Program Committee Meetings Considered Harmful

Robbert van Renesse
Department of Computer Science, Cornell University

Abstract

This paper discusses various disadvantages of having a program committee meeting.

1 Introduction

Many computer systems conferences, and certainly the predominant ones, have a program committee (PC) that meets face-to-face to discuss which of the submitted papers should be accepted to the conference. They do so based on paper reviews generated before the meeting. Of the many aspects of the paper selection process for a conference that may be scrutinized, this position paper considers the need for a PC meeting itself. A PC meeting certainly serves one or more purposes. However, for the sake of discussion the paper shall take mostly a negative position, and discuss various disadvantages of having a face-to-face PC meeting. Some of these disadvantages also carry over to PC meetings that are held by phone conference.

At the time of writing this position paper, the author served on 28 program committees, (co-) chairing 5 of these.

2 Disadvantages of PC meetings

2.1 Unproductive meeting dynamics

Discussion in a meeting can be a great thing. PC members, removed from the many distractions of a normal work day, are forced to participate fully and give a paper and its reviews serious thought. Unfortunately, discussion is not without problems. Instead of expertise it is sometimes the discussion skills of a PC member that determine whether s/he can sell a valid or invalid point about a paper. Also, group discussion can easily take an overly negative turn, all but ruling out that the paper be accepted. Finally, these meetings almost always end in a time crunch in which many hasty decisions are made with insufficient discussion.

2.2 Skewed Representation

PC chairs often insist that attending the PC meeting is a requirement of serving on the PC. This requirement leads to a skewed representation of experts in the field. People that are geographically close to the PC meeting location are more likely to commit to attending the meeting, and thus agreeing to server on the PC, than people that live far away. People that decide to attend in any case and make the long trip may be so tired at the meeting that the quality of their contribution is below their usual standard.

2.3 Harm to the environment

Many of the PC members fly to a PC meeting. According to the British Royal Commission on Environmental Pollution the “fuel use and carbon dioxide emissions per passenger-kilometre for a fullyloaded [sic] cruising airliner are comparable to a passenger car carrying three or four people” (page 30, the Environmental Effects of Civil Aircraft in Flight, March 2007). Thus a single trip to a PC meeting may generate more carbon emissions than the PC member generates the whole year driving a car around. Besides carbon dioxide and other emissions, airplanes condense moisture in the upper atmosphere that also, on balance, trap heat and contribute to global warming.

2.4 Non-optimal Productivity

Traveling to a PC meeting, attending, and returning home, takes many of the PC members two to three days. It is useful to consider how such time may be used if there were no PC meeting.

Say that a PC member can review 3 papers per day. If we conservatively estimate that two days could be freed up for say, 10 people, and one day for 5 people, then these people could generate an additional 75 reviews if the PC meeting were canceled. It would be possible to add a last round of reviewing in order to improve the quality of ranking for contentious papers. Also, the additional reviews may improve the quality of those papers that get accepted.

Another possibility is to charge PC members with constructing a summary review for each paper. This would force a PC member to read all reviews for a paper, carefully construct a summary, and send it around to other members for approval. This process would likely lead to on-line discussions of contentious papers. Such discussions would be held concurrently and could be allocated more time than a discussion at a PC meeting. The resulting summary is intented to send a more consistent message to the authors than a collection of individual reviews.

2.5 Non-optimal Use of Money

For each person traveling to the PC meeting we could save enough money to send a student to the conference who otherwise could not attend. It is possible that sending more students to the conference is more worthwhile than the benefits a PC meeting can provide to the conference.

3 Alternative

So if there is no face-to-face PC meeting, what is a good alternative for making the final selection of papers? For completeness sake, I briefly offer a proposal here, realizing that better solutions may be possible. In this proposal, there would be a small group of two or three PC co-chairs that oversee the entire reviewing process. (More than one is necessary to deal with conflict-of-interest and so on.) After the reviewing cycle, the chairs use the review scores to generate an initial ranking. The chairs may fiddle with how the scores are used to generate the ranking, as it is difficult to generate a formula a priori that is guaranteed to produce a satisfying result.

PC meeting or not, chairs should carefully track contentious papers during the reviewing phase. Papers may be contentious for different reasons: technical quality, contributions to the field, quality of writing, and so on. In such cases chairs should invite additional reviews. In my opinion, reviews should not be shared between PC members during the reviewing cycle as much as possible, but in contentious cases PC chairs may deem it necessary to make the contents of the conflicting reviews available to respective members of the PC. Chairs are ultimately responsible for deciding if a paper is acceptable or not.

