Principled Methods to Improve Peer Review

Nihar B. Shah
Machine Learning Department and Computer Science Department
Carnegie Mellon University
nihars@cs.cmu.edu

ABSTRACT

There is an urgent need to improve peer review, particularly due to the explosion in the number of submissions especially at ML and AI venues. Peer review faces a number of challenges including noise, calibration, subjectivity, and strategic behavior. This paper presents a survey of our recent works towards addressing these challenges. Our works take a principled approach to tackle these issues, towards developing an algorithmic toolkit for improved peer-review processes. Our algorithms focus on achieving objectives of fairness, accuracy, and robustness in these goals. We supplement our algorithms with strong theoretical guarantees as well as empirical evaluations on conference data. The ideas, results, and insights of this work as applicable broadly to a variety of applications beyond peer review.

1. INTRODUCTION

Peer review is a cornerstone of academic practice today and also for years to come [1]. The peer review process is highly regarded by the vast majority of researchers and considered by most to be essential to the communication of scholarly research [2; 3; 4]. However, there is also an overwhelming desire for improvement [4; 2; 5].

An empirical evaluation of the peer-review process was recently preformed in a remarkable experiment conducted by the program chairs of the Neural Information Processing Systems (NeurIPS) 2014 conference [6]. Here, 10% of the submissions were assigned to two independent committees, each tasked with the goal of accepting 22% of the papers. It was found that 57% of papers accepted by one committee were rejected by the other. Such a high level of inconsistency is a major concern, particularly due to the widespread prevalence of the Matthew effect ("rich get richer") in academia [7]. Indeed, various past studies show that small changes in peer review quality can have extensive consequences not only for the submitted papers but also for the career trajectories of the authors [8; 9].

The following quote from Drummond Rennie, in a Nature commentary titled "Lets make peer review scientific" [10], provides an apt summary of the state of peer review today:

> *"Peer review is touted as a demonstration of the self-critical nature of science. But it is a human system. Everybody involved brings prejudices, misunderstandings and gaps in knowledge, so no one should be surprised that peer review is often biased and inefficient. It is occasionally corrupt, sometimes a charade, an open temptation to plagiarists. Even with the best of intentions, how and whether peer review identifies high-quality science is unknown. It is, in short, unscientific."*

The need to improve peer review is particularly urgent due to the explosion in the number of submitted papers in various fields. Conferences in machine learning and artificial intelligence are experiencing a near-exponential growth in the number of submissions, but a significantly slower growth in terms of the number of expert reviewers. The increase in number of submissions is also large in many other fields beyond computer science: according to McCook [11] *"Submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint"*.

Despite the importance of peer review and the urgent need for improvements, there is surprisingly little work on principled approaches towards addressing these problems particularly at scale. The goal of this research is to address these important and challenging problems in peer review in a principled and practical manner.

For concreteness, we consider the setting of peer review in conferences, where a set of papers are submitted at a given time and must be evaluated within a strict time frame. That said, a number of ideas, results, and insights in this work generalize to various applications such as crowdsourcing, A/B testing, peer grading, recommender systems, hiring, college admissions, and many others.

This paper presents a survey of some of our recent works addressing the issues of noise, miscalibration, subjectivity, and strategyproofness. In these works, our focus is on the following objectives:

- **Fairness:** Treat all papers as equal as possible.
- **Accuracy:** Maximize correctness of decisions.
- **Robustness to modeling assumptions:** Recognizing that humans are inherently complex, make minimal assumptions on how people behave.

Our works provide both theoretical guarantees as well as empirical evaluations towards these goals.

2. SUBJECTIVITY

Joint work with R. Noothigattu and A. Procaccia [12].

