Towards a Model of Computer Systems Research

Thomas Anderson
Department of Computer Science and Engineering
University of Washington

This paper develops a model of computer systems research as a way of helping explain to prospective authors the often obscure workings of conference program committees. While our goal is primarily descriptive, we use the model to motivate several recent changes in conference design and to suggest some further potential improvements.

1 Introduction

Peer to peer systems have become a popular area of distributed systems research over the past few years, reflecting both the widespread use of systems like BitTorrent [5] and the emerging commercial potential of peer to peer systems for a variety of applications well beyond file sharing [4, 12, 18, 19, 20]. This line of research has resulted in substantial progress in understanding system behavior, as well as better techniques for constructing more robust, more scalable, and more efficient systems. For example, workloads, churn, and available resources are all heavy tailed, and this is fundamental to understanding aggregate system behavior in practice [5, 15]. Modeling peers as rational, sometimes altruistic, and occasionally byzantine agents [1] is essential to building systems that are both more robust and more efficient [14, 12]. And randomness is widely used in peer-to-peer systems as a fundamental design technique [18, 5].

In this paper, we turn our attention to another peer to peer system that has received less attention from the systems research community: the systems research community itself. Our approach is intentionally tongue in cheek, but we observe many similarities, at least on the surface, between peer to peer systems and the systems research community. For one, there is relatively little central coordination! Rather, progress is reached through the mostly independent actions of individual researchers, interacting mainly through the conference publication system. [1] Citations, and in all likelihood research reputation as well, is heavy tailed [16]. As any program committee knows all too well, authors are often rational, sometimes altruistic, and occasionally byzantine [7]. And while randomness in conference program committee behavior might be considered undesirable, some have suggested that it may dominate many decisions in practice [8].

Our goal is primarily descriptive. In our experience, many students and even faculty find decisions made by conference program committees to be, well, inscrutable [17]. Speaking as someone who has both authored many papers and served on many program committees, the feeling is often mutual: reviewers often worry researchers are intentionally gaming the system. An additional goal of this paper is to shine a light on the potential for what economists would call *perverse* incentives, where the conference review process causes misdirected effort that might be avoided with a better designed mechanism. On the other hand, we also urge caution: seemingly intuitive changes to regulatory mechanisms often yield the opposite of the intended effect. We give an example of one such pitfall below.

Our model has three parts taken directly from the peer to peer literature: randomness, heavy tailed distributions, and incentives. We discuss these in turn, concluding with some modest suggestions for improving conference design that we hope might better align author and conference incentives. Since each of the elements of our model has been observed before with respect to research publications, we focus most of our discussion on the interplay between these elements.

2 Randomness

The task facing a technical conference program committees is easier said than done: under tight time constraints, select a small number of meritorious papers from among a much larger number of submissions. Authors would like a predictable and correct outcome, and they become legitimately upset when their papers are declined while

[1] In computer systems research, conferences, rather than journals, are the main way that research results are disseminated and peer reviewed.

