Response to PLOS ONE decision letter

Luke Holman and Claire Morandin

Comments from the Editor

Generally speaking, this manuscript brings a relevant contribution and is technically sound, but some additional effort is needed in terms of framing and result interpretation. In addition to comments made by Reviewer 1 (especially the one related to the meaning of 'preference'), please address the following issues:

1.- The motivation and contributions of this manuscript against the state of the art should be stated in a clearer way in the introduction. The article proposes a methodology to control the "Wahlund effect" in gender studies applied to scientific collaboration. What are the aspects of scientific collaboration potentially producing such an effect and, consequently, being the focus of this article? (e.g. different career stages and countries, heavily gender-biased disciplines ..).

Firstly, many thanks for your feedback and your attention to our manuscript, and this resubmission.

As noted by Reviewer 1, we need to be quite careful speculating as to the causes of 'gender homophily', because we have no direct data on these causes. The main purpose of this paper was to establish where homophily really exists as has been claimed (and measure it precisely in many fields of STEM), or whether it disappears once one accounts for statistical artefacts (what we call the Wahlund effect). There are many reasons why same-gender collaborations might be disproportionately common (besides artefacts), and we discussed many of these in the paper (a few in the Introdution, and several more in the Discussion). The most obvious one is that people prefer to work with others who are similar to themselves, and/or they prefer to avoid opposite-sex collaborators. It's also not clear whether this hypothetical preference is predominantly expressed by men, by women, or both. We could get into this more if you like, but we note that reviewer 1 already thought our cautious treatment was a little too much, and they disliked the paragraph in which we speculated about the individual-level processes that drive the population-level patterns of gender assortment that we observed. We are quite happy with the balance between speculation and caution as it is, but we are open to specific suggestions.

2.- Please, motivate the need to explore the correlation between α' and the number of authors.

We have now expanded this section to provide some motivation, which was previously left until the Discussion. Basically, we are trying to extract information about the processes that cause non-random assortment by gender, and responding to a colleague's comments on an earlier draft which specifically requested this figure. Here is the revised section of the Results:

One possible explanation for this finding is that 2-authors papers might be more likely to have an author list that is evenly split between career stages (e.g. a postgraduate student and their supervisor), increasing the chance that the authors are mixed gender (see Figure 6). The result also suggests that the processes responsible for gender homophily are similar in small (e.g. 3 author) and larger (5+ authors) collaborations (and across disciplines where small versus large collaborations are the norm).

3.- The article reports a statistically significant linear relationship between standardized journal impact factor and α'. However, there is a lot of noise and the slope of the corresponding regression is approx. 0.012. This looks like a rather weak relationship. Does it really worth being stressed so much?

We don't think that we stress this result much. The impact factor result is only discussed in two very cautious sentences in the Discussion (cautious words shown in bold):

We also **found a little evidence** that gender homophily is detrimental to research quality, in that high-impact journals **tended to have** weaker homophily. **Assuming** that papers published in high-impact journals are of higher average quality **(which is contentious; [67])**, our results **provide non-experimental support for the hypothesis** that mixed-gender teams produce better research than single-gender teams [42–48].

Please let us know if any of that still seems excessive upon a second reading. Also, a regression slope of -0.012 is actually quite a lot, for this dataset. The mean α' across journals is about 0.07, and so this slope means that increasing the journal impact factor by one standard deviation above the mean decreases α' from about 0.070 to 0.058; in proportional terms, this is 0.058 / 0.07 = 83% of the mean. The R^2 is 4%, meaning that if we know the journal impact factor, we can explain about 4% of the variance in α', which is not huge, but is perhaps bigger than one would expect given the large number of unmeasured covariates that also affect the two variables being correlated here. At any rate, we are very upfront in saying the relationship is noisy, we give all the statistics and confidence intervals needed to interpret the result, and we are careful to remind readers that correlation is not causality.

4.- What does 'significant heterogeneity' mean in line 141? I guess they are located above the 95% CI, but please specify.

Line 141 (in the original draft) actually says 'significant heterophily', where heterophily is our term for an excess of opposite-sex coauthor pairs relative to the randome expection (defined in the Introduction). We define 'significant' heterophily as the case where our measure of assortment by gender, α, is negative and has 95% confidence limits that lie entirely below zero. We explain this definition of statistical significance in the Methods (page 15).

