Identifying Reviewer Disciplines and Their Impact on Peer Review

Chenyue Jiao, Erzhuo Shao, Louis Shekhtman, Satyaki Sikdar

Peer review is crucial to the research enterprise, serving as quality control and subjecting research to rigorous standards (Bornmann, 2011). Given the rise in interdisciplinary and multidisciplinary research, there has been an increasing need to obtain the perspectives of multiple peer reviewers, each with unique expertise on different aspects of an interdisciplinary research topic (Baker, 2019; Bammer, 2016). Anecdotally, researchers performing interdisciplinary work using a particular disciplinary framework have often commented on the challenges that arise during peer review, especially from reviewers coming from a different disciplinary training which might use distinct methods, approaches, and standards of evidence (Carrel et al., 2022; Lamers et al., 2021; Langfeldt, 2006; Teplitskiy et al., 2018). In this project, we aim to explore how reviewer disciplines relate to the peer review process and if reviewers from particular disciplines tend to be more critical of papers applying a different disciplinary approach.

METHODS

We begin by leveraging a unique dataset of open reviews from Nature Communications to answer these questions. We began by downloading the full texts and reviews of 20,510 papers and focused on the first round of reviews, which likely contained the most substantial comments. Overall, our final dataset contains 12,587 papers with 35,198 reviews. To determine the disciplines associated with these papers and reviews, we used the OpenAlex concept tagger[1], which, given a title and abstract, tags a paper with a set of relevant concepts from a set of 65,027 concepts from WikiData. Notably, these concepts are organized hierarchically, with 19 concepts at level 0 which are broad fields like Physics or Mathematics, to level 1 concepts that are more specific such as Statistical Physics or Geometry, and so on going down. We successfully tagged 2,894 papers and 8,033 reviews, specifying the concepts they are most related to. We also view the scores defining the extent to which each paper is associated with each concept as an embedding vector, allowing us to measure the similarity between the vector of concepts for a particular review and the paper on which it is commenting. Finally, to understand the sentiment of the review we feed each paper into OpenAI's ChatGPT API and provide it with the text of the peer review and the prompt "*The above is the peer review of a research paper. Grade the attitude of this reviewer (from 1 to 5)*" to have ChatGPT rate the sentiment of each review five times. We set the reviewer score as the average of these five scores.

[1] https://github.com/ourresearch/openalex-concept-tagging

RESULTS

We constructed a network of the level 0 concept tags cited in papers and reviews. We linked these concepts with a directed link from the main reviewing concept to the main paper concept and determined the mean sentiment of that link (scale from 1-5 from ChatGPT). In Fig. 1, we visualize this network, where Political scientists write primarily positive reviews. At the same time, Engineers are relatively more negative on Physics papers; Economists are more likely to rate Biology negatively. Computer scientists, on average, rate Engineering papers more favorably than papers on Medicine.

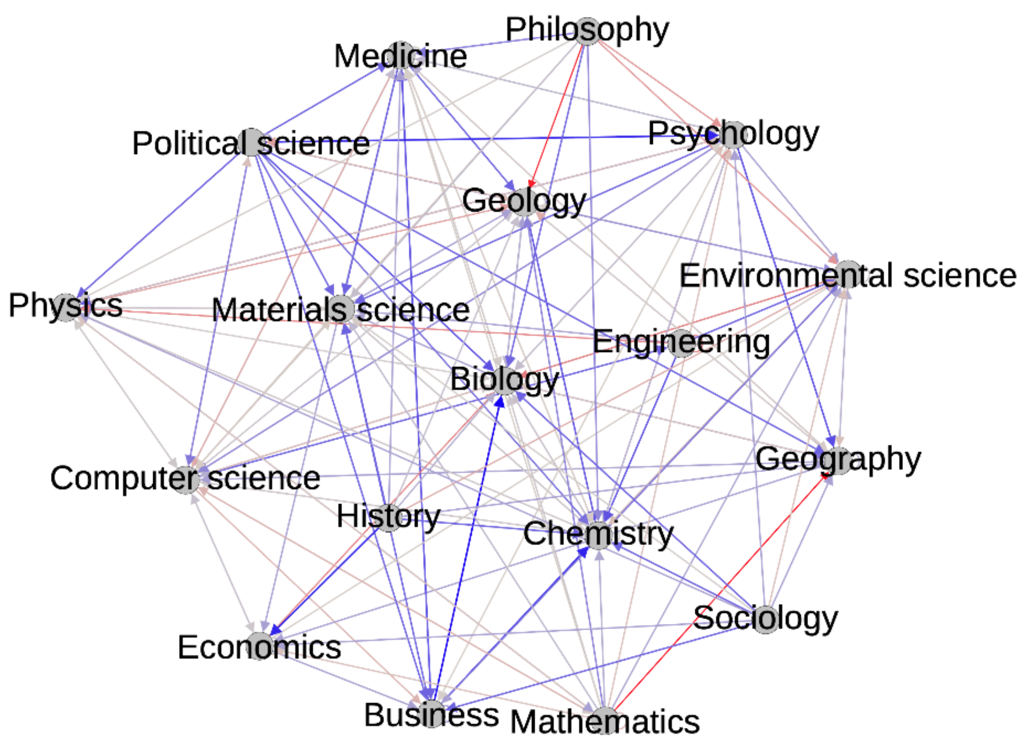

Figure 1: Network of highest-level concepts with links pointing from the reviewers' concept to the papers' concept. Link colors represent the relative rating of the reviewer, with blue links reflecting more positive reviews, gray links reflecting relatively more neutral reviews, and red links reflecting more positive reviews.