It is known that different reviewers have different, subjective opinions about the relative importance of various criteria in

judging papers [13; 14; 15; 16; 17]. On the other hand, in order to ensure fairness, every paper should ideally be judged by the same yardstick. For instance, suppose three reviewers consider "improvement of at least 10%" as most important, whereas most members of the community have a high emphasis on "novelty". Then a highly novel paper that yields a 5% improvement over the state of the art may be rejected if reviewed by these three reviewers but would have been accepted by any other set of reviewers. Indeed, as revealed in the survey [18], more than 50% of reviewers say that even if the community thinks a certain characteristic of a manuscript is good, if the reviewer's own opinion is negative about that characteristic, it will count against the paper; about 18% say this can also lead them to reject the paper.

2.1 Problem setting

In this work, we first provide a framework for a principled aggregation of subjective opinions in peer review. Let $\mathcal{R}$ denote the set of all reviewers and $\mathcal{P}$ denote the set of all papers. Each reviewer i reviews a subset of papers, denoted by $P(i) \subseteq \mathcal{P}$. Each reviewer assigns scores to each of their papers on k different criteria, such as novelty, experimental analysis, and technical quality, and also gives an overall recommendation. For simplicity (and rescaling as necessary), we assume that all scores lie in the interval $[0, 1]$. We denote the *criteria scores* given by any reviewer i to any paper $j \in P(i)$ as $\mathbf{x}_{ij} \in [0, 1]^k$, and the *overall recommendation* as $y_{ij} \in [0, 1]$.

We further assume that each reviewer has a monotonic function in mind that they use to compute the overall recommendation for a paper from its criteria scores. By a monotonic function, we mean that given any two score vectors $\mathbf{x}$ and $\mathbf{x}'$, if $\mathbf{x}$ is greater than or equal to $\mathbf{x}'$ on all coordinates, then the function's value on $\mathbf{x}$ must be at least as high as its value on $\mathbf{x}'$. We let $\mathcal{F}_k$ the set of all coordinate-wise non-decreasing functions from $\mathbb{R}^k$ to $\mathbb{R}$.

Inspired by empirical risk minimization, we compute the function in $\mathcal{F}_k$ that minimizes the $L(p, q)$ loss [19; 20; 21; 22] on the data. In more detail, given hyperparameters $p, q \in [1, \infty]$, we compute

$$\widehat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}_k} \left\{ \sum_{i \in \mathcal{R}} \left[\sum_{j \in P(i)} |y_{ij} - f(\mathbf{x}_{ij})|^p \right]^{\frac{q}{p}} \right\}^{\frac{1}{q}}. \quad (1)$$

This learned mapping $\widehat{f}$ represents the community's weighting of the different criteria. Once the function $\widehat{f}$ has been computed, we propose to apply it to every review, that is, for every reviewers i and papers a to obtain a new overall recommendation $\widehat{f}(\mathbf{x}_{ij})$.

Under this framework, the question that then arises is: What values of p and q to use?

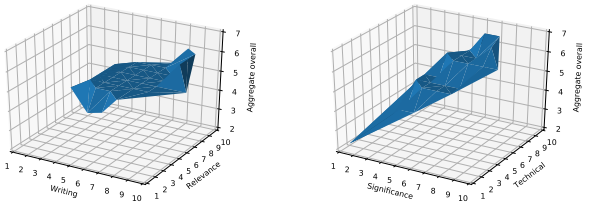
2.2 Solution

We take an axiomatic approach based on social choice theory. We impose three simple requirements (or "axioms") on the learning algorithm: (1) *Consensus*: if all reviewers map some $\mathbf{x} \in \mathbb{R}^k$ to the same $y \in \mathbb{R}$ then the learnt mapping must have $\widehat{f}(\mathbf{x}) = y$. (2) *Efficiency*: if a paper A is "obviously" better than paper B from the review data then the learnt mapping must respect that. (3) *Strategyproofness:* No reviewer should be able to bring the learnt mapping closer to her/his own subjective preference by manipulating the provided reviews.

Recall that our goal is to reduce the space of possible choices of $p, q \in [1, \infty]$ by means of imposing these axioms, thereby then leading to a principled choice. On the other hand, it is often the case in social choice theory [23; 24; 25] that imposition of only a few such natural axioms leads to results of non-existence of any solutions. For our problem at hand, we show that surprisingly, the three aforementioned natural axioms are satisfied by exactly one choice of the hyperparameters.