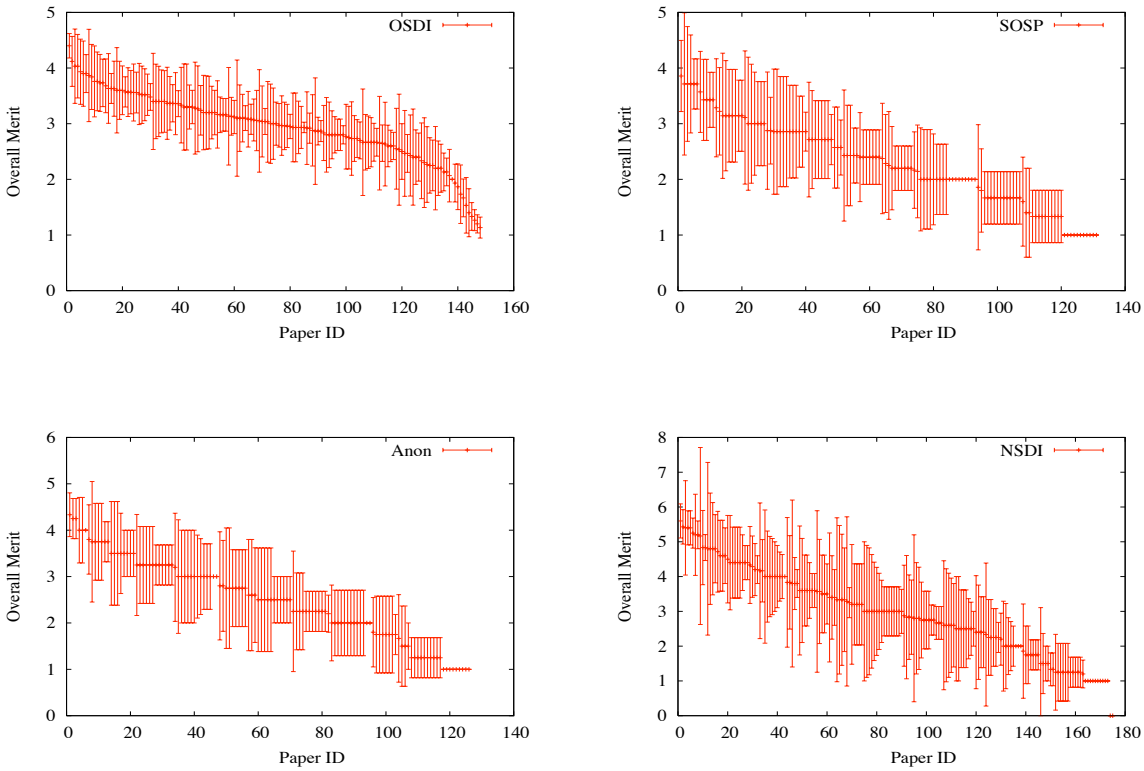

Figure 1: Mean evaluation score with standard deviation, for each paper submitted to four recent systems research conferences.

"obviously" worse papers are accepted. While one might ascribe this to author bias (everyone believes their own paper is above average) [7], authors with multiple submitted papers have an even better measure of predictability: did the PC ranking match their own? Often the answer is no.

How can this be? In computer systems research, individual reviewers differ significantly on the very fundamental issue of what is merit: how much to weight various factors such as importance of the topic, uniqueness and creativity of the solution, thoroughness of the evaluation, and effectiveness of the presentation [17, 2]. Some reviewers penalize good ideas that are incompletely evaluated, as a spur to encouraging authors to complete their work prior to publication; others do the opposite, as a way to foster follow-up work by others that may fill in the blanks. Some reviewers are willing to accept papers that take a technical approach that they personally disagree with, as long as the evaluation is reasonable; others believe the role of a program committee should attempt to prevent bad ideas from being disseminated.

Even if reviewers could somehow agree on all these factors, the larger the program committee, the harder it is to apply a uniform standard to all papers under review. Systems research conferences have seen a rapid increase in the number of papers submitted; while we might conceivably try to prevent this by raising the cost of submission, research has found that the rate of production of scientific research papers has been doubling every fourteen years for the past several centuries [6]. Thus it seems unlikely that we will see a moderation in the rate of increase in the production rate of systems research papers anytime soon. To deal with this, either the workload of a given program committee member, or the size of the program committee, or the number of conferences, must increase. Or all three. One hopes not exponentially.

