WSDM 2021 Tutorial on Systemic Challenges and Computational Solutions on Bias and Unfairness in Peer Review

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

ABSTRACT

Peer review is the backbone of scientific research. Yet peer review is called “biased,” “broken,” and “unscientific” in many scientific disciplines. This problem is further compounded with the near-exponentially growing number of submissions in various computer science conferences. Due to the prevalence of “Matthew effect” of rich getting richer in academia, any source of unfairness in the peer review system, such as those discussed in this tutorial, can considerably affect the entire career trajectory of (young) researchers.

This tutorial will discuss a number of systemic challenges in peer review such as biases, subjectivity, miscalibration, dishonest behavior, and noise. For each issue, the tutorial will first present insightful experiments to understand the issue. Then the tutorial will present computational techniques designed to address these challenges. Many open problems will be highlighted which are envisaged to be exciting to the WSDM audience, and will lead to significant impact if solved.

CCS CONCEPTS

• **Information systems**; • **Computing methodologies** → *Artificial intelligence*; *Machine learning*; • **Human-centered computing** → *Collaborative and social computing*;

ACM Reference Format:
Nihar B. Shah. 2021. WSDM 2021 Tutorial on Systemic Challenges and Computational Solutions on Bias and Unfairness in Peer Review. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM ’21), March 8–12, 2021, Virtual Event, Israel.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3437963.3441660

1 MOTIVATION

Peer review is a cornerstone of academic practice today and also for years to come [45]. 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 [39, 41, 70]. However, there is also an overwhelming desire for improvement [39, 54, 70].

WSDM ’21, March 8–12, 2021, Virtual Event, Israel.

ACM ISBN 978-1-4503-8297-7/21/03.
https://doi.org/10.1145/3437963.3441660

The following quote from Rennie [46], in a Nature commentary titled “Let’s make peer review scientific” 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.”*

Problems in peer review have consequences much beyond the outcome for a specific paper or grant, particularly due to the widespread prevalence of the Matthew effect (“rich get richer”) in academia [37]. As noted by Triggle and Triggle [65] *“an incompetent review may lead to the rejection of the submitted paper, or of the grant application, and the ultimate failure of the career of the author.”* (See also [55, 63].)

The importance of peer review and the urgent need for improvements, behooves research on principled approaches towards addressing problems in peer review, particularly at scale. In this tutorial, we outline several directions of research on this topic, and also highlight important open problems that we envisage to be exciting to the community.

2 OUTLINE OF THE TUTORIAL

The tutorial will broadly cover six topics.

(1) *Demographics:* We will first discuss biases due to demographics. We will begin by discussing a remarkable semi-randomized controlled trial [64] at the WSDM conference in testing for biases in single-blind (versus double blind) review. We will then demonstrate [56] – by showing certain issues in the experimental setup and tests from [64]– how one must be careful in running any experiments or statistical tests for biases in any such reasonably complex problem setting. We will subsequently discuss a framework for such problems in the context of peer review [56], and also present general principles that can be applied to other sociotechnical systems. Finally, we will discuss testing of biases using the *text* of the reviews [36]. Auxilliary references: [8, 22, 43, 71].
(2) *Miscalibration:* Even in the absence of any demographic bias, there is unfairness due to miscalibrations of individuals – e.g., some reviewers may be strict, lenient, extreme, moderate etc. [53]. We will discuss the complexity of human miscalibration [7], followed by three key approaches towards solving this problem: Model-based approaches [4, 16, 19, 34, 44, 48], ranking-based

approaches [2, 17, 21, 38, 40, 47], and a model-free rating-based approach [68]. Auxilliary references: [52, Section 3.3], [13].

(3) *Dishonest behavior:* Since conference peer-review is competitive, some participants gain advantage by gaming the system, thereby rendering the application unfair for other honest participants. We will first detail an insightful experiment [6] on the behavior of human evaluators in competitive environments. We will then overview a popular algorithmic building block designed to prevent dishonest behavior in certain settings of human-provided evaluations [1, 3, 12, 15, 24, 26, 29, 72]. We will then evaluate a variant of this building block in the context of peer review on data from ICLR 2017 and 2018 [72]. Auxilliary references: [5, 11, 23, 25, 30, 35, 58].

(4) *Assignment of reviewers:* The assignment of reviewers to papers is known to be one of the most important parts of evaluation, and contributes significantly to the noise in the reviews. We will detail the current methods of assigning reviewers to papers in major ML/AI conferences [9]. We will then highlight problems of unfairness in these assignment procedures, for instance, that they are unfair interdisciplinary papers (both in theory and practice). We will subsequently present alternative assignments with theoretical guarantees [18, 28, 57], and empirical evaluations on CVPR 2017, CVPR 2018, MIDL 2018 [28] and ICML 2020. Auxilliary references: [10, 14, 20, 33, 62].

