POPL'12 Program Chair's Report

(or, how to run a medium-sized conference)

Michael Hicks

Department of Computer Science
University of Maryland, College Park, USA

1. Introduction

It was a pleasure and a privilege to serve as the program committee (PC) chair of the 39th Symposium on the Principles of Programming Languages (POPL). This paper describes the review process we used, why we used it, and an assessment of how it worked out.[1]

We made some substantial changes to the review process this year, most notably by incorporating a form of double-blind reviewing. These and other changes were made in an attempt to improve accepted paper quality, as well as to improve review quality and fairness (both of which ultimately support paper quality).

Much of this paper argues in favor of these changes based on principle, i.e., why one might think the process should increase quality. Ideally we could also evaluate the process directly, i.e., by showing that this year's program was better than it would have been under a different review process. Unfortunately, I think it would be very difficult to efficiently evaluate a review process directly (e.g., by having two committees and two review processes on the same papers). As such, I exercised a more tractable alternative: I polled the authors and reviewers to report on their experience, and to see whether that experience convinces them that the process has merit. In most cases, the answer was "yes."

In detail: I invited the review committee to provide their opinions about the proposed review process, and alternatives, in February 2011, and then asked them about it again after all decisions had been made, in October 2011. Of the 26 members on the PC, 19 responded to the first survey, and 25 responded to the second survey. Of the 60 of the external review committee (ERC), 48 responded to the first survey, and 50 responded to the second survey. Finally, I asked the 494 authors (of 205 submitted papers) their opinion of the process in October 2011 and 275 of them responded. I greatly appreciate the time the authors and committee members took to fill out these surveys.

Summary data from all of the surveys (including anonymized textual comments), my slide presentation from POPL'12 with graphs illustrating trends from this data, and some scripts and other code is available at my website [3]. I present overall results from the surveys in the context of discussion about the process throughout Sections 2–5.

[1] Note that this article expands on the foreword I wrote for the proceedings, with some parts of that foreword reproduced verbatim.

1.1 Recommendations

Here I list each of the things we did for this year's POPL, and my recommendations for whether to do them again:

Do use *light double blind reviewing* (DBR), in which authors make few changes to blind their paper, there are few limits to post-submission dissemination, and authors' names are revealed to a reviewer at the time the review is submitted

Do use an *external review committee* ERC to augment the program committee (PC) to provide expert reviews, employing ad hoc external reviewers when necessary, and to review PC submissions

Do assign a *guardian* per paper to ensure it receives high-expertise reviews

Do ask reviewers to provide two separate *preference* and *likely-expertise* scores when bidding for papers

Do use a global paper assignment algorithm, such as the min-cost max-flow algorithm I used

Do allow supplementary material separate from the main submission (enforcing a hard page limit on the latter)

Do allow a four-day period during which authors may respond to their reviews

Do hold an in-person meeting of PC members to decide papers (all ERC discussion is electronic)

Do allow two weeks of electronic discussion about papers prior to the PC meeting (to determine which papers should be discussed)

Do discuss papers in quasi-random order at the PC meeting

Do not require supplemental material be made available only after a review is submitted; make it available (blinded) during the initial review of the paper

Do allow non-blinded supplementary material (e.g., code, URLs to demos) to be made available after a review is submitted in addition to blinded material available before

Do not use several fine-grained scores on review forms; use *Overall merit* and *Expertise*, and possibly *Confidence*

Do use the HotCRP conference management system (but with a different paper assignment algorithm)

As the PC Chair I had several activities particular to me, and I would do them again:

Do check that author-entered conflicts of interest are valid before bidding on papers begins

Do read all of the reviews, to make sure they are clear about the reasons for their overall assessment, to be fair to the authors, and to ensure informed discussion later

Do avoid reviewing papers, leaving this job to the PC/ERC so as to maintain anonymity

Do read papers on occasion to help un-wedge discussions

There are some things we did not contemplate initially, but after the fact I think they might be a good idea:

Do encourage ERC members to identify PC papers needing more than electronic discussion, and use conference calls to decide their fate

Do employ a *scribe* to summarize the discussion of each paper at the PC meeting, to be added in some form to the paper's reviews

1.2 Outline

The remainder of this document expounds upon these recommendations. The next three sections focus on the most significant elements of the POPL'12 process: light double blind reviewing (Section 2); the external review committee for handling most expert reviews and PC submissions (Section 3); and guardians, one assigned to each paper, to share the responsibility with the me of ensuring the paper receives sufficiently expert reviews (Section 4). Section 5 considers other aspects of the review process, such as how I chose the PC, how we handled author response, etc.

