Before and After WOWCS: A literature survey, A list of papers we wish had been submitted

Jeffrey C. Mogul
(HP Labs, Palo Alto, Jeff.Mogul@hp.com*)*

Tom Anderson
(University of Washington, tom@cs.washington.edu*)*

Abstract

The Workshop on Organizing Workshops, Conferences, and Symposia for Computer Systems (WOWCS) was organized to "bring together conference organizers (past, present, and future) and other interested people to discuss the issues they confront." In addition to the position papers submitted to the workshop, the WOWCS program committee has collected a bibliography of previous publications in this area. We also list some topics about which we wish we had received submissions, but did not; these could be good topics for future articles.

1 Introduction

Computer systems researchers place an unusually high value on conference publications, to the point that these no longer take second place to journal publications. This puts pressure on conference chairs and committees, who must handle large numbers of submissions and generate detailed, well-reasoned reviews and acceptance decisions on tight deadlines. Yet there is relatively little "institutional memory" or written folklore on how to organize computer systems conferences, and many policy issues require repeated community or program committee (PC) discussions.

The April 2008 Workshop on Organizing Workshops, Conferences, and Symposia for Computer Systems (WOWCS) brought together conference organizers (past, present, and future) and other interested people to discuss the issues they confront. The workshop had several goals:

- **To discuss (and perhaps settle) certain contentious policy issues**, such as whether single-blind or double-blind reviewing is the best policy;
- **To preserve folklore and experience in written form**, such as how to choose PC members and other volunteers;
- **To evaluate tools and techniques for conference organizers**, such as review-management software.

The workshop attracted a moderate number of submissions from a variety of authors with significant experience in running PCs for computer systems conferences and workshops. These position papers, available online at http://www.usenix.org/events/wowcs08, present a wide variety of viewpoints, but do not cover the full range of possible topics. (That range would have been impossible to cover in a one-day workshop, in any event.) Also, while no workshop with the same goals as WOWCS had been held before, there are many previous publications (formal and informal) on these topics.

This article is our attempt to summarize the previous publications, and to list some topics that have not been discussed in writing (or at least, have not been satisfactorily resolved).

We somewhat fuzzily restrict our focus to "computer systems" publications, and to conferences rather than journals, since we believe that such events often require different handling than journals or events in other fields.

1.1 Context

Peer-reviewed scientific publication is an inherently contentious topic, because by its nature there are winners (papers published in prestigious venues) and losers (papers that are not published, or are published late or in low-prestige venues). The goals of the community are sometimes in conflict with the goals of individuals. The community values the advancement of a shared, tested base of knowledge and practice; individuals value their careers and self-esteem.

While computer systems research may be less contentious than fields such as medicine, where lives or commercial success depend directly on the results of peer review, we all seem to care a lot about the review process.

Given what appears to be an increasing number of computer systems conference papers to review each year, we also have to respect the practical limits on how much time individuals are willing and able to invest in the review process.

These issues have led to a lot of innovation with conference review processes, although there has not always been quantitative analysis of whether the results meet our goals.

But what are our goals? Even though we all share the aim of a fair, efficient process that results in the "best" conference programs, that still leaves points of contention: how to balance fairness vs. efficiency vs. conference quality, and how to define what counts as "goodness" for a conference paper – how we balance novelty, rigor, utility, and clarity.

We do not all have the same criteria for evaluating papers. For example, do we value technical rigor more or less than novelty? How necessary is it that an idea be practically implementable? How do we balance papers with new ideas against papers that validate prior work? Do we reject a paper with a half-baked execution of a good idea, hoping to get a better paper later, or accept it hoping to foster further work by others? How important is it to construct a broad program for a given conference, or to ensure that the topic boundaries between conferences are reasonably clear?

In short, what do we hope to achieve by the innovations in the review process? We do not presume to answer that question in this article, but we encourage others to be clear about their goals when proposing or evaluating review-process innovations.

2 Past publications about conference policies and methods

Many people have already published advice about, analysis of, or problems with computer science publication. Some of these have been peer-reviewed, although many have been published informally. Although quite a few papers have been published on these topics, they tend to be scattered around a large set of publication venues, and so most readers are likely to be aware of only a few of these.

This section summarizes some of the previous publications; we did not attempt to find them all.

2.1 Best practices in general

In 2005–2006, ACM and IEEE formed an ad hoc "Health of Conferences Committee." HCC's goals were "to collect the best practices onto a web page so that conference organizers can see innovative ways to cope with the demands of paper submissions, refereeing, and presentations, as the number of papers increase [sic]." Their results are available on a Wiki [19], which covers topics including tracking reviews across conferences; the use of two-phase reviewing (or "quick rejection"); the value of allowing author rebuttals to reviews before the final decision; double-blind submissions; whether conferences should grow to increase acceptance rates; the use of hierarchical program committees (although this does not seem to have covered the "heavy + light" model used recently by several conferences); ways to encourage wilder papers; co-locating workshops with conferences; and a few other topics. (For some reason, SIGOPS [listed here usually as "SIGOS"] contributed very little to this activity, and SIGCOMM contributed less than many other SIGs.)

Fred Douglis, in his role as Editor-in-Chief of *IEEE Internet Computing*, wrote several editorials: one on how to deal with misbehaving authors (in particular, those committing self-plagiarism and those who submit similar papers to multiple venues) [9], and the other on how to deal with misbehaving reviewers (who submit reviews late, never, or badly) [10].