Anyone who has served on a top tier program committee understands the result: altogether too much randomness in the outcome. Figure 1 shows the mean and standard deviation (square root of the variance) among review scores for papers submitted to four recent first-tier systems conferences: OSDI 2006, SOSP 2007, NSDI 2008, and another that shall remain anonymous. Papers are ranked by average review score, with error bars for the standard deviation among scores for each paper. Each of the conferences accepted between 25-30 papers. In most cases, reviewers were permitted to revise their scores after reading other reviews, but few chose to do

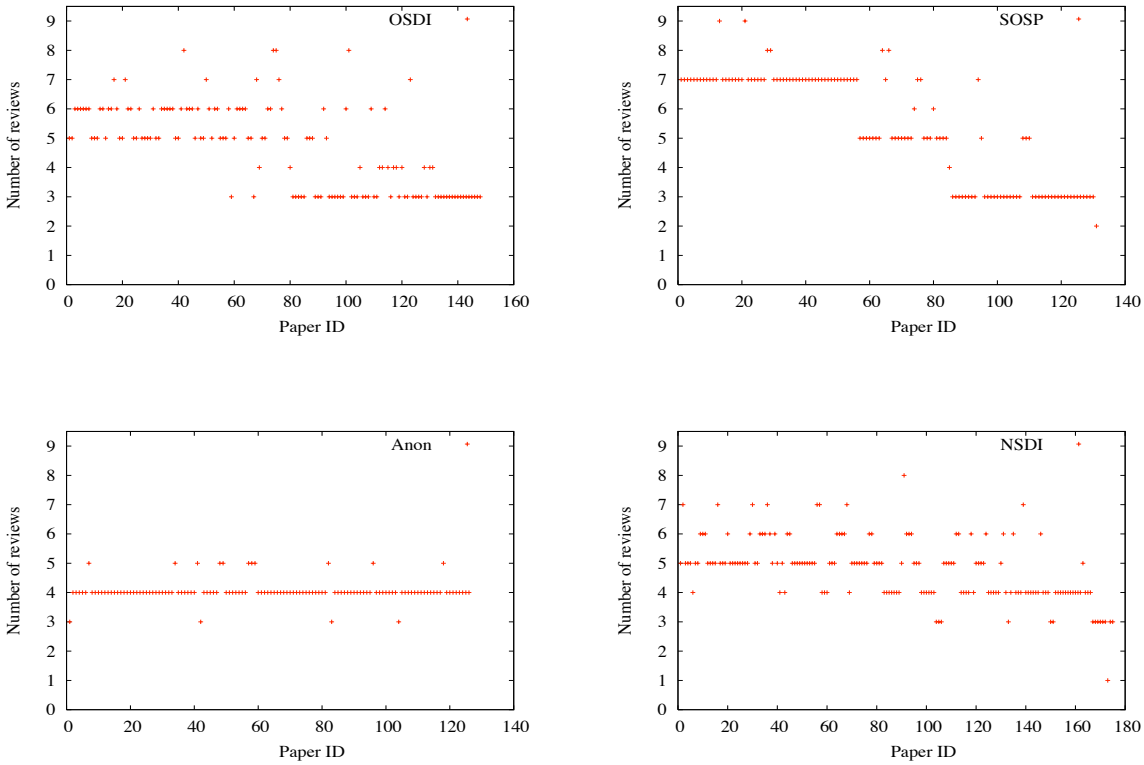

Figure 2: Number of paper reviews, for each paper submitted four recent systems research conferences. Papers are sorted by mean score, as in Figure 1.

so. Hence Figure 1 largely reflects the underlying variance in reviewer opinion, rather than the consensus that emerged from the program committee meeting. Figure 2 reports the number of reviews for each paper, ranked in the same order as Figure 1; if the review scores for a particular paper are iid around a common mean, one would need to quadruple the number of reviewers to halve the standard deviation, not a particularly welcome scenario.

As the figure shows, the variance in reviewer scores is far larger than the difference in the mean score, for a broad range of papers around the boundary between accept and reject. All four conferences held a conventional program committee meeting; papers were accepted after a discussion, not solely based on the mean score. While it might seem intuitive that the program committee discussion adds value, some caution is merited. When forced to make a choice between nearly equivalent options, the human brain will make up reasons post hoc for why the differences were in fact significant [9]. Several times over the past few years, a comparison between the programs chosen by a shadow PC and the real PC have shown a remarkable lack of congruence. One could explain this as due to the relative lack of experience of the shadow PC, but an equally plausible explanation is simply random chance [8].