2. Light double blind reviewing

In a single-blind review (SBR) process, the authors do not learn the identity of the reviewers, but the reviewers know the authors' identity. In a double-blind process, the reviewers do not know the identity of the authors—hence *double* blind. The intention of DBR is to increase the fairness of reviewing and the quality of the accepted papers by avoiding initial, perhaps unconscious bias for or against a paper based on its authorship. A reviewer who picks up a paper known to be written by a famous author and/or institution may grant more to it ("Joe wrote it so I'm sure it must be good") than to a paper by unknown authors from an obscure institution ("I've never heard of this place—do they even have reasonable PL researchers there?"). If this bias is strong enough, lesser papers may be accepted over ones of higher quality, or certain population segments, such as women, may be discriminated against [7]. After consulting with the POPL Steering Committee and past Chairs of other conferences, and reading some relevant literature (e.g., Snodgrass [6] and McKinley [4]), I decided that DBR could have a positive effect and thus it was worth trying.

While DBR aims to improve fairness and quality, it complicates the process of writing and reviewing papers. For example, some blinding processes require the authors change their paper in certain ways that make the authorship less apparent, but also may weaken the overall paper. The authors may be asked to remove personal judgments gained from past experience that motivate the current work ("when we worked on the WizWoz system, we discovered that ..."). Or, they may be asked to change the names of systems known to be developed by a select group of possible authors, with the consequence that reviewers may think that the authors do not know their own past work ("this work seems surprisingly similar to the WizWoz system, which the authors seem unaware of ...")! Authors are also often asked not to share drafts of their paper for comment or otherwise talk about their work while the paper is under review, potentially inhibiting scientific progress, and even complicating interviewing for a job.

To mitigate these disadvantages, we employed a "light" form of DBR, which relaxes the blinding actions of both authors and reviewers.

2.1 Light blinding

Blinding is light in that authors make only two identity-masking changes to their paper: the must redact their names from the front page and cite their own work in the third person (e.g., not "We build on our previous work [8]" but rather "We build on the work of Bailey et al. [8]"). More draconian changes to obscure likely authorship, such as altering the names of well-known systems, are not required.

Most forms of post-submission dissemination are permitted. For example, authors are permitted to post their paper on their web page, and to share the paper with those not on the POPL committee. (They may also share it with those on the committee with whom they have a conflict of interest, since those members cannot review the paper anyway.) Authors are also permitted to give talks about their work, e.g., for job interviews. On the other hand, authors are asked not to deliberately subvert the aims of blinding, e.g., by e-mailing committee members their unblinded paper or broadcasting it to a major mailing list, e.g., the TYPES list.[2] In short, the goal is to support a reviewer who does not wish to know the identity of the authors, but at the same time not unduly hinder scientific progress.

[2] Having previously broadcast a paper on a major mailing list would not preclude submission.

For a reviewer, the conference management software reveals a paper's authors immediately upon submitting a review, as opposed to later in the process. The idea is that by the time a review is submitted, initial bias has been mitigated, so revealing author identity should not negatively impact fairness. This approach confers several advantages. First, reviewers can solicit additional expert reviewers, if needed, since by knowing the authors they can (more easily) avoid soliciting those with conflicts of interest. Second, reviewers can adjust mistaken assessments that hinge on author identity, e.g., criticisms of similarity to the authors' own prior work. Finally, revealing identities before the PC meeting avoids certain abuses. For example, it avoids the possibility that a program committee member will know a paper's authors (e.g., due to outside information) and advocate for the paper on that basis (e.g., due to a personal friendship), but not appear to have any bias since other reviewers would not know the paper's authorship.

2.2 Survey results: overall effectiveness

In each of the three surveys I conducted, the majority of respondents preferred light DBR over traditional SBR, with the total average being 70% in favor.[3]

Preferred process		Light DBR		SBR	
PC+ERC	pre-review	67%	43	33%	21
	post-review	70%	47	30%	20
PC	pre-review	78%	14	22%	4
	post-review	92%	22	8%	2
ERC	pre-review	63%	29	37%	17
	post-review	58%	25	42%	18
Authors	all	70%	173	30%	73
	non-reviewers	72%	152	28%	60

(The first three rows come from the two reviewer surveys, and the last row comes from the author survey which included some reviewers). One interesting trend in this data is that while the PC was more favorable to light DBR after having undertaken it, the ERC was less favorable. Since there is no guarantee the respondents to the first and second reviewer surveys are the same, this difference could easily be in the margin for error. Indeed, by and large, reviewers said they viewed DBR more favorably when asked about it specifically in the second survey. In particular, the entire PC was more favorable to DBR and 75% of the ERC was more favorable, with 84% more favorable overall.

Change in opinion about DBR	Improved a lot		Improved a little		Degraded a little		Degraded a lot	
PC+ERC	23%	15	61%	39	11%	7	5%	3
PC	49%	11	51%	13	0%	0	0%	0
ERC	10%	4	65%	26	17%	7	7%	3

[3] In this table and all tables I present, the percentage of respondents is given prominently, with the count of respondents adjacent to it. Note that not all respondents answered all questions, and non-answers are not interpreted as abstention since such an interpretation was not obvious; e.g., in many cases, survey takers seemed to just have quit the survey at a certain point.

