Crowdsourcing Scientific Paper Review*

WORKING PAPER

Atish Das Sarma[†] Luca de Alfaro[‡]

June 9, 2011

1 Motivation

The current system of peer-reviewed publications is governed by a cycle — write a paper, submit it, have it undergo review, and be published in an archival journal or conference — that was shaped by the medium used for disseminating the results. Printing and shipping journals or conference proceedings is expensive, and takes time; this gave rise to a publishing process with long lead times and a review process centered on selecting, rather than ranking or commenting. Information nowadays can be disseminated immediately essentially at no cost, and furthermore, in a manner that makes it open to social interaction: in blogs, wikis, forums, social networks, and other venues, people can both share and comment on information. As the medium and its capabilities have changed, so, we argue, should the process of scientific peer reviewing and publishing.

The current process has several drawbacks. One of the most salient is the delay imposed on the dissemination of results. In a typical computer science conference, six months may elapse from submission to publication in the proceedings, and this assumes that the conference deadline came just when the paper was ready for submission, and more importantly, that the paper was accepted. To avoid this delay, many authors submit the paper to open repositories such as Arxiv[1] at the same time as they submit it to a conference or journal. While this makes it available to

[*] The views expressed in this paper are those of the authors, and they do not necessarily reflect those of the insitutions where they are employed.

[†] Google, Inc. Email: dassarma@google.com

[‡] Google, Inc. and University of California, Santa Cruz. Email: luca@dealfaro.org

[1] http://arxiv.org/

other researchers, a submission to Arxiv does not come with any of the mechanisms that advertise a conference or journal publication to the wider community. Submitting to Arxiv is not an alternative for submissions to conferences or journals with double-blind review policies, and citing works submitted in Arxiv, but not yet peer-reviewed, is generally not a welcomed practice in science.

A related issue is the one of selection. Conferences and journals were born at a time when paper, printing, and storage in libraries, were expensive resources. The current process of scientific reviewing, consequently, aims at deciding which papers to accept for publication, and which to reject. Correctness is only one factor in such a decision: commonly, there are many more correct submissions (in the sense of exempt from scientific errors) than can be accepted, and the decision to accept or reject is motivated by judgements on the significance of the submissions. The paper acceptance process is thus of necessity an uncertain process, where a demarcation line needs to be drawn among papers of fairly similar apparent significance. Papers that present correct results, but which do not make the cut, are subjected to a delay as they are re-submitted to different journals of conferences. The process is slow and wasteful of resources.

One last drawback of the current process is that papers are not presented to readers in the context of the accumulated knowledge and judgement. While this shields papers from potentially irrelevant reviews, this also means that insightful observations from readers and researchers cannot help to understand papers and put them in context.

2 Design Principles

We envision an open, on-line system, where authors can publish their papers, and where anyone can comment on them, providing insight and distributing praise to the worthy ones. This would serve the scientific community as a whole, by making the dissemination of results more open, predictable and less subject to delays, and by helping researchers view papers in light of the accumulated knowledge and wisdom. At the core of our proposal are the following design principles.

No delays to publication. While papers that have just been submitted are unreviewed, this should not prevent their circulation. Authors can make their unreviewed papers available to all even now by posting them on their home page, or uploading them to Arxiv, but we advocate here that it should be possible to make papers immediately available to the public in the same system where they will be reviewed and ranked.

One natural objection is whether making papers available immediately deprives readers from the quality guarantee conferred by a formal process of paper review. We believe that the benefits of the prompt communication of scientific results far outweigh the drawback of circulating papers in various stages of review. The status of a peer reviewed paper is often assumed by people not familiar with the process to be a seal of approval that guarantees the correctness of the results. In reality, errors in scientific papers are not always discovered by the conference or journal review committees to which the papers are submitted: more often, the errors are discovered by the authors themselves, or by people who try to use or extend the papers results. Only papers that are widely read, and whose results are extensively used, can be trusted to be highly likely to be correct.

Publication as a beginning rather than an end. Scientific review and discussions of a paper or a field is a continuous process; publications should not be considered as a goal or end of a project, and should rather be thought of as opening new avenues for discussion and future collaboration. The current notion of accepting papers suggests a culture by which a paper is finalized, and the authors consider it a suitable end goal or closure of a project. In an ideal world, the acceptance or approval of a paper should only stimulate further work and open new doors in a continuous fashion. Thus, we believe that the notion of a peer-review as a prerequisite to dissemination should be replaced by the gradual degree of approval that a paper receives as reviews and comments accumulate. The system should offer a collaborative platform for discussion. The discussions and reviews would be also of value to researchers new to the field, as they could offer a guide to the contributions presented in the papers, put the papers in perspective, and help steer the researchers to the papers to be read first.