If we cannot eliminate randomness in the paper selection process, we should at least actively manage it. At SIGCOMM 2006, the author and Nick McKeown put into practice a review process that aimed at cost-effectively using scarce reviewing cycles. Papers at either end of the quality spectrum were reviewed less often than papers at the margin. Papers with high variance in review score were automatically given additional reviews, while those with no variance were not. One particularly controversial paper received nine reviews (and a half hour discussion at the PC meeting) before being accepted, others were rejected after two reviews, or accepted after four. Further, by seeking reviews whenever there was variance, we were able to assign reviewers as reviews rolled in, rather than having to hold up each phase of assignments for the inevitable PC laggard. In a committee with fifty members, someone will always be late.

Before reviewing started, we engaged the program committee in a collective discussion as to how to weight the various factors of merit discussed above, and we care-

fully chose the questions on the review form to reinforce that social consensus. Cognitive experiments suggest that value judgments can be significantly influenced by subtle reminders [9].

To manage workload and improve confidence intervals, we had a relatively large program committee split between "heavy" and "light". All program committee members were assigned approximately 15 papers to review, and essentially asked to bin their set of papers into strong accept, marginal accept, marginal reject, and reject, proportionately to those numbers in the overall program. (In this fashion, we hoped to force reviewers to make a judgment call – would they include this paper in the program? Otherwise a large number of papers would end up in the marginal camp, and no information would be conveyed by the review score.) Initial paper assignments were done randomly among program committee members, among those qualified in the paper topic area, to further improve the confidence intervals. We used external reviewers only to provide missing expertise.

Based on score, variance, and an email consensus, we pre-accepted nearly half the papers at the conference prior to the program committee meeting, so that we could focus the in-person discussion on precisely those papers for which the answer was least clear. The light program committee was not asked to travel; the heavy program committee met in person to consider the papers at the margin. To improve consistency, each paper under discussion was read by at least a quarter of the heavy PC, hence the term, "heavy PC".

Readers are invited to judge for themselves whether the quality of the program differed significantly from other iterations of the same conference. What we did find, however, was somewhat surprising: this process nearly drove the heavy PC insane. The difference between an accepted and a rejected paper is hugely important to the authors, and yet, in the end, there is very little rational basis to decide between papers at the margin. Explaining why this may be so is the topic of the next section.

3 Zipf

In this section, we consider how merit is distributed among the papers submitted to a conference. To make the discussion concrete, Figure 3 draws two plausible alternate models. In the first model, we assume a conference in which thirty papers are accepted, all accepted papers have the same value, and no rejected paper has any value. In the second model, we assume paper value is distributed according to a zipf curve, $x^{-1/2}$ for the top sixty submitted papers, with the other papers having zero value. The curves are scaled to have the same total value.

There is a widespread recognition that simple paper counting is an invalid way to determine the impact of a researcher's career. After all, there are so many publication venues, that surely some venue will publish virtually any valid paper. The length of a CV does not indicate much of anything, beyond effort. Nevertheless, with respect to any single conference or journal, especially well-known ones, paper counting is a widespread practice. This is for good reason: with no additional information about the papers other than their titles, all the papers accepted at a given conference are equivalent, and all rejected papers are unknown. Program committees encourage paper counting by providing no ranking information among accepted papers, except in some cases to identify a small number of award papers. By default, then, this leads to the valuation function as drawn in Figure 3: a step function where papers that are accepted are all equally valued, and papers that are rejected are, for all practical purposes, worthless.