2.2 Single-blind vs. double-blind reviewing

Most scientific reviewing is *blind*, in the sense that authors do not know who the reviewers are. In *double-blind* reviewing, reviewers are not supposed to know who the authors are, either. The goal of double-blind reviewing is to increase the assurance to all authors that the PC is doing its best to be fair: to avoid favoritism, revenge, or *status bias*, where reviewers put less value on papers from authors or institutions with lower status.

In theory, double-blind reviewing should improve fairness. In practice, there are some concerns about how well this works: reviewers often can guess the authorship of papers, and other PC members can guess who wrote a paper when conflicted PC members are kicked out of the PC meeting.

Single-blind reviewing has some potential advantages for the review process. When the authors are known, reviewers are better able to evaluate the work in context: compared to what has been published by the same authors before, does the paper under review add anything new? Also, less-experienced authors sometimes seem to have trouble anonymizing their submissions without damaging them. Finally, single-blind reviewing reduces the logistical challenges for PC chairs.

The SIGMOD community has published several articles on this topic. Since the SIGMOD conference has been double-blind since 2001, while SIGMOD 1994–2000 and the VLDB conference 1994–2005 were single-blind, this provided a data set to partially evaluate the effects of double-blind reviewing. Samuel Madden and

David DeWitt [17] published an analysis concluding that "double-blind reviewing has had essentially no impact on the publication rates of more senior researchers in the database field." However, Anthony Tung [29] looked at the same data set using a different statistical analysis, and came to the opposite conclusion.

Richard Snodgrass, in his role as Editor-in-Chief of *ACM Transactions on Database Systems* (TODS), wrote an editorial analyzing the published literature on the effects of double-blind reviewing [27]. He noted that, in a previous experiment not based on computer science literature, almost half of the authors of double-blinded papers could be guessed by the referees. He concluded that the existing studies showed enough of a status bias against "those in the gray area: neither at the top ... nor at the bottom" to justify a double-blind policy for TODS. However, most of the existing studies cover fields other than computer science, and it is possible that the level of bias varies between fields.

The debate around single-blind vs. double-blind reviewing may reflect different ideas about goals. For example, the use of double-blind reviewing might lead a PC to fail to realize that a submission is too similar to a prior publication by the same authors; this happens more often than one would like. How do we balance the risk of undetected cheating against the risk of status bias?

2.3 Opening up the review process

While reviewer identities are typically hidden from authors (and the world at large), to encourage greater honesty, some fear that certain reviewers abuse this anonymity. This has led to several kinds of experiments with the review process.

One such approach is "open reviewing," in which the reviewer names are revealed to authors (at the end of the process). The main goal of open reviewing is to increase the accountability of reviewers for their reviews, which might lead to better reviews and better choices when recruiting PC members. Some versions of open reviewing also publicize the non-anonymous reviews of accepted papers. Michalis Faloutsos, Reza Rejaie, and Anirban Banerjee described an experiment with open reviewing at Global Internet '07 [11]. They view this experiment as a success, based on feedback from authors and reviewers, although there were some complications.

Fabio Casati, Fausto Giunchiglia, and Maurizio Marchese go even further. They analyzed the ills of the existing process, and proposed simply eliminating the model of using pre-publication reviews to decide what gets published [6]. Instead, they propose that all papers (good and bad) are immediately published online, and then the community somehow manages to decide which of these papers have value. They suggest a process similar in some aspects to the PageRank algorithm.

Similar to open reviewing is the use of "public reviews," where a member of the PC publishes a signed review along with each published paper, to provide context that readers might otherwise lack. Public reviews can capture some of the commentary about papers by experts, which otherwise is not easily available. They also provide a way for a PC member to editorialize about a paper in ways that are not appropriate for authors to do, and they can help demystify the reasoning behind the PC's decisions. Public reviews, unlike open reviews, are not intended either for helping the PC's decision-making process or the author's paper-revision process. SIGCOMM has experimented with public reviews (e.g., HotNets 2004 [3], SIGCOMM 2006, and the *Computer Communication Review* newsletter).

The 2007 Passive and Active Measurements conference (PAM) experimented with author ratings of the reviews they received. The PC chair, Konstantina Papagiannaki, reported on the experiment and drew some conclusions: authors whose papers are rejected do not always give the reviewers bad scores; authors prefer longer reviews; authors prefer reviews with clear justifications for the reviewer's decision [20]. However, because this experiment was double-blind, reviewers did not know which of their reviews were taken badly, and so did not find the results very useful.

2.4 Advice to reviewers

The review process depends on a constant supply of willing and competent reviewers. Most of us learn this task on the job, but (especially for conference reviewing, where deadlines are usually tight and there is no editor to intermediate between the authors and the reviewers) newer reviewers often need written advice.

In 1990, Alan Jay Smith wrote a widely-circulated article explaining "The task of the referee" [26]. At about the same time, Ian Parberry wrote "A Guide for New Referees in Theoretical Computer Science" [21]. We could not find a subsequent formally-published paper on the same topic, perhaps because Smith's article was definitive.

However, plenty of people have posted advice on the Web, sometimes specific to a single conference. Most of these otherwise provide little of novelty. One that stands out, partly for its focus on computer systems conferences, is Timothy Roscoe's [22]. In particular, he explains why and how reviewers should avoid the snarky tone sometimes taken by overloaded reviewers who have put up with more mediocrity than they can tolerate. Mark Allman has also offered advice to reviewers [2], including suggestions for how reviewers should respond to papers with ideas that they do not like, using the slogan "review papers, not ideas."

