Paper Rating vs. Paper Ranking

John R. Douceur
Microsoft Research
One Microsoft Way
Redmond, WA 98052-6399
425-706-8827
johndo@microsoft.com

ABSTRACT

Within the computer-science community, submitted conference papers are typically evaluated by means of rating, in two respects: First, individual reviewers are asked to provide their evaluations of papers by assigning a rating to each paper's overall quality. Second, program committees collectively rate each paper as being either worthy or unworthy of acceptance, according to the aggregate judgment of the committee members. This paper proposes an alternative approach to these two processes, based on rankings rather than ratings. It also presents experiences from employing rankings in PC discussions of a major CS conference.

Categories and Subject Descriptors

K.7 [**Computing Milieux**]: The Computing Profession.

General Terms

Human Factors.

Keywords

Program committee process, paper rating, paper ranking.

1. INTRODUCTION

When an academic journal receives a submission, the journal asks reviewers to assess the quality of the paper relative to an established quality bar for the journal. The bar is determined by the selection of papers that have appeared in previous volumes of the journal. Once the reviewers have judged a submission to be above bar, the manuscript will be published, either in the next issue or – in the event that a particularly large set of high-quality submissions is received in a brief interval – in a shortly following issue. If, over time, the backlog of accepted-but-not-yet-published papers continues to grow, the journal's editors may ask future reviewers to raise their standards for subsequent papers. However, submission quality need have no immediate effect on the acceptance bar.

By contrast, academic conferences typically have a target number of papers to accept, or at least a target range. Therefore, the quality bar is at least somewhat dependent on the quality of the submissions to that particular year's conference, rather than strictly by the conference's history. Conferences have no freedom to delay the effect of current submission quality on the acceptance bar. Decisions must be made about whether to accept or reject each submitted paper, in light of the space budget of the conference.

Since a reasonable decision about each paper cannot be made in isolation from decisions about other papers under consideration, two common practices in program selection are highly suspect:

- in the reviewing process, asking reviewers to render a judgment about whether a submitted paper meets the conference's quality bar
- in the PC meeting, making accept/reject decisions individually for each paper

Neither of these practices is reasonable given that the bar is not known a priori. Moreover, by employing these common practices, conference organizers incur two significant problems:

- conflating reviewers' standards of stringency and leniency with the reviewers' judgment of the merits and weaknesses of each paper
- psychologically entrenching early acceptance decisions based upon insufficient information

Herein, we propose that both of these problems could be avoided by employing rankings rather than ratings for both individual reviewer assessments and program-committee discussion.

1.1 Assumed goal

We are assuming that the goal of a program committee is to ensure that every accepted paper is of higher quality than every rejected paper.[1] Though ideal, this goal is absurd in at least three respects: First, no objective standard of quality exists, so the goal is not well formed. Second, even if we assume that the opinions of PC members serve as an acceptable metric for evaluating paper quality, there may be differences of opinion among members regarding judgments of quality. And third, it is not generally possible to eliminate cases in which two reviewers disagree about which of two papers should exclusively be in the accepted set [4].

Despite the impossibility of the idealized goal, there are efforts we could take toward minimizing violations of this goal. In this, we are aided by the fact that there is often a sizeable set of papers that could plausibly end up on either side of the acceptance decision.

1.2 Scope of proposal

The key issue this paper addresses is that, under the present system, the subjective judgment of a paper's quality is bound up with an additional subjective determination of whether that quality is above or below the bar for acceptance.

This paper does not attempt to address any of the underlying causes of reviewer subjectivity, such as:

- emphasis preference – Reviewers may differ on the importance of aspects of a paper, such as novelty, completeness, extent of evaluation, currency, conference topicality, and clarity.
- topic interest – Reviewers have differing areas of interest; what is boring to one is engrossing to another.
- qualifications – Reviewers differ in their technical ability to adequately assess papers on various topics.

[1] This may not be strictly true, insofar as PCs may wish for balance among multiple subject areas and may thus tolerate lower-quality papers on subjects with lesser representation. In such cases, the recommendations of this paper could be applied within each subject area.

- defaults – Reviewers differ in their judgments of what to do with a paper they don't fully understand; whereas some are inclined to be charitable, others tend to be ruthless.

