Distributed peer review of academic papers

Paolo Giordano

University of Vienna

The basic idea of this work is to study feasibility, features and consequences of a more objective, sure and qualitatively better review process for articles, project proposals, contributions for conferences. The basic idea is to develop a collaborative network where anyone (belonging to a certain community) can upload a paper and anyone can review a paper (different from its own). The obtained system must be self-organized, that is we must have a system of rewards that naturally conducts the participants to behave in a good way, and a system of exclusions that forces the exit from the community of bad behaving people. It would be better to prove its self-organization and stability using a mathematical model and not stating it as an opinion.

The first problem is if the system is stable with respect to intelligent tricks: does there exist a way to acquire reward from the system, without doing any real good work? It would be a great goal to insert AI-based methods in the above mentioned model so as to prove that the system is stable with respect to this type of tricks.

1. The first step is to create an identification process to form a community of reviewers and authors. E.g. identifying a person m needs to request for a registration to a specific committee providing a link to a web page of a University or research center or office, etc. At this link, the committee can find the scientific description of the applying person m. To speed up the identification process, each recognized center can have an internal committee. In this way, only the quality of the internal committees have to be identified. The certification process has to include also the initial expertness $\varepsilon_0(m)$ of m as a reviewer before her entering into the distributed peer-reviewing system. In case she cannot prove her affiliation to a research institution, she has to provide some other proof of her interest to be inserted into the community, and of the

community to accept the new member (like a list of papers, or patents, etc[1].) Of course, after the identification process, the applicant receives the permission to upload her papers or to write reviews of uploaded papers. This identification process aims at avoiding the creation of false reviewers, actually corresponding to the same person.

2. At the end of each review of a paper p done by a member m of the community, we will have a score $q(p, m) \in [0, 1]$, called the *quality of the paper p with respect to the opinion of the member m*. This score can change in time (only if there is a change of the corresponding review or a change of the paper p). In any case, with the symbol $q(p, m)$ we will always mean the last evaluation of the paper p by the member m.

3. Each member of the community can also make an evaluation of a review. This evaluation is given by a multi-choices questionnaire with fixed questions so as to facilitate the evaluation of a review and to have common criteria. One of these questions can be "The reviewer proposed so many ideas that she must be inserted in the list of authors of the paper", so as to stimulate the preparation of good reviews. The final result of an evaluation will be a number $e(r, m) \in [0, 1]$ for each review r and each member m. It's important to have a review made both of qualitative assessments (text) and numerical values (as in review of proposals for research projects). The text permits to evaluate the quality of the review and the corresponding numerical evaluation. The numerical values permit to compute $e(r, m)$.

4. We have to define an indicator $\varepsilon(m)$, called the *measure of expertness of m*, for each member of the community m. The aim of this indicator is to measure its expertness, and it serves as a reward for the behavior of the member m inside the community; for these reasons it must increase in case of good behavior and decrease in case of bad behavior of m. The measure of expertness must have, at least, the following characteristics:

   (a) $\varepsilon(m)$ has to be proportional to the mean of the evaluations of the reviews done by m in the last Y years (we can think e.g. $Y = 4$). In this way the more other members of the community judge that the reviews of m are good, and the more she will be considered expert. We can precise this requirement on $\varepsilon(m)$ by introducing

[1] Always remember the case of A. Einstein or similar cases.

the *mean evaluation of a review* r. Let $P(r)$ be the set of all the members n of the community ($n \neq m$) that evaluated the review r. If the cardinality $|P(r)| > T$, where T is a threshold value, then we set

$$\bar{e}(r) := \frac{1}{\sum_{n \in P(r)} \varepsilon(n)} \cdot \sum_{n \in P(r)} e(r,n) \cdot \varepsilon(n). \tag{0.1}$$

Only if the cardinality $|P(r)| > T$, we can consider the average (0.1) as meanigful, otherwise we will consider $\bar{e}(r)$ as undefined. We can e.g. try some first experiment by setting $T = 2$. Note that in (0.1), each evaluation $e(r,n)$ is weighted with the expertness measure $\varepsilon(n)$ of the author n of the evaluation $e(r,n)$ of the review r. In this sense the definition of $\varepsilon(m)$ is recursive: to define $\varepsilon(m)$ we already need a definition of $\varepsilon(n)$ of others members of the community. Hence we can define:

$$\varepsilon(m) := \frac{1}{|R^{\mathrm{d}}_Y(m)|} \cdot \sum_{\substack{r \in R^{\mathrm{d}}_Y(m) \\ |P(r)|>T}} \bar{e}(r) \tag{0.2}$$

where $R^{\mathrm{d}}_Y(m)$ is the set of reviews done by m in the last Y years . We can set $\varepsilon(m)$ as undefined if $|R^{\mathrm{d}}_Y(m)| = 0$. This also represents a certain force to realize a certain minimum number of reviews in Y years. At the entering of m into the system, we create a fictitious r with $R^{\mathrm{d}}_Y(m) = \{r\}$, $|P(r)| > T$, with $\varepsilon(n) = 1$ and $e(r,n)$ equals to the initial value of expertness $\varepsilon_0(m)$ of m for all $n \in P(r)$ (only one fictitious review r evaluated as $e(r,n) = \varepsilon_0(m)$ by each $n \in P(r)$, so that $\varepsilon(m) = \varepsilon_0(m)$).