THEOREM 1. *$L(p, q)$ aggregation, where $p, q \in [1, \infty]$, satisfies consensus, efficiency, and strategyproofness if and only if $p = q = 1$.*

We also perform an empirical analysis using our framework (Equation 1 with $p = q = 1$). We employ a dataset of reviews from IJCAI 2017. First, we observe that writing and relevance do not have a significant influence (Figure 1(a)). Really bad writing or relevance is a significant downside, excellent writing or relevance is appreciated, but everything else in between in irrelevant. Second, technical quality and significance exert a high influence (Figure 1(b)). Finally, we compute the overlap between the set of top 27.27% papers selected by our method with the actual 27.27% accepted papers. We find that the overlap is 79.2%, which we think is quite interesting — our approach does make a significant difference, but the difference is not so drastic as to be disconcerting.



(a) Varying 'writing quality' and 'relevance' (b) Varying 'significance' and 'technical quality'

3. MISCALIBRATION

Joint work with J. Wang [26].

It is well known [27; 28; 29; 30; 31; 32] that the same rating score may have different meanings for different individuals. For instance, if reviewers are asked to provide scores in the interval $[0, 1]$, some reviewers may be lenient and usually provide scores greater than 0.5 whereas some others may be strict and rarely give scores greater than 0.5. Or some reviewers are more moderate whereas others provide scores at the extremes of the allowed interval. Such mismatches cause additional difficulty in the final acceptance decisions as well as lead to unfairness [33]: *"the existence of disparate categories of reviewers creates the potential for unfair treatment of authors. Those whose papers are sent by chance to assassins/demoters are at an unfair disadvantage, while zealots/pushovers give authors an unfair advantage."*

In the literature, there are two popular approaches towards this problem miscalibration. The first approach [34; 35;

36; 37; 38; 39] is to make simplifying assumptions on the nature of the miscalibration, for instance, assuming that these miscalibration is linear or affine. The research following this approach designs algorithms to learn "parameters" of the miscalibration. However, it is known that such assumptions are frequently violated (see [29] and references therein). Then these algorithms can be "significantly harmful" in practice [40]. A second approach [41; 31; 27; 30; 42; 28] towards handling miscalibrations is to either directly elicit rankings from reviewers or convert the scores into rankings. This approach is often believed to be the only resort when the underlying calibration functions may be arbitrary. The question that we thus ask is whether this folklore belief is true — if the miscalibration functions can be arbitrary (or adversarially chosen) then is there any algorithm based on ratings that can perform better than using the rankings alone?

3.1 Problem setting

For brevity, consider a simplified setting with two reviewers and two papers. The two papers have some "true" scores $x_1^* \neq x_2^* \in [0,1]$ that are a priori unknown. Each reviewer $i \in \{1,2\}$ has a "miscalibration" function $f_i : [0,1] \to [0,1]$ which is a priori unknown. The miscalibration function means that if a reviewer i reviews a paper with true score $x^* \in [0,1]$, then the reviewer reports a score i provides the score $y = f_i(x)$. The functions f_1 and f_2 are arbitrary (and can be chosen by an adversary) — the only constraint we impose on these functions is strict monotonicity.

Now suppose one different paper each is assigned to each reviewer uniformly at random. Let y_1 and y_2 respectively denote the scores for paper 1 and paper 2 given by their respective reviewers. Then given the data comprising (y_1, y_2) and the knowledge of who reviewed which paper, the goal then is to identify the paper with the higher true score.

Observe that each reviewer reviews only one paper, and hence using a "ranking" elicited from each reviewer is vacuous. Thus any ranking-based algorithm can attain a probability of success no more than 0.5. The key question now is whether it is possible to do any better using ratings.

3.2 Solution

We show that surprisingly, counter to the popular belief, there exists an algorithm which performs strictly better than rankings even if the miscalibration of ratings is arbitrary or adversarially chosen.