(5) *Subjectivity:* It is well known that subjective opinions of individual reviewers leads to unfairness to some participants. We will first discuss this problem of "commensuration bias" [27, 32]. We will then describe an algorithmic framework designed using machine learning and social choice theory to mitigate this bias in the context of peer review, and also describe experiments in IJCAI 2017 [42].

(6) *Norms and policies:* The presentation will conclude with a discussion on driving actual policy change. This will include experiments conducted in the peer-review process of ICML 2020 on (i) reviewer bias due to knowledge of previous rejections [61], (ii) herding effects in reviewer discussion [59], and (iii) an experiment with novice reviewers [60]. The findings from these experiments are interesting on their own, and inform the policies that are set by the community. We will also discuss policy-related issues pertaining to gender skew in conference paper awards and the need for transparency [67], and biases due to alphabetical orderings [66].

The tutorial will also discuss a number of open problems in each of the aforementioned topics, as well as overarching issues such as how to measure the quality of a peer review process [31, 52, 60, 69].

3 CONCLUSIONS

There are many sources of systemic biases and unfairness in peer review. The need to improve peer review is important and urgent for scholarly research to thrive. The current research on peer review has only scratched the surface of this important and urgent problem domain. There are lots of open problems which are exciting, challenging, impactful, and allow for a broad spectrum of theoretical, applied, and conceptual contributions.

ACKNOWLEDGMENTS

This tutorial is supported by NSF CAREER Award CIF: 1942124. A version of such a tutorial was also presented at AAAI 2020 [49], SIGMOD 2020 [50] and TheWebConf 2020 [51].

REFERENCES

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

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

[3] Haris Aziz, Omer Lev, Nicholas Mattei, Jeffrey S Rosenschein, and Toby Walsh. 2016. Strategyproof Peer Selection: Mechanisms, Analyses, and Experiments.. In *AAAI*. 397–403.

[4] Yukino Baba and Hisashi Kashima. 2013. Statistical Quality Estimation for General Crowdsourcing Tasks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

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

[6] Stefano Balietti, Robert L Goldstone, and Dirk Helbing. 2016. Peer review and competition in the Art Exhibition Game. *Proceedings of the National Academy of Sciences* 113, 30 (2016), 8414–8419.

[7] Lyle Brenner, Dale Griffin, and Derek J Koehler. 2005. Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes* 97, 1 (2005), 64–81.

[8] Amber E. Budden, Tom Tregenza, Lonnie W. Aarssen, Julia Koricheva, Roosa Leimu, and Christopher J. Lortie. 2008. Double-blind review favours increased representation of female authors. *Trends in Ecology and Evolution* 23, 1 (2008), 4 – 6. https://doi.org/10.1016/j.tree.2007.07.008

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

[10] L. Charlin, R. S. Zemel, and C. Boutilier. 2012. A Framework for Optimizing Paper Matching. *CoRR* abs/1202.3706 (2012). arXiv:1202.3706 http://arxiv.org/abs/1202.3706

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

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

[13] Wenxin Ding, Nihar B., and Weina Wang. 2020. On the Privacy-Utility Tradeoff in Peer-Review Data Analysis. In *AAAI Privacy-Preserving Artificial Intelligence (PPAI-21) workshop*.

[14] T Fiez, N Shah, and L Ratliff. 2020. A SUPER* Algorithm to Optimize Paper Bidding in Peer Review. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

[15] Felix Fischer and Max Klimm. 2015. Optimal impartial selection. *SIAM J. Comput.* 44, 5 (2015), 1263–1285.

[16] 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 Explor. Newsl.* 11, 2 (May 2010), 63–67. https://doi.org/10.1145/1809400.1809413

[17] Yoav Freund, Raj D. Iyer, Robert E. Schapire, and Yoram Singer. 2003. An Efficient Boosting Algorithm for Combining Preferences. *Journal of Machine Learning Research* 4 (2003), 933–969. http://www.jmlr.org/papers/v4/freund03a.html

[18] N. Garg, T. Kavitha, A. Kumar, K. Mehlhorn, and J. Mestre. 2010. Assigning Papers to Referees. *Algorithmica* 58, 1 (01 Sep 2010), 119–136. https://doi.org/10.1007/s00453-009-9386-0

[19] Hong Ge, Max Welling, and Zoubin Ghahramani. 2013. A Bayesian model for calibrating conference review scores. http://mlg.eng.cam.ac.uk/hong/nipsrevcal.pdf

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

[21] 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. 2009. Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? *International Business Review* 18, 4 (2009), 417–432.