Rank rather than select. When a paper is submitted to a conference of a journal, the question of whether to accept it or reject it typically revolves on the relevance of the paper, rather than on its correctness. After the papers that are clearly flawed are eliminated, there are invariably too many papers to fit in the conference or journal format; the committee must then select the papers to accept on the basis of their quality. The committee thus essentially performs a ranking task, applying then a binary threshold dictated by conference or journal constraints. For the rejected papers that were indeed correct, this process results in an unjustified delay to publication; as these are typically resubmitted, the work that went into ranking them is also wasted.

This summary of the current review process is greatly simplified. In truth, there are many conferences and journals, with different typical quality levels, and authors choose the venue where to submit the paper in order to compromise between the prestige of the venue, and the probability that the paper is accepted. Nevertheless, the process is wasteful of time and work. We believe it would be better to use the reviews and comments for ranking, rather than for selection. There would be no need to artificially set a cut-off line; all papers would be ranked and available on-line.

Once a ranking of the papers were available (even if approximate), journals and conferences could use the ranking for selection purposes. For example, a conference could gather people interested in a particular field, and allocate paper presentation slots to the 30 highest-ranked papers of the year, and poster presentation space to the next 50 highest-ranked; a journal or book editor could similarly publish (and distribute to libraries in archival form) the best 50 papers of each year. Certainly many users of the system could use the ranking for selection purposes, but the main goal of the system would be to generate a ranking, not a selection.

Tailored to scientific communities. There cannot be a single ranking of scientific papers, as papers in different disciplines can hardly be compared. Furthermore, each discipline has a different set of expert reviewers. We believe that a crowdsourced system for paper reviews should be equally tailored to these communities, operating on one of these at a time. The precise granularity of these communities is a matter of debate. The guiding principle must be the existence of a community of experts that is numerous, active, and with sufficiently long-lived interest. Within ACM, for example, there are special-interest groups (SIGs) on various topics, ranging from embedded software, to graphics, to databases, and more; we believe that those groups could be of the proper granularity to form the basis of the process of crowdsourced paper review we envision.

No double-blind review. In a double-blind review, the authors submit an anonymous paper which gives ideally no hint of the identity of the papers authors. If the paper is accepted, the authors submit a second version, which includes the author list, for publication. Double-blind review has some advantages: in particular, it can offer a better guarantee of impartiality than a review process where the reviewers are aware of the identity of the author. Nevertheless, we do not propose using double-blind reviews, for two reasons. Most importantly, a paper submitted to a double-blind review process cannot also be circulated by the author. Thus, the double-blind review process imposes by its very nature a delay in the dissemination of results. Furthermore, material that has been previously circulated in the form of technical reports cannot be meaningfully submitted to a double-blind review process. We believe that these limitations are too strict, and pose too great an impediment to the free circulation of ideas. We believe that it is worth exploring approaches to fair paper review and ranking that do not rely on double-blind reviewing.

Fairness through transparency and reputation. The concern of achieving fairness is paramount in the design of a review system. Insofar as possible, we believe the best approach consists in achieving fairness through a mix of transparency, and reputation systems. Transparency, or the possibility to see other participants actions in the system, is a very valuable tool in making people responsible for their actions, and in engendering trust in the system. Full transparency is usually not possible: usually, some balance has to be struck between transparency, and guaranteeing people the ability to express their opinions on papers without fear of reprisal, but we believe that on balance, we should strive to make the workings of the system as transparent as possible. Reputation is another central notion. Not all people are equally expert in all subjects, and this is all the more true in focused fields of science. There needs to be a mechanism to weigh opinions differently, and to grant more weight to people that have demonstrated greater expertise. When the system is opened for use, it will be fundamental to prime the reputation system with reputation values derived from the real-world expertise and achievements of scientists. Reputation systems can also provide incentives towards constructive behavior, which is key to a properly functioning system.

3 Proposed Design of a Crowdsourced Review System

We propose here a concrete, if high-level, design guided by the above design principles. We focus on the design of the comment and ranking system, and we do not delve on other important aspects, such as the archival storage of papers and data. While those are essential concerns, they are somewhat orthogonal from the task of designing a review and ranking system.