The Neural Information Processing Systems (NIPS) conference's Web site has a detailed essay [8] on the evaluation criteria used for the 2006 conference, including somewhat different criteria for different subfields. This serves both to guide reviewers and also to help authors write better papers.

2.5 Shadow PCs

One interesting approach to training reviewers, and to vetting them for future PC service, is to have them serve on a "shadow PC." A shadow PC has access (subject to the authors' permission) to the papers submitted to a conference, and goes through the same review process as the regular PC, but does not have any effect on the final outcome. (Actually, since shadow-PC reviews are usually returned to the authors, these may serve to improve the final papers.) The shadow PC members may also learn things about the review process that will help them improve their own future submissions.

SOSPs in the 1980s and 1990s allowed some graduate students to informally review papers, but the first formal shadow PC that we know of was for NSDI 2004. Other recent systems conferences (SIGCOMM 2005, SOSP 2007) have also run shadow PCs. Anja Feldmann wrote a detailed report on the SIGCOMM 2005 experience [12]. Between Feldmann's experience and that of NSDI 2004, which ran five distinct shadow PCs [31], it seems that shadow PCs seldom pick anywhere near the same program as the regular PC. It is not clear how much of the difference is due to the greater experience of the regular-PC members, and how much is due to the randomness of the process.

2.6 Advice to authors

There is a lot of published advice to the authors of scientific papers. This article is not the place for a comprehensive review of that literature. However, several of these have expressed specific complaints about the quality of papers submitted to systems conferences, and illuminate some of the problems that conference committees are facing.

Roy Levin and David Redell wrote about the somewhat disappointing quality of submissions to the Ninth SOSP, which they co-chaired. They also gave advice to authors of subsequent systems papers [16]. More recently, Mark Allman wrote a plea to authors based on his own struggles trying to review badly-written submissions to SIGCOMM 2001 [1].

Although it is not at all specific to computer systems conferences, every scientific author should read George Gopen and Judith Swan's classic article on "The Science of Scientific Writing" [13]. They describe not how to write an entire paper, but how to write sentences and paragraphs that readers (and overburdened reviewers) can understand. Too many authors clearly have not learned this skill.

Finally, Tomáš Grim reports on the results of a study on scientific authorship that simply cries out for replication among the computer systems community [14].

2.7 Review-management software

There are lots of review-management systems available, both open-source and for-profit. A conference chair must choose one such system and stick with it for the duration; someone who has not chaired a conference recently may not have a good basis for making this choice. In the absence of a "Consumer Reports" guide to the relative merits of review-management systems, people usually get advice from other recent chairs, or use what they have used before.

The ACM SIG Governing Board Executive Committee decided in 1998 to attack this problem. Rick Snodgrass published a summary of 19 systems used by a variety of SIGs, including details of which features each system supported, and comments on the stability and usability of some of these systems [28]. In the intervening decade, we know of no other published comparisons, while the set of review-management systems has changed dramatically (and the ones that have survived since 1998 probably have evolved).

2.8 Reviewing of extended versions of workshop papers

Groups like SIGCOMM and SIGOPS have created a variety of workshops as a way to encourage the publication of preliminary or highly speculative work. Often the best of these short papers become longer, more polished submissions to regular conferences. While we do not want to publish the same paper twice, we also do not want to discourage people from writing workshop papers by preventing them from later publishing an overlapping full paper at a prestigious conference.

The usual approach is to look for "adequate" or "significant" additional content in the final paper, and if so, for the conference PC to evaluate the full paper's entire contribution, not just its new material.

This test becomes more complicated with double-blind reviewing, since if the workshop and conference papers were written by different authors, the conference paper's authors should not get the credit for the ideas in the workshop paper. SIGCOMM has debated this issue, and adopted an advisory policy whereby review and discussion of a follow-on paper provisionally assumes that it shares authorship with the prior workshop paper. In a final phase, a provisionally-accepted paper is rejected if

the authorship does not sufficiently overlap [25]. (This approach unavoidably inverts double-blind reviewing's normal assumption that the reviewers have absolutely no idea who wrote the paper.)

3 Papers we wish someone had submitted to WOWCS

Many of the issues that led to the creation of WOWCS remain unresolved or unaddressed. These topics have come up in past discussions over drinks, over meals, and in hallways, but potentially could benefit from more careful, written treatment.

We crudely divide these topics into policy issues, metrics, preservation of folklore and experience, and tools and techniques. The division is artificial, since many issues cover several of these categories. For example, any given review-management system inevitably embodies decisions about policies and metrics.

3.1 Policy issues

These issues represent fundamental policy choices, and many are problems for the community to resolve, not just for a single PC.

Double-blind vs. single-blind reviews While other communities have dealt with this issue (especially SIGMOD [17, 29, 27]), the computer systems community has not resolved it yet. SOSP and SIGCOMM are double-blind; OSDI, NSDI, and the annual USENIX conference are single-blind.

Since OSDI and SOSP alternate years, attract approximately the same kinds of papers and authors, and are now regarded as of roughly equal quality, they potentially offer a data set that would allow us to evaluate whether double-blind reviewing serves a useful purpose.