Comments from Reviewer 1

This study investigates correlations among the genders of authors of papers, using authorship data extracted from the Pubmed database. Prior studies have already shown that there is homophily among authors, i.e. that publications include more authors of the same gender than expected under random assortment of authors. However, homophily may be the result of hidden subdivision of the data, for instance if different subfields have different male:female ratios. The authors call this a "Wahlund effect", by analogy with the effect of population subdivision on homozygozity in population genetics.

The study is overall well-written and well-referenced, as far as I can tell. I would also like to congratulate the authors for providing the entire R code behind their study, in an amazingly well-commented file. This is still too rare and I surely hope that this will become a standard.

That said, I have a few comments on the manuscript. They are presented below as they appear in the manuscript, and I highlight just below the most important ones.

You have our sincere thanks for your time and the very useful feedback, as well as these compliments.

Main comments

1) It is an issue common to all these "big data" studies, but the limitation should be better highlighted in the study. You are considering the lists of authors of articles in different journals, but you do not have information on individuals themselves. On of the main issues to identify the production of individuals is the presence of homonyms, and different ways to identify an author in a paper (e.g., F. Lastname vs Firstname Lastname). Consequently, the results are about the correlation among genders of authors of papers, but not directly about individual preference. For instance, you may have some individuals who are over-represented in your dataset (very prolific authors – or powerful heads of institutions adding their names everywhere), and their presence on different papers/journals may be counted as independent units – while they are not. I am aware that the problem cannot be solved with ISI or Pubmed data, but it really needs to be acknowledged.

We did not knowingly make any claims that pertain to individuals: we are well aware that because our study focuses on population-level patterns of assortment by gender, we can only speculate about the individual-level processes that produce the population-level patterns. Therefore, the fact that we cannot (or rather, did not try to) identify specific individuals does not compromise the paper's results and conclusions. We agree that some uses of the word 'preference' in our manuscript could be misleading in this respect, and so in the revision, we have replaced the word in key place with more neutral/agnostic terms (e.g. 'non-random assortment').

I do not think that the authors can conclude about individual preferences. As explained

above, the study is not at the scale of individuals, and I fear that the over-representation vs under-representation of some authors may play a part in the results. However, it is entirely fair to conclude about list of authors being more same-gendered than expected by chance.

See previous comment; we agree that we cannot make firm individual-level inferences. However, we believe that the population-level patterns almost certainly result from individuals' decisions and choices (even though these choices are constrained, e.g. by the shortage of women in supervisory roles, minor taboos against opposite-sex collaboration in some countries, etc), and so we believe it is important to speculate about the causal processes (while making it clear what is speculation and what is demonstrable fact). In the revision, we also take extra care to avoid possible sources of reader confusion, given that this is a contentious topic and the paper is quite statistics-heavy. For example, we point out that individual people with all-male collaborators are not necessarily picking their collaborators by sex, since thousands of people would have all-male collaborators even in a world with totally random assortment (especially in mostly-male fields).

In addition, I'd advise against the use of "preference/preferentially", because it may imply a choice. The correlation can however be a by-product of other factors, and the data are not about actual preference (while it would for instance if the authors had been surveyed).

Agreed: this word can denote active/conscious choice of collaborators based on gender, which is only one possible mechanism that might produce homophily. This word does not properly capture our intended meaning (i.e. "greater frequency than expected by chance"), and so we have re-written the manuscript with this in mind.

2) It is almost impossible to properly understand the part on theoretical explanations with the information provided in the manuscript. While it is nice to provide the R code, the reader should be able to have a rather precise idea of what you did without having to dive into the R code. The corresponding subsection in the Methods should be expanded and much more detailed.

We have now expanded the methods, and greatly expanded the supplementary material associated with Figure 6. See our response to this point in the response letter below.

3) An important motivation for the study was to control for a Wahlund effect in this kind of study. It would be great to actually illustrate this feature by providing the estimate for alpha' when all journals are pooled together (otherwise, this seems to be only a replication of previous studies).

Firstly, we took the reviewer's recommendation and calculated α for our entire PubMed dataset (all journals and all 15 years; n = >3 million papers, >16 million authors), using the same methods as were used for a single journal. We found that $\alpha = 0.126$, which is almost twice as high as the median α for individual journals ($\alpha = 0.070$, shown in top left of Figure 2). This suggests that indeed, α

can be inflated if one lumps together data from very different research disciplines and time periods, as shown in Figure 1. However, α remained positive even though we controled for the Wahlund effect far more stictly than in all previous studies, suggesting that those studies are likely to be broadly correct (though they might over-estimate the strength of homophily). We now discuss this new result in the first paragraph of the Results section.