3.1 User reputation system

Although the system we propose might be workable without the support of a user-reputation system, we believe that the use of such a system would be highly beneficial. We propose to associate with users reputation levels; these reputation levels would be displayed in a leader-board for each community (reputation in different communities would be mostly independent). This reputation need not be fine-grained: we think that a simple division in four groups, such as no-star, bronze-star, silver-star, and gold-star could be used. The purpose of this reputation system is both to organize the user interactions, and to reward users for their work. We believe that such a system can be primed by an analysis of the previous contributions to a community (as measured, for instance, by the number of papers published, the amount of times a user has served in a program committee, and so forth); we will describe later how this reputation can be updated.

3.2 Paper submission

In our proposed design, users can upload papers, or references to papers uploaded elsewhere, e.g., to Arxiv. Once a paper (or its reference) is uploaded, it becomes visible to users, and will be subject to reviews and comparisons as described below. For this process to work properly, it is important that users are not able to change the submissions under review. For this purpose, the system might enforce that old revisions cannot be deleted, and new revisions can be uploaded at most at a certain rate (once a month, perhaps).

In our proposed design, users can upload papers, or references to papers uploaded elsewhere, e.g., to Arxiv. Once a paper (or its reference) is uploaded, it becomes visible to users, and will be subject to reviews and comparisons as described below. For this process to work properly, it is important that users are not able to change the submissions under review. For this purpose, the system might enforce that old revisions cannot be deleted, and new revisions can be uploaded at most at a certain rate (once a month, perhaps).[2]

Any web system where it is possible to upload content or URLs is abused by spammers and vandals. Given the high intrinsic value of papers, measured in the amount of effort they take, we believe that effective strategies can and should be developed to fight such spam. For instance, we might allow only authors with some reputation to submit papers directly; other authors would have to submit to a ante-chamber, visible only to authors with reputation, where the papers can be approved. Making visible the name of the paper approvers would help eliminate spam, and would be similar to the way in which members of communities and clubs can invite new members.

3.3 Review actions: comments and comparisons

Once papers are uploaded, users can take two main types of actions: they can comment on a paper, or they can compare two papers. Comments are used to share insights and opinions about papers, but they do not, per se, contribute to ranking the papers. Comparisons are used to rank the papers.

[2] Arxiv uploads cannot be later deleted; therefore, submitting links to papers uploaded to Arxiv would satisfy the requirements of the proposed system.

We believe that users should be able to either comment on a paper, or compare the paper to other papers, but not do both for the same paper. In this way, a user who believes that a paper is faulty must choose in which way to affect the ranking of the paper: directly, by voting the paper down in a comparison with another paper, or indirectly, by writing an unenthusiastic comment that might lead others to vote down the paper in comparisons. Users must either rank, or convince others: splitting these two powers helps limit the influence of any individual on the system.

3.3.1 Comments

Comments can be very useful, as they allow the insights and opinions of expert users to be shared, but they can also be very damaging to the papers and to the quality of the overall system, when they are offensive, careless, or uninformed.

Anybody can comment on papers. Comments can be either signed, or anonymous; if the comment is signed, the reputation of the author writing it appears besides the name. The author can reply to each comment, and the author of the comment cannot reply in turn — the author gets the last word. This gives authors at least a minimum degree of control on the comments that appear beside their papers. Comments can be deleted by a sufficient number of deletion votes by users with reputation (for instance, we may require two gold-level user votes, or three silver-level user votes, to delete an anonymous comment); to prevent abuse, the original comments, and the users who voted for their deletion, are visible to all authors of sufficient reputation. Comments can be voted up or down according to their usefulness, similarly to how replies can be voted up or down on stackoverflow.com.

There should be some means for users to request others to comment, in a similar fashion to how it is possible now to request reviews from experts on a paper.

3.3.2 Comparisons

We propose that paper rankings be based on comparisons. We believe that comparisons are a more reliable way of producing rankings, than absolute ratings. Comparisons contain more information than ratings that only assign one of a few discrete scores to a paper. Furthermore, comparisons side-step the problem of calibrating the ratings to the average technical quality of papers submitted to particular interest areas (thereby resulting in automatic normalization).

We propose that comparisons work in two steps. First, users can compare pairs of papers, specifying for each pair which paper they believe is more interesting. In the second step, these comparisons are sent for approval to established, high-reputation members of the community. These members can see the two papers being compared, and the users comparing them, but crucially, not the outcome of the comparison: the approval is based on whether the user is judged to be sufficiently knowledgeable in the subject area to provide a comparison. Further, whenever needed, certain paper comparisons are explicitly sought. This two-step process provides to authors the very important guarantee that their papers have been ranked by users who are knowledgeable in the particular topics of the paper. We believe that the proper matching of papers, and users providing comparisons, can be best judged by field experts, rather than algorithms, and we also believe that authors would prefer, at least initially, such a human oversight over the ranking process.

