

# SODA PC meetings

Claire Mathieu\*

October 6, 2008

Report on the SODA 2009 electronic PC meeting. Advice for future SODA PC chairs.

## Pre-submission work

First task is to form a committee. I got as much help as I needed from previous chairs, David Johnson, and SIAM. Besides the obvious priority of balancing research areas, I tried to balance experienced vs. junior researchers, and also tried to fight gender imbalance. This happens in early January. Most junior people accept the charge but among the most senior people, the odds are about fifty-fifty, so this takes time. Goal: have a committee formed by the business meeting of SODA one year before the conference.

Innovations should be thought through before that time so that they can be announced at the business meeting.

Second task is to choose a submission server. I went with Easychair. Setting it up is very easy.

Third task is to communicate with SIAM about the content of their web page with the SODA announcements.

Fourth task is to choose invited speakers. This should happen in May-June, before the hard work starts, and is the first opportunity for the committee to have a discussion. The rule is that they must be approved by SIAM before invitations can be issued, however I only learned it after the fact. I think I invited 5 people before 3 accepted. My plan was to have one person in algorithms, one in discrete math, and one in a completely different area or even discipline.

## Getting submissions in

For 2009, I asked for authors to submit a title and short abstract one week before the full submission deadline. Pros: forces people to have their result at least one week ahead of time, so they spend at least one week writing it up. Enables PC members to bid on submissions before the submission deadline, this speeds things up significantly after the submission. Cons: More time pressure, further delays between results and conference; a little bit more hassle for each

---

\*Soda 2009 PC chair

submitting author because of the double deadline. In addition a few people ( $< 5$ ) were unable to submit altogether because they accidentally missed the abstract deadline. The deadline must be no earlier than a few days after the FOCS submission results are out. Paul Beame requested that in the future, no deadline related to SODA happened earlier than July 1.

I put the deadline on a Friday (just before a long weekend) so as to encourage people to spend time with their family or at least away from work during the weekend. I put the deadline at 3pm so that tech and admin staff would still be around at the moment of the deadline in case of emergency technical problems. I recommend a Friday 3pm deadline. (I was actually a bit lenient with the exact time and turned off the server about half an hour to an hour after the deadline.)

An innovation this year was that authors had to submit complete proofs. The paper length was still limited to 12 pages but they could add an appendix, with no constraints on its length (and to be read or not, at the discretion of the committee), and the full proof had to be there. Everybody complied with just one or two exceptions, nobody complained, and it was extremely useful in addition to (I hope) giving increased respectability to SODA for people outside the field.

About 550 abstracts were submitted, of which about 460 became actual submissions a week later (a few more than 460, then there were some withdrawals during July and we ended up with 458 submissions.) The Easychair submission server basically worked well. The server was down for a couple of hours during the day of the submission, and sometimes slow, but no big deal. Right after the submission deadline, I had the bids of the PC members already since they had looked at the abstracts the previous week, so I ran the Easychair automatic assignment program and got a good assignment of papers to PC members quickly and painlessly. Within 24 hours of the deadline, we were ready to get to work evaluating the submitted papers!

Each paper was assigned to 3 PC members, and each of the 27 PC members, including myself, had about 50 papers to read. I wish the assignment had been such that for each cluster of closely related papers, one PC member had the whole batch, but there was no way to ensure that at that early stage.

## Evaluation

### Phase 1: Individual reading

I gave the PC about 5 weeks to turn in their reports. The majority of the PC members used subreviewers for the majority of the submissions in their batch. There was great response from the research community, with most people accepting requests to review submissions.

If I was doing it again, I would not give myself any paper to evaluate. Instead, I would spend a few minutes on each submission, to try to make clusters of related submissions, and to get a global sense of the whole set of submissions instead of having just a partial view. At 15 minutes per paper, times 460 submissions, this would already require 115 hours, that is, 23 hours per week

for 5 weeks. And comparing papers to one another take much more time... Also, I spent most of my time after this first phase managing the global decision process and had little time to look back at the submissions in my own batch, so I think that they might have gotten somewhat shortshrifed.

The scores ranged from -3 to +3, with 0 representing a "borderline" assessment. I gave instructions to the PC members with a guideline target distribution of scores for their batch. This was a mistake since some people adhered to my guidelines strictly while others interpreted it as mere advice, and in the end it only created some confusion. I should have made it clear that it was purely indicative.

I did not at that stage give instructions about how much to penalize for poor quality of writing. It would have been better if I had. (My eventual guidelines: Do not penalize poor writing in the appendix. As for the main body of the submission, do not give much penalty to writing issues which can be fixed in a few days, but do give a heavy penalty to papers that would require several weeks of work to be in good shape.)