Secondly, our study is not a simple replication of previous work, whatever the value of α' pooled across all the journals. To our knowledge it is the only study to explain that statistical artefacts can theoretically produce 'spurious' gender homophily (and probably did, in some of the previous studies). Additionally, ours is the largest study of gender homophily, and the broadest in the set of research disciplines it covers. Also, we selected PLOS ONE partly because its editorial policy does not penalise replication studies or papers with non-novel results.

Comments as they appear in the manuscript

(I am also mentioning typos since PLOS One papers are just put in the PLOS format but not properly copy-edited)

Title and Abstract

Flags raised while reading: - Meaning of preferentially: actual preference (active behavior) or other factors also playing a role? (see point 1) above)

See our response above

- "High impact factor" is this a proxy for quality? (the point is addressed later in the discussion)

No; we agree journal impact factor is a poor proxy for quality. A LaTeX error in the original submission prevented part of the Discussion from appearing: it was supposed to read "Assuming that papers published in high-impact journals are of higher average quality (which is contentious; [16])...". Originally, the "contentious" part did not format properly.

Introduction

- line 62 please spell out "m" (not sure it's international notation since m is meters)

We now write 'million' instead of m.

- line 77 please spell out FDR the first time you use it (since the methods are currently at the end of the manuscript)

Done

- line 89 Author order differs across disciplines, e.g. it is alphabetical in maths. Please specify that you focus on life-sciences journals.

We already specify that PubMed focuses on life sciences journals, as the reviewer notes elsewhere. Also, the presence of some alphabetically-ordered papers in the set would not alter the fact that the average seniority of last authors is greater than that of first authors, in most/all of the journals we studied.

Methods

- line 265 Please specify how gender was assessed (even though this may already be in [5]). In particular, was it done automatically or manually? There may be issues with automatic gender assignment for authors of different geographic origins.

We prefer to leave the full details to our previous paper, which has extensive descriptions, two different quality controls, and the full R code that was used to programmatically identify gender using authors' first names. Given our sample size (36 million authors), it hopefully goes without saying that automatic methods were used, but we now point this out explicitly in the Methods. Our earlier paper has extensive discussion of the geographic origin issue, and success rates for each of >100 different countries (almost all of which are over 70%, with China and Korea being significant exceptions among the major research-producing nations).

- line 271 Maybe it is clearer to say what you kept rather than what you discarded (avoid double negations)

Fixed, thanks.

- lines 296-303 Please better explain where the effect comes from: you cannot be your own co-author. If there are m men and f women, a man can be a co-author with (m-1) men while women can be a co-author with m men. If you could be your own co-author, there would be 10 possibilities in your 2-author examples, and alpha would be 0.

Changed as suggested, thanks.

- line 322 Now it's PLOS (not PLoS anymore)

Changed throughout

- line 326 cf. point 3) above

We addressed this by calculating α across the whole of PubMed; see above.

- line 352 log (no capital L)

Done

- line 362 Why would senior authors be the main drivers? It could as well be junior authors who chose their advisor...

Indeed, the purpose of the first author vs last author analysis was to see if we can get some clues as to whether homophily is driven mostly by senior or junior scientists: we outlined both possibilities in this section. I think you were misled because the two sentences beginning on line 362 use the word "Assume" rhetorically; the sentences basically say, "Assuming senior researchers drive the results, we predict X. This is because of Y." We do not start from the assumption that senior researchers are responsible for homophily. We also began that paragraph by pointing out that the 'senior researchers drive homophily' hypothesis is already predominant in the literature (though to our knowledge it has not been effectively tested before).

- lines 397-407: Where do your data come from? actual proportions of women among early vs late career researchers?

There is no data - it's a theoretical model, as explained in the section referenced here (e.g. the title says "Theoretical expectations"). We investigated all combinations of values from 0% - 60% women, which spans >95% of journals on PubMed (see Holman et al. 2018).

Results

- Figure 2: – please use transparent shade so that the curve behind remains visible when it is lower than the curve in the front.

This is a stacked density plot, so there is no curve behind the other one (the blue is stacked over the white). We have clarified this in the figure legend.

– Right-hand side figure: what do the dots correspond to? a single journal?