These sources of subjectivity present challenging problems, but they are beyond the scope of this paper.

2. RATING PROCESSES AND PROBLEMS

Typically, reviewers are asked to rate each paper with ratings such as "strong accept", "weak accept", "weak reject", and "strong reject". Then, in the PC meeting, the committee collectively assigns a rating to each paper, based largely on the ratings provided by individual reviewers. The rating categories are similar, although they are characterized differently, such as "accept", "accept if room", "accept as poster", and "reject". Such ratings cause problems in both phases of the review process.

2.1 Rating-based reviews

As described above in §1.2, reviewers may have many axes of difference in the way they evaluate papers. But even if two reviewers happen to have the exact same emphasis preferences, topic interests, qualifications, and defaults, they might give drastically different ratings to a paper, because of differences in how stringent or lenient they tend to be. In practice, this means that a paper reviewed by a stringent reviewer will receive a less favorable rating than a paper of comparable quality reviewed by a lenient reviewer (cf., §5.3).

Some program chairs have attempted to neutralize these tendencies by tagging each rating with a percentile range, such as "strong accept (top 10%)", "weak accept (top 25% but not top 10%)", etc. However, anecdotal evidence suggests that many reviewers discard these prescriptions in favor of the direct interpretations of each rating.

It might be possible to enforce a curve on ratings with sophisticated conference-management software that evaluates how well a reviewer's ratings fit the curve intended by the program chair. Imagine a dialog box that tells a reviewer, "You have strongly accepted 30% of the papers you reviewed. The overall acceptance rate for this conference will be approximately 12%. For randomly assigned papers, there is less than a 2% probability that the selection of your papers is skewed enough to warrant this discrepancy. Are you confident of your recommendations?"

If we were to take such an approach, we would have to answer the question of what to do when a reviewer insists on submitting off-curve ratings. If the software allows this to happen, then willful reviewers will easily circumvent this hypothetical safeguard. But if the software does not allow off-curve ratings to be submitted, we risk annoying reviewers, who might then decide not to submit any review because they feel themselves over-constrained.

2.2 Rating-driven PC meetings

The focus of a PC meeting (whether electronic or in-person) is to judge each submitted paper as either above or below the bar for acceptance. However, conferences typically have both a limit on the number of accepted papers and a (not necessarily official) quota to fill. For the count of accepted papers to fall within this target range, the quality bar must be set according to the quality distribution of submitted papers. However, this distribution is unknown until the committee has had the opportunity to discuss a significant fraction of papers.

This presents a Catch 21.[2] One cannot discuss whether to accept a paper without first determining where the bar is, but one cannot determine where the bar is without first discussing a representative sample of papers. However, this is exactly what is called for by a process of sequential discussions on the acceptability of each paper in turn.

This situation gives rise to a dynamic that is likely to be familiar to anyone who has ever served on a PC: Early in the PC meeting, members maintain a very high standard for papers, rejecting good ones for fairly minor reasons. Later on, as it becomes clear that the quota will not be met, members start becoming looser about what they consider acceptable. Eventually, someone notes that the committee is accepting papers that are notably weaker than papers it had earlier rejected. This observation prompts earlier rejections to be revisited in light of the revised bar.

However, strong empirical evidence from psychology [3] shows that once a person renders a judgment on the desirability of an item, his opinion becomes reinforced, which strongly biases future judgments about the same item. Thus, even though a prematurely rejected paper may be brought up for reconsideration, it will generally not receive as much leniency as a paper that had not been tarnished with an early negative judgment. As Triesthof famously quipped, "You never get a second chance to make a first impression."

Note that this problem occurs irrespective of the order in which papers are discussed. Therefore, it cannot be fixed by modifications to the paper-discussion order, such as discussing high-variance papers first.

3. PROPOSAL: RANKING

We propose that the problems enumerated in §2 could be avoided by basing reviews and PC discussion on rankings instead of ratings. Although rankings could be applied to reviews without applying them to PC meetings, or vice versa, the full benefits of ranking are only obtained when implemented together.

3.1 Ranking-based reviews

For many years, college admission boards have faced the problem of varying stringency among high schools in judgments of students' grades. The widely adopted solution to this problem is for colleges to judge students by their class rank instead of by their GPA. In fact, the recent trend among some high schools of not reporting class rank has led college boards to complain that this reduces their objective information on students' academic performance [5].