## **Phase 2: Rejecting 180 papers and accepting 45 papers**

From the beginning of this phase until the end, for the PC chair this is basically a full-time job.

The second phase took about a week:

I gave PC members who were late a few more days to complete their last few reports. I asked for some additional reviewers for some papers that had only two reviews. (Every accepted paper got three reviews.) I was very impressed by how responsive the additional referees were when I asked them for a last-minute review. The vast majority agreed!

I read all reviews which had a wide discrepancy in the scores. In many cases, the reason was that one reviewer said "the result is already known" or "the paper is missing an important set of bibliographical references" or "the argument is flawed" or "the authors did not cite my wonderful work". I brought these criticisms to the attention of the PC members. They were resolved by discussion, by sending excerpts of one review to another subreferee and getting their reaction, or by contacting the authors with a question. I asked the PC member who had found the error to phrase a message for me and sent it myself to the author, to protect their anonymous status. Authors are amazingly responsive to such messages: I always got a response within 24 hours! Usually it was an acknowledgement that their proof was indeed wrong. At the end of that phase, all these obvious factual inconsistencies were resolved one way or another.

The confidence level of the reviews ranged from 0 to 4. I looked at the total confidence of the reviews and encouraged getting higher confidence reviews, particularly for papers that might end up in the accept list. (Every accepted paper and almost all rejected papers got reviews with total confidence at least 5.)

In parallel, I started marking the obvious rejects and obvious accepts using the following rules.

First round: Obvious reject: a paper with maximum score is less than or equal to 0. Obvious accept: a paper with minimum score greater than or equal to 2.

Second round: Obvious reject: a paper whose average score is less than or equal to 0 and maximum score less than or equal to 1. Obvious accept: a paper with average score greater than or equal to 2, and minimum score greater than or equal to 1.

At the end of that easy phase, we had about 180 rejects and 45 accepts. If there had been a physical PC meeting, all this could have been done prior to the meeting.

### Phase 3: Getting to 105 accepted papers and 80 undecided papers

SODA has room to accept at most 135 papers. As it turned out, when ranking papers by average scores, the 135th highest was among a big group of submissions with average score 1.3.

I announced that, within a few days, I would accept every paper with average score  $> 1.3$  and no negative score, and reject every paper with maximum score  $\leq 1$ . The PC started an active distributed online discussion, discrepancies magically disappeared, and things converged nicely. At the end of that surprisingly successful phase, we had a total of 105 accepted papers, and 80 undecided papers still to deal with.

One issue during that phase was the huge discrepancy between PC members' distributions of scores. In fact, a "1" for one PC member can be a "2" for another even though they really mean the same thing. It is normal that the distributions differ, maybe even by a lot since we all have different areas of specialty and some bid for what appear to be the potential best paper whereas others choose some evident bad papers in their bids to have a batch that's easier to deal with. But discrepancies are worrisome given that decision at that stage is based on average score. It is not clear how to address this. How about taking a small purely random sample of the papers and distributing it at random among the PC to check for inconsistent scoring? In our case, someone wrote a script to compute normalized scores and ranked the submissions by average normalized score. I compared that ordering to ours, and to my surprise and relief, there was great agreement - namely, if we consider submissions as belonging to 3 groups: A accept (top 100), B borderline (next 35 for normalized scoring, 80 still undecided for our scoring), C reject (the rest), then the ranking by average score and the ranking by average normalized score had switches between A and B and between B and C, but no switches between A and C. So - discrepancies in PC members' distributions are maybe not such a big deal after all!

(Here is Warren Schudy's suggestion about this: "To solve the problem of PC members grading differently, I suggest using a probabilistic model such as the following. Assume that

$$\text{Evaluation}(\text{paper } p, \text{ reviewer } r) = \text{papervalue}(p) + \text{reviewerbias}(r) + \text{noise}(p,r)$$

where  $\text{papervalue}(p)$  is uniformly distributed over  $(-4,4)$ ,  $\text{reviewerbias}(r)$  is normally distributed with mean 0 and standard deviation around 1, and each

noise(p,r) is normally distributed with mean 0 and standard deviation around 1.5. Use linear least squares to find most likely values for the unknown random variables  $\text{papervalue}(p)$  and  $\text{reviewerbias}(r)$  using the known evaluations.

You could also model the reviewers as having individual variance as well, but that would make it a non-linear least squares problem, hence harder to solve.

