Towards More Constructive Reviewing of SIGCOMM Papers

Jeffrey C. Mogul
JeffMogul@acm.org

ABSTRACT

Many people in CS in general, and SIGCOMM in particular, have expressed concerns about an increasingly "hypercritical" approach to reviewing, which can block or discourage the publication of innovative research. The SIGCOMM Technical Steering Committee (TSC) has been addressing this issue, with the goal of encouraging cultural change without undermining the integrity of peer review. Based on my experience as an author, PC member, TSC member, and occasional PC chair, I examine possible causes for hypercritical reviewing, and offer some advice for PC chairs, reviewers, and authors. My focus is on improving existing publication cultures and peer review processes, rather than on proposing radical changes.

1. INTRODUCTION

Scientific publication benefits from peer review, which not only prevents the dissemination of "wrong" results,[1] but also encourages authors to write better papers. The prospect of having our work judged by unknown experts, and in competition with other papers, should encourage us to build the strongest possible written case.

The progress of any scientific field depends on both incremental work ("normal science," in Kuhn's words [5]) and highly innovative work (though this does not always go as far as a true paradigm shift).

Unfortunately, reviewers do not always tolerate either normal, incremental work, finding it too boring to publish, or highly innovative work, finding it too risky to publish. In the worst case, this can turn into a sort of subconscious game, which reviewers win by finding reasons to reject a paper.

Of course, not all papers should be accepted. Conferences and journals have limited time slots and page budgets, and many papers just are not ready to publish. But the vast majority of CS papers[2] are not actually wrong in some fundamental way. Instead, most papers are wrong in many small ways, giving reviewers many excuses to be critical.

In too many cases, CS reviewing has become hypercritical. Jeffrey Naughton's keynote at ICDE 2010 [8] has been widely cited as an articulation of this problem. Naughton points out:

- "Reviewers hate EVERYTHING!" – Even for conferences with hundreds of submissions, the PC finds almost no papers that all reviewers like.
- Researchers rely on reviews for training, on what papers to write, and how to evaluate others – "Receiving dysfunctional reviews begets writing dysfunctional reviews."
- Given the belief that CS reviewing is more hypercritical than in other fields, "Funding agencies believe us when we say we suck."

Naughton, in his keynote, attempted to diagnose the problem and propose some solutions. Bertrand Meyer contributed a CACM blog posting [6] reinforcing Naughton's messages, pointing out that for NSF proposal evaluations, "the average grade of computer science projects is one full point lower than the average for other disciplines." Brighten Godfrey, in a blog post [4], made additional suggestions. Many others in our field have contributed to the discussion.

In this article, I offer some suggestions based on my personal experience. Although I have done some journal reviewing (and a little journal-editing), most of my experience has been with conferences: as an author, reviewer, chair, and steering committee member. For better or worse, conferences are the dominant form of publication for SIGCOMM and the other areas I work in. Therefore, I will focus on ways to reduce hypercriticality in conference reviewing.

This paper is a slightly revised version of my position paper for the Dagstuhl Perspectives Workshop on Publication Culture in Computing Research[3], which took place in November, 2012.

2. ROOTS OF THE PROBLEM

Inevitably, as scientists and engineers, our inclination is to uncover the roots of the problem at hand. People have offered lots of explanations for hypercriticality in reviewing, including

- an overload of submissions (especially those that keep getting rejected and resubmitted without substantial improvement);
- a tendency towards "science envy," given that computer science's status as an actual science has occasionally been questioned, in spite of efforts by Peter Denning [10] and others. This envy may lead to the hope that perhaps, if we only insisted on more rigor, real scientists would take us seriously.
- the self-reinforcing role modelling that Jeffrey Naughton pointed out: young CS authors who receive nasty reviews may internalize nastiness as the norm. Those who succeed in spite of nasty reviews become the next generation of reviewers, and may not recognize the harm that hypercriticality

[1] we hope; see Fang *et al.* [3]

[2] By "CS," do I mean Computer Science or Computer Systems? My expertise is in computer systems, and I believe that systems researchers often behave differently, but I will let the reader decide how to interpret "CS" in this paper.