Obviously, a step function is a poor approximation of the underlying merit of a set of submitted papers, and we will argue later that using a step function, even in part

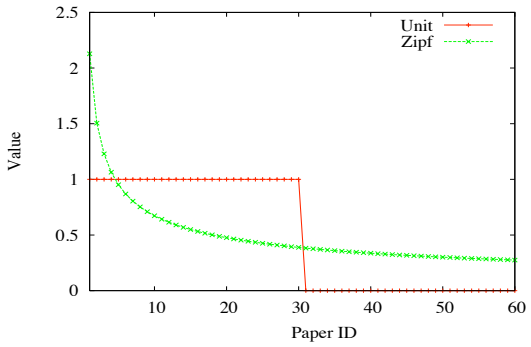

Figure 3: Two alternate models of research value. Both curves have the same aggregate value.

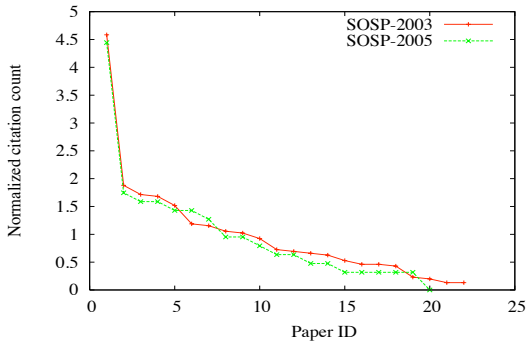

Figure 4: Normalized citation count for two recent SOSP conferences.

or by default, may incent researchers inappropriately. In earlier work, we showed that the step function reward curve used in BitTorrent makes it particularly vulnerable to stragetic client behavior [14].

But first, what might we use instead? A full characterization of research value is beyond the scope of this paper, but we believe a zipf distribution may be a reasonable approximation. In a zipf distribution, if x is the rank of an item, $f(x) = x^{-\alpha}$, where α can vary between 0 and 1. Zipf distributions have been found to model many apparently different structures, such as the frequency of words in books, the size of cities, the popularity of movies and web pages, and so forth.

More specifically for this paper, a zipf distribution captures the intuitive notion that most papers submitted to a particular conference have something useful to say. Papers at the very top of the accepted list are often quite a bit better than the others (there's even a phrase for this: "clear accept"). But there are not many clear accepts, and for the remainder, there is precious little difference between accept and reject. As shown in Figure 3, with $\alpha = 1/2$ and thirty accepted papers, assuming the program committee was perfect in its judgment, the difference between the worst accepted paper and the best rejected paper is approximately 1%. There is no reviewing system that I know of that can reliably distinguish that level of difference (or anywhere close to it), and so in practice, given the randomness discussed in the previous section, the best rejected paper may be quite a bit better than the worst accepted one.

We note that citations in scientific research in general [16] and computer science in particular [3, 13], are distributed according to a zipf curve, with $\alpha = 1/2$, at least for the top 10-20% of the total universe of published research papers. Since most readers will find that claim implausible, we illustrate it using two recent SOSP conferences. In Figure 4, we plot the normalized citation count (that is, the citations to a specific paper, divided by the total citations to all papers appearing in the conference) for all papers published at SOSP 2003 and SOSP 2005. Note that the program committees for those conferences most likely (and quite rightly) did not select papers for their citability; thus the papers that were published are a selection out of the total universe of systems papers. It is the total universe of systems papers that is likely to be zipf. Nevertheless, on this one metric, it is interesting that there is a 5-10x difference between papers appearing at the same conference.

In our view, citation counts do not represent all, or even most, of the intrinsic value of research papers. Rather our point is simply that a zipf curve is a more realistic model of research value than a unit model that considers only the fact of publication at a specific venue. Citation counts typically mix all sources of citations together, regardless of the merit of the conference or the depth of the contribution implicit in the citation. More fundamentally, citation counts favor the early over the definitive. To take an extreme example, the most referenced computer systems research paper ever published (according to citeseer [11]) is the initial TCP congestion control paper [10]. While that paper is influential, it would be hard to argue that it is more meritorious than every other systems paper that has ever been published. Rather, the TCP paper, like a large number of widely cited papers, was (i) early, (ii) left ample room for others to innovate, and (iii) was in a research area that had a low barrier to entry for other researchers (in this case, because of the widely used simulation package, ns2). Only some of those three characteristics could be considered inherently valuable.