[22] Shawndra Hill and Foster J. Provost. 2003. The myth of the double-blind review? Author identification using only citations. *SIGKDD Explorations* 5 (01 2003), 179–184.

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

[24] Ron Holzman and Hervé Moulin. 2013. Impartial nominations for a prize. *Econometrica* 81, 1 (2013), 173–196.
[25] Steven Jecmen, Hanrui Zhang, Ryan Liu, Nihar B. Shah, Vincent Conitzer, and Fei Fang. 2020. Mitigating Manipulation in Peer Review via Randomized Reviewer Assignments. In *NeurIPS*.
[26] Anson B Kahng, Yasmine Kotturi, Chinmay Kulkarni, David Kurokawa, and Ariel D. Procaccia. 2017. Ranking Wily People Who Rank Each Other. *Technical Report* (2017).
[27] Steven Kerr, James Tolliver, and Doretta Petree. 1977. Manuscript characteristics which influence acceptance for management and social science journals. *Academy of Management Journal* 20, 1 (1977), 132–141.
[28] Ari Kobren, Barna Saha, and Andrew McCallum. 2019. Paper Matching with Local Fairness Constraints. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
[29] David Kurokawa, Omer Lev, Jamie Morgenstern, and Ariel D Procaccia. 2015. Impartial Peer Review. In *IJCAI*. 582–588.
[30] Michèle Lamont. 2009. *How professors think*. Harvard University Press.
[31] N. Lawrence and C. Cortes. 2014. The NIPS Experiment. http://inverseprobability.com/2014/12/16/the-nips-experiment. [Online; accessed 11-June-2018].
[32] Carole J Lee. 2015. Commensuration bias in peer review. *Philosophy of Science* 82, 5 (2015), 1272–1283.
[33] Cheng Long, Raymond Wong, Yu Peng, and Liangliang Ye. 2013. On Good and Fair Paper-Reviewer Assignment. In *Proceedings - IEEE International Conference on Data Mining, ICDM*. 1145–1150.
[34] R. S. MacKay, R. Kenna, R. J. Low, and S. Parker. 2017. Calibration with confidence: a principled method for panel assessment. *Royal Society Open Science* 4, 2 (2017). https://doi.org/10.1098/rsos.160760
[35] Michael J Mahoney. 1977. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research* 1, 2 (1977), 161–175.
[36] Emaad Manzoor and Nihar B Shah. 2021. Uncovering Latent Biases in Text: Method and Application to Peer Review. In *AAAI*.
[37] Robert K Merton. 1968. The Matthew Effect in Science. *Science* 159 (1968), 56–63.
[38] Ioannis Mitliagkas, Aditya Gopalan, Constantine Caramanis, and Sriram Vishwanath. 2011. User rankings from comparisons: Learning permutations in high dimensions. In *Allerton Conference on Communication, Control, and Computing*. 1143–1150.
[39] Adrian Mulligan, Louise Hall, and Ellen Raphael. 2013. 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 (2013), 132–161.
[40] Sahand Negahban, Sewoong Oh, and Devavrat Shah. 2012. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*.
[41] David Nicholas, Anthony Watkinson, Hamid R Jamali, Eti Herman, Carol Tenopir, Rachel Volentine, Suzie Allard, and Kenneth Levine. 2015. Peer review: still king in the digital age. *Learned Publishing* 28, 1 (2015), 15–21.
[42] Ritesh Noothigattu, Nihar Shah, and Ariel Procaccia. 2020. Loss Functions, Axioms, and Peer Review. In *ICML Workshop on Incentives in Machine Learning*.
[43] Kanu Okike, Kevin T. Hug, Mininder S. Kocher, and Seth S. Leopold. 2016. Single-blind vs Double-blind Peer Review in the Setting of Author Prestige . *JAMA* 316, 12 (09 2016), 1315–1316. https://doi.org/10.1001/jama.2016.11014 arXiv:https://jamanetwork.com/journals/jama/articlepdf/2556112/jld160026.pdf
[44] S. R. Paul. 1981. Bayesian methods for calibration of examiners. *Brit. J. Math. Statist. Psych.* 34, 2 (1981), 213–223.
[45] Simon Price and Peter A Flach. 2017. Computational Support for Academic Peer Review: A Perspective from Artificial Intelligence. *Commun. ACM* 60, 3 (2017), 70–79.
[46] Drummond Rennie. 2016. 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 (2016), 31–34.
[47] Milton Rokeach. 1968. The Role of Values in Public Opinion Research. *Public Opinion Quarterly* 32, 4 (1968), 547–559. https://doi.org/10.1086/267645
