Paper and Proposal Reviews: Is the Process Flawed?

Henry F. Korth (Lehigh University), Philip A. Bernstein (Microsoft), Mary Fernandez (AT&T Labs-Research), Le Gruenwald (National Science Foundation), Phokion G. Kolaitis (IBM Almaden Research Center and UC Santa Cruz), Kathryn McKinley (University of Texas at Austin), Tamer Özsu (University of Waterloo)

Abstract

At the 2008 Computing Research Association Conference at Snowbird, the authors participated in a panel addressing the issue of paper and proposal reviews. This short paper summarizes the panelists' presentations and audience commentary. It concludes with some observations and suggestions on how we might address this issue in the near-term future.

1. Introduction

Every two years, the Computing Research Association (CRA) hosts a conference for chairs of computer-science and computer-engineering departments and directors of industrial and government computer-science research labs from across North America. We proposed a panel on paper and proposal review processes—a hot topic for this audience. There is a proliferation of experiments with new review processes and publication venues in most computer science fields, which affect how to evaluate publication record for promotions. Moreover, there is a pervasive sense of unease within these communities about the quality and fairness of the review process and whether our publication processes truly serve the purposes for which they are intended. The goal of the panel as stated in the CRA Snowbird program was:

The review process for computer science publications and proposals is crucial to the health of our field, especially for new researchers seeking to establish themselves in the field. Current and past processes have been criticized for a variety of reasons, including timeliness of decisions; fairness, especially to "outsiders;" and openness. The responses have included turnaround time guarantees and process changes. Some journals and conferences have moved to double-blind reviewing, though not without strong opposition. NSF moved some time ago from a journal-style review process to doing most reviews via panels that meet physically in one location. Meanwhile, conference program committees have moved in the opposite direction. Many do not meet physically and instead use an asynchronous on-line process. This panel will discuss the concerns that have led to change, the degree to which process changes have addressed these concerns and/or created new problems, and what further steps ought to be taken from here.

That the panel was well-attended despite competition from several excellent concurrent sessions points to the importance the Snowbird attendees placed on the review process and its quality. The community has shown increasing interest in establishing high quality review process. As examples, USENIX held a workshop (http://www.usenix.org/event/wowcs08/) on this topic in April 2008 and during questions, the panel audience offered well thought-out and insightful commentaries.

2. Panelist Presentations

The panelists attempted to survey the range of problems and proposed solutions. In this paper, we shall summarize each panelists' remarks and some of the key comments from the audience. For brevity, we omit the full details of the panel presentations and instead point the reader to the panel slides online at http://www.cra.org/Activities/snowbird/2008/agenda.html.
We conclude with some reflections of what we learned from this panel.

Hank Korth

Recently, the database research community has adopted a number of changes to the paper-review process with the goal of improving accuracy, fairness, speed, and efficiency. While these changes have been well intentioned, many in our field view at least some of the changes as a step

in the wrong direction. More generally, there remains a pervasive sense that serious problems remain. In preparation for the panel, we reviewed processes in various subfields of computer science and found that concerns in the database research community are indeed widespread across computer science.

Kathryn McKinley

The review process determines the progress and direction of our field (see SIGPLAN Notices 2008: Editorial: http://www.cs.utexas.edu/users/mckinley/notes/blind.html). Double-blind reviewing, in-person program committee meetings, review panels, and author response all offer important advantages despite several objections that have been raised to each one. All of these approaches entail more work for reviewers and, especially for double-blind reviewing, for authors, but the benefits outweigh the costs. Several specific studies were noted that show nepotism and gender biases are problems when publications and applications are not "blinded."
Also see the session slides from "Practical Solutions to a Continuing Problem: Sexual Harassment and Gender Discrimination" (http://www.cra.org/Activities/snowbird/2008/agenda.html).

Le Gruenwald

The number of proposals to NSF Division of Information and Intelligent Systems (IIS) has more than quadrupled in the past 10 years. To control this growth, pre-proposals and limits in the number of submissions per principal investigator have been adopted. Reviews are normally done via in-person panels at NSF. There have been some combinations including in-person panelists, ad-hoc reviewers, teleconference panelists, and/or videoconference panelists. NSF faces a challenge in getting enough panelists from both academia and industry, especially due to its strict conflict of interest rules. It would be helpful if academia had a way of providing rewards for this sort of professional service that go beyond the modest consideration it currently receives.

