SJS The Self Journal of Science

Essay and Opinion

Peer Evaluation of Science
Version 1

Thomas Arildsen [1]

1. Department of Electronic Systems - Aalborg University

Made public on Sep, 28th 2015 under Creative Commons 4.0 Attribution License
Reviewed and discussed at http://www.sjscience.org/article?id=401

Abstract This is a proposal for a system for evaluation of the quality of scientific papers by open review of the papers through a platform inspired by StackExchange. I have chosen to publish this proposal on SJS since this is a platform that comes quite close to what I envision in this proposal. I hope that readers will take the opportunity to comment on the proposal and help start a discussion about it.

Peer Evaluation of Science

Thomas Arildsen

2015-9-28

1 Introduction

Researchers currently rely on traditional journals for publishing their research. Why is this? you might ask. Is it because it is particularly difficult to publish research results? Perhaps 300 years ago, but certainly not today where anyone can publish anything on the Internet with very little trouble. Why do we keep publishing with them, then? - they charge outrageous amounts for their services in the form of APCs from authors or subscriptions from readers or their libraries. One of the real reasons, I believe, is prestige.

The purpose of publishing your work in a journal is not really to get your work published and read, but it is to prove that your paper was good enough to be published in that particular journal. The more prestigiuous the journal, the better the paper, it seems. This roughly boils down to using the impact factor of the journal to evaluate the research of authors publishing in it (bad idea, see for example Wrong Number: A closer look at Impact Factors). It is often mentioned in online discussions how researchers are typically evaluated by hiring committees or grant reviewers based which journals they have published in. In Denmark (and Norway - possibly other countries?), universities are even getting funded based on which journals their researchers publish in.

I think the journal's reputation (impact factor) is used in current practice because it is easy. It is a number that a grant reviewer or hiring committee member can easily look up and use to assess an author without having to read piles of their papers on which they might have to be experts. I support a much more qualitative approach based on the individual works of the individual researcher. So, to have any hope of replacing this practice, I think we need to offer a quantitative "short-cut" that can compete with the impact factor (and H-index etc.) that say little about the actual quality of the researcher's works. Sadly, a quantitative metric is likely what hiring committees and grant reviewers are going to be looking at. Here I think a (quantitative) "score" or several such scores on different aspects of a paper accompanying the (qualitative) review can be used to provide such an evaluation metric. Here I am going to present some ideas of how such a metric can be calculated and also some potential pitfalls we need to discuss how to handle.

I believe that a system to quantify various aspects of a paper's quality as part of an open review process could help us turn to a practice of judging papers and

their authors by the merits of the individual paper instead of by the journal in which they are published. I also believe that this can be designed to incentivise participation in such a system.

Research and researchers should be evaluated directly by the quality of the research instead of indirectly through the reputation of the journals they publish in. My hope is to base this evaluation on open peer review, i.e. the review comments are open for anyone to read along with the published paper. Even when a publisher (in the many possible incarnations of that word) chooses to use pre-publication peer review, I think that should be made open in the sense that the review comments should be open for all to read after paper acceptance. And in any case, I think it should be supplemented by post-publication peer review (both open in the sense that they are open to read and also open for anyone to comment - although one might opt for a restriction of reviewers to any researcher who has published something themselves as for example Science Open uses).

What do I mean by using peer review to replace journal reputation as a method of evaluation? This is where I envision calculating a "quality" or "reputation" metric as part of the review process. This metric would be established through a quality "score" (could be multiple scores targeting different aspects of the paper) assigned by the reviewers/commenters, but endorsed (or not) by other reviewers through a two-layer scoring system inspired by the reputation metric from StackExchange. This would, in my opinion, comprise a metric that:

1. specifically evaluates the individual paper (and possibly the individual researcher through a combined score of her/his papers),

2. is more than a superficial number - the number only accompanies a qualitative (expert) review of the individual paper that others can read to help them assess the paper,

3. is completely transparent - accompanying reviews/comments are open for all to read and the votes/scores and the algorithm calculating a paper's metric is completely open.

I have mentioned that this system is inspired by StackExchange. Let me first briefly explain what StackExchange is and how their reputation metric works: StackExchange is a question & answer (Q&A) site where anyone can post questions in different categories and anyone can post answers to those questions. The whole system is governed by a reputation metric which seems to be the currency that makes this platform work impressively well. Each question and each answer on the platform can be voted up or down by other users. When a user gets one of his/her questions or answers voted up, the user's reputation metric increases. The score resulting from the voting helps rank questions and answers so the best ones are seen at the top of the list.

2 The System

A somewhat similar system could be used to evaluate scientific papers on a platform designed for the purpose. As I mentioned, my proposal is inspired by StackExchange, but I propose a somewhat different mechanism as the one based on questions and answers on StackExchange does not exactly fit the purpose here. I propose the following two-layer system.

- First layer: each paper can be reviewed openly by other users on the platform. When someone reviews a paper, along with submission of the review text, the reviewer is asked to score the paper on one or more aspects. This could be simply "quality", whatever this means, or several aspects such as "clarity", "novelty", "correctness". It is of course an important matter to determine these evaluation aspects and define what they should mean. This is however a different story and I focus on the metric system here.
- Second layer: other users on the platform can of course read the paper as well as the reviews attached to it. These users can score the individual reviews. This means that some users, even if they do not have the time to write a detailed review themselves, can still evaluate the paper by expressing whether they agree or disagree with the existing reviews of the paper.
- What values can a score take? We will get to that in a bit.

How are metrics calculated based on this two-layer system?

- Each paper's metric is calculated as a weighted average of the scores assigned by reviewers (first layer). The weights assigned to the individual reviews are calculated from the scores other users have assigned to the reviews (second layer). The weight could be calculated in different ways depending on which values scores can take. It could be an average of the votes. It could also be calculated as the sum of votes on each review, meaning that reviews with lots of votes would generally get higher weights than reviews with few votes.
- Each author's metric is calculated based on the scores of the author's papers. This could be done in several ways: One is a simple average; this would not take into account the number of papers an author has published. Maybe it should, so the sum of scores of the author's papers could be another option. Alternatively, it might also be argued that each paper's score in the author's metric should be weighted by the "significance" of the paper which could be based on the number of reviews and votes on these each paper has.
- Each reviewer's metric is calculated based on the scores of her/his reviews in a similar way to the calculation of authors' metrics. This should incentivise reviewers to write good reviews. Most users on the proposed

platform will act as both reviewers and authors and will therefore have both a reviewer and an author metric.

3 Which Values Can Votes Have?

I propose to make the scores of both papers (first layer) and individual reviews (second layer) a ± 1 vote. One could argue that this is a very coarse-grained scale, but consider the option of for example a 10-level scale. This could cause problems of different users interpreting the scale differently. Some users might hardly ever use the maximum score while other users might give the maximum score to all papers that they merely find worthy of publication. By relying on a simple binary score instead, an average over a (hopefully) high number of reviews and review endorsements/disapprovals would be less sensitive to individual interpretations of the score value than many-level scores.

4 Conclusion

As mentioned, I hope the proposed model of evaluating scientific publications by accompanying qualitative reviews by a quantitative score would provide a useful metric that - although still quantitative - could prove a more accurate measure of quality of individual publications for those that need to rely on such a measure. This proposal should not be considered a scientific article itself, but I hope it can be a useful contribution to a debate on how to make peer review both more open and more broadly useful to readers and evaluators of scientific publications.

I have chosen to publish this proposal on SJS since this is a platform that comes quite close to what I envision in this proposal. I hope that readers will take the opportunity to comment on the proposal and help start a discussion about it.