THEOREM 2. *There is a computationally-efficient, randomized algorithm which succeeds with probability* strictly *greater than* 0.5 *for any miscalibration functions and any true scores of the two papers.*

The proposed algorithm is simple to describe. Then the algorithm declares the paper with the higher reviewer-provided-score as "better" with probability $\frac{1+|y_1-y_2|}{2}$.

While our algorithm is randomized, we also show that every deterministic rating-based estimator fails to improve upon rankings. Finally, we extend our positive results from the 2×2 setting to more general application settings.

4. NOISE

Joint work with I. Stelmakh and A. Singh [43].

Data from people is often noisy due to lack of expertise. In peer review, the assignment of the reviewers to papers determines the expertise of the reviewer who will review any paper. It is also known [44; 14] that unique and novel works, particularly those interdisciplinary in nature, face significantly higher difficulty in gaining acceptance. A primary reason for this undesirable state of affairs is the absence of sufficiently many good "peers" to aptly review interdisciplinary research [45]. Our focus is thus to design algorithms to assign reviewers to papers that can help combat this noise. Indeed, the importance of the reviewer-assignment stage of the peer-review process cannot be overstated: quoting [46], "one of the first and potentially most important stage is the one that attempts to distribute submitted manuscripts to competent referees."

4.1 Problem setting

An assignment algorithm takes as input a "similarity" $s_{ij} \in [0,1]$ between every reviewer i and paper j. A higher value of the similarity indicates a better review by the reviewer for that paper. The popular approach to assigning reviewers to papers is to maximize the similarity of the assigned reviewers *summed across all papers and reviewers*. This approach is followed in [47; 48; 49; 50], conference management systems such as EasyChair, HotCRP, and the Toronto Paper Matching System or TPMS [51] used in all top AI and ML conferences.

The issues discussed above strongly motivate the dual goals of the reviewer assignment procedure we consider in this paper — fairness and accuracy.

We consider the notion of max-min fairness [52; 53; 54; 55]. In our context, max-min fairness posits maximizing the sum-similarity for the paper having the least qualified reviewers. It guarantees that no paper is discriminated in favor of luckier counterparts — even the most idiosyncratic paper with a small number of competent-enough reviewers will receive as good treatment as possible.

A main goal of the conference peer-review process is to select the set of "top" papers for acceptance. Thus, it is important that the assignment of papers to referees is built to achieve the accuracy of the final decisions. However, all prior works on paper assignment problem known to us (such as the references above) focus on developing algorithms that optimize the assignment for certain deterministic objectives. In contrast, we take the first approach to connect the quality of the assignment to the accuracy of the entire conference peer-review process.

4.2 Solution

We first show that popular assignment approach discussed above can be quite "unfair". For instance, consider a setting with 3 papers and 3 reviewers, where each reviewer must review 1 paper and each paper must be reviewed by 1 reviewer. Consider similarities given by:

	PAPER a	PAPER b	PAPER c
REVIEWER 1	1	1	1
REVIEWER 2	0	0	1/5
REVIEWER 3	1/4	1/4	1/2

The popular assignment approach will assign reviewers 1, 2, and 3 to papers a, b, and c respectively. Under this assignment, paper b is assigned a reviewer with insufficient expertise to evaluate the paper.

We present a novel algorithm called PeerReview4All to assign reviewers to papers. Our algorithm is based on a construction of multiple candidate assignments which cater to different structural properties of the similarities and a judicious choice between them provides the algorithm appealing properties.

We then analyze PeerReview4All in terms of its fairness and statistical accuracy. As an example of fairness, for the similarities in the table above, PeerReview4All assigns reviewers 1, 2, and 3 to papers a, c, and b respectively. PeerReview4All thus ensures that every paper has a reviewer with similarity at least 1/5. This "fair" assignment does not discriminate against the disadvantaged paper b (and a) for improving the review quality of the already benefiting paper c.