Class rank is immune to grade inflation. Analogously, a reviewer's ranking of a set of papers is immune to the reviewer's standards for acceptance. Papers reviewed by a stringent reviewer will not suffer unfairly in comparison to those reviewed by a lenient reviewer.

Some PC chairs have attempted to circumvent differing standards by normalizing reviewers' ratings. However, if the number of rating choices is too small, they may contain too little information to discern the reviewer's relative opinion of papers. On the other

[2] Almost, but not quite, a Catch 22.

hand, if the number of rating choices is very large, this is really just a poor way of collecting rankings, since psychological evidence suggests that experts are better at rendering comparative judgments than absolute ones [6, 7]. In addition, if ratings are explicitly bound up with decision intentions (such as “accept”, “weak reject”, and so forth), this may still incur the problems described in §2 above.

It may not be necessary to restrict reviewers to a total ordering. Perhaps a reviewer could be allowed to indicate that two or more papers are of equal rank, which would provide some additional reviewer flexibility. More generally, reviewers could specify the size of the quality difference between adjacent papers. This could be a simple set of choices such as “marginal”, “significant”, or “dramatic”. Empirical evaluation could help determine whether this freedom would tend to be abused by reviewers who hate (or love) every paper they read.

3.2 Ranking-driven PC meetings

A ranking-driven PC meeting can cleanly separate two processes that are tightly coupled – and conflated – in a rating-driven PC meeting: the judgment of paper quality (relative to other papers) and the determination of where to set the bar for acceptance.

Prior to the meeting, the chairs establish a straw-man global ranking. A simple method for producing such a ranking is, for each paper, convert each reviewer’s rank to a numerical score, and average the scores of all reviewers. Then, sort the papers according to their average scores. It remains to be seen what function would be best for the rank-to-score conversion, but it seems likely that the function should be nonlinear: There is probably more quality difference between papers ranked 1 and 2 than there is between papers ranked 10 and 11. The function should perhaps also account for the reviewer’s self-assessed confidence rating. A similar procedure is currently used in many PC meetings for determining a rough ranking for paper discussion order; however, since discussion is focused on individual accept/reject decisions, rather than changes to the rank order, the problems enumerated in §2.2 remain.

The PC meeting then proceeds in two phases. In the first phase, the committee debates the relative ranks of papers. A typical instigating comment might be, “I thought paper 384 was far better than paper 721, but it is ranked three slots lower.” For pairs of papers which no single reviewer has reviewed, it is still possible to have an intelligent discussion among the reviewers of each paper: “Paper 219 has a really solid evaluation. Would the reviewers of the papers ranked above it please comment on the quality of the evaluation sections?”

Since the lower and upper bounds of the acceptable paper count are usually established beforehand, it is straightforward to avoid debating the relative ranks of any set of papers that are all well above or well below the cutoff range. The rankings of such papers, relative to each other, will not affect their ultimate acceptance or rejection. Papers whose ranks are within or near the cutoff range are the best candidates for intense debate, and so should be discussed first.

Divergent opinions could give rise to ordering cycles, but such cycles highlight papers that are important to thoroughly debate and/or to solicit additional reviews.

In the second phase, the committee establishes the cutoff point for paper acceptances. This decision could be based on a number of factors:

- The quality of papers within the target range might suggest that the bar should gravitate toward the upper or lower end of the range.
- The PC might be inclined to be generally lenient, or to be generally stringent.
- If the papers within the target range have some particularly desirable property, such as a fresh topic, the bar should perhaps go below them.
- A large gap in the assessed quality of adjacent papers may indicate a good cutoff point.
- If a short-distance cycle remains in the final ranking, the cutoff point should be positioned so that no cycle spans the cutoff, if possible.

The main benefit of a ranking-based discussion is to avoid prejudicing the committee’s judgments, but it has another benefit as well. A particular PC member may be especially dominant or persuasive, and in a rating-based system, he can intimidate or cajole the few other reviewers of a paper into accepting his view. However, in a ranking-based system, if any member wishes to significantly raise or lower the rank of a paper, he will have to argue against the reviewers of many other papers. This will decrease the unwarranted influence of fearsome or charismatic members.