Nevertheless, the difference in opinion between the PC and ERC is curious. I believe part of the explanation is due to whether blinding was (ever) effective for ERC reviewers, compared to PC reviewers. In a fourth survey[4] I asked reviewers how often they guessed a paper's authorship right, and how often they guessed it wrong, and also, when they did not have a specific guess about the authors, whether they were surprised at who they were.

Outcomes (per reviewer)	Guessed (at least one) correctly		Guessed (at least one) incorrectly		No guess, but surprised		No guess, authors unknown	
PC	100%	21	86%	18	90%	19	90%	19
ERC	81%	33	22%	9	39%	16	61%	25

Only 22% of ERC members guessed at least one of their papers incorrectly, while 86% of PC members made an incorrect guess. The explanation is probably due to ERC members reviewing fewer papers, and more likely being an expert reviewer. As a result, the average ERC member rarely felt that blinding made a difference, whereas the average PC member was affected by the surprise of a wrong guess to the point they wondered about their objectivity. Text comments from the post-review survey support this explanation; e.g.:

> On two submissions, if I had known the authors, I would have started with a higher opinion of the submissions than was justified. I would have realized before long, but the DBR saved me the time of realizing that people I respect could have done better.
>
> I was really surprised by authors in a handful of cases. My reviews might have been biased if I'd known authorship up front.
>
> I thought I would be able to easily guess who the authors of various papers were. I turned out to be mistaken in many cases, and as a result, I stopped thinking about authors altogether. I think this is a good outcome.
>
> As a reviewer, my opinion [of DBR] improved because there was one instance where I was truly surprised. The degree of surprise suggested to me that I might have held some unconscious bias for the authors had I known their identity in advance. ...

The following table counts the outcomes of particular reviews reported by PC and ERC members; of the 812 reviews submitted by PC and ERC members, the table recounts outcomes of 540 of them.

[4] This was a separate survey from the three I listed above; 21 PC members responded, while 41 ERC members responded. I had asked them to keep track of the answers to these questions before they began reviewing.

Outcomes (per filed review)	Guessed correctly	Guessed incorrectly	No guess, but surprised	No guess, authors unknown
PC+ERC	42% 228	12% 66	11% 58	35% 188
PC	41% 159	13% 50	10% 38	36% 142
ERC	46% 69	11% 16	13% 20	30% 46

Overall, when the reviewers felt they could guess the paper's authorship (first two columns in the table) they were right 77% of the time, with the PC right 76% of the time, and the ERC right 81% of the time. This is consistent with Snodgrass' summary of past studies of blinding efficacy [6].

2.3 Survey results: impact on authors

Above, I redacted the last line of the last quote; it reads "As a writer, I didn't like it because the paper I wrote built upon a previous paper I wrote." In short, while DBR may be useful, it imposes the cost on the authors that they must understand the blinding instructions and change their paper accordingly. DBR also places limits on dissemination of papers under review. The question is, how significant are these costs?

When I surveyed the authors, I asked them about the difficulty of the blinding instructions; 98% (224 out of 229 respondents) said they were easy or mostly easy to understand. I also asked about particular changes that authors made to their papers. The authors that responded represented 138 of the 205 submitted papers, and the changes required to the papers were organized as follows (a paper could have 0 or more of the changes listed in the first four columns):

Number of papers affected (138 of 205 papers covered)				
Citations 3rd person	Redacted qualitative judgment	Omitted text or other references	Anonymized citations	Only "easy" changes
73% 101	16% 22	24% 33	27% 37	53% 73

The last column shows that 53% of papers required no changes at all or *only* modification of citations to be in the third person for the authors' own prior work; thus 47% of the papers required at least one of the other three changes. I was surprised at the high percentage of anonymized citations (27% of papers) since the instructions indicated that this should be done only in very rare circumstances. (Authors may have misinterpreted this question to mean citing one's own work in the third person, making it "anonymous.")

I also asked the authors whether they felt their submission was hurt by the changes they made to it.

Impact of blinding changes on paper (226 respondents)				
Improved substantially	Improved slightly	No change	Hurt slightly	Hurt substantially
1% 1	1% 2	79% 178	18% 41	2% 4

80% of the respondents felt either the paper improved or was not impacted, while the other 20% felt it was hurt, though few thought it was hurt substantially. Interestingly, the authors' judgment did not correlate with a paper's final decision. I correlated these answers with decisions on the authors' papers, and 18% of authors whose papers were (all) rejected deemed their papers as having been hurt, while 17% of authors whose papers were (all) accepted also deemed them as having been hurt. From various conversations I got the feeling that some authors did not trust that the rules should be implemented exactly and made more changes to their paper, to hide their identity, than were actually required. I expect that over time authors would gain trust and follow the instructions more precisely, and thus feel less like their papers were being negatively impacted.

Finally, I asked the authors whether they changed their post-submission behavior so as not to violate the rules of dissemination. I aggregated the results by paper; in the following table, the first three columns are not mutually exclusive.