For analyzing statistical accuracy, we consider a popular statistical model [34; 56; 57] which assumes existence of some true objective score for every paper. We also propose a new model that incorporates subjectivity in the reviews. For both models, we analyze the minimax risk, studying the loss in terms of "incorrect" acceptance decisions

Theorem 3. PeerReview4All *is max-min fair up to a constant factor, and is minimax optimal up to constant factors under both objective and subjective-score models.*

PeerReview4All thus simultaneously achieves fairness and statistical accuracy. Interestingly, our results suggest that *fairness is the right proxy towards statistical accuracy.*

5. STRATEGIC BEHAVIOR

Joint work with H. Zhao, Y. Xu and X. Shi [58].

Peer-review is susceptible to strategic manipulations. A reviewer may be able to increase the chances of acceptance of their own submissions by manipulating the reviews (e.g., providing lower scores) for other papers. A recent empirical study [59] examined the strategic behavior of people in competitive peer review, and concluded that *"...competition incentivizes reviewers to behave strategically, which reduces the fairness of evaluations and the consensus among referees."* See [60] for more anecdotes. As Thurner and Hanel [61] posit, even a small number of selfish, strategic reviewers can drastically reduce the quality of scientific standard.

It is thus highly important to protect peer review from any possible strategic manipulations. We define strategyproofness in terms of a "conflict graph", which is a fixed graph given to us. A conflict graph is a bipartite graph with all reviewers and papers as its vertices, and has an edge between a reviewer vertex and a paper vertex if the reviewer has a conflict with the paper. Examples of conflicts include authorship conflicts (e.g., the reviewer is an author of that paper), institutional conflicts, etc. Now strategyproofness means that no reviewer must be able to influence the final ranking of her/his conflicted papers by manipulating the reviews that she/he provides.

A number of past works [62; 63; 64; 65; 66; 67; 68] consider designing strategyproof procedures of "peer grading" in MOOCs and classrooms. There are two key differences between these peer-grading settings and the peer-review setting we consider. First, the peer grading setting involves conflict graphs of degree at most 1, that is, every reviewer conflicts with at most one paper and every paper has at most one author. On the other hand, even if one considers only authorship conflicts in conference peer review, every author may submit multiple papers and any paper may have multiple authors, thus requiring strategyproofness with respect to more general graphs. Second, these prior works do not account for "heterogeneity" in the papers and reviewers with the motivation that all students in peer grading take the same course. On the other hand, conference papers and reviewers are more diverse in terms of their expertise and subject matter. Hence any peer-review framework must have significant flexibility to accommodate the various intricacies. These differences make the peer-review setting strictly more general and significantly more challenging.

5.1 Problem setting

We now present our framework for strategyproof peer review. There are two design elements in this framework: an assignment of reviewers to papers, and an algorithm to aggregate the reviews to yield a final ranking of the papers. These must be designed to meet strategyproofness. In addition, we require the algorithms to satisfy "unanimity" which is popular in social choice theory [23] as a basic requirement for any algorithm. Unanimity in our context necessitates that if the papers can be partitioned into two sets such that for all pairs of papers p_1 in the first set and p_2 in the second set, all reviewers reviewing both p_1 and p_2 rank p_1 higher than p_2, then the final aggregate must also rank p_1 higher than p_2.

5.2 Solution

Our solution first assigns papers to reviewers in which one can use any assignment algorithm subject to the constraint that *there is no path in the conflict graph between any assigned reviewer-paper pair.* This assignment requires that the conflict graph can be partitioned into large enough disconnected components. Once the reviews are in, our framework allows for any arbitrary aggregation within each partition. Given the rankings of the papers in each partition, we finally interleave papers in a simple alternating fashion to obtain the final ranking.

Theorem 4. *Our algorithm guarantees strategyproofness and unanimity.*

We complement our positive results with negative theoretical results where we prove that under slightly stronger requirements, it is impossible for any algorithm to be both strategyproof and efficient.

Finally, we perform an empirical analysis using data from ICLR 2017. We show that the condition on the partitioning of the graph indeed holds in ICLR 2017. We further demonstrate a simple trick to make the partitioning method more practically appealing under conference peer-review settings.

