review submission j.

Each agent $i \in N$ has a preference ranking over the agents in $N \setminus \{i\}$ with respect to her ℓ-th submission, denoted by $\sigma_{i,\ell}$. Let $\sigma_{i,\ell}(i')$ be the position of agent $i' \in N \setminus \{i\}$ in the ranking. An agent i prefers her submissions $p_{i,\ell}$ to be reviewed by i' rather than i'', if $\sigma_{i,\ell}(i') < \sigma_{i,\ell}(i'')$. Typically, each paper receives multiple reviews; hence, we need to define the preferences of agents over sets of reviewers. When agent i prefers (resp., weakly prefers) her ℓ-th submission to be reviewed by the set of agents S rather than the set of agents S', we denote it by $S \succ_{i,\ell} S'$ (resp., $S \succeq_{i,\ell} S'$). We assume that this extension from preferences over individual agents to preferences over sets of agents satisfies the following very natural property.

Definition 2.1 (Order Separability). Let $S_1, S_2, S_3 \subseteq N$ with $|S_1| = |S_2|$. If for each $i' \in S_1$ and each $i'' \in S_2$, it holds that $\sigma_{i,\ell}(i') < \sigma_{i,\ell}(i'')$, then $S_1 \cup S_3 \succ_{i,\ell} S_2 \cup S_3$.

To extend this to preferences (resp., weak preferences) of the agent over assignments, denoted by $\succ_i$ (resp., $\succeq_i$), we need to collate her preferences across all her submissions. We simply require that the collated preference extension satisfies the following natural property.

Definition 2.2 (Consistency). Let R be an assignment, $\hat{R}$ be an assignment restricted over $N' \subseteq N$ and $P' = \cup_{i \in N'} P_i'$, where $P_i' \subseteq P_i$ for each $i \in N'$, and $i \in N'$ be an agent. If $R^p_{p_{i,\ell}} \succeq_{i,\ell} \hat{R}^p_{p_{i,\ell}}$ for each $p_{i,\ell} \in P_i'$, then we must have $R \succeq_i \hat{R}$.

In this work, we are interested in finding assignments such that no subset of agents has an incentive to deviate with any subset of their submissions and implement a restricted assignment that each deviating agent prefers. Formally:

Definition 2.3 (Core). An assignment R is in the core if there is no $N' \subseteq N$, $P_i' \subseteq P_i$ for each $i \in N'$, and assignment $\hat{R}$ restricted over N' and $P' = \cup_{i \in N'} P_i'$ such that $\hat{R} \succ_i R$ for each $i \in N'$.

We show the following result.

Theorem 2.4. *There exists a polynomial time algorithm that returns an assignment in the core.*

We also show that there are instances where no assignment in the core can provide an approximation better than $\Omega(n)$ with respect to utilitarian social welfare. Moreover, we show that it is NP-hard to find an assignment in the core with maximum utilitarian social welfare.

Alg.	USW	ESW	α-Core	Pr
PRCore	0.98 ± 0.04	0.07 ± 0.04	1	0
TPMS	1.23 ± 0.04	0.07 ± 0.04	1.984 ± 0.32	1
PR4A	1.06 ± 0.04	0.07 ± 0.04	1.456 ± 0.02	1

Table 1: Results on CVPR

3 EXPERIMENTS

Here, we empirically compare our algorithm, called PRCore, with two other famous algorithms that aim to achieve different objectives. The most known objective, as it is used at the Toronto Paper Matching System (TPMS) [1] is the maximization of the utilitarian social welfare. We denote the algorithm which computes such an assignment as TPMS. A different objective that was introduced by Stelmakh et al. [11] is to maximize the egalitarian social welfare. Stelmakh et al. [11] also considered the extended leximin version of this objective where subject to maximize the minimum utility of all papers, they aim to maximize the second minimum utility of all papers, and subject to that they aim to maximize the third minimum utility of all papers and so on. The algorithm that achieves this objective is called PeerReview4All (PR4A).