(b) Since the number of authors can be very high, they can decrease the expertness by a great number of bad evaluations. For this reason, only one review evaluation coming from one of the authors is allowed.

5. Another proposal for the definition of $\varepsilon(m)$ could be:

$$\varepsilon(m) := \mu \left[\frac{1}{|R^{\mathrm{d}}_Y(m)|} \cdot \sum_{\substack{r \in R^{\mathrm{d}}_Y(m) \\ |P(r)|>T}} \bar{e}(r) \right]$$

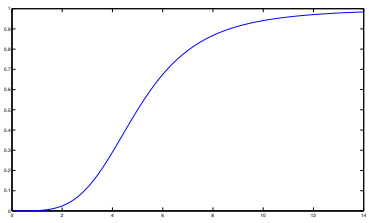

Figure 1: Example of a threshold-saturation function

where $\mu : \mathbb{R} \longrightarrow [0,1]$ is a threshold-saturation function (e.g. a Hill function) as in Fig. 1:

The use of this type of functions permits to take into account that only after a suitable (low) value of their arguments (threshold effect) we can have a real contribution to the value of $\varepsilon(m)$, and after a suitable (high) value of their arguments (saturation effect) the value of $\varepsilon(m)$ cannot really change. In this way we are also sure that we will always have $\varepsilon(m) \in [0,1]$. For the reader with some knowledge of fuzzy logic, the function μ can also be interpreted as membership functions of suitable fuzzy sets or as fuzzy truth values of suitable fuzzy sentences.

6. evaluations of reviews are signed

7. the idea is to facilitate an agreement, as far as possible, between a negative review and a negative evaluation of that review: the author will try to be kind with the (possibly anonymous reviewer, who knows the name of the evaluator of her review) because she is a potential reviewer also of her future papers. The reviewer will try to be kind because the author can evaluate her review.

8. A quantitative evaluation of a paper, project proposal, contribution to a conference, etc. can be defined using the weighted sum

$$\bar{q}(p) := \frac{1}{\sum_m \varepsilon(m)} \cdot \sum_m q(p,m) \cdot \varepsilon(m)$$

where each evaluation $q(p,m)$ is weighted with the expertness measure $\varepsilon(m)$ of the author of the review[2]. We are not saying that e.g. a

[2]We can impose $q(p,m) = 0$ if m is one of the authors of the paper p

paper has to be accepted using only quantitative evaluations, like the previous one, but that these kind of quality indicators can help the evaluation process, exactly as the citations index can help to evaluate a list of papers, but cannot substitute the judgment of an expert.

9. The values of $\varepsilon(m)$, the statistical distributions of $q(p, m)$ and of $e(r.m)$ are public and can be used to create committee of reviewers for conferences, to help in the acceptance of papers for journals or to create a better CV. In particular the above mentioned distributions can be useful to measure quantitatively the level of selection of a conference or of a journal, because from them, with m belonging to the reviewers committee, we can estimate what will be the probability to have a contribution accepted for the conference, journal or for funding.

From several points of view, the measure of expertness is similar to the impact factor already used for the evaluation of journals (and researchers).

To obtain a good value of $\varepsilon(m)$ the member m must write reviews with a high mean evaluation in the last Y years;

Let's note that:

- The measure $\varepsilon(m)$ doesn't depend on the papers written by m and by their mean evaluation $\bar{q}(p)$. On the one hand, usually, a good reviewer is also a good producer of papers. On the other hand this permit to have a figure of those experienced people that are able to produce good reviews, but that produce few papers per year.

- It seems to me that it is not important to identify the discipline in which m has to be considered expert. Indeed if she has a good measure $\varepsilon(m)$, then there exists a sub-community of researchers that evaluates positively her work. In other words she is considered an expert by this sub-community, and hence about a certain discipline: that of the sub-community.

- The names of the members can remain anonymous because the important key point is the quality of their reviews and the public availability of $\varepsilon(m)$.

- What prevents m to prepare a bad review? Other members of the community, and surely the authors of the paper, will write a bad evaluation of this review and hence this will decrease the value of $\varepsilon(m)$.

- What prevents m to prepare a too much good review? Other members of the community will write a bad evaluation of this review and hence this will decrease the value of $\varepsilon(m)$.

- What prevents m to prepare a bad review evaluation or a a too much good review evaluation? Evaluations of reviews are always signed. This is a strong disincentive to write such evaluations.

Of course the system of distributed reviewing will work only if there is a sufficiently big number of member that review and evaluate each single uploaded paper[3]. On the one hand this depends on the success and diffusion of the use of the measure of expertness. On the other hand, always using the mathematical simulation model, one can understand when the number of reviewers of a given paper is so small that the measure $\varepsilon(m)$ becomes inaccurate. The impossibility to do this estimate is a measure of small, close sub-community of self-referential researchers.

[3]The same type of objection could be do to Wikipedia at its beginning.