Peer Feedback for Cambria Naslund and Vicki Yang

From: David M. Liu

Hi Cambria and Vicki,

Congratulations on writing a succinct, cohesive paper. I found it impressive that the two of you were able to survey an entire field within the scope of a class project. Specifically, I found your group’s work to be a win for “Big Data” methodology. Reading the paper helped me better appreciate all of the data collection efforts orchestrated to make this project possible. Because of your successful data collection leveraging APIs and databases, I believe that your findings are interesting to a wide body of researchers – sociologists and chemists alike.

Below, I have included specific feedback points that I think could bolster this article even further. Overall, the setup and results were very clearly explained and laid out. The majority of my comments are aimed at better defending these results against potential inquiries. I would also note that these comments are not provided in a strict order:

My first comment is actually a question. I am curious which audience this paper is targeted towards? If I assume that this paper is targeted towards sociologists, I would recommend further justifying why the field of chemistry was studied. At the moment, the motivation seems to primarily stem from the fact that Foster et. al had conducted a study. Instead, I would be curious if a STEM field like chemistry has been generally regarded as “prone” to gender citation bias. Perhaps there are previous qualitative works in which female chemists have voiced concern. In either case, more motivation behind choosing chemistry would help guide the assumed target audience of sociologists.

One of my favorite sentences in the article was actually in the abstract when you two state, “While our research cannot refute the existence of gender-based disparities in the sciences…” This wording really helped me digest the results. Perhaps it is because we just took a year of statistics, but these results seem analogous to “failing to reject the null hypothesis”, where the null hypothesis is an unbiased scientific community. In other words, the findings do not argue that there is definitively no bias in chemical citations but that the collected data does not show any.

A suggestion I have regarding methodology is to cross reference the gender assignments from genderize.io against other sources of data. My main concern is that the missingness from the assignments, even if small, is correlated with the dependent variable (citation counts). For the sake of example, as partially acknowledged in the paper, it is possible that names of foreign descent are disproportionately missing gender assignment but that gender bias is also particularly significant within populations of foreign authors. As a suggestion, I would consider cross referencing against the websites for researchers for which genderize.io is uncertain or unable to provide a label. I am not sure it would work, but you could try running an image classifier on the headshots for the authors. Or more likely, I would manually assign the gender if there are only a few ambiguous cases.

My next two comments are statistical. First, I appreciate that the model specification in equation 2. I think it was smart to not only insert a dummy variable for gender but to also create interaction terms between gender and strategy. Even though the result is to fail to reject regardless, I believe this model specification provides more depth to the finding, stating that bias does not appear to exist even within strategy groups.

My second statistical comment is to consider using year fixed effects in the model. In other words, instead of a single linear term for year in equations 2 and 3, I would consider a dummy variable for each year. Or if there are too many years, perhaps a dummy variable for each decade. I say this because it is possible that the relationship between citations and time is not linear. I can imagine a scenario in which citation counts in chemistry are sinusoidal, perhaps general interest in chemistry is cyclic and there have been peaks and troughs in overall citation counts.

Again, I think your results are accessible and interesting to a wide population. Even your methodology will likely be useful to researchers who may be interested in extending your approach to a different area of science. Great work!

Best,

David