Scaling Internet Research Publication Processes to Internet scale

Jon Crowcroft
The Computer Laboratory, University of Cambridge
Cambridge, UK
jon.crowcroft@cl.cam.ac.uk

S. Keshav
University of Waterloo
Waterloo, ON , Canada
keshav@uwaterloo.ca

Nick McKeown
Stanford University
Stanford, USA
nickm@stanford.edu

1. INTRODUCTION

The reviewing process used by most computer systems conferences originated in pre-Internet days. In this process, authors submit papers that are anonymously reviewed by program committee (PC) members and their delegates. Reviews are single-blind: reviewers know the identity of the authors of a paper, but not *vice versa*. At the end of the review process, authors are informed of paper acceptance or rejection and are also given reviewer feedback and scores. Authors of accepted papers use the reviews to improve the paper for the final copy, and the authors of rejected papers use them to revise and resubmit them elsewhere, or withdraw them altogether.

Recently, some conferences have tried two minor innovations. With double-blind reviewing, reviewers do not know (or, at least, pretend not to know) the authors. And, with "shepherding", a PC member ensures that authors of accepted papers with minor flaws make the revisions required by the PC.

Surprisingly, the advent of the Internet has scarcely changed the reviewing process. Everything proceeds as before, except that papers and reviews are submitted online or by email, and the paper discussion process, at least for second-tier conferences and workshops, does not require the physical presence of the PC. A naive observer, seeing the essential structure of the reviewing process preserved with such verisimilitude, may come to the conclusion that the process has achieved perfection, and that is why the Internet has had so little impact. Such an observer would be, sadly, rather mistaken.

We argue that the current reviewing process is fundamentally flawed, with at least five systemic problems that undermine the integrity of the process (Section 2). A game-theoretic modeling, presented in Section 3, demonstrates that these visible symptoms are due to inherent conflicts in the underlying incentive structure. Understanding the incentive structure allows us to design several incentive-compatible mechanisms, presented in Section 4, that remedy nearly all the problems we identified.

2. PROBLEMS WITH THE CURRENT REVIEW PROCESS

The review process for computing systems conferences suffers from at least the following five problems:

- **A rapid increase in the total number of papers**: Reasons include the ever-increasing number of active researchers on the planet, particularly with Korea, India, and China emphasizing fundamental research; the lack of disincentives for submission; the serial 'recirculation' of minor variants of the same paper to different conferences; the publish-or-perish imperative; and the tenure and merit systems in some cultures. Given that the number of reviewers is not growing at the same rate, this increases average reviewer workload.

- **Skimpy reviews**: Some reviewers do a particularly poor job, giving numeric scores, but no justification. Authors of rejected papers feel that their hard work has come to nought. Ironically, even authors of accepted papers feel cheated.

- **Declining paper quality**: Anecdotally, there is a perceptible decline in the quality of the average submitted paper. Some fraction of papers submitted are not quite ready, but are sent in anyway to elicit feedback so that the authors save time (and student authors save advisor cycles) and the real intent is to submit the paper elsewhere.

- **Favouritism**: PC members are thought to favour certain 'famous' researchers, so that even mediocre work from these researchers may sometimes be accepted. Moreover, there is always the fear that PC members may unfairly favor papers from their own circle of friends and collaborators.

- **Overly negative reviews**: There is a well known tendency for systems people to review negatively. This is perhaps because some reviewers are very competitive and feel that their own work is so much better. Another reason could be that reviewers who are also authors submitting work to the same conference may give overly negative reviews to make their own paper look good in comparision. A more benign explanation is simply that, as a community, we love to *debug*, and nothing is more fun than finding bugs in someone else's work. Unfortunately, the conference system, so much more popular than journals in our community, does not permit bug fixing, only reporting [1]. Ideally, a good idea with fixable bugs should be reported, and who can truly claim their own work has no bugs? Nevertheless, it appears to us that typical high quality events are rejecting about as many papers that are of good quality as they accept. In some cases, a rejected paper may in fact be head-and-shoulders above several accepted papers.

These problems are inter-related. The increase in the number of papers leads, at least partly, both to a decline in paper quality and a decline in the quality of reviews. It also leads to an ever-increasing variance in paper quality. Similarly, as the acceptance rate of a conference declines, there is a greater incentive for reviewers to write overly negative reviews and favor their friends.

What is needed are incentive mechanisms that preventing favoritism, match the number of PCs, PC members and reviewers dynamically to the current size of the research community, decrease the number of poor quality papers, and provide transparent reports on the reputations of the stakeholders (conferences, PCs, reviewers, authors, publishers). The design of such mechanisms is the subject of the remainder of the paper.

3. THE PAPER PUBLISHING GAME

Paper reviewing and publishing can be viewed as a game. There are three players in this game, who have the following incentives (pragmatically, if cynically defined):

- **Authors** Authors want to get published, (or, at least, get feedback on their work). They also don't want to be roped into becoming reviewers.