Phil Bernstein

The review process is, in some ways, like grading of student papers. Hardly anyone likes to be reviewed (or graded), hardly anyone likes to do a lot of reviews (and no one likes grading), authors often find reviews to be unfair or "random" but, on average, we think the best researchers (and students) get the best reviews (and grades). However, just as students may "game" the system to get better grades, some uncreative researchers game the system by writing well-formed but uninspiring papers that get excellent reviews. Why does this happen? The heart of the problem is that there are too many borderline papers and only a fraction can be accepted. Choosing that fraction is a random process.

Fewer people complain about the journal review process than the conference review process, presumably because journals offer two rounds of review. But they don't offer an associated presentation slot. These differences are historical and artificial. So, why not have both? That is, a conference proceedings becomes a journal with two rounds of review. Or an existing journal is linked to a conference and guarantees a presentation slot to authors. The program committee determines the length of the presentation: full, short, or poster. This might make journals more desirable, since authors are visible as presenters at conferences. Or it might de-value journals, since conferences offer all of the advantages of journal except space for long papers. Perhaps journals will find a new mission, such as more project summaries and surveys. These changes might force us once again to educate academic tenure committees.

Phokion Kolaitis

Over time, conferences have become more important than journals in computer science. The community had to work hard to make the case that promotion and tenure committees should assign (at least) as much weight to conference publications as they do to journal publications. The 1999 CRA Best Practices Memo entitled "Evaluating Computer Scientists and Engineers for Promotion and Tenure" (http://www.cra.org/reports/tenure_review.html) stated the case eloquently and was widely adopted. In recent years, however, we have been witnessing the proliferation of workshops that take on several features of conferences, such as large program committees and some sort of published proceedings, but, at the same time, have rather short review periods. In the span of just one week in June 2008, more than twenty calls for papers for workshops were posted at *dbworld* alone. This state of affairs blurs the distinction between workshops and conferences, and creates additional difficulties in evaluating the scholarly work of computer scientists and

engineers. Many conferences have adopted duplicate submission policies regarding workshop publications. It is time for the community to take a stand on workshop publications. Workshops are not mentioned in the CRA Best Practices Memo. We should not move to make workshop proceedings rise to the status of conference proceedings; instead, we should encourage workshops to be true workshops again with only informal proceedings that do not conflict with strict duplicate-submission policies for conferences.

Mary Fernandez

The CRA Best Practices Memo states "*Publication in the prestige conferences is inferior to the prestige journals only in having significant page limitations and little time to polish the paper. In those dimensions that count most, conferences are superior*". However, page limits force authors to sacrifice completeness, clarity, or both. A pledge to include everything in a technical report is not always kept. Reviewers suffer from these compromises and have trouble understanding and/or believing the results, leading to exhaustion and cynicism. The journal review process is better, but relatively few journal papers are being written. This lack of reproducibility is growing worse because others, including scientists in other fields, depend on our results (as in the partial replacement of wet labs by virtual computation labs). Why should they trust us if we can't trust ourselves?

We should link each conference to an efficient journal, such as the new VLDB e-Journal (http://www.jdmr.org) as a means to allow authors to be more thorough and reviewers to have greater focus and investment in the outcome. The result should be improved scholarship.

Tamer Özsu

We have a fundamental problem in how we conduct experiments and how we report them. Our students (and perhaps we, ourselves) do not know how to run experiments. Many of our experiments are not repeatable: setup is not properly described, source code is not available, and data sets are not available. The results often fail to report confidence intervals. Experimental repeatability is a fundamental feature of scientific research, and we need to find ways of ensuring that experimental results that we report are meaningful; many of them are not. Where intellectual property issues permit, data sets should be made available publicly. Conference papers should focus on experimental setup and on stating what experiments would be interesting to run and why rather than trying to give "full" experimental results that are never complete and usually not repeatable (many times because the experimental setup is not properly described). As a result of refocusing conference papers, it should be possible to reduce their page limits.