6. DISCUSSION

The need to improve peer review is important and urgent for scholarly research to thrive. We are developing a toolkit of algorithms – with provable guarantees – towards this goal. In addition to the works surveyed in this paper, our work has addressed other problems in peer review including developing tools to test for biases [69], and methods for more efficient use of human time and effort [70], with several others underway. Finally, we are also conducting outreach with an aim to drive positive policy change – please see `researchonresearch.blog`.

7. REFERENCES

[1] Simon Price and Peter A Flach. Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM*, 60(3):70–79, 2017.

[2] Adrian Mulligan, Louise Hall, and Ellen Raphael. Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the Association for Information Science and Technology*, 64(1):132–161, 2013.

[3] David Nicholas, Anthony Watkinson, Hamid R Jamali, Eti Herman, Carol Tenopir, Rachel Volentine, Suzie Allard, and Kenneth Levine. Peer review: still king in the digital age. *Learned Publishing*, 28(1):15–21, 2015.

[4] Mark Ware. Peer review: benefits, perceptions and alternatives. *Publishing Research Consortium*, 4:1–20, 2008.

[5] Richard Smith. Peer review: a flawed process at the heart of science and journals. *Journal of the royal society of medicine*, 99(4):178–182, 2006.

[6] N. Lawrence and C. Cortes. The NIPS Experiment. `http://inverseprobability.com/2014/12/16/the-nips-experiment`, 2014. [Online; accessed 11-June-2018].

[7] Robert K Merton. The Matthew effect in science. *Science*, 159:56–63, 1968.

[8] Warren Thorngate and Wahida Chowdhury. By the numbers: Track record, flawed reviews, journal space, and the fate of talented authors. In *Advances in Social Simulation*, pages 177–188. Springer, 2014.

[9] Flaminio Squazzoni and Claudio Gandelli. Saint matthew strikes again: An agent-based model of peer review and the scientific community structure. *Journal of Informetrics*, 6(2):265–275, 2012.

[10] Drummond Rennie. Make peer review scientific: thirty years on from the first congress on peer review, drummond rennie reflects on the improvements brought about by research into the process–and calls for more. *Nature*, 535(7610):31–34, 2016.

[11] Alison McCook. Is peer review broken? submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint about the process at top-tier journals. what's wrong with peer review? *The scientist*, 20(2):26–35, 2006.

[12] Ritesh Noothigattu, Nihar Shah, and Ariel Procaccia. Choosing how to choose papers. *arXiv preprint arxiv:1808.09057*, 2018.

[13] Kenneth Church. Reviewing the reviewers. *Computational Linguistics*, 31(4):575–578, 2005.

[14] Michèle Lamont. *How professors think*. Harvard University Press, 2009.

[15] Von Bakanic, Clark McPhail, and Rita J Simon. The manuscript review and decision-making process. *American Sociological Review*, pages 631–642, 1987.

[16] Mohammadreza Hojat, Joseph S Gonnella, and Addeane S Caelleigh. Impartial judgment by the gatekeepers of science: fallibility and accountability in the peer review process. *Advances in Health Sciences Education*, 8(1):75–96, 2003.

[17] Michael J Mahoney. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research*, 1(2):161–175, 1977.

[18] Steven Kerr, James Tolliver, and Doretta Petree. Manuscript characteristics which influence acceptance for management and social science journals. *Academy of Management Journal*, 20(1):132–141, 1977.

[19] Alireza Rahimpour, Ali Taalimi, and Hairong Qi. Feature encoding in band-limited distributed surveillance systems. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 1752–1756. IEEE, 2017.

[20] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha. R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd international conference on Machine learning*, pages 281–288. ACM, 2006.

[21] Deguang Kong, Chris Ding, and Heng Huang. Robust nonnegative matrix factorization using l21-norm. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 673–682. ACM, 2011.

[22] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint 2, 1-norms minimization. In *Advances in neural information processing systems*, pages 1813–1821, 2010.

[23] Kenneth J Arrow. A difficulty in the concept of social welfare. *Journal of political economy*, 58(4):328–346, 1950.