[3] `http://www.dagstuhl.de/12452`

causes to other young researchers "who may be questioning whether they belong in the field," as Anne Condon has suggested [2].

I have heard many senior committee members and experienced PC chairs express a concern that junior reviewers have, as a broad generalization, more bias towards harshness. This might be a real effect, which diminishes with experience. Or, perhaps, PC chairs do tend to learn to avoid recruiting overharsh reviewers, once they have shown their tendencies.

However, as scientists we must also admit that we will probably never have the time and energy to collect sufficient data to fully resolve the question of root causes. The speculation is useful, because it leads to proposals for solutions, but we ultimately can only test the solutions, not the hypotheses about causes.

3. EVOLUTION OR REVOLUTION?

We seem to be faced with a buggy culture in CS. We would like to fix this culture, but changing a culture gradually can take a long time.

We could "solve" the problem of hypercritical reviewing via a number of radical approaches. For example, we could eliminate peer review altogether, and use a crowd-sourced approach to establishing the validity of CS publications. As I wrote, very few CS papers are actually wrong, and (unlike, say, clinical studies of medication effects), the consequences of disseminating a "wrong" CS paper are probably not dire.

My sense is that, in spite of the apparent attractiveness of some proposals for radical change in CS reviewing, the collateral damage could be quite high (for example, to the tenure prospects of a generation), especially without a unified governing body to control how the change is effected.

Lacking a benevolent dictator, we may need to accept that enlightened attitudes arrive slowly. But at least we can start trying to guide the reviewing culture in the direction we want, through good leadership and role models. (As Gandhi never actually said, "be the change you wish to see in the world."[7])

4. PROCESS ADVICE

As I have thought about what to do about negativity in CS reviewing, and especially in my occasional role as a PC chair or co-chair, I have tried out a number of processes designed to encourage more useful reviewing. (I am indebted especially to Greg Minshall, John Byers, and Martin Arlitt for their contributions to my thinking, although many other people have helped.)

4.1 Process points for PC chairs

PC chairs have the primary responsibility for controlling the review process, and so most of my advice is for chairs.

Choosing a PC: The solution to the hypercriticality problem (and many others) starts with choosing PC members. Chairs must balance many criteria, including topic expertise, geographic/gender/organizational diversity, etc. There is often some pressure, especially for a conference that is not yet established as top-tier, to choose widely respected people; many authors base their evaluation of a CFP on whether they know and respect the PC members.

But chairs should not ignore the fact that respect usually comes from good research results, and not always from good reviewing (especially when reviews are anonymous). When considering reviewers who have a reputation for being hypercritical, chairs should think about whether the reviewers' prestige or expertise compensates for their potential negativity. If such reviewers are added to a PC, the chairs bear some responsibility for keeping a careful eye on their reviews.

PC chairs must also consider the size of their PCs. An undersized PC puts a large workload on the reviewers (and may discourage some good but busy people). However, small PCs can encourage better behavior, by building a better sense of shared purpose; by exposing each reviewer to a wider set of submissions; and especially by engaging a larger subset of the PC in the final decisions on papers, which prevents one naysayer from dominating. My sense is that oversized PCs, on the other hand, lead to poor reviewing. (On the other other hand, if the PC is so small that their workload becomes unsustainable, review quality will decline.)

Some conferences have experimented with a "heavy/light" structure, where only a subset of the PC attends the meeting and makes the final decisions. While this approach can spread the workload and broaden the expertise of the PC, it might also lead light-PC reviewers to be over-critical (or perhaps over-positive) in their reviews, knowing that they will not be at the meeting to defend them.

Managing the review process: The next task of the PC chair(s) is to guide the reviewers to do the best possible reviews. There are several ways to guide reviewers away from hypercriticality:

- PC chairs need to set the right expectations. Telling reviewers simply "here are your assigned papers; read them and return your scores by the deadline" is not enough guidance. Tell the reviewers what kinds of papers you hope the conference will accept, and remind them that excessive negativity helps nobody.

- The review form can be designed to elicit positive comments, not just negative ones. For example, it has become common to provide separate text boxes for (1) a concise summary of the paper and the major reasons for the scores, (2) a short summary of the strengths of the paper, and (3) a short summary of the paper's weaknesses, in addition to numeric scores and detailed comments. Asking for these three summaries can guide reviewers to be explicit about a paper's good aspects, as well as its problems.