- **Reviewers/PC members** Reviewers want to minimize their work (for instance, by giving scores, but no justifications), while trying to reject papers that compete with their own papers, and accepting papers from their friends (and/or "famous" researches). They want to reject clearly bogus papers that would embarrass them. Finally, they want to get the prestige of being in the PC.

- **The conference/PC Chairs/research community/paper readers** These stakeholders want to have the highest quality slate of papers, while trying to include fresh ideas, and providing some sense of coverage of the field.

Interestingly, the problems described in Section 2 arise because the existing paper reviewing process does not explicitly address these contradictory incentives. There is no explicit incentive for authors to become reviewers or for authors to limit the number of papers they submit, or to submit good quality papers. There is no check on reviewers who write skimpy reviews [2], are overly negative, or play favorites. No wonder the system barely works!

A1	Authors should not submit poor quality papers
A2	Authors should become reviewers
R1	Reviewers should submit well-substantiated reviews
R2	Reviewers should not favour their friends
R3	Reviewers should not favour 'famous' researchers
R4	Reviewers should not trash competing papers
C1	The conference should maintain paper quality while covering the field and both fresh and well-substantiated ideas

Figure 1: Table 1: Mechanism Goals

4. MECHANISMS FOR INCENTIVE ALIGNMENT

Our goal, then, is to design mechanisms that directly address and balance the incentives laid out above. That is, it is incentive compatible, given the mechanisms, to do the Right Thing, which we summarize in Table 1.

If A1 and A2 are met, then the overall paper quality increases and the review workload reduces. If R1-R4 are met, review quality increases and charges of favouritism decrease. Finally, C1 is a top-level goal of the whole community.

We now describe some mechanisms to achieve these goals. They are based on a combination of peer pressure and a virtual economy. Our proposals include some steps that have been tried by some brave conference PC chairs, and others which are novel and would need experimentation and experience.

4.1 Author incentives

Our first mechanism addresses A1 using peer pressure. It requires the conference to publish not only the list of accepted papers, but also, for each author, the author's acceptance rate for that conference. For example, if an author were to submit two papers and none were accepted, the conference would report an acceptance rate of 0, and if one was accepted, the author would have an acceptance rate of 0.5. Because no author would like to publicly be seen to have a low acceptance ratio, we think that this peer pressure will enforce A1. As a more extreme measure, to prevent 'recirculation', the conference could publish both the authors and the titles of rejected papers.

Our second set of mechanisms address A2 by raising the prestige of reviewing. Some possible mechanisms are to have a 'best reviewer' award for the reviewer with the best review score [3]. Conference organizers can also publicly thank reviewers during the conference, or give them a discount in the registration fee.

A more radical step would be to solve A1 and A2 simultaneously by means of a virtual economy, where tokens are paid for reviews, and spent to allow submission of papers[4]. Assuming that each paper requires three reviews on average, we arrange for a reviewer to get one token per review (independent of the conference), and that authors pay three tokens to submit a paper (to any conference). However, au-

[1] One reviewer suggested that conference organizers suggest some papers be re-submitted after revision. This seems plausible at first glance, but the thought of having to wait a year to publish a conference paper seems excessive.

[2] Other than the slight risk of embarrassment at the PC meeting.

[3] See Section 4.2 for details on review scoring.

[4] The reader may note the analogy with fair decentralised matching of input data rates to link capacity in communication networks. Also, we have been informed by one of the reviewers of this paper that this scheme was first suggested by Jim Gray, though we cannot find a citation to this work.

thors of papers that are accepted would be refunded one, two, or all their tokens depending on their review score. Authors of the top papers would therefore incur no cost, whereas authors of rejected papers would have spent all three of their tokens. Note that the question of what to do with tokens in the situation where conferences reject good papers (defined as papers that get good reviews) but have insufficient space should not arise, because authors of papers with good reviews, even if the paper is ultimately rejected, would get their tokens refunded. Clearly, this scheme forces authors to become reviewers, and to be careful in using the tokens thus earned, solving A1 and A2.

We note that we obviously need to make tokens non-forgeable, non-replicable, and perhaps transferrable. E-cash systems for achieving these goals are well known - they merely need to be adopted to a non-traditional purpose.

We now describe three refinements of this basic scheme. First, authors could be permitted to incur a limited debt for a limited time period. Thus, they may submit a paper without sufficient tokens, but if it gets bad reviews, they will find they cannot submit any more papers for a while, or until they submit some reviews.

Second, from time to time, the size of the research community in a given area actually grows (significantly in the case of research areas of global interest). We need to enlist young, energetic and skilled PC members to match this legitimate growth. We also need the equivalent downsizing procedures. Students of the market will recognise this as inflationary and deflationary economics, sometimes implemented by printing, and withdrawing cash from the economy. We anticipate that we can increase and decrease token supply, in response to inflationary and deflationary tendencies, by regulating the number of tokens required for submitting a paper, using fractional tokens if necessary.