Two kinds of experiments might be worth performing:

- **Using externally-available data**: Analyze the sets of published authors in SOSP and OSDI, to test whether the conferences differ on the fraction of papers they take with "junior" authors – where "junior" could be defined as N years since graduation, or as "having $< N$ prior publications", or perhaps based on citation histories, for various N. This might depend on whether one looks at the most-junior or least-junior author for a co-authored paper, etc. Similarly, one could look at the set of author institutions.

  This might be a good exercise for some first-year or second-year OS students, since it would force them to become familiar with titles and authors of a decade's worth of papers.

- **Using confidential data**: Looking only at published papers does not reveal whether a PC has a (perhaps unconscious) status bias against certain authors or institutions. Ideally, one could look at the behavior of a PC with respect to the submitted papers. This would, of course, require careful observance of the confidentiality of rejected papers, and thus would probably require the active participation of the PC chairs for recent conferences. The other problem is that the relevant data might no longer be available for older conferences, or it might be tedious to get it into a canonical format for analysis.

The tricky aspect of this analysis is that OSDI and SOSP, while perhaps of equivalent quality, might not have equivalent PC mind-sets. For example, it is plausible that authors from low-status institutions have difficulty getting their papers accepted by SOSP not because of any status bias (since SOSP is double-blind) but because there is a certain style of paper that SOSP tends to prefer – and authors from low-status institutions do not have peers who can help them cast their papers in this style.

When should "open reviews" be used? Faloutsos *et al.* described the use of open reviews in the context of a fairly small event. It would be useful to have a more comprehensive discussion of the circumstances in which open reviews would be appropriate, as well as of the potential drawbacks from this innovation. To the extent that reviewers and authors attempt to game the system, open reviewing will change the game-theory rules and could create new incentives for misbehavior. For example, will junior reviewers avoid making negative comments about papers written by senior authors? Will open reviewing lead to more log-rolling (i.e., sets of people covertly agreeing to give good reviews to each other's submissions)?

CS-wide citation reporting and indexing Citation indices are a well-established mechanism for evaluating the impact of papers, authors, institutions, and conferences – in fields other than computer science. We do not have a good track record in this respect. The major science-wide indices, such as the *Science Citation Index* (SCI) and *Scopus* seem to provide only random coverage of CS literature (or so it seems to one who has looked up his own papers in these). CiteSeer provides good coverage of CS, but relies on automatic extraction of citations and does not have access to papers held inside walled gardens. Google Scholar has similar limitations.

The ACM Digital Library has good coverage of ACM and IEEE citations, partly because they now insist on receiving citation meta-data along with ACM-published

papers. However, they do not include citations from papers published by other organizations, such as USENIX, and the meta-data for some of their older articles may include OCR errors.

Issues that the computer systems community ought to address include:

- Are citation counts a useful metric in the first place? A high citation count can imply that a paper introduced an important new idea, or provided definitive results, but it can also result from negative evaluations, perfunctory mentions, or log-rolling [18].
- Does computer systems need a comprehensive citation index? (Or, does computer science as a whole need this?)
- How can we collect the necessary meta-data? Should all future CS publications (in any venue) require submission of appropriate meta-data in a standardized format? How do we create the meta-data for older publications?
- Who gets to control the index contents? Can they charge for using it, and if not, who has the incentive to maintain it?
- What analyses could be made using this information?
- Which citation-based analyses are of value to the community, and which are either mere noise, or actually counterproductive?

Travel reduction "Symposium" derives from a Greek word meaning "to drink together." Physical togetherness is one of the main reasons why we attend conferences; we know that the informal interactions are often more important than the paper presentations (since, one assumes, the main long-term benefit from the accepted papers is what appears in the proceedings). In spite of impressive advances in the state of teleconferencing, there is still no real substitute for physical meetings.

Unfortunately, physical togetherness means physical travel, and travel means wasted time, global warming, and significant expenses. (Travel expenses, once food and hotel costs are included, account for the majority of explicit conference-related spending.) As the number of conferences increases, as global warming has become more pressing, as travel budgets are being cut, and as air travel hassles multiply, one has to ask whether and how we ought to optimize the travel burden of conferences.

For example, should we be co-locating events more often and more carefully? Should we kill off certain conferences that fail to provide the community-building benefits of primary events, or convert them to journal-like publications? Should conference organizers refrain from putting conferences in "interesting" places, and instead aim to optimize the overall sum (or median, or 90th percentile) of travel costs and of carbon emissions?

A modest proposal: we should normalize an author's citation impact based on his or her carbon impact. This might discourage the practice of submitting a paper to a second-rate conference merely because its likely acceptance would justify a trip to some sunny beach resort.

Decoupling publication from presentation The norm in our community is that if a PC accepts a paper, it is both presented at the event and published in the proceedings. We break this coupling for special cases, such as posters, usually to provide advance exposure for work in progress and for students. Conference programs also often include a few invited speakers or keynotes, who present work that has not been peer-reviewed.

Even for full, peer-reviewed papers, this coupling does not always make sense. Some papers are worthy of publication, but make for really boring talks. Especially with the use of online publication, which allows for more proceedings pages without killing as many trees, a conference could accept more papers for publication than for presentation.

For example, the Neural Information Processing Systems (NIPS) conference decouples paper acceptance from the presentation decision [30]. The most interesting papers are presented at length, but many or most papers are presented only as posters, with 45-second "spotlight" presentations as brief advertisements for the posters.