Rationale: the approach you used, linearly rescaling each reviewer to have common mean and variance has a potential problem that you noticed: some reviewers may have picked unusually good or bad papers to review. The above model only concludes a reviewer is biased to the extent that they rank papers higher or lower than the other reviewers did who reviewed the same papers.”)

During that phase, I did send email to all PC members with each person’s average score, with the normalized ranking of all papers, and with the number of papers headed for acceptance in each person’s batch. This was for information only, but enabled gentle pressure on people whom others felt were overly generous or overly stingy in their scoring.

We also were worried about identifying and comparing related submissions. I started doing this myself, but it was a daunting task (way too long for a single person, even full time). To centralize the information, I created a ”paper comparison” web page where each PC members could post a short list of papers in similar topics (For example: ”There are x submissions on algorithms for graph coloring. My opinion is the following partial ranking among those:  $x_1 > x_2 = x_3 > x_4$ ”). This was extremely useful and we should have started it earlier and developed it more.

Here were the topics of those clusters: Fixed-parameter tractability, bandit problems, secretary problems, applied topology and meshing, sublinear algorithms and property testing, streaming algorithms, Nash complexity, sponsored search auctions, clustering, classical scheduling, succinct data structures, regularity lemma papers, packet scheduling, MCMCs and random walks, graph coloring, linear algebra, higher dimensional geometry problems, lift-and-project, hashing data structures, embeddings, analysis of data structures, belief propagation. Clusters typically had 3 to 6 papers.

Since there was beginning to be some competition between papers, we started seeing occasional tension between PC members who have differing perspectives and different styles. I remember how quickly flame wars have developed in some past electronic PC meetings that I have attended. I strongly advise the PC chair to pick up their phone and talk to PC members to understand the issues at the first sign of tension. As it happened, this year I was lucky to have a pleasant PC, but I was always vigilant on that point.

#### **Phase 4: Accepting 30 among the 80 borderline papers**

During that phase, I looked at the clusters defined on the ”paper comparison” page. I assigned each cluster to a PC member and asked them to come up with a total ordering of those submissions, post it on the web page, and to check that our accept/reject decisions would be consistent with that ordering. (Again, at first I tried to do it myself but it was a daunting task. Just a single comparison

between two related papers that get similar scores can easily take one or two hours, particularly if those papers were not in your batch!)

This last phase took about 5 days. It was disorganized, time-consuming and except for the above checks, the outcome was pretty arbitrary. At that point we were under time pressure. The goal was to choose 30 papers from among the 80 remaining undecided papers. We took the following three steps.

1. First, each PC member “championed” a paper that then automatically went on the accept list. I asked PC members to use that with discretion: a few people who already had many accepted papers in their batch refrained from championing while a few people who had very few accepted papers in their batch championed two or even three papers. (In the end the number of accept papers in PC members’ batch varied from 7 to 21, a huge difference!)
2. Second, we now had 5 slots left. I asked people to nominate some of the remaining undecided papers and took a vote. We used sequential proportional approval voting. At that stage I think that it is fine to inject some of the PC chair’s personal priorities: I mentioned the couple of papers which, in the remaining group, had an all-student authorship, to make sure that people would pay attention to them... (Another PC chair could have highlighted submissions from minorities, or from unusual countries, or from a small institution that does not usually submit to SODA, or whatever.)

For the sequential proportional approval voting protocol, see near the bottom of <http://www.nationmaster.com/encyclopedia/Proportional-approval-voting> . In this system, voters give a set of “approved” candidates. Voters are allowed to approve of as many candidates as they wish. Candidates are elected one-by-one until all seats are filled. The first candidate elected is simply the one approved by the most voters. After that, voters are re-weighted and the new most-approved candidate is added to the pool of winners. A voter’s weight is  $1/(1+m)$ , where  $m$  is the number of already-elected candidates that voter approves of.

Sequential proportional approval voting has the advantage that if 2/3 of the people are algorithms people and prefer algorithms papers, and 1/3 are in discrete math and prefer discrete math papers, then the resulting accepts will also be 2/3 algorithms, 1/3 discrete math: thus minority opinions do get a voice. Also, at that stage I felt that every one had had a chance to push for whatever paper they felt really strongly about, so I wanted each vote to have equal weight, which is another property of sequential proportional approval voting.

The main disadvantage that I see is that PC members had difficulty understanding the voting rules. The rules are very very simple, but this was new to all of them, so they had to read the instructions, which is always the difficult part, isn’t it. Plus, at that stage some committee fatigue was beginning to show (including on my side, with some confusing messages.)