Regarding journals versus conferences: journal first round review times are now competitive with conference review times, and they can be reduced further. We should move to online, article-based publishing to reduce delays as compared with our current off-line issue-based mode of publication. Having (for the most part) convinced tenure committees about the value of our conferences, we now need to convince ourselves that journals are equally valuable and important venues to publish *fuller* results (including fuller experimental results).

3. Audience Commentary

At the end of the panelists' presentations, various members of the audience offered comments, other issues in reviewing, and descriptions of how various subfields in computer science are handling the issue of reviewing. We list many of the comments here (with the caveat that not all have been verified by us independently).

- ACM SIGCOMM *Computer Communication Review* published an article related to this panel (J. C. Mogul and T. Anderson, "Open Issues in Organizing Computer Systems Conferences", Vol. 38, Issue 3). Related to this is a recent USENIX workshop. Papers and slides from that workshop appear at http://www.usenix.org/events/wowcs08/tech
- ACM TODS has a good discussion of double-blind reviewing on its Web page http://tods.acm.org/editorial.pdf
- CHI offers a presentation slot to authors who have published in ACM Transactions on CHI; others in the audience recommended this practice.
- Several comments were made about the reviewing process. Selection of papers was described as a "beauty contest" in which the most attractive papers are chosen rather than the most interesting work. Reviewing

should focus on the contribution of the paper, why it is important, why one should believe it.

- Face-to-face program committee meetings produce better results.
- Panels (as used by NSF) are subject to influence by one strong-willed panelist, which may lead to "randomness" of the results. Others pointed out that program directors have input that can mitigate this concern. Is it better to have more funded proposals at smaller amounts versus more of a "winner take all" approach?
- The purpose of conference papers should be the benefit of the research community, not the authors. Low acceptance rates and need for an acceptance for some to get travel money are a harmful combination. We should emphasize more papers rather than better papers. In many Physics conferences, presentations are only 12 minutes long.
- Several members of the audience expressed concern about getting good reviewers. There should be some value associated with getting a good reputation as a reviewer. In various subfields, some people gain a reputation as good PC members, get asked to multiple committees, and as a result PC membership is prestigious. NSF does not have the same level of reputation process. NIH rewards panelists with some relief from proposal deadlines.
- NSF review panels tend to be conservative in looking at each proposal rather than seeking a portfolio that includes riskier proposals. Conservative panelists can make it harder for trailblazing research to be funded.
- A further issue in experimentation is the phenomenon of 10K datasets being used to study petabyte-sized problems.
- In 2008, the SIGMOD program committee convened a trial sub-committee to evaluate repeatability of experimental results in submitted papers. In 2009, this trial will continue on a voluntary basis. Authors may submit their experimental results to a standing committee, who will evaluate results for repeatability and give them a "stamp of approval". Other communities have made similar efforts to measure the quality of experimental results.

4. Conclusion

As the importance of top conferences in the tenure and promotion process is being more widely recognized and accepted, there are efforts emerging to make the conference review process more journal-like (e.g. two rounds of review with author feedback). However, given the page limits, the resulting paper is necessarily incomplete. While such papers indicate true academic achievement and thus represent a valid benchmark for tenure, they lack the level of detail that permits readers to gain a deep understanding of the work and to repeat experiments.

It was clear from the reaction to the panel that concerns with the reviewing process cut across many, if not all, fields of computer science. While numerous changes are being tested, there is a larger concern about how new types of publication will be interpreted by tenure-and-promotion committees, many of whose members may not be familiar with the norms of our field. Much can be learned from the variety of experiments, but this same variety may create career-management issues for academics.

Despite a broad recognition of the importance of the issues discussed, there was no clear conclusion in terms of next steps. There is substantial support for "out of the box", novel, and, perhaps, risky experiments in the review process and the mode of publication. However, these novel approaches are met with concerns from some, especially as regards explaining to tenure committees (usually consisting of mostly or entirely non-computer-science faculty). There is disagreement on whether the CRA Best Practices Memo needs updating, and, if it does, when that should happen.

The database research community has taken a strong leadership role in its experiments, including double-blind reviewing, considering ways to ensure the repeatability of experiments, and the VLDB e-journal. Each of these has led to healthy debate and discussion. It is clear from this panel session that the database research community is not alone in its interest in testing alternative review processes and modes of publication.

We look forward to continued consideration of these issues.