This approach has risks. For example, the process for deciding which papers get presented might be biased against non-native speakers (which is painfully evident even in double-blind reviews).

3.2 Metrics

Many of the questions that we would like to resolve depend on new or better metrics for what we do as a community.

How do we quantify the merit of a conference? This might seem like an odd question, but there are a number of contexts where it is useful to compare the value or merit of a conference against others. These include: which conference should I submit this paper to? Is it worth my time to be on the conference PC? Is it worth my time to attend this conference? Should a sponsor (e.g., a corporation or a government agency) be willing to help cover the costs, and for how much? This metric might also help to evaluate a young author's publication record, before enough time has passed to see subsequent citations.

For some of these questions, citation-count impact could be the right merit. For others, it probably isn't. For example, a potential sponsor might want to contribute money to help launch a new event, long before there is enough history to evaluate its citation-count impact. CiteSeer rates venues using a widely-used "impact factor," described as "the average citation rate, where citations are normalized using the average citation rate for all articles in a given year, and transformed using $ln(n+1)$ where n is the number of citations" [15].

There is some controversy over the value of the impact factor metric [32]. Bollen *et al.* assert that the widely-accepted Thomson Scientific ISI Impact Factor is biased towards popular journals, rather than prestigious ones [5]. They suggest that a weight PageRank-style metric would favor high-prestige journals. Their paper includes an analysis of computer science journals, but not conferences. It might be interesting to apply their analysis to both journals and conferences in CS, to test how the best CS conferences compare to the best journals.

Do PCs tend to favor PC-authored papers? Conference PCs try to scrupulously avoid giving PC-authored submissions an unfair advantage. In addition to the usual conflict-of-interest rules, PCs often decide to "set a higher bar" for these papers, so as to avoid the perception of favoritism. But are PCs falling victim to unconscious biases?

One might test this question by looking at the citation-index impacts of papers published in a set of not-too-recent conferences (to the extent that this data is available) and checking whether the PC-authored papers, as a set, rank higher or lower than the others. Naturally, one would assume that PC members are drawn from the best in the community, and so ought to have a better record than average. Given this assumption, it might be necessary to compare the relative impacts of these authors' papers in the conferences where they were on the PC, and in the other conferences where they published.

How random are PC decisions? Many authors are mystified by PC decisions, and many PC members do have the impression that some of these decisions are random. In the very few cases where we have some independent decisions from shadow PCs, there is some evidence to support the random-decision hypothesis [12]. Of course, shadow-PC members do have less experience than regular PC members.

It is probably not worth the effort to conduct the obvious experiment to test this question, which is to constitute two equally-qualified PCs that simultaneously and independently evaluate the submissions to a conference, and then compare their decisions to see how well correlated they are. Is there a feasible way to test this, and, if so, what would we do with the result? Or should we simply accept some randomness as a fact of life?

How big is the rejected-paper tumbleweed? When a conference rejects a paper, it does not just go away. The paper usually gets resubmitted to a different conference, with (one hopes) improvements based on the reviewer comments. Sometimes this works well, but often the same papers bounce from conference to conference, and many of them seem not to improve much on each hop. This leads to excessive load on reviewers, and quite possibly reduces the attention that can be paid to more innovative submissions.

Several of the WOWCS papers propose ways to deal with this problem, such as establishing a repository of prior reviews. But nobody really knows how big the problem is; we generally find out about the resubmissions by accident, when someone has served on multiple PCs and spots a familiar submission.

It would be useful to measure the frequency at which rejected papers are resubmitted, the distribution of how many times a paper is reviewed until it either goes away or gets published, and whether the typical paper's trajectory is downward (that is, the authors keep aiming lower until it is accepted) or upward (the paper actually does improve). Automated techniques might be necessary to measure this, using textual similarities between submitted papers to track the life of a given paper. Without a good data set spanning many conferences, this might be impossible. John Douceur has suggested that it might be possible to analyze the extensive database of the EDAS review-management system [24] to obtain this kind of information, although many top-tier computer systems conferences have not used EDAS.

Is there a correlation between PC size and conference impact? Some PC chairs prefer to have relatively small PCs, which makes decision-making more coherent, but places a huge load on each PC member. Other PC chairs prefer to have very large PCs. (More generally, PC sizes seem to be more or less constant for a given conference.)

Anecdotal evidence suggests that conferences with very large PCs tend to have relatively low merit. Perhaps these conferences simply make poorer decisions, or perhaps people serving on these PCs put in less effort because less is demanded of them. It would be useful to know whether there is a real correlation (positive or negative) between PC size and conference merit, and if so, what might be causing it.

Does overlapping membership between PCs decrease diversity? Some PC members reappear year after year,

either on the same conference's PC or on the PCs of related conferences. Part of this is inevitable; competent, willing PC members are in short supply. Overlap between subsequent PCs of a given conference provides institutional memory; overlap between PCs of related conferences helps to calibrate our approach to submissions that might fall in the gaps. But balanced against these benefits is the perception, and perhaps the reality, that a small group of "permanent PC members" have excessive influence, and may be biased against outsiders or heterodox ideas. On the other hand, excessively-experienced reviewers might become so jaded as to give too much credit to papers with novel ideas.

To evaluate the overall effect of PC membership overlap on authorship diversity, one could look at the results of a PC's decision process to see whether high-overlap PC members gave higher or lower scores to submissions from authors (or institutions) that had not previously published in that venue. Similarly, if the reviews include a score for novelty, one could test whether high-overlap PC members favored papers with higher or lower mean novelty scores.