Also, I used a google spreadsheet to implement it, and that wouldn't scale very well.

3. Thirdly, we now had a list of 135 tentatively accepted papers. I asked people to propose swaps between accepted and rejected papers. The papers in the swapping pair did not need to be related; it just had to be a pair such that the PC member felt that the program would be better as a result of the swap. I viewed this as a chance to achieve some better balance between areas, correcting for under- and over-represented research areas. In the end we settled on three swaps.

We were short on time and on energy during that phase. It also coincided with the beginning of the semester for me so I wasn't able to give it as much attention as previously. Moreover, I think that having a three-step process is an overkill, and that we would have been more efficient with a simpler process. (I could have stopped at 130 accepted papers...) I would advise the future chair to plan something easier, and, if he or she is in academia, to make sure that decisions are finalized before the first day of the semester.

### **Conflicts of interest**

PC members mark the paper on which they have a conflict of interest. There is a problem with papers that have several PC members with a COI: such papers typically do not get evaluated by the PC members closest to their research area and most likely to get excited by the results, and so they may get evaluated more harshly, and more importantly, if they are still undecided in the last phase then they are very unlikely to be championed or voted for or proposed for a swap. I think that it is a real handicap for those papers but do not know how to address it.

### **Getting results out**

I wanted to send comments back to the authors at the same time as the notification. I asked PC members to proofread and clean up their reviews, checking privacy issues, urbanity of writing style, additional information coming from the comments and discussions. I assigned each paper to one PC member who was responsible to verify that the reviews were appropriate feedback to the authors and who was encouraged, when he felt that might be useful, to add a summary review recapping the evaluation of the PC in one or two sentences. There was a little bit of confusion in implementing this, as well as some tiredness, and we were short on time, so this was done imperfectly. Also, I forgot to tell PC members to check that all of their subreviewers had been duly entered into the system, so there are probably a couple who will have been forgotten and will not be acknowledged in the proceedings.

Nevertheless, because of the multiple rounds of refereeing and the excellent response rate to requests for reviews, overall I think the reviews are remarkably

detailed. They were also greatly helped by the presence of the full proofs in the appendices.

## **Post-meeting work**

Voting for best paper and best student paper, preparing a schedule, and putting together a booklet of abstracts of all accepted papers.

## **Easychair**

Easuchair worked well for the basics: getting the submissions in, the reviews entered with comments on each submission, and the results out. It also has a good automatic algorithms for creating the assignment of papers to PC members according to their bids. I think that it also offers facilities to create proceedings. So, there was no essential problem.

Beyond that, Easychair was rather poor. All the information was there but only accessible in a few pre-set ways, and it was very difficult to gather information in a different way; some things I thought should have been easy but turned out to be impossible. During the evaluation process, I had some ideas about how to proceed that were impossible to put into action because of the features of Easychair. Doing divide-and-conquer (having subcommittees evaluate subsets broadly filtered according to area), for example, was not possible. PC members also had some frustrations.

In short, Easychair was ok and did the job, it was reliable for all the essentials, but it was not comfortable and far from ideal.

Here are some comments from individual PC members.

## **Feedback from individual PC members**

1. I would love to see an EasyChair page that does not automatically close after a given time period.
2. I would really appreciate that the SODA Easychair URL be given in the signature of every email sent by the PC chair.
3. Easychair options should be expanded to include categories for: voting for best paper, voting for best student paper, voting on a subset by that SPAV scheme (with a link to the definition of SPAV), and swapping exercise, so that we don't have to work with "fake" submissions like 553 and 554.

Easychair: I would be happy to see a reminder: it is not a program you run by yourself; rather, it is centralized and run by some organization; this raises the question of the confidentiality of the informations. I'm glad we used EasyChair, but I feel this is an important drawback.

Score distribution guidelines: I think it's useful to give a guideline for the score distribution at the beginning, maybe along the lines: "The batches of two PC members may be of different quality. Yet, I expect that, on an average batch, the scores will be distributed as follows...". If you hadn't given any guideline at all, my guess is that the score discrepancy would have been even higher.

I have one very minor suggestion for improving the Easychair interface. Currently, when the review of a paper is displayed, there is only a very thin white line separating different sections of the review. Given the color scheme of the background, it's quite hard to see this thin line. In particular, when reading a review it's awfully hard to see where the public part of the review ends and the "PC only" section begins. It would be good to change the layout and/or color scheme so that it's easier to see the separation between sections of a review.

Seeing each PC member's distribution of scores, along with average and standard deviation, is a basic feature that should be added.