The PC chairs would then make the ordered list, along with notes on how the list was generated, as well as all reviews, available to the PC members and allow them to send rebuttals over a period of, say, two or three days, adjusting reviews and scoring as necessary. The PC chairs then select the highest ranked papers and put the result up for 2/3 majority vote (no further discussion allowed at this point). It is my expectation that in most cases the membership will adopt the proposal unanimously, and that voting down a proposal would be an extremely rare event.

4 Evaluation

Being in systems I feel compelled to add an evaluation. Obviously, such a evaluation has to make many assumptions and simplifications, and there is not much objective data available to support these. Nonetheless, it is useful and entertaining to consider what such an evaluation might look like.

Say there are N submissions, out of which n should be accepted. Unknown to the PC, each paper p has a *true value* $f(p)$. For simplicity, assume that no two papers have the same true value. We say that a paper p *is better than* another paper q if $f(p) > f(q)$ (*i.e.*, p is better suited to the conference than q). Papers are ranked based on their true value. Perhaps non-intuitively, we consider that the rank of the best paper is 0, while the rank of the worst paper is $N - 1$. We indicate by p_i the paper with rank i. For example, p_0 is the best paper.

It is convenient to use probabilities for true values, and thus we will interpret $f(p)$ as the probability that a paper is accepted. The probability of the best paper should normally be close to 1, indicating that it should be accepted with high probability, while the probability of the worst paper should be close to 0. The probabilities should be independent. The expected number of papers accepted should be n, and thus $\sum_p f(p) = n$. Also, the paper on the edge of being accepted or rejected should have a 50% probability of getting in, and thus $f(p_n) = 0.5$.

The PC members only know N and n. Each review r of p binds a *perceived value* $g(p,r)$ to a paper p, which is also a probability. We assume that with an increasing number of reviews the average or perceived values tends towards the true value of a paper. Formally, if $\mathcal{R}_p$ is the set of reviews of p, then

$$\lim_{|\mathcal{R}_p|\to\infty} \frac{\sum_{r\in\mathcal{R}_p} g(p,r)}{|\mathcal{R}_p|} = f(p) \qquad (1)$$

For this particular evaluation we will assume that the author of a review selects a score from five options: strong accept, weak accept, borderline, weak reject, strong reject. We model a knowledgeable reviewer as follows: For a paper p, the reviewer gets a weighted coin which he gets to flip four times. The coin is weighted by $f(p)$, unknown to the reviewer. With probability $f(p)$ the coin comes up heads. The reviewer assigns the following score based on the outcome of this experiment:

# heads	score	meaning
4	1	strong accept
3	0.75	weak accept
2	0.5	borderline
1	0.25	weak reject
0	0	strong reject

The reader may verify that this function satisfies Eq. 1 (the score is the average outcome of four binomial experiments), and thus an infinite number of reviewers would together converge towards the true value of a paper by averaging their scores.

We will compare two approaches for generating a conference program. One mimics a PC chair selection, while the other mimics a selection based on discussion at a PC meeting.

For the PC chair selected program, the chairs average the scores of the reviews of each paper, and use this to sort the papers. Note that the average scores are not necessarily unique and more than one paper may end up with the same score, so the order is partial. Using randomization among papers with the same score, the chairs converts this list into a totally ordered list and accept the first n papers.

In a PC meeting, the whiteboard is often divided into columns, with the leftmost column being definitely accept and the rightmost column being definitely reject. During the meeting, these columns are populated with paper identifiers. We mimic this process here, using five columns, one for each score. The PC members that reviewed a paper get to vote which column a paper goes into. PC members that assigned a score of 0.5 (borderline) sit out. A member that scored 1.0 gets two votes in favor, a member that scored 0.75 gets one vote in favor, a member that scored 0.25 gets one vote against, and a member that scored 0.0 gets two votes against. The votes are tallied and the paper is assigned an aggregate score, identifying one of the columns, as follows (approximately modeling PC meeting dynamics):

- 1.0 if nobody voted to reject; or else
- 0.0 if nobody voted to accept; or else
- 0.75 if the majority of votes are in favor of acceptance; or else
- 0.25 if the majority of votes are in favor of rejection; or else
- 0.5

As with the chair-selected option, this induces a partial ordering that is essentially reconciled through randomization when, at the end of the day, the exhausted PC members under direction of the exhausted PC chair start moving paper identifiers between the second and third columns, more-or-less randomly, in order to end up with n papers in the leftmost two columns.