Is there a correlation between number of papers accepted and diversity? The PC co-chairs of SIGCOMM 2006 (Tom Anderson and Nick McKeown) expanded the program to include more papers than at any prior SIGCOMM. Their claim was that this added diversity of authorship and ideas to the program [4]. That year's conference did have high diversity in terms of author institutions and geographic locations. How significant is this effect? Is it a more useful way to increase diversity in computer systems conferences than, for example, double-blind reviewing?

Do overall scores predict what gets accepted? A typical procedure for a conference is to ask reviewers to assign a set of scores to each paper, then ignore all of those scores except the "overall" rating, and to rank the papers based on the mean of that rating. If the conference uses multiple rounds of reviewing, only papers scoring above a threshold make it into the next round. Generally, the set of papers discussed at the PC meeting is also determined by a threshold score, and often the order in which papers are discussed depends on their scores.

Given the difficulty of getting reviewers to assign consistent scores, and the difficulty of encoding multiple criteria (technical quality; novelty; presentation quality; suitability for the conference) into a single score, one might wonder whether this procedure generates the right outcome.

Ideally, one would want to compare the review-process scores (normalized for the conference's scoring system) of each accepted paper with its citation-index impact after several years. This is probably infeasible.

As a weak substitute, perhaps one could ask PC chairs for recent events to supply bit-vectors where the index of a bit corresponds to a paper's score-based rank, and the value of that bit is either "accepted" or "rejected." A collection of such bit vectors, while revealing nothing confidential, could lead to some interesting analyses. For example, one could plot the CDF of papers accepted at or below a given rank, as a way to measure the effectiveness of the scoring function.

Note that there are alternatives to the traditional mechanism. For example, OSDI 2006 did not allow reviewers to report an overall score. Instead, the PC co-chairs synthesized an overall score from a weighted combination of scores for technical quality, novelty, and presentation quality, thus removing from each reviewer the power to decide which of these aspects to value more highly. Possibly, therefore, if we could collect and analyze multi-component "score-vectors" for a set of conferences, we could establish whether PCs are favoring novelty over rigor, or vice versa – and whether papers selected based on novelty ultimately had a higher or lower impact than papers selected based on rigor.

3.3 Preserving folklore and experience

PC chairs typically learn their roles partly from observing other chairs while serving as PC members, partly by asking for help from other PC chairs, and partly by making their own mistakes. It would be helpful to have a written handbook for PC chairs, but not much of this exists.

The WOWCS workshop has established a Wiki, at `http://wiki.usenix.org/bin/view/Main/Conference/CollectedWisdom`, for PC chairs to share this kind of information. As of this writing, anyone can create an account on the USENIX Wiki and then contribute their own wisdom. We expect this Wiki to represent a range of opinions, possibly contradictory, about how to organize conferences; it is not meant to define universal norms.

This section lists some of the questions that could be answered in the future.

What is the best structure for a large PC? Given the need to balance reviewer load against the ability to have well-informed discussions in a PC meeting, and the large number of paper submitted to prestigious conferences, what is the best way to structure a PC? Use a small PC and torture the members with too many reviews? Use a huge PC and not get much coherence? Use a hierarchical PC, in which reviewers report to track chairs, and the track chairs make decisions without having read many of the papers?

Several conferences (e.g., SIGCOMM and SOSP) have recently experimented with a "heavy + light" model for their PCs. This practice started with SIGCOMM 2006 [4]. In this model, some PC members (the "light PC") review a modest number of papers, usually in the earlier phases, but do not attend the PC meeting. "Heavy PC" members review more papers, often focussed on the later phases, and do attend the PC meeting. This practice seems to be a good compromise between reviewer load (even the "heavy" members have a lower load than on a monolithic PC) and informed discussions (since the papers that are likely to be discussed in the PC meeting have been reviewed by a decent number of "heavy" members). Also, "light" members may be more useful than external reviewers, since they are chosen more carefully, and do enough reviews to provide calibration. However, this approach still requires some PC members to accept a relatively heavy load.

How to choose PC members and other volunteers Some of the WOWCS papers addressed how to avoid choosing certain people as PC members, based on their past dereliction of duty. We have relatively little shared wisdom about how to choose novice PC members, even though the steady-state process clearly requires an influx of new, competent people.

It is one thing to ask a junior researcher (such as a grad student) to serve as external reviewer for a paper. If the review appears to be out of line with reality (and many young reviewers seem unusually harsh), the PC can choose to ignore that review. It is much harder for a PC, or a PC chair, to decide to ignore another member of the PC – once someone is on the PC, the assumption is that his or her input has to be respected. Therefore, PC chairs are reluctant to invite people they don't know to join their committees. Are there ways we can develop useful information about potential PC members before they have ever served on a PC?

How to handle suspected author misbehavior Not all authors conform to community norms, and it usually falls to PC chairs to enforce these norms when violations are suspected. Sponsoring organizations generally have clear procedures for cases of plagiarism or self-plagiarism, but deciding what constitutes self-plagiarism is sometimes a judgement call [7, 23].

Other misconduct may fall into grayer areas. For example, we are all aware of authors who publish papers at the borderline of a "least publishable unit" (LPU) of novelty. This might be distinct from self-plagiarism, but it is still a burden on the community. Should conferences simply reject these papers, or should they do more to discourage it?