Change in post-submission dissemination actions (138 papers)			
Did not send to PC/ERC member	Did not announce paper to public forum (list/blog)	Did not offer to give talk at PC/ERC institution	No change
32% 4	48% 2	23% 32	49% 67

In total, 51% of papers' post-submission dissemination actions were impacted. One question is whether the authors were being unnecessarily conservative in their reading of these instructions, out of fear of having their paper rejected for violating the rules. When I asked the authors directly whether they felt uneasy about disseminating their paper, most said 'no':

Did you feel uneasy about disseminating your paper? (228 respondents)		
No	Yes, a little uneasy	Yes, quite uneasy
65% 150	29% 64	6% 14

Interestingly, when I correlated these results, I found that feeling more or less uneasy did not impact actual dissemination actions. Again, perhaps over time authors would become more comfortable with the rules.

2.4 Assessment

Ultimately, we would like to be able to assess final outcomes. Does (light) DBR lead to more papers being accepted fairly, i.e., are fewer papers rejected because their authors happen to be women or minorities, or are simply unknown or affiliated with unknown or lightly regarded institutions? Answering this question is very difficult. Comparing to prior POPLs would be difficult because they involved different committees and different submissions. Using two different committees to review the same papers, one committee using SBR and one using light DBR, would help, but such an experiment would be incredibly costly. Even if it were not, one might guess that two different committees using the *same* process on the same papers would come to different outcomes—how to factor out this difference?

Nevertheless, my overall feeling is that light DBR was worth it. It was not too much work on the authors, essentially no extra work for the committee, and just a little more work for the Chair (dealing with COIs, see Section 5.3). Though some post-dissemination actions were restricted, the fact that papers seem to have been more carefully and fairly reviewed strikes me to be a net positive, as authors receive better feedback and the committee's judgments of acceptance are better informed. Of more than three hundred members of the POPL community I surveyed, 70% seem to agree that light DBR was worth it.

3. External review committee

A significant drawback of typical DBR is that it can complicate the process of finding expert reviewers. POPL, like other flagship conferences, has become so broad that it is unlikely that a standard 20-25 person program committee can review each paper with the requisite level of expertise. Thus it has become common to seek expert reviews from outside the committee. In a traditional DBR process, the task of finding external experts falls to the program chair: to avoid soliciting reviewers with a conflict of interest one must know a paper's authorship, and only the chair knows the authors. As McKinley pointed out, for a conference with 200+ submissions, like POPL or PLDI, this can be a big job [4].

Her suggested solution, originally proposed and piloted by Steve Blackburn for ISMM'08 and now regularly employed by SIGPLAN conferences such as ASPLOS, PLDI, and ISMM, was to use an external review committee (ERC).[5] Essentially an ERC is just an additional program committee whose members are asked to perform fewer reviews. ERC members bid for and are assigned papers along with the regular PC, and the process takes into account potential conflicts of interest in the same way. As Aiken points out [1], even in a single-blind review process it is advantageous to have a large, diverse body of committed reviewers on hand rather than try to find them while "on the clock."

Another advantage of an ERC is that it can be used to review PC submissions. The alternative of allowing PC members to review each others' papers can lead to trouble at the in-person PC meeting. For one, a paper's author may be able to influence his potential reviewers, which seems unfair to non-PC papers. Or, a PC member may become upset at how his paper was reviewed and/or whether it is accepted, creating tension at the meeting. Mooly Sagiv's POPL'11 Program Chair report [5] hints at these problems: "My biggest mistake by far was the way PC papers were handled which led to unnecessary rejections and caused some really bad feelings among all of us on the PC. In retrospect, it would have been better and fairer if I had assigned the papers to external experts outside the PC and discussed them prior to the PC meeting with the external experts not involved the PC members at all and announce the results after the meeting."[6] By contrast, with an ERC, all PC members know that no one present at the meeting reviewed their paper and no amount of influence can change the paper's outcome, which is determined by the ERC electronically and publicized after the PC meeting.

A final advantage of the ERC compared to ad hoc external experts is that ERC members "have more context" when judging papers because they have reviewed several of them, rather than just one. I had thought that there was roughly an "absolute bar" by which a submitted paper could be judged, but what I found is that each reviewer's notion of that bar is different. Moreover, even an individual's notion is a bit fuzzy, and must be made more solid through comparison to other papers. Having even just three or four papers can be helpful in calibrating a reviewer's sense of quality.

For all of these reasons I decided to go with an ERC for POPL. We ended up with a PC of 26 researchers who reviewed 20–23 papers each, while the ERC consisted of 60 members who reviewed an average of 3–4 papers each, with a maximum of 6. Ultimately, we completed 852 total reviews for 205 papers: 559 were performed by the 26 PC members (20–23 each); 253 were performed by the 60 ERC members (2–6 each); and 40 were completed by outsiders. (Twelve additional researchers assisted PC members with their reviews, so a total of 50 outsiders helped with reviewing.)

3.1 Assessment

On the post-review survey I asked about the utility of the ERC. The overall response was very much in favor.