[24] Allan Gibbard. Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society*, pages 587–601, 1973.

[25] Mark Allen Satterthwaite. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 10(2):187–217, 1975.

[26] Jingyan Wang and Nihar B Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *AAMAS*, 2019.

[27] Ioannis Mitliagkas, Aditya Gopalan, Constantine Caramanis, and Sriram Vishwanath. User rankings from comparisons: Learning permutations in high dimensions. In *Allerton Conference on Communication, Control, and Computing*, pages 1143–1150, 2011.

[28] Ammar Ammar and Devavrat Shah. Efficient rank aggregation using partial data. In *SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, 2012.

[29] Dale Griffin and Lyle Brenner. *Perspectives on Probability Judgment Calibration*, chapter 9, pages 177–199. Wiley-Blackwell, 2008.

[30] Yoav Freund, Raj D. Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

[31] Anne-Wil Harzing, Joyce Baldueza, Wilhelm Barner-Rasmussen, Cordula Barzantny, Anne Canabal, Anabella Davila, Alvaro Espejo, Rita Ferreira, Axele Giroud, Kathrin Koester, Yung-Kuei Liang, Audra Mockaitis, Michael J. Morley, Barbara Myloni, Joseph O.T. Odusanya, Sharon Leiba O'Sullivan, Ananda Kumar Palaniappan, Paulo Prochno, Srabani Roy Choudhury, Ayse Saka-Helmhout, Sununta Siengthai, Linda Viswat, Ayda Uzuncarsili Soydas, and Lena Zander. Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? *International Business Review*, 18(4):417–432, 2009.

[32] Nihar B Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. Design and analysis of the nips 2016 review process. *arXiv preprint arXiv:1708.09794*, 2017.

[33] Stanley S Siegelman. Assassins and zealots: variations in peer review. special report. *Radiology*, 178(3):637–642, 1991.

[34] Hong Ge, Max Welling, and Zoubin Ghahramani. A Bayesian model for calibrating conference review scores, 2013.

[35] R. S. MacKay, R. Kenna, R. J. Low, and S. Parker. Calibration with confidence: a principled method for panel assessment. *Royal Society Open Science*, 4(2), 2017.

[36] Yukino Baba and Hisashi Kashima. Statistical quality estimation for general crowdsourcing tasks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013.

[37] S. R. Paul. Bayesian methods for calibration of examiners. *British Journal of Mathematical and Statistical Psychology*, 34(2):213–223, 1981.

[38] Magnus Roos, Jörg Rothe, and Björn Scheuermann. How to calibrate the scores of biased reviewers by quadratic programming. In *AAAI Conference on Artificial Intelligence*, 2011.

[39] Peter A. Flach, Sebastian Spiegler, Bruno Golénia, Simon Price, John Guiver, Ralf Herbrich, Thore Graepel, and Mohammed J. Zaki. Novel tools to streamline the conference review process: Experiences from SIGKDD'09. *SIGKDD Explor. Newsl.*, 11(2):63–67, May 2010.

[40] John Langford. ICML acceptance statistics, 2012. `http://hunch.net/?p=2517` [Online; accessed 14-May-2018].

[41] Milton Rokeach. The role of values in public opinion research. *Public Opinion Quarterly*, 32(4):547–559, 1968.

[42] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, 2012.

[43] Ivan Stelmakh, Nihar Shah, and Aarti Singh. PeerReview4All: Fair and accurate reviewer assignment in peer review. In *Algorithmic Learning Theory*, 2019.

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

[45] Alan L Porter and Frederick A Rossini. Peer review of interdisciplinary research proposals. *Science, technology, & human values*, 10(3):33–38, 1985.

[46] Marko A Rodriguez, Johan Bollen, and Herbert Van de Sompel. Mapping the bid behavior of conference referees. *Journal of Informetrics*, 1(1):68–82, 2007.

[47] Cheng Long, Raymond Wong, Yu Peng, and Liangliang Ye. On good and fair paper-reviewer assignment. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 1145–1150, 12 2013.