Many authors blatantly violate submission-format rules, whose purpose is to limit the length of the papers that reviewers must read. There is some controversy over whether we should even have such rules, but there are two good arguments in their favor:

1. Overburdening the reviewers does not help the system as a whole, and violates the implied contract between the PC chair and the people who volunteered their time for the PC. This can lead to reviewers failing to finish their reviews on time.

2. While not always true, generally a concise presentation of an idea is more valuable to readers, and more likely to be carefully reviewed, than a bloated presentation. Length limits force authors to make some attempt at concision.

Given that format rules are usually stated in the CFP, failure to enforce them not only adds to reviewer loads, it also biases the process against authors who are scrupulous about obeying, and who therefore have to work harder than the authors who ignore the rules.

Many conferences, however, do not enforce these rules at all. When we enforced them for OSDI 2006, we decided to limit our sanctions to six papers which contained substantially more text (through violations of margins, font-size, and line-spacing rules) than the others, rather than kicking out the much large set of papers that had minor violations of the rules. We also informed the OSDI audience that we had asked authors to withdraw their papers for this reason.

Abuse of author-declared conflicts Review-management systems often let authors declare a set of reviewers that should be considered conflicted for their paper. In most cases, this is simply an expedient way to populate the conflict matrix, rather than having the PC chair look at each author list manually to guess at conflicts (which might not be apparent if you don't know the authors and their past affiliations).

However, some have speculated that authors could bias the review process in their favor by declaring bogus conflicts with reviewers they don't like or trust. Or an author could simply declare conflicts with all reviewers known to have expertise on the topic matter, hoping to "snow" the other reviewers with a good story.

This abuse could be hard to detect, especially in a double-blind process where the PC chair cannot ask the reviewers whether they believe a conflict is legitimate. And it could become a significant problem with open reviews, since authors would learn quickly which reviewers to avoid. Authors could also blackball PC members who had been on a previous PC that rejected the same paper, so as to avoid detection of a lack of improvement.

How to handle suspected reviewer misbehavior Fortunately, most reviewers follow ethical rules, even though they do always not get their reviews done on time, or with enough detail. However, some ethical transgressions do take place, including:

- Violations of confidentiality – especially troubling is when a reviewer takes advantage of something they learned from a submitted paper to advance his or her own work – and doubly troubling if the submission was rejected.
- Attempts to guess at the authorship of double-blind submissions, beyond accidental discoveries when checking for related prior work.
- Explicit bias for or against authors.
- Log-rolling among reviewers.

PC chairs need to be willing to handle these problems, but it is not always clear what the right approach should be.

How to get a PC meeting to finish its job on time PC meetings for healthy conferences usually run late, but they cannot run forever; sooner or later, the PC evaporates, with some members leaving to catch flights, and others losing their mental presence for lack of food or sleep.

If the PC chairs manage their time well, the last part of the PC meeting is an interminable discussion about a few papers for which consensus cannot be achieved. In this case, the chairs must simply find a way to resolve the lack of consensus.

If the chairs manage their time badly, the meeting may end with papers being rejected simply because there was no time to discuss them, or with contentious papers ending up accepted or rejected in a hurried process.

Keeping a PC meeting on schedule is a difficult process, since one ought not simply allocate a fixed amount of time for each discussion. (However, it can be useful to limit the initial discussion for each paper, and then return to contentious papers only after all have been discussed at least once). It would be helpful to document techniques that work.

When, why, and how to shepherd Many conferences assign shepherds to accepted papers, sometimes to enforce a conditional acceptance, and mostly to help the authors improve the paper. Shepherding almost always significantly helps a paper, and authors are often grateful for the help. But it does place an additional burden on PC members who have already done a lot of work. When is shepherding worth the effort? When should acceptance be conditional, and what does the PC chair do if a shepherd refuses to accept a paper (or an author refuses to cooperate with the shepherd)? What are best practices for shepherds to follow?

One question that often comes up is whether it is appropriate for the shepherd to insist, on behalf of the PC, for the authors to do new work, rather than to simply improve the presentation of the submitted work. Also, how far should a shepherd go towards, in effect, becoming a co-author of the paper? Is it ever appropriate for a shepherd to be listed as a co-author on the published version?

3.4 Tools and techniques

We rely on software systems to manage the review process, especially for conferences that get lots of submissions and that generate lots of reviews.

Reviews of review-management software As explained in Section 2.7, we lack information about the relative merits of review software systems, and it would be useful to have some reviews of these systems, written by people with experience using more than one.

This should not necessarily lead to every conference using the same software. Different tools might be optimized for different purposes and use models. (For example, some review-managed systems are "hosted" services; others must be installed, run, and managed by a conference volunteer or organization staff member.) But it might be worth some consolidation in this market, to avoid wasted software effort, to improve the ability of chairs to share expertise, and to simplify the analysis of submission data to improve conference design.

Proposals for new or improved review-management features Probably every PC chair has wanted features that the review-management system does not provide. Often we are driven to make things work using spreadsheets or shell scripts. If we are lucky, we get to convince the developers of our particular system to add our favorite feature, but it would be more useful to have feature proposals that all developers were aware of.

4 Summary

When WOWCS was first proposed, there was some concern that there would neither be enough to discuss, nor enough people interested in the discussion. While WOWCS was not a large workshop by any standards, we received interesting submissions from a variety of experienced researchers, many other people expressed regret that they could not attend, and we found no lack of things to discuss. Given the extensive list in this article of topics that might be subjects for future publications, we would not be surprised to see a second WOWCS, if people can

be convinced to organize the event and to write the papers.