Yes – we now state this in the figure legend.

- Do you identify individual authors/ Control for the fact that some are more prolific than others?

No, but this would not affect the veracity of our findings. Let's imagine that 90% of all papers in the world had one very prolific individual in their author lists. Let's assume this author is a man, though the argument would be the same if they were a woman. The existence of this prolific man would increase the proportion of male co-authors for both sexes, since most people have co-published with him. However, it would not affect our test statistic α', because the proportion of male coauthors would be elevated equally for men and women (and so α' would not change). You might ask, what if the super-author has preference for one gender of coauthors – wouldn't that skew the results somehow? But then, α' would be working as intended: it would correctly record that homophilous (or heterophilous) co-authorships were more common than expected under random assortment by gender. This highlights a feature of α' noted earlier by the reviewer: it is a population-level measure, so we cannot make statements about individuals, e.g. we don't know if everyone is a little bit homophilous, or

if there are a few prolific and very homophilous people (though the former is probably closest to the truth, since super-prolific authors with atypical levels of homophily are presumably the exception not the rule).

Additionally, the basic version of α, proposed in a short pre-print by Bergstrom et al., would be downwardly-biased (basically, because our hypothetical super-author cannot co-author papers with himself). We developed α' in order to get rid of this issue, though in practice the two versions of alpha were close to identical (see Figure S8), except for a few very small subsets of the data, so the basic α is fine most of the time.

- Theoretical expectations: it is not possible to understand what has been done in this section. The methods section about this part is too limited; the reader is asked to go and check an annotated R code; we should be able to have at least an idea of what has been done. As a result, it is not really possible to fully understand Figure 6, and in particular, the differences between columns in the figure. Also, we'd need details about the choices made regarding the list of authors, such as the number of authors.

We have expanded the methods in the paper so that this part is clear without reading the supplement, as well as expanding the supplement to make it legible to people who have never seen R code. We also added much clearer variable names to the R code, and added extra comments to the code to make it easier to follow. For your convenience, here is the revised paragraph of the methods describing this theoretical exercise:

We can think of no tractable method of controlling for this issue using our dataset, which contains no information on career stage. Therefore, we instead decided to derive theoretical expectations for α when there is a difference in gender ratio across career stages, in order to determine if and how this effect should alter our inferences. For simplicity, our calculations assume there are only two career stages ('early-career' and 'established'), though we expect that the general conclusions would also apply to a multi-tier career ladder. Under the null model that gender has no causal effect on collaboration, we calculated α for various combinations of the four free parameters in our simple model. These parameters are: the gender ratio among early-career researchers (x-axis of Figure 6), the gender ratio among established researchers (y-axis of Figure 6), the frequency of within- versus between career stage collaborator pairs (rows in Figure 6), and lastly the frequency of within-stage collaborations that are between two early-career researchers as opposed to two late-career reserachers (columns in Figure 6). When these four parameters are specified, one can easily calculate the relative frequencies of collaborator pairs that involve two men, two women, or a man and a woman. In short, if we have specified the frequency of women at both career stages, as well as the frequency of the three possible types of collaboration with respect to career stage (early-early, early-established, and established-established), then we can calculate the frequency of collaborators pairs comprising two women, or a woman and a man, and thus find α (see the Online Supplementary Material for the annotated R code).

Discussion

- line 173 : "We found evidence that researchers preferentially publish with same-gendered coauthors" I do not think that this conclusion is adequate given your data. You found evidence that there is a correlation between the genders of authors, but your data does not provide evidence regarding the behavior of individual authors. Even if there were also such a correlation at the individual level, your data does not allow you to identify the cause of this correlation – a preference (implying some choice) or a by-product of other constraints.

We originally used the word 'preference' a bit incautiously, and now use terms like 'assortment' that are more agnostic about the mechanism. We do think individual-level preferences are likely to be partly responsible, however, and so we discuss preference in a few places, while spelling out that it is one possible interpretation for the results.

- line 176: "Across the life sciences" note that Pubmed is a more limited database than e.g. Web of Science, even for biology... so some disciplines are not covered, the main focus is on biomedical sciences.

We are aware that PubMed (and even Web of Science, which is not open-access) does not cover the entire life sciences literature, but we think the statement is accurate as written (since 'across' does not imply 'all'). PubMed does span essentially all major fields of the life sciences, even if its coverage of fields and journals is not 100%. See Figures 1 and 2 in Holman et al. 2018, and the accompanying web app, which lists data for >6000 journals indexed by PubMed (these include essentially all of chemistry, biology, psychology, biophysics, etc in addition to things related to medicine and health).