Use an ERC?	In favor		Against	
PC+ERC	89%	56	11%	7
PC	73%	16	27%	6
ERC	98%	40	2%	1

Interestingly, the ERC was highly in favor of itself. This response contradicts some claims that I have heard that being on an ERC is not worth it, since it is more work than being an ad hoc external reviewer, but the extra recognition in the proceedings is not much of a benefit.

I also asked specifically about the roles the ERC takes, and whether another mechanism might have worked better. With regard to handling expert reviews, the response was very much in favor of what we did:

Handling non-PC member expert reviews (62 PC+ERC respondents)							
ERC+outsiders		Outsiders only		ERC only		Other	
82%	51	14%	9	2%	1	2%	1

Finally, I asked how PC submissions should be handled, with the preference again in favor of the process we used, compared to using outsiders only, allowing the PC to review

[5] This acronym is sometimes expanded to *extended* review committee, which is also apropos since an official committee is no longer "external."

[6] Mooly points out that you do not need an ERC to do this; you could also ask ad hoc external reviewers.

its own submissions but holding them to a higher standard, or to disallow PC submissions entirely:

Handling PC submissions (56 PC+ERC respondents)			
ERC + outsiders	Outsiders only	PC can review (higher standard)	No PC submissions
71% 40	4% 2	14% 8	11% 6

Interestingly, when I broke down the responses according PC/ERC membership, I found that the PC respondents were more likely to suggest the PC should not review its own papers (18% to 4%) and even that they should be able to submit papers at all (18% to 5%). Perhaps PC members can more easily imagine, having recently met in person, that judging each other's papers could lead to bad feelings just as Mooly had suggested.

4. Guardians

In (hopefully rare) circumstances, the PC and ERC may not have "the right person" to review a particular paper. Moreover, the imperfect process of assigning reviewers to papers may fail to assign the right person even when he/she is on the committee, e.g., because of an unfortunate choice of bids, or because a particular reviewer is overloaded. Since these circumstances do arise, I expected I would need to occasionally solicit an outside reviewer or shuffle review assignments. However, I worried that because there are many sub-communities with which I am not intimately familiar, I would have trouble identifying some experts without help.

My solution to this problem was to specifically assign a *guardian* to each paper.[7] The guardian is chosen as a likely expert among the initially assigned reviewers for the paper, and is responsible for submitting a review mid-way through the review period. If the guardian's review is not expert-level (most were), he or she could request an additional review, either from the ERC or PC, or from an outside reviewer. By asking a likely-expert to get his/her review in early, I ensured that I had time to take advantage of this expertise before the review period was over, if it was needed. I also generally encouraged reviewers to submit reviews as they completed them, rather than waiting until the end, so that I had a pretty good picture of where we stood on particular papers throughout the review process. In total, we received reviews or feedback from 50 outside reviewers, and we occasionally shuffled ERC and PC assignments to better take advantage of available expertise.

4.1 Assessment

On the post-review survey, 93% of respondents were in favor of using guardians, with a few hedging this recommendation to only DBR processes (which arguably might have more need of guardians because in an SBR process an external review can be solicited at any time during the process):

Assign per-paper guardian (62 PC+ERC respondents)		
Both for SBR and DBR	For DBR only	Do not use guardians
87% 54	6% 4	6% 4

5. Remaining details of reviewing

Here I present a few other details of the process I used, including how I chose the PC and ERC, how I handled conflicts of interest (COI), how I performed paper assignment, and how we incorporated additional author feedback and organized our discussion about papers.

5.1 Choosing the PC and ERC

Before being asked to be PC Chair of POPL I had a fairly good knowledge of several of POPL's sub-communities, but there were several other sub-communities I knew very little about. Thus it was a bit of a daunting prospect to find good reviewers across the whole space of possible submissions. Here were the steps I took.

First, I downloaded DBLP metadata about the authorship of papers appearing in the past five POPLs and in conferences whose areas overlap with POPL, including CAV, ECOOP, ESOP, FSE, ICFP, ICSE, ISSTA, LICS, OOPSLA, PLDI, and VMCAI. Combining all of this I ended up with the recent publication history of 3259 authors. Then I wrote some Ruby scripts to cull down this list to a more manageable number. For example, I filtered out authors who had no POPL papers (this is POPL after all!), whose most recent paper in any of the above venues was more than two years back (approximating recent research inactivity), and whose earliest paper was only three years back (likely a student or a new entry to the area). I also filtered out anyone who had been on the POPL committee in the past three years. At this point I could browse the data and get a sense of the activity of relevant researchers and their particular sub-areas. Many of these people I knew, and that personal knowledge was important in my decision to invite them. But many I did not know at all, or had heard of only peripherally. I continued to play with and augment the data I had by visiting people's web pages, looking at their papers, and consulting with the steering committee.

Eventually I had a diverse group of 26 PC members. A few people turned me down, on several occasions because they did not want to travel to an in-person PC meeting. Most of these agreed to be on the ERC, and of course I invited many more to comprise the 60-person ERC. I cannot say enough about the members of POPL'12 PC and ERC: they were simply fantastic, on the whole submitting carefully written, thoughtful, detailed reviews, on time.