Third, to deal with cold-start, new entrants (graduate students, employees at research labs, junior faculty) should start with a clean slate, with an allocation such as one token a year, that must be used or they lose it. They can always do some reviewing to earn more tokens.

We recognize the regulating the economy is not trivial. Over-damping the system would lead to conferences with too few papers, or too few reviewers. Under-estimating the value of tokens would only slightly mitigate the current problems, but would add a lot of expensive overhead in the form of these mechanisms. Moreover, it is not clear how this system can be implemented. Indeed, even if it was, it would not be obvious how it can be bootstrapped, or whether it would have unintended consequences. One possible technique would be to start by publishing signed reviews and rely on technologies such as Citeseer, Google Scholar etc, as we describe below in more detail.

We stongly believe that, in general, transparency is essential, so the choice of authority and the auditing of such a system would be crucial design decisions.

4.2 Reviewer incentives

We first discuss dealing with R1 and R4. We propose that authors should rate the reviews they receive for their papers, while preserving reviewer confidentiality. Average (non-anonymized) reviewer scores would then be circulated amongst the PC. No PC member wants to look bad in front of his or her peers, so peer pressure should incentivize R1 and R4 (PC collusion will damage the conference reputation).

A more radical alternative is to have conference publish scores for pseudonymous reviewers, i.e. reviewers with numeric or alphanumeric ids. Again, the idea is that a reviewer, even one only with a pseudonym, does not want to look bad in front of his or her community, thus encouraging good behavior and enforcing R1 and R4. Unfortunately, this means that only a pseudonym gets public reputation, not the reviewer's real name.

An even more radical alternative is for reviews to be openly published with the name of the reviewer. The idea is that reviewers who are not willing to publish a review about a paper are perhaps inherently conflicted and therefore should not be reviewing that paper. Of course, there is a danger that public reviews will be too polite, but this will no doubt sort itself out over time. The advantage of using true identities ("verinyms"), is that this handles R1, R2, and R4.

We can imagine combining the second and third mechanisms. A reviewer could have both a verinym and one or more pseudonyms. If one is tenured, comfortable, and thick-skinned, one can use one's verinym for everything. But, when one is not too sure, one uses one of one's pseudonyms.

Finally, we think R3 is an unsolvable problem. Note that public reviews can actually encourage deferral to "famous" researchers (i.e. trashing R3). Though double-blind reviewing mitigates it, the intrinsic problem is that it is hard to separate one's prejudices in favour of a person from the text-as-written.

4.3 Conference incentives

The PC is responsible for C1, and ultimately, it determines the stature of the conference. That ought to be, in the long term, self correcting: bad conferences will attract no papers and no reviwers, and will die out, as they ought.

5. A GRAND UNIFIED MECHANISM

A deeper examination of the incentive structure suggests that perhaps the real problem is that too much of the work of submitting and selecting papers is hidden. What if the entire process were made open, transparent, and centralized? The goal would be to have a standard way for members of the community to review and rank papers and authors both before and after publication, in a sense adding eBay-style reputations to Google Scholar or arXiv. All papers and reviews would be public and signed, with either pseudonyms or verinyms. This system, would, in one fell swoop achieve many simultaneous goals:

1. Readers can draw their own conclusions (and tell the world) about the quality of papers published by an author. This would incentivize authors not to submit bad papers (achieving A1).

2. Community members who publish often and review rarely would be exposed, achieving A2.

3. We would see the reviews and the names of the reviewers alongside the paper, incenting R1, R2, R3, and R4.

4. We get to see whose opinions correlate well with our own to help decide what papers to read

5. There is a good chance that very good papers that end up as technical reports or in smaller, less well-known conferences, are raised to the top by popular acclaim.

6. The system would allow continued discussion and feedback about papers even after they have been published (1) to help others (busy senior people, and new people not knowing where to start), and (2) to give a chance for others to pitch in and debate.

We believe that the academic community as a whole is crying out for such a system. We realize such a system can also be gamed. As with e-cash, the hardening of reputation systems to resist collusion and other attacks is well known, and we merely need to import the appropriate machinery.

6. CONCLUSIONS

We have identified the underlying incentive structure in the paper publishing game, shown where these incentives lead to bad behavior, and proposed several mechanisms that incentivize authors, reviwers, and the community to do the Right Thing. There are several feedback processes in the schemes described above, and each has a coupling factors, which needs to be determined. It is important to get these right in a robust way, and this is the focus of our future work.

In general, the entire purpose of our proposals is to shift the operating point of the community so that we cease wasting cycles on submitting, resubmitting, and reviewing weaker work, and that we provide reasons for people to become better authors and reviewers. We hope that at least some of our proposals will make their way into future conferences and workshops.

7. ACKNOWLEDGMENTS

We would like to acknowledge the anonymous reviewers who helped improve the quality of this paper.