- line 181: "Typical gender ratio" Please give a range (it seems low in the example)

To support this statement, we cited Holman et al. 2018, which draws on data from the author lists of >10 million research papers, and shows that many fields have at least this great a shortage of women authors. See Figures 1 and 2 in that paper for the percentages (there are >100, for many diverse fields, so it would not be helpful to give a one-size-fits-all estimate without 'getting into the weeds' about why we picked that particular estimate, which is not relevant for the point being made on line 181).

- line 181: "there was no inflation of α in highly multidisciplinary journals" One of the aims of your study was to control for Wahlund effect. Maybe you could show the value of alpha' if you group all journals, to better motivate the need to control for the effect of "subdivision" by journal/disciple?

As suggested, we now calculate α for every suitable PubMed paper in our dataset (3.7 paper, 16m authors), ignoring the journal and publication date, to see if it inflates α as we expected. Reassuringly, the α for PubMed as a whole is 0.126, which is almost double the median α calculated on 1-year datasets for individuals journals (i.e. our main result, shown in Figure 1), supporting our theoretical

argument that ignoring the Wahlund effect does add a non-trivial amount of inflation to α.

- Line 246 "our results provide non-experimental support for the hypothesis that mixed-gender teams produce better research than single-gender teams" is better said this way than in the abstract

We feel that both phrasings are carefully chosen, and accurately reflect what we found in the study. For comparison, the Abstract equivalent of this sentence reads, "Interestingly, journals with a high impact factor for their discipline tended to have comparatively low homophily, as predicted if mixed-gender teams produce better research." Note that: 1) the word 'Interestingly' reflects our feeling that this result is an unexpected curio that requires follow up study (although there are already some experimental studies examining the effects of diversity on the success of team-based projects, which we cite), 2) the word 'tended' stresses that this is not a hard-and-fast finding, but rather a noisy correlation, and 3) the word 'predicted' stresses to the reader that this result is consistent with the theory but does not prove it (that is, we do not commit the fallacy of 'affirming the consequent').

- line 257: "Given that many collaborative research projects unfortunately involve a gendered division of labour [61], working with a same-gendered colleague may provide exposure to new parts of the research process, and (especially for the minority gender) a welcome change of pace." Not sure I understand (/agree with) the logic behind this last sentence. Do you mean that if you collaborate with a same-gendered colleague, you may end up doing things that are different than if you collaborated with a different-gendered colleague? This sounds a bit odd to me... A different type of interaction, different level of implication, sure. But if I am a lab tech, working with a male or female PI won't really make me change the type of work that I do – or will it? And I am not sure I understand what is alluded to by "welcome change of pace"...

You mention the right answer here – our intended meaning was "a different type of interaction". For women working in a field where women are the minority (e.g. mathematics or surgery), the majority of their professional interactions will be with men, and so working with another woman would be a (perhaps welcome) novelty, given that men and women differ consistently in various traits (and that some people sometimes respond differently to male-female collaborator pairs relative to same-sex pairs, e.g. by making gender-based assumptions about who did what towards the project).

Supplementary

- Kudos for providing the code! Rmarkdown is wonderful (show/hide comments is also a great feature!).

Thanks, we agree it's an excellent tool for presenting the details of a coding-intensive analysis and maximising its transparency, reproducibility, and potential

to be adapted and re-used by other researchers.

- Could there be a pdf output as well? and a table of contents and lists of tables/figures? To avoid having to use ctrl-F to find the supplementary figures...

If the MS is accepted, PLOS will ask us to upload all the supplementary figures and tables one by one, giving a clickable route to each one of the supplementary figures in the paper itself, and so it's not necessary to make a table of contents for our HTML supplement. As for making it into a PDF, this doesn't seem useful (since it will contain the same information, but will lose all the nice features of HTML), but we're happy to do it if asked to by the copy editors. This will simply involve opening the HTML document using a web browser, then saving it as a PDF; we can do this if asked, or leave it up to any readers who need a PDF copy (note that HTML can be opened using any web browser, and is likely to be just as future-proof as PDF).

- Fig S2 seems to be missing.

Thanks, we fixed a typo and the figure now shows up.