[48] Magnus Roos, Jörg Rothe, and Björn Scheuermann. 2011. How to Calibrate the Scores of Biased Reviewers by Quadratic Programming. In *AAAI Conference on Artificial Intelligence*.
[49] Nihar B Shah and Zachary Lipton. 2020. Fairness and Bias in Peer Review and other Sociotechnical Intelligent Systems (AAAI 2020 Tutorial Syllabus). AAAI 2020 tutorial.
[50] Nihar B Shah and Zachary Lipton. 2020. SIGMOD 2020 Tutorial on Fairness and Bias in Peer Review and Other Sociotechnical Intelligent Systems. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 2637–2640.
[51] Nihar B Shah and Zachary Lipton. 2020. TheWebConf 2020 Tutorial: Fairness and Bias in Peer Review and other Sociotechnical Intelligent Systems. TheWebConf 2020 tutorial. (2020).
[52] 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.
[53] Stanley S Siegelman. 1991. Assassins and zealots: variations in peer review. *Radiology* 178, 3 (1991), 637–642.
[54] Richard Smith. 2006. Peer review: a flawed process at the heart of science and journals. *Journal of the royal society of medicine* 99, 4 (2006), 178–182.
[55] Flaminio Squazzoni and Claudio Gandelli. 2012. Saint Matthew strikes again: An agent-based model of peer review and the scientific community structure. *Journal of Informetrics* 6, 2 (2012), 265–275.
[56] Ivan Stelmakh, Nihar Shah, and Aarti Singh. 2019. On Testing for Biases in Peer Review. In *NeurIPS*.
[57] Ivan Stelmakh, Nihar Shah, and Aarti Singh. 2019. PeerReview4All: Fair and Accurate Reviewer Assignment in Peer Review. In *Conference on Algorithmic Learning Theory*.
[58] Ivan Stelmakh, Nihar Shah, and Aarti Singh. 2021. Catch Me if I Can: Detecting Strategic Behaviour in Peer Assessment. In *AAAI*.
[59] Ivan Stelmakh, Nihar Shah, Aarti Singh, and Hal Daumé III. 2020. A Large Scale Randomized Controlled Trial on Herding in Peer Review Discussions. *arXiv* (2020).
[60] Ivan Stelmakh, Nihar Shah, Aarti Singh, and Hal Daumé III. 2021. A Novice-Reviewer Experiment to Address Scarcity of Qualified Reviewers in Large Conferences. In *AAAI*.
[61] Ivan Stelmakh, Nihar Shah, Aarti Singh, and Hal Daumé III. 2021. Prior and Prejudice: The Novice Reviewers' Bias against Resubmissions in Conference Peer Review. In *CSCW*.
[62] 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*.
[63] Warren Thorngate and Wahida Chowdhury. 2014. By the Numbers: Track Record, Flawed Reviews, Journal Space, and the Fate of Talented Authors. In *Advances in Social Simulation*. Springer, 177–188.
[64] 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
[65] Chris R Triggle and David J Triggle. 2007. What is the future of peer review? Why is there fraud in science? Is plagiarism out of control? Why do scientists do bad things? Is it all a case of: "All that is necessary for the triumph of evil is that good men do nothing?". *Vascular health and risk management* 3, 1 (2007), 39.
[66] Jingyan Wang and Nihar Shah. 2018. There's Lots in a Name (Whereas There Shouldn't Be). Research on Research blog. https://researchonresearch.blog/2018/11/28/theres-lots-in-a-name/.
[67] Jingyan Wang and Nihar Shah. 2019. Gender Distributions of Paper Awards. Research on Research blog. https://researchonresearch.blog/2019/06/18/gender-distributions-of-paper-awards/.
[68] Jingyan Wang and Nihar B Shah. 2019. Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings. In *AAMAS*.
[69] Jingyan Wang, Ivan Stelmakh, Yuting Wei, and Nihar Shah. 2021. Debiasing Evaluations that are Biased by Evaluations. In *AAAI*.
[70] Mark Ware. 2008. Peer review: benefits, perceptions and alternatives. *Publishing Research Consortium* 4 (2008), 1–20.
[71] Thomas J. Webb, Bob O'Hara, and Robert P. Freckleton. 2008. Does double-blind review benefit female authors? *Trends in Ecology and Evolution* 23, 7 (2008), 351 – 353. https://doi.org/10.1016/j.tree.2008.03.003
[72] Yichong Xu, Han Zhao, Xiaofei Shi, and Nihar Shah. 2019. On Strategyproof Conference Review. In *IJCAI*.