The PC meeting model is hard to analyze mathematically, so I wrote a simulator. We use $N = 150$ (# papers submitted) and $n = 25$ (# papers accepted). In order to run the simulation the true value of the papers needs to be fixed, satisfying the various conditions above. We used the following simple function:

$$f(p_i) = \begin{cases} 1 - \frac{i}{2n} & \text{if } i < 2n \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

As desired, the true value of the best paper is $f(p_0) = 1$, the true value of the borderline paper $f(p_n) = 0.5$, while all papers ranked worse than $2n$ have a true value of 0. (Experimentation with other such functions yielded similar conclusions.)

As a metric for quality, I measured in each experiment both the worst paper that got accepted in an experiment, and the best paper that got rejected. Figure 1 shows the results of 100 experiments in which voting is used to select papers, and 100 experiments in which averaging (by the PC chairs) is used.

We can make the following observations. First, with the chosen settings six reviews is approximately what is needed before a paper can be confidently accepted—generating a few more reviews does not add much fidelity, and generating many more reviews is likely not an option.

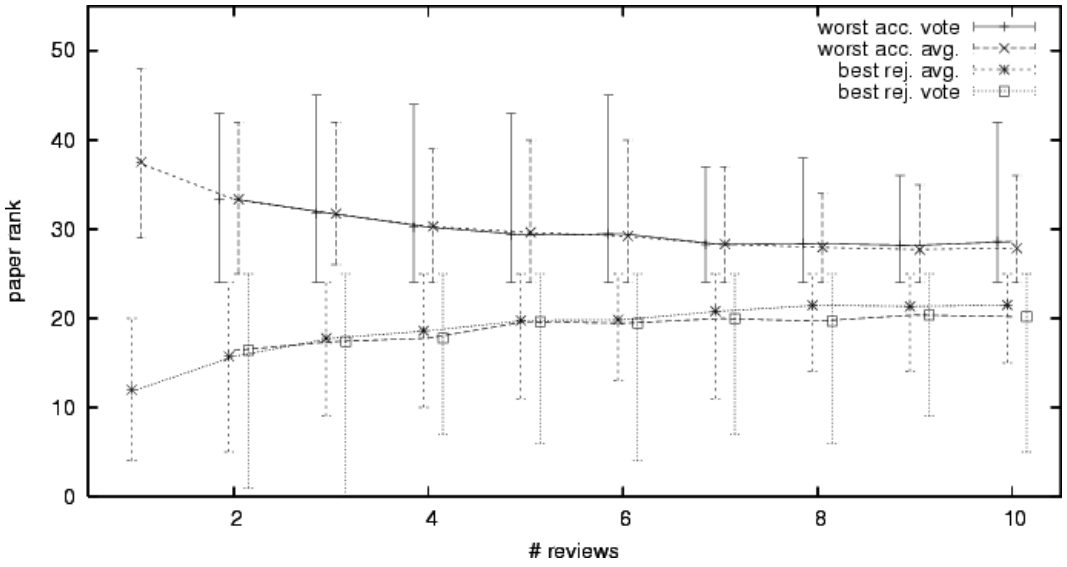

Figure 1: The top two curves show the worst paper accepted for scoring based on voting and on averaging, averaged over 100 experiments, as a function of the number of reviews available. The bottom two curves show the average best paper rejected. The error bars show the minimum and maximum among 100 experiments.

Second, on average the PC chair-based selection is slightly better than the PC meeting-based selection. However, more importantly, the voting process used in a PC meeting is more likely to make big mistakes than simple averaging, as judged from the error bars. For example, with 6 reviews the worst paper accepted by one of the experiments had a true ranking of 45 with the voting scheme, but only 40 for the averaging scheme. More alarmingly, the best paper that got rejected had a true ranking of only 4 for voting, and should definitely have been accepted. With averaging the best paper that got rejected was ranked 13. (Perhaps we should skip the number 13 for ranking.)

5 Conclusion

So how serious am I about all this? Enough so that I feel that the necessity of PC meetings should be considered and discussed at the workshop. Frankly, I would be nervous to be the first person to experiment with doing away with a PC meeting for a major conference, but does anyone strongly believe that a PC meeting improves the quality of the reviewing process sufficiently to offset the disadvantages I have presented?