I have made available all of the scripts I used to download and manipulate the DBLP data [3].

5.2 Bidding and paper assignment

It is typical for reviewers to express preferences about which papers they would like to review (i.e., "bid" for them), to

[7] The idea for guardians was based on an idea suggested to me by Stephanie Weirich.

assist the program chair in assigning papers to reviewers. I did two things that are a bit different than recent processes I have experienced.

First, I initiated paper bidding only after the full paper was submitted. In recent years, in an attempt to shorten the review period, conferences have asked that abstracts be submitted first, and then final papers a week later, with bidding only taking the abstract into account. I find this practice problematic for two reasons. First, an abstract is not very much information, which reduces the quality of the bids. It could be a reviewer knows the application area very well, but a cursory glance of the paper might have told him that the methods being used are outside his comfort zone. I encouraged reviewers to take bidding very seriously, i.e., actually looking at the submitted paper, since well-informed bidding should lead to a better assignment, which leads to better reviews and decisions. The second problem with bidding only on abstracts is that it wastes reviewers' time if many submitted abstracts never materialize as actual papers. This year, we received 238 "registrations" of papers about four days before the deadline (for handling conflicts of interest, see below), but only 205 materialized. I have heard of past POPLs in which as many as 60 abstracts never became real papers.

The second change I made was to use two-dimensional bids: reviewers indicate how much they would like to review a paper (the *preference*) and also what level of expertise they would expect to have if asked to actually perform a review (the *relevance*). This approach allows interested outsiders to have a chance at reviewing a paper (high preference, low relevance) while ensuring that doing so does not eliminate necessary expertise (e.g., at least one high relevance reviewer should be assigned). It also gives a way for an expert to express disinterest while still expressing that he is relevant to the paper should an interested expert be unavailable. Relevance scores were also the basis for picking guardians when papers were assigned.

To take this information into account, I worked with my colleague Samir Khuller and his student Matt McCutchen to adjust their neat paper assignment algorithm to my process. The Haskell source code for this algorithm, including extensions to support POPL'12, is available at `https://mattmccutchen.net/match/`. I give a brief description of it below. I also had to make changes to the HotCRP software itself to support two-dimensional bids; an undergraduate at Maryland, Jamie Salts, programmed most of the changes. Finally, I wrote some scripts to interface the matching algorithm with HotCRP's text-file-based paper assignment interface [3].

Algorithm Let N be the number of papers and P be the number of reviewers. Suppose that each paper needs q reviews, so a total of qN reviews need to be generated. Ideally, from the perspective of the papers, we would like to assign each paper the q most qualified reviewers for the paper. Of course, this could lead to a load imbalanced solution where the load on some program committee members is very high, and the load on others is low. On the other hand, we could insist on a perfectly load balanced solution in which the number of papers assigned to each program committee member does not exceed $L = \lceil qN/P \rceil$. However, this may lead to a solution which is not optimal from the perspective of the papers. Therefore we introduce a *load tolerance* factor C, to allow each reviewer to be assigned up to $L + C$ papers. For POPL'12 we set $C = 2$.

We formulate the assignment problem as a min-cost max-flow problem, where each unit of flow represents one review. In the construction, there are nodes that represent reviewers and nodes that represent papers, so a flow that can pass from the source s through a node for reviewer i and then a node for paper j to the sink represents an assignment of reviewer i to paper j. The construction is somewhat involved in order to incorporate all the desired incentives. As some examples: (1) each reviewer has a zero-cost, capacity-L edge from the source so he/she can review (up to) L papers, along with C additional edges of increasing cost to allow overload; (2) each paper has a zero-cost edge of capacity q to the sink to ensure it receives q reviews (the construction ensures the maximum flow will saturate these edges); (3) there are multiple nodes per paper, where each represents a review at a particular level of expertise—there is a "bonus" (unit-capacity, negative-cost) edge to incentivize expert reviews.

The algorithm also turned out to be useful for assigning guardians: the potential guardians for the paper were the PC members assigned to review it; one of them should be chosen ($q = 1$); and we want to maximize the relevance and preference in the assignment, as usual, without overloading a particular PC member (I believe I used $C = 3$ here).

More details about the algorithm are given in a paper that accompanies its source distribution. As a general point, the thing I liked most about the algorithm is that it globally considers all information in making a choice, whereas other assignment algorithms (e.g., the one used for HotCRP) take a greedy approach which can lead to poor assignments.

5.3 Conflicts of Interest

When using SBR, reviewers can identify potential conflicts of interest with a paper's authors during bidding. In a double blind process, the authorship is not known, so this method will not work. One standard approach is for authors to specifically identify, at the time they submit their paper, those reviewers with whom they perceive there to be a conflict. For example, if author Bob is in the same Department as reviewer Alice, Bob will select Alice's name from the PC members listed on the submission form when submitting the paper. When Alice bids for papers Bob's paper is not shown.