One implication of a zipf distribution of merit for conference submissions is that, for a popular conference, the aggregate value of the rejected papers may be comparable to that of the aggregate value of the accepted ones. Zipf is a heavy-tailed distribution, which means there is significant area under the tail of the curve. Of course, the result is somewhat different if we consider value per square inch of paper!

All this might give support to advocates of conferences with parallel tracks. After all, why not simply accept all valid papers, and run as many parallel tracks as necessary? All things being equal, this would maximize the information content of the conference, compared to one that picked an arbitrary threshold. However, from the audience perspective, the information content of a multi-track conference is strictly less than a single track one. If, as is standard practice, the conference organizers distribute papers among the various tracks according to topic, rather than according to value, the typical conference attendee is faced with the conundrum that the best papers at the conference are spread thinly across all sessions. Multitrack conferences also seem to run afoul of incentives, our next topic.

4 Incentives

We next turn our attention to the role of incentives in conference design, particularly the interplay between incentives, randomized selection, and heavy-tailed merit. Clearly, a full investigation of researcher incentives is beyond the scope of this paper. It seems likely that no two researchers share the same set of motivations, other than that it is unlikely to be monetary reward! For the purposes of discussion, we assume researchers are motivated in part by research recognition, recognition is given in part by publication at prestigious venues, and recognition is unit-value based on the venue (the average of the true merit of all papers that appear there). Of course, this model is unrealistic in that most researchers are not

primarily motivated by the mere fact of publication. We further assume merit is heavy-tailed among the universe of publications in a particular research area, and authors are aware of the relative merit of their own work – another dubious assumption!

Under these assumptions, it is easy to see why it is rare to find high prestige multi-track conferences. Authors of better papers would have an incentive to send their papers to a more selective single-track conference, and if one didn't exist, they would have an incentive to band together to create a better alternative for their papers. (Alternatively, we can assume conference organizers have an incentive to attract high merit papers. Essentially, at a multitrack conference with unit-value, the good papers are cross-subsidizing the reputations of the relatively worse ones. As with any cross-subsidy, there is an incentive for the subsidizers to avoid the tax. Moreover, once such a single-track conference was successfully created, authors of other papers would be incented to send their papers to the single-track conference, as they would benefit by association. Under the unit value model, it is better to be a worse paper at a good conference than a good paper at a worse conference.

By contrast, a stable situation is the one we often see in practice. First, there is often a high-prestige single track conference, with award papers, to capture and reward the top end of the zipf distribution. Since that inherently leaves a large number of unpublished papers of significant value (nearly equal to the intrinsic value of the median paper at the high prestige conference!), those authors would be incented to send their papers to a different venue. The second conference is more usefully multi-track, as there is less difference between successive papers as we go down the distribution. In a zipf distribution with $\alpha = 1/2$, the difference between paper 30 and paper 100 is the same as the difference between paper 8 and paper 30.

5 Improvements

We conclude with an observation. It should be the goal of conference design to encourage authors to complete their work to the point of diminishing returns. Great ideas should be thoroughly fleshed out, while mediocre ones should be quickly documented, allowing the authors to move on to greener pastures. However, under the current system of multiple conferences per year per area, unit value reward for a specific conference, and noise in the evaluation system, authors have an incentive to do considerably less than this ideal.