4. CHALLENGES

A basic implementation challenge is that existing conference management software is designed to operate on ratings. Modifying this software to operate on rankings could require substantial reworking.

The ranking-driven PC meeting begins with a straw-man global ranking. If this ranking is poorly established, it may result in a very inefficient meeting. In the general case, it is impossible to convert a set of individual rankings into a global ranking [2]. However, we have no need for an optimal – or even consistent – global ranking. The initial ranking need only be good enough to avoid wasting time comparing papers of wildly different quality.

It is possible that there may be insufficient information for the PC to determine the relative ranking of certain pairs of papers. We suspect that, in practice, this will not be a common occurrence, because the transitivity of partial ordering will inform the relative ranks of most papers unless their levels of judged quality are very similar. If there are two papers that seem to settle near each other as the ranking is adjusted, and if no single reviewer is familiar with both papers, and if the papers lie near a likely cutoff point for acceptance, then this is a strong indication that a reviewer of each paper should be appointed to read the other paper and make a solid comparison.

Perhaps the biggest challenge with the ranking-driven PC meeting is that it complicates anonymous reviewing for papers submitted by PC members. In rating-driven meetings, a single paper is discussed at a time, so any members with conflicts-of-interest can step out of the room. In a ranking-driven meeting, multiple papers will be in discussion concurrently, since their relative merits are under consideration. Although a discussion of the merits and demerits of anonymous reviewing is beyond the scope of this paper, a simple way of addressing many conflicts is for PC members not to leave the room unless they are actually authors of the paper under discussion. We are aware of at least one top-tier conference's PC meeting that required conflicted non-authors merely to refrain from discussion, rather than to leave the room. Another alternative is to conduct PC meetings online (cf., §5), which allows the chair significant freedom in determining which PC members should be involved in which discussions.

5. EXPERIENCES

The author recently had an opportunity to test paper ranking in the context of managing PC reviews and discussions for a major computer-science conference.

5.1 Context

The author was asked to serve as an organizer for ICDCS 2009, in the capacity of vice chair in charge of the OS & Middleware track. (Hereafter, the author will be referred to as "the chair".)

This track had 67 submissions. The track's 20 PC members were instructed to select approximately 6 papers for acceptance to the conference, and to select approximately 6 more papers for further consideration by the conference's organizing committee.

The conference – and thus all tracks thereof – used ratings as the official form of reviewer input, on a scale of 1–5.

Reviewers diverged widely in their overall ratings of submissions. The ratings of one reviewer had a mean value of 1.9, whereas another reviewer had a mean rating of 3.4. As described in §2.1, ratings make it unclear whether the former reviewer had received a dramatically worse batch of papers than the latter, or whether the former reviewer was merely stricter in assessing paper quality. (§5.3 presents evidence that these differences in mean ratings are entirely due to differences in reviewer's standards.)

Reviewers also varied widely in how much of the rating scale they used. Although one reviewer's ratings had a standard deviation of 1.5, another's had a standard deviation of 0.7. Such limited use of the range provided limited information about the latter reviewer's relative opinion of papers.

It was thus difficult to discard a large fraction of the papers based solely on reviewer ratings, so over half of all papers (35 out of 67) were discussed. The PC meeting was conducted online over a span of 10 days, and each paper was discussed only among that paper's reviewers.

For 12 of the papers, discussions resolved to a clear consensus for rejection. For 7 more papers, the chair imposed a conclusion of rejection, based on the tenor of the discussion relative to other discussions in the chair's global view.

For two of the remaining papers, discussions reached a consensus for acceptance. However, because the discussions were online and only among each paper's reviewers, this conclusion was reached in the absence of a global view of paper quality, so the soundness of the decisions was slightly uncertain. (See §5.3 for a discussion of the fate of these two papers). For the remaining 14 papers, discussions did not reach consensus.

5.2 Use of rankings

At this point, the chair opted to employ rankings for the 16 papers that had not been decidedly rejected. The chair emailed each reviewer with a request to provide relative rankings for the subset of papers with which each reviewer was familiar. Most reviewers replied with total orderings, such as: E → O → L → N → P, where each letter indicates a particular submission and each arrow indicates a "better than" relationship. Three reviewers replied with partial orderings, such as: A → (B, D) → C. The chair interpreted this as: A → B → C and A → D → C.