We further explored the cosine similarity of the concept vectors of the reviews to those of the papers as a proxy for the semantic distance. When we examined how this semantic distance correlated with the review ratings in Fig. 2, we did not find that reviews conceptually closer to the papers were more positive. Instead, the effect was insignificant and suggested that a lower semantic distance led to decreased reviewer scores.

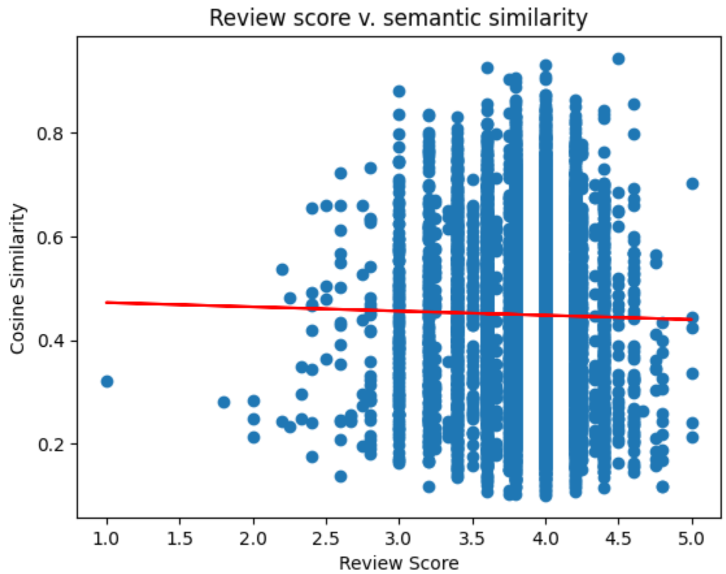

Figure 2: Scatter plot of ChatGPT assigned reviewer score v. cosine similarity between a paper and its reviews for 8,038 reviews of 2,894 papers. The OLS line is drawn in red, with slope -0.008 and intercept 0.48, and p = 0.144 shows a slightly negative trend.

CONCLUSIONS

Our results identify network pairs of review and paper concepts, identifying relatively more positive and more negative ratings of papers by reviews. Several of these pairs correspond to prior expectations, such as engineers' disapproval of many physicists' work and economists' criticisms of biology studies. At the same time, a broader regression between reviewer scores and the cosine similarity between concepts in papers and reviews did not yield a significant result and suggested a negative correlation, implying that reviews that use concepts more similar to a paper are less favorable.

These findings have limitations and require more comprehensive work to understand how reviewer discipline affects reviewer commentary on papers from other disciplines. First, we only have papers that were accepted at Nature Communications. It can be expected that rejected papers had more negative reviews and perhaps more reviews from disciplines distinct from those of the paper. Furthermore, our approach to understanding the sentiment of the paper relies on ChatGPT assigning a rating on a scale between 1 and 5. It is biased towards providing more positive reviews; the mean score centers around 4. Finally, the interpretation of these results has many confounders as OpenAlex is classifying the topics noted in the review, which may or may not reflect the reviewers' actual disciplinary training. Further work and collection of additional datasets of full peer reviews with actual reviewers' names from interdisciplinary venues could improve these results.

REFERENCES

1. Baker, N. (2019). A network of science: 150 years of Nature papers. *Nature*. https://doi.org/10.1038/d41586-019-03325-6
2. Bammer, G. (2016). What constitutes appropriate peer review for interdisciplinary research? *Palgrave Communications*, *2*(1), 16017. https://doi.org/10.1057/palcomms.2016.17
3. Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, *45*(1), 197–245. https://doi.org/10.1002/aris.2011.1440450112
4. Carrell, S., Figlio, D., & Lusher, L. (2022). *Clubs and Networks in Economics Reviewing*. National Bureau of Economic Research. https://doi.org/10.3386/w29631
5. Lamers, W. S., Boyack, K., Larivière, V., Sugimoto, C. R., van Eck, N. J., Waltman, L., & Murray, D. (2021). Investigating disagreement in the scientific literature. *ELife*, *10*, e72737. https://doi.org/10.7554/eLife.72737
6. Langfeldt, L. (2006). The policy challenges of peer review: Managing bias, conflict of interests and interdisciplinary assessments. *Research Evaluation*, *15*(1), 31–41. https://doi.org/10.3152/147154406781776039
7. Teplitskiy, M., Acuna, D., Elamrani-Raoult, A., Körding, K., & Evans, J. (2018). The sociology of scientific validity: How professional networks shape judgement in peer review. *Research Policy*, *47*(9), 1825–1841. https://doi.org/10.1016/j.respol.2018.06.014