Consider the marginal incentive for an author of a research paper. By marginal incentive, we mean the incremental benefit to the author of putting additional effort into improving a particular paper. After all, the author does have a choice: put more effort into an existing project, or start a new one. Some authors make a virtue of this, by going after "low hanging fruit." (To exploit the analogy, cognitive experiments indicate that the higher the fruit, the tastier it will be [9].) Suppose we posit a noiseless system with a single conference, unit reward for getting papers into the conference, and perfect information, e.g., knowledge by the author of the threshold to be applied by the program committee. The author's marginal incentive in such a system would be an impulse function: do no work unless the idea is above threshold, and then do only enough work to push it above the threshold for publication. Some might say that SOSP prior to 1990 was such a system. The conference met once every two years (reducing the opportunity for retries), there were few alternate venues of comparable merit, the threshold was high (simulation studies need not apply), and much of the community could guess whether a particular paper idea had a chance. The result was a relatively small number of high quality submissions.

Uncertainty changes the equation. In most cases, an author can't perfectly predict the threshold a particular program committee will use, and noise in the evaluation system makes the outcome for a particular paper near the margin quite unpredictable. Thus, the marginal incentive for an author, considering a single conference at a time, is a smeared impulse function: increasing as it approaches the likely threshold, and then falling off as the paper becomes increasingly likely to be accepted. Oddly, the *more* randomness in the evaluation and the less predictable the outcome, the more incentive authors have to work harder. Slot machines work on a similar principle.

In practice, another effect seems to matter. In many areas of computer systems research, we have prestigious conferences every six months. This is a direct consequence of increasing competition – as we observed, the universe of papers is increasing exponentially, and authors of good papers at weak conferences have an incentive to try to create higher prestige venues for their papers. The multiplicity and frequency of venues provides authors an alternate strategy to compensate for unpredictable program committees: submit papers initially with the minimal amount of work needed to be competitive. If accepted, move on. If rejected, double the amount of work and repeat. Thus the more competition there is for conferences, the more conferences we have, and the more incentive authors have to stop when their papers are merely mediocre. From a game theory perspective, it would be ineffective for a program committee to attempt to reject papers that are great ideas, but incompletely executed: in that case, authors would have the incentive to work only on those research ideas that are obviously mediocre.

Most of these counterproductive incentives would dis-

appear if we were able to set rewards to model the underlying value of the work. Provided noise is a second order effect, authors would be incented to complete the work to the point of diminishing returns. In fact, authors might be incented to hold back papers until they were ready, or equally valid, publish a short form of the research idea quickly followed by a later, more thorough evaluation.

We therefore make the modest suggestion that conferences publish the rankings of papers by the program committee, along with the confidence intervals. Public reviews, instituted by the author and Nick McKeown for HotNets 2004, are a step in the right direction of making the decision process of the program committee more apparent to the audience, but they do not go far enough. With public reviews, the program committee publishes the rationale for why they accepted each paper, along with an assessment of how the paper fits into the broader research context and the paper's strengths and weaknesses. However, public reviews impose a high overhead on conference program committee, since the public review needs to be carefully authored; as a result, public reviews are unlikely to be done as a regular practice.

Publishing paper rankings and error bars might be accomplished by asking each program committee member to rank order the set of papers they reviewed. The rank and error could be automatically computed from this data. It seems implausble to ask a program committee, struggling to reach consensus on each accept/reject decision, to also reach consensus on the rank ordering of papers. Note that this extra information can align incentives regardless of how value is distributed among papers. If randomness in the review process meant that a program committee's internal ranking was unrelated to a paper's merit, that would be useful for the research community to know. If merit was more uniformly or exponentially distributed instead of heavy tailed, the community could adapt by placing greater or less weight on the rank.

Finally, it seems likely that a program committee's initial assessment of a research paper's merit may bear only a small relationship to its long term merit. This seems inevitable given the time constraints of the conference evaluation system. We might instead turn to journals to be the ultimate arbiter of research quality, but in a fast changing field such as computer systems research, more attention is naturally paid to the most recent results. This leaves most journal papers as "write once, read none," and does not address the issue of heavy-tailed merit, even among journal papers.