  Some people prefer review forms with just a few score fields (such as "overall merit" and "reviewer confidence"). It may be true that no other fields contribute to the final decisions. However, by asking reviewers for separate scores on other aspects (such as technical quality, novelty, and presentation quality), the PC chairs can guide them to think explicitly about these aspects, rather than simply focussing on whether they want to accept the paper or not. (Also, I suggest avoiding a "reviewer expertise" score in favor of "reviewer confidence" – non-expert reviewers can be quite confident in their reviews, and vice versa.)

  If scores are constrained to be integers, they should provide enough dynamic range. Five levels is not enough! since most reviewers seem to avoid handing out the highest possible score, and only really bad papers get the lowest score, reviewers often pick an arbitrary integer score in the middle, then encode their actual feelings in "comments for the PC," which sometimes are overlooked in the ranking process.

- The PC chairs should be familiar with the papers, probably to the extent of reading each submission (but not necessarily writing full reviews). If the conference has co-chairs, they can split this task. This is a lot of work, but it allows the chairs to prevent papers from being rejected too hastily.

  It also helps chairs when they are trying to make judicious decisions about which papers to promote to a second (or

third) round of reviewing, to decide which papers might need an external-expert review, and when they are trying to lead the PC meeting towards making decisions on the final set of papers.

- More reviewers per paper generally improve the chances that an innovative or unusual paper will find at least one advocate on the PC. Since reviewer time is precious, and large PCs are bad, this leads to a multi-round review process. Using multiple review rounds allows the chairs to assign a larger number of reviews to those papers that are plausibly acceptable, instead of spreading the reviews equally across all submissions.

  In the first round, each paper should get three reviews, if possible, so that a decision against promoting a paper to the second round can be made with sufficient evidence (but PC chairs should never look at just the scores when making this decision). One can then assign two or three additional reviews (or more for contentious papers), without forcing the PC to write five or six reviews of every submission.

- The PC chairs should encourage online discussions among reviewers, at the end of each round ("should this paper be promoted?") and prior to the PC meeting. This allows contentious issues to be aired, and perhaps resolved, with the luxury of more time than is available during the meeting. HotCRP, for example, provides a nice facility for this kind of discussion.

- Despite everyone's best intentions, sometimes reviewers do slip up and put inappropriate remarks into their reviews. Or, they write comments such as "the authors are obviously unfamiliar with the wide literature on the subject" without providing any specifics (authors really hate this, and for good reason). PC chairs should skim the reviews at the end of each round, not only because it helps them decide which papers to promote, but also to give reviewers suggestions about improving tone and content. They should also check that "comments to the PC" are consistent with the numeric scores, if the scores put the paper at risk.

Managing the PC meeting: PC meetings cost a lot of time, and face-to-face meetings cost actual money. But well-managed PC meetings can create social mechanisms that reduce hypercriticality, especially if PC members are encouraged to call it out, in a friendly way, when they see it. This may be one of main benefits of a face-to-face meeting, since tricky interactions are usually easier when people can see each other's faces, or can hold hallway conversations.

PC chairs must manage the meeting for multiple goals, aside from the obvious one of picking a good set of papers. Meetings that get out of control can lead to a set of decisions that work against papers that start out with inappropriately low review scores, or with hypercritical reviews.

Several methods can produce better results:

- Time management: we hold meetings so that we can discuss a much larger number of papers than can be accepted; including more papers in the discussion reduces the effects of noisy scores, but creates time pressure. When some discussions drag on too long, other papers get short-changed. PC chairs should ruthlessly control the time allotted for discussions, stamping out "me too" comments and rat-holes.

  Some (perhaps most) papers will need a second round of discussion; the time budget should allow for that.

- "Identify the champion"[9]: Someone should speak in favor of each paper, and should be encouraged to defend it, with evidence, against unwarranted criticism.

  Ideally, each paper to be discussed at the meeting has been assigned to a "discussion lead," who (*before* the meeting) prepares a very brief summary of the pros and cons, so that the discussion starts out without a lot of confusion, and people who have nothing to add do not waste time by adding it.