This approach can have false positives and false negatives. Bob may fail to mark Alice as conflicted, in which case Alice may end up bidding on and ultimately reviewing Bob's paper, only to discover when the review is done that there is a conflict. Conversely, Bob may mark a non-conflicted, but highly qualified reviewer Charlie, thus preventing Charlie from reviewing his paper. The result is a possible subversion of the peer review process; e.g., Bob could know that Charlie is unlikely to find his paper convincing and thus blackballs his review by marking him as "conflicted."

To avoid these problems, I asked reviewers to list, in the submission system account metadata, all of their institutional and personal conflicts of interest in advance of seeing any submitted papers. Reviewers sometimes maintain such lists anyway because they may be required when submitting grant proposals. Then I had authors "register" their paper four days before the submission deadline. For this I only asked for a title, authors, affiliations, and conflicts (an abstract was optional). I then manually cross-referenced the author and reviewer affiliations and listed conflicts. When conflicts were missing, I added them. When conflicts seemed spurious, I queried the reviewers and if they agreed, I asked authors as to why they had marked the person in conflict. On a couple of occasions, the authors admitted to attempting to prevent a non-conflicted reviewer; most often they were honest mistakes. The most difficult case is handling conflicts due to personal friendship; if Bob and Charlie are friends, Bob may think himself unable to review Charlie's paper, so he marks Charlie as conflicted for his paper, but in fact the feeling is not mutual.

This process of handling conflicts was time consuming, taking 8-10 hours of work over a few days, and I believe it was the largest cost of the DBR process I used that would not be incurred by an SBR process.

5.4 Author response and paper discussions

At the end of the main review period (10 weeks total), authors were given four days to respond to their reviews. I released all of the reviews to the authors, including the numeric scores, so authors knew which reviewers were most and least favorable. I did not place any limit on the length of the response, but encouraged authors to be brief, since it would increase the chances that reviewers read and considered the response. Hard limits on response length discourages authors from quoting reviewer comments and responding to them directly, which I find is what reviewers most want to see. On the other hand, authors can go overboard; in one case the author response was 6600 words! I told reviewers to apply their own judgment as to what constituted a fair reading of a response.

Following the author response we had two weeks of electronic discussion. The goal for non-PC papers was to come up with a list of papers to discuss at the meeting. We used A–D ratings for a paper's *Overall merit* score, and any paper with an A would be discussed, along with papers whose average score was roughly a B. During the discussion, people often adjusted their scores, both up and down, and eventually we had 84 non-PC papers to discuss over two days.

During the PC meeting, we discussed papers in quasi-random order, as recommended by Kathleen Fisher [2] (and employed prior to her published note by several others). The idea is that discussing papers in order of average merit score (e.g., highest to lowest) creates an implicit expectation that later papers are not as good as earlier ones. Randomizing the order mitigates this problem. On the other hand, I reordered some papers to group similar sets of conflicted PC members, to reduce traffic in and out of the room.

PC papers were decided entirely through electronic discussion amongst the ERC. In most cases, electronic discussion was sufficient to reach a consensus. In one case, the reviewers held a conference call to discuss the paper.

5.5 Changes and additions

I tried two new things that I would not do again, involving review forms and supplemental material. It also occurred to me that I should have employed *scribes* at the PC meeting to summarize paper discussions to be conveyed to the authors.

Review forms. I added four additional numeric scores to the review form: *novelty*, *importance*, *conviction* (or "convincingness"), and *clarity*. The idea is that these fields roughly represent an accepted basis for judging a paper, e.g., according to guidelines from many journals that I checked. As such I thought it would make sense to give them explicit scores in the review form to, if nothing else, remind reviewers that they ought to be taken into account.

Some reviewers were very positive about these scores. The main problem with them is that they are very subjective and sometimes hard to determine, and they ended up not being very relevant to me as PC Chair. They were also confusing to authors, who might see several of these scored highly but the overall score being quite low. Obviously there is no direct function from them to the merit score. On the other hand, overall merit and expertise scores are both extremely useful to the PC Chair to determine which papers should be discussed and whether more reviews might be needed, respectively. I would also consider adding a *confidence* score to identify how well a reviewer understood a paper, whether or not he or she is expertly informed about the area or related work, since this information might also help determine whether further reviews are needed.

On the post-review survey, 25 of the respondents (58%) were either ambivalent or opposed to the additional numeric scores, as opposed to 18 (42%) in favor of them. On the other hand, 73% of respondents were in favor of an expertise score, and 59% were in favor of a *confidence* score. Only two respondents (5%) were in favor of a *quirkiness* score, used in some recent conferences (not POPL'12).

Supplemental material. Authors could submit supplemental material, such as proofs or the code of their implementation, but this material could be non-anonymous, so we only made it available after a review was submitted. The idea was that this would be less work for authors, and it would emphasize to reviewers that the main paper is what is under review, and not the supplemental material.

On the other hand, several reviewers were very hampered by this policy, and in some cases just submitted "stub" reviews to gain access to supplemental technical reports with proofs. Given that it is not hard to anonymize this material, and that reviewers can be more confident in reviews if given access to it, it seems sensible to make this material available pre-review. At the same time, material that is less easily anonymized, e.g., URLs to authors' home institutions to point to code or web demos, can be made available after a review is submitted.