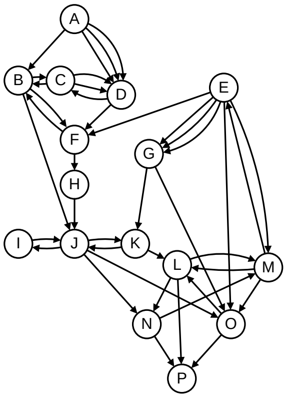

Figure 1: Multi-graph of reviewer rankings

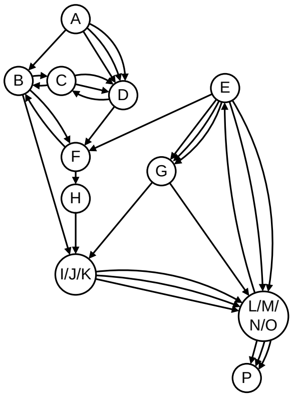

Figure 2: Multi-graph with several nodes merged

Figure 1 shows a multi-graph that aggregates all reviewers' orderings. Each directed edge corresponds to a "better than" relationship expressed by one reviewer. Thus, the count of arrows between pairs of nodes indicates the degree of consensus about the relative ranking of the nodes. Fortunately, the resulting graph is connected; if it had not been, the chair would have had to solicit additional reviews to ensure that all papers were comparable.

The nodes in Figure 1 are sorted to provide a generally downward flow. However, the multi-graph contains several cycles, which reveal some significant disagreement among reviewers' rankings.

To provide some clarity, the chair merged several papers into two clusters. According to the collective opinion of two reviewers, papers I and J are each better than the other, as are papers J and K, so all three were merged into a single node. Papers L, M, N, and O have an even more complicated relationship: L and M are each better than the other, and there is a cycle L → N → M → O → L, so all four were merged. In both cases, all directed edges not internal to the cluster were preserved across the transformation.

Figure 2 shows the result. This is somewhat clearer, but it is still difficult to work with, because it includes two complex cycles: The first involves nodes E, F, G, H, I/J/K, and L/M/N/O. The second involves nodes B, C, D, and F. The chair dealt with these as follows:

In the first cycle, the flow capacity from E to L/M/N/O is 5, which greatly outweighs the flow capacity of 1 in the reverse direction. Thus, it was decided to break the cycle by striking the edge from L/M/N/O to E. Absent this edge, papers L, M, N, O, and P are clearly ranked at the bottom, so they were rejected.

The second cycle lacks any obvious asymmetries in flow capacity that would justify striking a particular edge. However, if all nodes in this cycle were to share the same fate, then the relative rankings would be irrelevant. Thus, it was decided to select papers A, B, C, D, E, and F as the top 6 papers for acceptance.

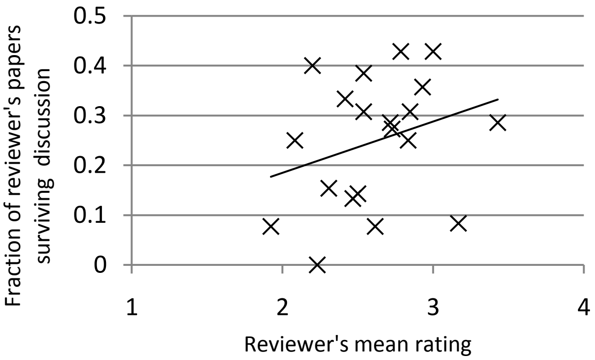

**Figure 3: Reviewers who gave high ratings did not get significantly more papers through discussions;
The linear fit has very weak correlation (R^2 = 0.088)**

This left the five papers G, H, I, J, and K for further consideration by the conference's organizing committee. Because the above process had established a partial order, the committee had some guidance about the comparative strengths of these papers. (At the committee meeting, all five papers were accepted.)