[48] L. Charlin, R. S. Zemel, and C. Boutilier. A framework for optimizing paper matching. *CoRR*, abs/1202.3706, 2012.

[49] Judy Goldsmith and Robert H. Sloan. The AI conference paper assignment problem. WS-07-10:53–57, 12 2007.

[50] Wenbin Tang, Jie Tang, and Chenhao Tan. Expertise matching via constraint-based optimization. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 34–41, Washington, DC, USA, 2010. IEEE Computer Society.

[51] L. Charlin and R. S. Zemel. The Toronto Paper Matching System: An automated paper-reviewer assignment system. In *ICML Workshop on Peer Reviewing and Publishing Models*, 2013.

[52] John Rawls. *A theory of justice: Revised edition*. Harvard university press, 1971.

[53] Ellen L. Hahne. Round-robin scheduling for max-min fairness in data networks. *IEEE Journal on Selected Areas in communications*, 9(7):1024–1039, 1991.

[54] Ron Lavi, Ahuva Mu'Alem, and Noam Nisan. Towards a characterization of truthful combinatorial auctions. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pages 574–583. IEEE, 2003.

[55] Thomas Bonald, Laurent Massoulié, Alexandre Proutiere, and Jorma Virtamo. A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queueing systems*, 53(1-2):65–84, 2006.

[56] M. McGlohon, N. Glance, and Z. Reiter. Star quality: Aggregating reviews to rank products and merchants. In *Proceedings of Fourth International Conference on Weblogs and Social Media (ICWSM)*, 2010.

[57] W. Dai, G. Z. Jin, J. Lee, and M. Luca. Optimal aggregation of consumer ratings: An application to yelp.com. Working Paper 18567, National Bureau of Economic Research, November 2012.

[58] Han Zhao, Yichong Xu, Xiaofei Shi, and Nihar Shah. On strategyproof conference review. In *IJCAI*, 2019.

[59] Stefano Balietti, Robert L Goldstone, and Dirk Helbing. Peer review and competition in the art exhibition game. *Proceedings of the National Academy of Sciences*, 113(30):8414–8419, 2016.

[60] Jef Akst. I hate your paper. many say the peer review system is broken. heres how some journals are trying to fix it. *The Scientist*, 24(8):36, 2010.

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

[62] Geoffroy De Clippel, Herve Moulin, and Nicolaus Tideman. Impartial division of a dollar. *Journal of Economic Theory*, 139(1):176–191, 2008.

[63] Noga Alon, Felix Fischer, Ariel Procaccia, and Moshe Tennenholtz. Sum of us: Strategyproof selection from the selectors. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 101–110. ACM, 2011.

[64] Ron Holzman and Hervé Moulin. Impartial nominations for a prize. *Econometrica*, 81(1):173–196, 2013.

[65] Felix Fischer and Max Klimm. Optimal impartial selection. *SIAM Journal on Computing*, 44(5):1263–1285, 2015.

[66] David Kurokawa, Omer Lev, Jamie Morgenstern, and Ariel D Procaccia. Impartial peer review. In *IJCAI*, pages 582–588, 2015.

[67] Haris Aziz, Omer Lev, Nicholas Mattei, Jeffrey S Rosenschein, and Toby Walsh. Strategyproof peer selection: Mechanisms, analyses, and experiments. In *AAAI*, pages 397–403, 2016.

[68] Anson B Kahng, Yasmine Kotturi, Chinmay Kulkarni, David Kurokawa, and Ariel D. Procaccia. Ranking wily people who rank each other. *Technical Report*, 2017.

[69] Ivan Stelmakh, Nihar Shah, and Aarti Singh. On testing for biases in peer review. In *ACM EC Workshop on Mechansim Design for Social Good*, 2019.

[70] T Fiez, N Shah, and L Ratliff. A SUPER* algorithm to optimize paper bidding in peer review. In *ICML workshop on Real-world Sequential Decision Making: Reinforcement Learning And Beyond*, 2019.