To better align long term researcher incentives, specifically to encourage researchers to continue to work on promising research avenues even after the initial publication of the work, we might add a step to re-rank the set of published papers after some period of time has elapsed. In many areas of computer systems research, this is already done through the increasing use of "Test of Time Awards", typically given to the retrospectively most important paper published at a specific conference ten years earlier. Since it is likely that the merit of research papers is still heavy tailed after the test of time, we suspect the research community would benefit from being given a ranking of papers, rather than information about only a single paper.

6 Conclusion

In this paper, we developed a model of computer systems conference publication based on randomness in paper evaluation, heavy-tailed merit, and author and conference incentives. We hope this model will be helpful in explaining to prospective authors the often obscure workings of conference program committees, and in suggesting ways to organize conferences to better align author and conference incentives.

Acknowledgments

The author would like to thank Arvind Krishnamurthy, Eddie Kohler, Frans Kaashoek, Jeff Mogul, Jon Crowcroft, Mike Dahlin, and Derek Murray for their help in assembling the data presented in this paper.

References

[1] A. S. Aiyer, L. Alvisi, A. Clement, M. Dahlin, J.-P. Martin, and C. Porth. BAR fault tolerance for cooperative services. In *SOSP*, 2005.

[2] M. Allman. Thoughts on reviewing. *SIGCOMM CCR*, 2008.

[3] Y. An, J. Janssen, and E. E. Milios. Characterizing and mining the citation graph of the computer science literature. *Knowl. Inf. Syst.*, 6(6):664–678, 2004.

[4] S. Buchegger and J.-Y. L. Boudec. A Robust Reputation System for P2P and Mobile Ad-hoc Networks. In *Second Workshop on the Economics of Peer-to-Peer Systems*, 2004.

[5] B. Cohen. Incentives build robustness in BitTorrent. In *Workshop on Economics of Peer-to-Peer Systems*, 2003.

[6] D. J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[7] G. Ellison. Evolving standards for academic publishing: A q-r theory. *Journal of Political Economy*, 2002.

[8] A. Feldmann. Experiences from the SIGCOMM 2005 European Shadow PC. *SIGCOMM CCR*, 35(3):97–102, 2005.

[9] C. Fine. A mind of its own: How your brain distorts and deceives, 2006.

[10] V. Jacobson. Congestion avoidance and control. In *SIGCOMM '88*, pages 314–329, 1988.

[11] S. Lawrence, C. L. Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999.

[12] R. Mahajan, D. Wetherall, and T. Anderson. Mutually Controlled Routing with Independent ISPs. In *NSDI*, 2007.

[13] V. Petricek, I. J. Cox, H. Han, I. Councill, and C. L. Giles. A comparison of on-line computer science citation databases. In *Ninth European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, 2005.

[14] M. Piatek, T. Isdal, A. Krishnamurthy, and T. Anderson. Do incentives build robustness in BitTorrent? In *NSDI*, 2007.

[15] M. Piatek, T. Isdal, A. Krishnamurthy, and T. Anderson. One hop reputations for peer to peer file sharing workloads. In *NSDI*, 2008.

[16] S. Redner. How popular is your paper? an empirical study of the citation distribution. *European Physical Journal B*, 4(2):131–134, 1998.

[17] S. Shenker, J. Kurose, and T. Anderson. Improving SIGCOMM: A Few Straw Proposals. In *www.sigcomm.org/admin/July2001RepFinal.pdf*.

[18] I. Stoica, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan. Chord: A scalable Peer-To-Peer lookup service for internet applications. In *SIGCOMM*, 2001.

[19] V. Vishnumurthy, S. Chandrakumar, and E. Sirer. KARMA: A Secure Economic Framework for Peer-to-Peer Resource Sharing. In *Workshop on the Economics of Peer-to-Peer Systems*, 2003.

[20] P. Yalagandula and M. Dahlin. A scalable distributed information management system. In *SIGCOMM*, 2004.