The paper rankings will be visible to all. We believe it is best not to present a detailed ranking (even though such a ranking would obviously be known to the system), but rather, an approximate ranking consisting, e.g., of the top 20 papers, top 50 papers, top 200 papers, and so forth. Such an approximate ranking corresponds better to the uncertainty with which the quality of a paper can be ascertained, and avoids at least some of the arguments that would arise from a list of papers presented entirely in ranking order. Of course, the dividing lines between paper groups would still fall somewhat arbitrarily, just as the decisions to accept or reject do presently, but at least all papers, regardless of their ranking, will all be published and visible.

While authors will not know who compared their papers to other papers, authors will be able to see the list of users who approved the comparisons. In this way, authors are provided with some degree of assurance that their papers have been ranked by domain experts. We believe that providing such an assurance will be crucial, if authors are to entrust their best papers to a collaborative ranking system. Currently, when authors choose to which conference or journal to submit a paper, they examine the composition of conference program committees and journal editorial board, looking for the presence of domain experts. In the open scientific review platform we envision, authors can see the list of users who are high enough in rank to act as comparison approvers, and they also see the users who actually approved comparisons for their papers. We think this last element of information is important, both in providing increased assurance to authors, and in ensuring that approvers perform due diligence before determining whether a user is qualified enough to compare two papers.

3.4 Paper assignment and incentives for comments and comparisons

In an ideal world, users would be interested enough in papers to provide a sufficient number of comments and comparisons, so that an insightful ranking of papers might emerge. In practice, this is unlikely to happen, unless there is some system of incentives, and some way of matching papers with potential reviewers. As a starting point, we propose the following mechanisms.

Users declare their set of conflict of interest users. The conflict relation is symmetrical (and reflexive!), and the list of conflicts of interest for each person is public. We believe that a public list of conflicts of interest helps avoid both omissions, and unjustified exclusions.

Users can then choose the list of papers they are willing to review and compare. They can do so either by selecting individual papers (they have access to the full papers), or using keyword / subject classifications. There are two ways in which papers are assigned for (solicited) review: by algorithms, and by other users. In addition to the solicited reviews, users are welcome to provide unsolicited reviews for papers. We propose to develop an incentive system where users can gain reputation when they compare papers, or comment on papers. In general, we would like to reward actions that provide early insight, and are later proven correct:

- Comments, especially of papers that have received only few comments when a new comment is made, and especially if other users find the comments useful or "vote them up".

- Comparisons, especially comparisons that provide information on the ranking of papers for which little information is available at the time the comparisons are made, and that are later believed to be in agreement with the majority of other comparisons.

Approving comparisons is a time-consuming and delicate task, as it involves matching the detailed topic of the paper, with the expertise of the users providing the comparisons. Since only reputed users can approve comparisons, we hope that being listed as an approver will be a mark of expertise that will be sought after, much as right now, membership in a program committee is a honor. We could provide leader-boards for the most active approvers.

We believe that the development of a reputation system will be both an important step, and a delicate one to incentivize users to provide good and sufficient number of reviews. Different communities of scientists may have different opinions on the amount of merit that different types of actions entail, and the design of the reputation system should be such that it can be tailored to the needs of individual communities. Nevertheless, we believe that there will be enough commonality in the notions of constructive behavior (if not in the precise merit ascribed to each action) to justify building an underlying system.

4 Facilitating the Transition to Crowdsourced Paper Review

Writing a technical paper requires a considerable amount of work. The venues where the papers are published — the journals and conferences where they have appeared — are used as a metric of academic accomplishment. Scientific review and publishing will adopt a crowdsourced model only if we can provide authors with assurance that their papers will be duly considered, and if we can provide quality metrics that can supplement the current ones. The importance of this cannot be overstated, since the careers and visibility of work of researchers depends on this.

Many details in the proposed design are aimed at giving authors the assurance that their papers will be duly considered. In particular:

- Authors can see the list of users who approved the assignment of their papers to other users who performed the comparisons.
- User reputation in the system will be primed using the reputation of researchers in real life (measured, for instance, by the number of papers published, the activity in conference review committees, and so forth).
- Authors always have the right-of-last-reply on comments on their papers.
- Spurios comments on papers can be deleted by votes of at least two high-reputation users.
- Papers are ranked, so that good quality papers are likely to be presented in the company of other quality papers.

The rankings produced by the system will be usable in measuring scientific accomplishment: just as now one can brag to have a paper published in a given journal, so will one be able to brag about having a paper among the top 50 in a field that year. Traditional publishers would also be welcome to select the top papers, and publish them with all due honor, in special editions which could be bought by libraries for truly archival storage.