  Assigning the most positive reviewer as discussion lead gives this person a chance to look over the reviews before the meeting, and to point out obvious flaws in the negative reviews.

- Defer negative decisions: while it is crucial to cut off inconclusive discussions, this does not mean that PC chairs should force a decision too quickly; when there is controversy, this usually leads to rejection. My experience is that a meeting which defers all rejections until after each paper has been discussed will make better decisions.

  When a discussion ends without an "accept" decision, the PC chairs should record the main argument against the paper (for example, bucketing such papers as "boring," "technically flawed," "immature," "risky," "accept-if-room," etc.). This allows subsequent discussions to restart with some context, and it allows the PC chairs to prioritize the discussions if time is short.

  Chairs should also encourage PC members to prepare for follow-up discussions (perhaps by re-reading the other reviews during a break, or through hallway discussions).

- Focus on understanding risk: Given that few papers that make it to the PC meeting are provably "wrong," most arguments boil down to the question of what kind of risk the PC is taking by accepting the paper.

  Since we want to encourage innovative work, PCs should not be totally risk-averse. Chairs should ask PC member, when necessary, whether they can characterize the risk as "good" ("this system design might not work in practice, but if it does, it's a real advance") or "bad" ("we don't have the expertise to know if this paper has a horrible security flaw"). PCs try to avoid embarrassment; they should, but not to excess.

- Ignore the scores: Tom Anderson [1] showed that review scores are very noisy, fit a Zipf distribution, and are hard to normalize across reviewers, which means that except for a few top-scoring papers, the mean scores for the papers being discussed during the PC meeting have essentially no meaning. (They have more utility for deciding which papers progress past the earlier gates in the review process, but even then must be used cautiously.)

  PC chairs should scold PC members who, during the meeting, make arguments of the form "paper X had a higher mean score than paper Y and therefore deserves to be accepted first." While this logic appeals to quantitatively-minded people, it usually lacks factual basis.

  One can legitimately use the scores of each individual reviewer to infer his or her personal ranking of papers, or to quickly infer which PC members are likely to be champions (or critics) of a given paper.

- Should a conference lean towards accepting "as many papers as will fit" or "only superb papers?" My personal bias is the

former, but in either case, the PC chairs should establish a shared understanding of the goal before the meeting starts. This avoids having different PC members working, unwittingly, at cross purposes.

- When a paper is rejected before the PC meeting, usually the primary rationale for rejection is apparent in the reviews. However, papers discussed at the PC meeting can end up rejected on other grounds. Therefore, the PC should provide written feedback if the reasons for rejection are not explicit in the reviews. Given such papers are likely to be re-submitted elsewhere, the PC might also add some suggestions for improvements.

Other mechanisms: Several other mechanisms, which have been used to improve the conference-review process, can reduce negativity in reviewing and decisions:

- **Rebuttals**: Several conferences allow authors to rebut reviews at some point in the process before the final decisions. Occasionally, a rebuttal can correct a reviewer's misimpression and save a paper. However, rebuttals create more work for reviewers, anxiety for authors, and headaches for PC chairs who have to check that rebuttals do not violate rules about what new information they can introduce.

- **Shepherding**: Many, if not most, of the leading Systems conferences assign a PC member to shepherd every accepted paper. Shepherding almost always improves a paper: the shepherd can insist that the authors meet the main reviewer concerns, and can provide an outsider's point of view. (Sometimes, papers are "conditionally accepted" subject to final approval by the shepherd.)

  PCs that use shepherds might be willing to take larger risks that a borderline paper will be improved during revisions.

- **Accept more papers**: If the problem is a lack of space for papers on the cusp of acceptance, accepting more papers should help. PC chairs can create more slots by reducing the time allotted for each talk (even by a few minutes – but not at the expense of Q&A time), or by eliminating panel sessions or keynotes.

- **Public reviews**: A few conferences (and some other publications, such as the SIGCOMM newsletter) have provided short "public reviews" with published papers. These generally place the paper in context, summarize the reviewers' reasons for accepting the paper, and alert readers to caveats about the paper.