Acknowledgments

We would like to thank a lot of people for helping us to write this article, including the other members of the WOWCS PC, the WOWCS authors, and especially Mary Baker, Ed Lazowska, Hank Levy, Mehul Shah, Dan Weld, and David Wetherall.

References

[1] Mark Allman. A Referee's Plea. http://www.icir.org/mallman/plea.txt, May 2001.

[2] Mark Allman. Thoughts on Reviewing. *SIGCOMM Comput. Commun. Rev.*, 38(2), April 2008. http://www.icir.org/mallman/reviewing-ccr-apr08.pdf.

[3] Tom Anderson and Nick McKeown. Public Reviews of Papers Appearing at HotNets-III. In *Proc. Third Workshop on Hot Topics in Networks*, Nov. 2004.

[4] Tom Anderson and Nick McKeown. Program Chairs' Message. In *Proc. SIGCOMM*, Sep. 2006.

[5] Johan Bollen, Marko A. Rodriguez, and Herbert Van de Sompel. Journal Status. *Scientometrics*, 69(3):669–687, 2006.

[6] Fabio Casati, Fausto Giunchiglia, and Maurizio Marchese. Publish and perish: why the current publication and review model is killing research and wasting your money. *Ubiquity*, 8(3), 2007.

[7] Christian Collberg and Stephen Kobourov. Self-plagiarism in computer science. *Commun. ACM*, 48(4):88–94, 2005.

[8] NIPS 2006 Program Committee. Nips paper evaluation criteria. http://nips.cc/PaperInformation/EvaluationCriteria, 2006.

[9] Fred Douglis. Collective Wisdom: A Modest Proposal to Improve Peer Review, Part 1. *Internet Computing, IEEE*, 11(5):3–6, Sept.-Oct. 2007.

[10] Fred Douglis. Collective Wisdom: A Modest Proposal to Improve Peer Review, Part 2. *Internet Computing, IEEE*, 11(6):3–5, Nov.-Dec. 2007.

[11] Michalis Faloutsos, Anirban Banerjee, and Reza Rejaie. You Must Be Joking: A Historic Open Reviewing at Global Internet '07. *SIGCOMM Comput. Commun. Rev.*, 37(3):79–82, 2007.

[12] Anja Feldmann. Experiences from the Sigcomm 2005 European Shadow PC Experiment. *SIGCOMM Comput. Commun. Rev.*, 35(3):97–102, 2005.

[13] George D. Gopen and Judith A. Swan. The Science of Scientific Writing. *American Scientist*, 78(6):550–558, Nov.-Dec. 1990. Available at http://www.americanscientist.org/template/AssetDetail/assetid/23947.

[14] Tomáš Grim. A possible role of social activity to explain differences in publication output among ecologists. *Oikos*, 117(4):484–487, 2008.

[15] Steve Lawrence, Lee Giles, and Kurt Bollacker. Estimated impact of publication venues in Computer Science. http://citeseer.ist.psu.edu/impact.html.

[16] Roy Levin and David D. Redell. An Evaluation of the Ninth SOSP Submissions; or, How (and How Not) to Write a Good Systems Paper. *ACM SIGOPS Operating Systems Review*, 17(3):35–40, July 1983. Available at http://www.usenix.org/events/samples/submit/advice.html.

[17] Samuel Madden and David DeWitt. Impact of double-blind reviewing on SIGMOD publication rates. *SIGMOD Rec.*, 35(2):29–32, 2006.

[18] National Science and Technology Council, Committee on Fundamental Science, Subcommittee on Research. Assessing fundamental science. http://www.nsf.gov/statistics/ostp/assess/, July 1996.

[19] ACM/IEEE Health of Conferences Committee. http://wiki.acm.org/healthcc, 2006.

[20] Konstantina Papagiannaki. Author feedback experiment at PAM 2007. *SIGCOMM Comput. Commun. Rev.*, 37(3):73–78, 2007.

[21] Ian Parberry. A Guide for New Referees in Theoretical Computer Science. *Information and Computation*, 112(1):96–116, 1994.

[22] Timothy Roscoe. Writing reviews for systems conferences. http://people.inf.ethz.ch/troscoe/pubs/review-writing.pdf, 2007.

[23] Pamela Samuelson. Self-plagiarism or fair use. *Commun. ACM*, 37(8):21–25, 1994.

[24] Henning Schulzrinne. EDAS: Editor's Assistant. http://edas.info/doc/.

[25] SIGCOMM. SIGCOMM Frequently Asked Questions (FAQ). http://www.sigcomm.org/faq.html.

[26] Alan Jay Smith. The task of the referee. *Computer*, 23(4):65–71, Apr. 1990.

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

[28] Rick Snodgrass. Summary of Conference Management Software. http://www.acm.org/sigs/sgb/summary.html, 1999.

[29] Anthony K. H. Tung. Impact of double blind reviewing on SIGMOD publication: a more detail analysis. *SIGMOD Rec.*, 35(3):6–7, 2006.

[30] Daniel S. Weld. Personal communication. 2008.

[31] David Wetherall and Tom Anderson. Personal communication (unpublished manuscript). 2008.

[32] Wikipedia. Impact factor. http://en.wikipedia.org/wiki/Impact_factor.