5.3 Analysis

Our data can somewhat address the question of whether reviewers who gave low ratings did so because they received poor papers or because they were merely stricter. Figure 3 is a scatter plot that shows one data point per reviewer. The reviewer's mean rating is indicated by the X axis, and the *survival fraction* of the reviewer's papers is indicated by the Y axis. The term "survival fraction" refers to the fraction of the reviewer's papers that survived the discussion process prior to the ranking step. These discussions were quite extensive, so survival is a good proxy for the overall quality of a paper, prior to imposing an arbitrary limit on the count of acceptable papers. Figure 3 shows that there is virtually no correlation (R^2 = 0.088) between a reviewer's mean rating and the quality of that reviewer's papers. This is consistent with Tom Anderson's observation that the variability between reviewers is often the dominant factor in a paper's acceptance decision [1].

As mentioned in §5.1, reviewer discussions concluded to accept two papers: E and I. Although the ranking process accepted paper E, it selected paper I for further consideration. The conclusions about paper I may have differed because the ranking process produced a global comparison, whereas the reviewer discussion lacked a global perspective: None of the reviewers of paper I reviewed a paper that was selected for acceptance.

One notable weakness in this use of rankings is that critical parts of the ordering are dependent on a single reviewer's comparison. In view of the significant number of conflicting opinions evident in the ranking graph, it seems likely that if we were to substitute even one reviewer with another, the partial order might change drastically. This problem could be partly addressed by the more open PC-meeting process described in §3.2.

During the PC's meta-discussion about the multi-graph, the chair anonymized the nodes. This was primarily done to avoid problems with conflicts of interest. However, it also had the effect of forcing PC members to focus on the process without getting distracted by their opinions of the papers. In particular, recall that a "better than" edge of one PC member was struck from the multi-graph because it was outweighed by countervailing edges from other members. The member whose edge was struck did not raise any objections; however, neither this nor any other member was aware of whose edge was being struck.

5.4 Conclusions

We can draw several positive conclusions from this experience:

- Discussions did not reach consensus on accept/reject decisions, yet ranking revealed a rough consensus about the comparative value of papers.
- Although ranking did not establish a strict order, it was sufficient to facilitate making decisions.
- Ranking successfully separated reviewers' acceptance standards from their judgments of relative paper quality.
- Ranking successfully separated the process of agreeing on relative merit from the process of setting the bar.
- Every accepted paper is better than every rejected paper, in the opinion of 19 out of 20 reviewers.

ACKNOWLEDGEMENTS

The author extends thanks to the PC members of WOWCS 2008, to John Chandy, to Jeff Mogul, and to Tom Anderson for their helpful comments on various versions of this paper. The author owes a deep debt of gratitude to Jeff Mogul for the inspiration and effort that went into creating WOWCS, and to the other WOWCS 2008 authors for the inspiring ideas in their papers. The author also thanks the various program chairs who have been daring enough to invite the author onto their program committees, and to the organizers who have been even more daring to have the author serve as chair. The author is delighted by Tom Anderson's choice to employ rankings for SOSP 2009; hopefully, this experiment will prove fruitful. The author owes very special thanks to the PC members of the OS & Middleware track of ICDCS 2009 for serving as guinea pigs in a live trial of paper rankings, and to the ICDCS 2009 program chair, Xiaodong Zhang, for allowing the author sufficient freedom to employ rankings in discussions.

REFERENCES

[1] T. Anderson. "Conference Reviewing Considered Harmful," *Operating Systems Review* 43(2), April 2009.

[2] K. J. Arrow. "A difficulty in the concept of social welfare," *Journal of Political Econom*, 58(4), August 1950, pp. 328–346.

[3] J. Brehm. "Post-decision changes in desirability of alternatives," *Journal of Abnormal and Social Psychology* 52, 1956, pp. 384–389.

[4] J. Duggan and T. Schwartz. "Strategic manipulability without resoluteness or shared beliefs: Gibbard-Satterthwaite generalized," *Social Choice and Welfare* 17, 2000, pp. 85–93.

[5] A. Finder, "Schools avoid class ranking, vexing colleges," *New York Times*, March 8, 2006.

[6] C. Spetzler and C.-A. Stäel von Hostein, "Probability encoding in decision analysis," *Management Science*, 22, 340–358, 1975.

[7] H. Wang, D. H. Dash, and M. J. Druzdzel. "A method for evaluating elicitation schemes for probabilistic models," *IEEE Trans. on Systems, Man, and Cybernetics – Part B: Cybernetics*, 32(1):38–43, February, 2002.