When asked about supplemental material on the post-review survey, 34 respondents (56%) were in favor of revealing supplemental material only after the review is submitted (and half of these felt this way even if using an SBR process), while 16 (26%) felt that anonymous supplemental material should be available before pre-review (with 6 of these thinking both pre- and post-review material should be made available). Despite being in the minority, I now side with the 26%—I think it makes sense for reviewers to be able to check proofs and other claims when first reading the paper and that the extra cost to authors (in terms of anonymizing that material) is low or nil. Interestingly, 11 respondents (18%) felt that we should accept *no* supplementary material at all—they emphasized that the submitted paper should be able to make its case in the twelve allotted pages.

Scribes. Authors often want to know how the assessment of their paper changed following the author response. In some cases, papers that looked like they would be accepted were not, and authors were perplexed as to what happened. In these cases, I ended up recapitulating the discussion, e.g., from electronic comments, and shared that with the authors, so they better understood the situation (oftentimes, a negative expert was able to convince the positive non-experts that they were missing something important about the paper). To make such recaps more systematic, I would recommend having a scribe at the PC meeting to record the discussion that takes place, and once the discussants approve it, it can be added to one of their reviews.

6. Conclusions

Ultimately I am happy with the review process used for POPL'12 and I would do it again with only slight changes. While I have not evaluated the process directly, e.g., by showing that this year's program is better than it would have been under an alternative process, the survey results show at the least that many of the authors and reviewers think the process has merit. Considering the major elements of the review process, of those I surveyed

- most (70% of more than 300 responding authors and reviewers) were in favor of light double blind reviewing;
- a great many (89% of 63 responding reviewers) were in favor of using an external review committee; and
- a great many (93% of 62 responding reviewers) were in favor of assigning a per-paper guardian to ensure expert reviews

Of the most contentious question as to whether to use double-blind or single-blind reviewing: it is clear that even light double blind reviewing has costs that some authors and reviewers find troublesome, but on balance I believe these costs are relatively low and the benefits to review quality and fairness are worth it. Nevertheless, I encourage further study that would attempt to quantify these costs against the overall benefits. Peer review is the foundation of the scientific process—it is a gateway for new ideas and the foundation of our trust in published results. It is essential that we invest the time to find the most effective process we can.

Acknowledgements My heartfelt thanks goes to the PC, ERC, and outside reviewers. Having read nearly all of their 852 reviews, I continue to believe that POPL reviewers are the best around. I thank Alex Aiken, Emery Berger, David Evans, Kathryn McKinley, Andrew Myers, Todd Mowry, and Benjamin Pierce for discussions about running a big conference and using (or not using) double-blind reviewing. I think David Wagner for sharing stories he had gathered about the effect of blinding; these ultimately made it on the POPL FAQ. I must also thank the POPL'12 General Chair, John Field, and the POPL Steering Committee (Tom Ball, Kathleen Fisher, Manuel Hermenegildo, Graham Hutton, Chandra Krintz, Jens Palsberg, Mooly Sagiv, and Philip Wadler) for their excellent advice and guidance throughout. Eddie Kohler provided and supported the HotCRP software we used to manage reviewing, and Jamie Salts helped program some important extensions to it. Samir Khuller and Matt McCutchen worked with me to modify their paper assignment algorithm to suit my committee structure. Greta Yorsh provided expert help in organizing the program. Khoo Yit Phang assisted me in analyzing the data from the surveys.

Thanks also to Emery Berger, Michael Clarkson, Kathleen Fisher, Jeff Foster, Khoo Yit Phang, Justin McCann, Kathryn McKinley, and Phil Wadler, and for helpful comments on drafts of this paper.

Finally, thanks goes foremost to the authors of submitted and accepted papers, who provided the material for the excellent program that comprises POPL'12. Keep up the great work!

References

[1] Alex Aiken. Advice for program chairs. *SIGPLAN Not.*, 46(4):19–25, April 2011.

[2] Kathleen Fisher. In support of (quasi-)randomness in the order of considering conference submissions at program committee meetings. *SIGPLAN Not.*, 46(4):17, April 2011.

[3] Michael Hicks. POPL'12 materials. `http://www.cs.umd.edu/~mwh/popl12-materials.html`, 2012.

[4] Kathryn S. McKinley. Improving publication quality by reducing bias with double-blind reviewing and author response. *SIGPLAN Not.*, 43(8):5–9, September 2008.

[5] Mooly Sagiv. POPL'11 program chair report. *SIGPLAN Not.*, 47(4), April 2012.

[6] Richard Snodgrass. Single- versus double-blind reviewing: an analysis of the literature. *SIGMOD Rec.*, 35(3):8–21, September 2006.

[7] Rhea E. Steinpreis, Katie A. Anders, and Dawn Ritzke. The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national emprical study. *Sex Roles*, 41(718), 1999.