  PCs sometimes resist accepting a paper because, while they believe the paper has merit, they fear "sending the wrong message." Public reviews can quell this fear, giving PCs more willingness to take risks.

While these techniques could help reduce negativity, it might be impossible to measure the effect in any rigorous way. PC chairs will need to use their discretion, especially since all these mechanisms add to the PC's workload.

4.2 Process points for steering committees

Just as respected researchers are not always great reviewers, respected researchers (and even great reviewers) do not always make great PC chairs. The committee that chooses the PC chairs for a conference is, through that choice, helping to set the tone for the review process, and should consider whether candidates for PC chair are sensitive to hypercriticality.

Steering committees should not micro-manage PC chairs, but still have some role in establishing overall guidance. For example, the SIGCOMM Technical Steering Committee recently issued a brief statement providing guidance to SIGCOMM reviewers on the topic of hypercriticality[4].The TSC encourages SIGCOMM community members to contribute their own suggestions towards this goal, and towards improving the SIGCOMM review process in general.

4.3 Advice for authors

Reviewers are not solely to blame for hypercriticality. Writing good reviews is a difficult job, especially for well-run conferences that place a heavy load on each reviewer. Authors often unwittingly make mistakes that increase the work that reviewers must do, and reviewers, who are human and unpaid, sometimes react by being critical.

Authors can annoy reviewers in many ways, including:

- Submitting your paper to the wrong conference, in the hope that sooner or later it will get past the reviewers.

- Forcing reviewers to decode your paper – scientific papers are not puzzles or mysteries. Many authors write unclear sentences and paragraphs, or use poor organization, or leave out key points because they assume that the reviewers know the same things as the authors do. Most reviewers will not have the time to read your paper twice, especially if you annoyed them the first time through.

  As a general rule: if three expert reviewers all misunderstand your paper, then the fault is yours, not theirs.

- Violating the format guidelines, making figures too small to read, using unnecessary Greek letters, etc. Even omitting page numbers can be annoying; this makes it hard to write a review that says "on page 7, your third paragraph fails to ...".

- Making stronger claims than the evidence supports, or criticizing prior work more than it deserves, or citing papers that you obviously have not actually read.

In the end, however, reviewers will always focus on the flaws in a paper. Authors should understand that the accepted papers get negative comments, not just the rejected papers, and should tolerate a certain amount of venting, especially if they have triggered it through annoyances.

5. REFERENCES

[1] Thomas Anderson. Conference Reviewing Considered Harmful. *Operating Systems Review*, 43(2), Apr. 2009.

[2] Anita Borg Institute. Senior Technical Woman: Anne Condon, Professor and Head of the Department of Computer Science, University of British Columbia. `http://anitaborg.org/news/archive/senior-technical-woman-anne-condon-professor-and-head-of-the-department-of-computer-science-university-of-british-columbia/`.

[3] Ferric C. Fang, R. Grant Steen, and Arturo Casadevall. Misconduct accounts for the majority of retracted scientific publications. *Proc Natl Acad Sci*, 109(42):17028–33, 2012.

[4] `http://www.sigcomm.org/conference-planning/sigcomm-program-bcp/reviewing`

[4] Brighten Godfrey. What's wrong with computer science reviewing? http://youinfinitesnake.blogspot.com/2011/08/whats-wrong-with-computer-science.html, August 2011.

[5] Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

[6] Bertrand Meyer. The Nastiness Problem in Computer Science. Blog@CACM: http://cacm.acm.org/blogs/blog-cacm/123611-the-nastiness-problem-in-computer-science/, August 2011.

[7] Brian Morton. Falser Words Were Never Spoken. *The New York Times*, page A23, August 30 2011.

[8] Jeffrey F. Naughton. DBMS Research: First 50 Years, Next 50 Years. http://pages.cs.wisc.edu/~naughton/naughtonicde.pptx, 2010.

[9] Oscar Nierstrasz. Identify the Champion. In N. Harrison, B. Foote, and H. Rohnert, editors, *Pattern Languages of Program Design*, volume 4, pages 539–556. Addison Wesley, 2000.

[10] Ubiquity staff. An Interview with Peter Denning on the great principles of computing. *Ubiquity*, 2007(June), June 2007.