We empirically compare PRCore with TPMS, which is widely used, and PR4A which was used in ICML 2020 [10]. While the latter does not explicitly take into account conflicts between reviewers and submissions, when a reviewer is the author of a submission, we set the corresponding score to be equal to a large negative number.

We use the dataset from the Conference on Computer Vision and Pattern Recognition (CVPR) that was also used by [3]. In all the experiments, we set $k_a = k_p = 3$. The similarity matrix was available, but not the conflict matrix. There are less reviewers than papers. We constructed an artificial author matrix, by matching a paper to the reviewer that has the highest score for it and is not assigned as an author to any other paper so far. By this way, 1373 out of 2623 papers were matched. To measure the performance of different algorithms with respect to the core, we consider multiplicative approximations. In particular, we say that an assignment is in the α-core, if there is no deviating coalition such that all the authors improve their utility by a multiplicative factor of α. We report USW, ESW and the value of α. Because the calculation of the core approximation requires much time, we subsample 50 papers from each database and report means and standard deviation over 100 runs. We also report the probability that a deviating coalition exists. Following Kobren et al. [3] and Stelmakh et al. [11], we run 4 iterations of PR4A (they actually run only one), which ensures that the four minimum scores are maximized.

In Table 1, we see the results. As expected, we see that TPMS achieves the highest USW while PR4A achieves the highest ESW. We see that the mutliplicate approximation of PRCore with respect to USW is around 1.3, but it seems to achieve a much better approximation with respect to ESW. On the other hand, TMPS and PR4A violate core with certainty and the average value of α is more than 1.4 for both algorithms. Therefore, PRCore seems to achieve good approximations with respect to both USW and ESW in the average case. On the other hand, methods that are widely used in practise, violate core very often.

REFERENCES

[1] Laurent Charlin and Richard Zemel. 2013. The Toronto paper matching system: an automated paper-reviewer assignment system. In *Proceedings of the ICML Workshop on Peer Reviewing and Publishing Models*.

[2] Donald Bruce Gillies. 1953. *Some theorems on n-person games*. Princeton University.

[3] Ari Kobren, Barna Saha, and Andrew McCallum. 2019. Paper matching with local fairness constraints. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1247–1257.

[4] Xiang Liu, Torsten Suel, and Nasir Memon. 2014. A robust model for paper reviewer assignment. In *Proceedings of the 8th ACM Conference on Recommender systems*. 25–32.

[5] David Mimno and Andrew McCallum. 2007. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 500–509.

[6] Raymond S. Nickerson. 2005. What Authors Want From Journal Reviewers and Editors. *American Psychological* (2005), 661—-662.

[7] Andi Peng, Jessica Zosa Forde, Yonadav Shavit, and Jonathan Frankle. 2022. Strengthening Subcommunities: Towards Sustainable Growth in AI Research. *arXiv preprint arXiv:2204.08377* (2022).

[8] Marko A. Rodriguez and Johan Bollen. 2008. An algorithm to determine peer-reviewers. In *Proceedings of the 17th ACM conference on Information and knowledge management*. 319–328.

[9] Nihar B. Shah. 2022. Challenges, experiments, and computational solutions in peer review. *Commun. ACM* 65, 6 (2022), 76–87.

[10] Ivan Stelmakh. 2021. Towards Fair, Equitable, and Efficient Peer Review. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 15736–15737.

[11] Ivan Stelmakh, Nihar B. Shah, and Aarti Singh. 2019. PeerReview4All: Fair and accurate reviewer assignment in peer review. In *Algorithmic Learning Theory*. PMLR, 828–856.

[12] Hong Diep Tran, Guillaume Cabanac, and Gilles Hubert. 2017. Expert suggestion for conference program committees. In *2017 11th International Conference on Research Challenges in Information Science (RCIS)*. IEEE, 221–232.

[13] G. David L. Travis and Harry M. Collins. 1991. New light on old boys: Cognitive and institutional particularism in the peer review system. *Science, Technology, & Human Values* 16, 3 (1991), 322–341.