Editor’s Message

It is hard to believe it is already July. July marks a few milestones: i) schools are over, ii) most of the paper submission deadlines for the year are behind us, iii) a lot, but not all, of the reviewing duty has been accomplished. July also marks another milestone for CCR and myself: the longest CCR issue I have had the pleasure to edit this year! This issue of CCR features 15 papers in total: 7 technical papers, and 8 editorials. The technical papers cover the areas of Internet measurement, routing, privacy, content delivery, as well as data center networks.

The editorial zone features the report on the workshop of Internet economics, and the workshop on active Internet measurements that took place in early 2013. It also features position papers that regard empirical Internet measurement, community networks, the use of recommendation engines in content delivery, and censorship on the Internet. I found every single one of them thought provoking, and with the potential to initiate discussion in our community. Finally, we have two slightly more unusual editorial notes. The first one describes the experience of the CoNEXT 2012 Internet chairs, and the way they found to enable flawless connectivity despite only having access to residential grade equipment. The second one focuses on the criticism that we often portray as a community in our major conferences, in particular SIGCOMM, and suggests a number of directions conference organizers could take.

This last editorial has made me think a little more about my experience as an author, reviewer, and TPC member in the past 15 years of my career. It quotes Jeffrey Naughton’s keynote at ICDE 2010 and his statement about the Computer Science community - “Funding agencies believe us when we say we suck.”.

Being on a TPC, one actually realizes that criticism is something that we naturally do as a community - not personal. Being an author, who has never been on a TPC, however, makes this process far more personal. I still remember the days when each one of my papers was prepared to what I considered perfection and sent into the “abyss”, sometimes with a positive, and others with a negative response. I also do remember the disappointment of my first rejection. Some perspective on the process could possibly be of interest.

Starting from the basics

The paper submission and evaluation process we have in our community is, at large, standard across conferences. In July 2009, Luigi Rizzo and myself published a CCR editorial note on how the SIGCOMM 2009 TPC functioned. We thought that this kind of note would be useful to students and researchers to understand what is happening in the deep end of our most cherished process. That editorial captured the logistics behind the selection process. Papers get submitted and assigned to a number of reviewers, reviewers enter their comments on the strengths, and weaknesses of the work, and quantify the overall quality of the paper along a number of dimensions, with one particular metric aiming to capture the “overall merit” of the work. That one metric is meant to capture the suitability of the submission for the conference. Sometimes, the numerical score is accompanied by a textual description such as “Strong Accept: top 10% of the papers”, or “Strong Accept: I will argue in favor of the submission during the TPC meeting”. Personally, I find it very difficult to place all submissions in a quality scale, and prefer the second approach, which is essentially what TPC chairs need in order to decide if a paper will go to the TPC meeting or not.

If the paper receives significant support during the reviewing period, it is brought to the TPC meeting. This is where the final decision is taken, based on arguments from both supporters and detractors.

From an objective process to a subjective task

The process of reading a paper and mapping your assessment to a numerical score is necessary in the program selection, but unavoidably subjective. The additional fact that our top scientific venues aim to publish papers that are not only technically correct, but also that address a real problem, that are tested in some fashion to give confidence on the feasibility of the approach, etc. makes this mapping even harder. And it is not unusual for papers to be “weaker” along one or more dimensions. Balancing the importance of that dimension within the overall assessment is a second subjective element in the evaluation.

Getting some perspective

In the past year I have been in the TPC of 4 of the most prestigious conferences in our field. I thought of going back and checking how the objective/subjective process of assigning scores to papers can be observed in the online reviewing system used - HotCRP. Luckily, all 4 systems are still up and running and I could capture all the scores for all the papers that were submitted to the conference, barring the papers that I had a “conflict”. In what follows
I will keep the conferences anonymous - not sure that the names add information to what follows.

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td># submitted</td>
<td>212</td>
<td>210</td>
<td>183</td>
<td>177</td>
</tr>
<tr>
<td># accepted</td>
<td>32</td>
<td>33</td>
<td>32</td>
<td>30</td>
</tr>
<tr>
<td>Acceptance rate</td>
<td>15%</td>
<td>15.7%</td>
<td>17.5%</td>
<td>17%</td>
</tr>
<tr>
<td># conflicts (papers that are not observable by me)</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>Papers with at least one “accept” score</td>
<td>103</td>
<td>103</td>
<td>112</td>
<td>98</td>
</tr>
<tr>
<td>Papers with at least one top score</td>
<td>8</td>
<td>15</td>
<td>39</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 1: Conference Statistics (3 conferences in 2012, 1 in 2013)

All conference review forms were focusing on similar aspects for their evaluation: strengths, weakness, and used a metric “Overall Merit” to capture the final recommendation of the reviewer. That metric ranged from 1 to 5 for C1 and C2, from 1 to 4 for C3, and from 1 to 6 for C4. The range is configurable by the system, and up to the discretion of the chairs. Scores above 3 (for C1 and C2), above 2 (C3), and above 4 (C4) were textually tagged as recommendations to accept.

Table 1 shows some high level statistics across the 4 conferences. The acceptance rates are 15-17%. Interestingly, close to 50% of the submitted papers received at least one positive review! Top scores are, however, harder to receive. Conferences that use 3 different scores (3, 4, 5 or 4, 5, 6) to encode accept recommendations tend to scare reviewers away from the top score. C3 that uses only two scores for accept recommendations appears to be more effective in “pushing” reviewers towards the end of the scale - to the point that some papers with top score were not accepted!

From submission to acceptance

The number of reviews each paper receives depends on how far it advances in the reviewing process. In the studied conferences, papers were reviewed in two rounds. The first round solicited 3 reviews, upon the completion of which some papers were rejected and others continued. Accepted papers received up to 6 reviews.

For each conference I computed the normalized overall merit for each review (since the ranges are different), e.g. scores above 0.5 imply a recommendation for acceptance.

Then I looked at the distribution of normalized overall merit across all the reviews received by all submitted papers. Figure 1 shows the main statistics of that distribution using a boxplot (min, 25th percentile, average, 75th percentile, max).

Figure 1: Boxplot of the distribution of all scores for submitted and accepted papers

- C1 and C2 tend to assign low scores to a large fraction of the submitted papers. 50% of all reviews are less than 0.3. Moreover, they tend to be very frugal with high review scores.
- C3 and C4 still assign low scores to a large fraction of the submitted papers but the distribution is pushed to the higher end.

When we look at the scores received by the accepted papers as a whole - the entire distribution of all the scores in reviews for finally accepted papers, we see:

- C1, C2, and C3, interestingly, accepted a paper that someone from the TPC had marked with the lowest possible score!
- All distributions are shifted to higher scores, but the average score still barely exceeds 0.5, for C1 and C2.

The previous statistics reflect the overall distribution of all the reviews (and different papers receive different number of reviews). Figure 2 presents the boxplots for the average score received by each submitted paper to each conference, and for those that finally got accepted. This kind of result was even more striking to me, and I feel could provide some perspective to graduate students.

Figure 2: Boxplot for the distribution of average score per submitted and accepted paper

- The distribution of average score for accepted papers in C1 is actually very low! The average score of an accepted
paper is below 0.5. The only explanation for this is that there was an advocate for those papers in the TPC meeting; someone that could make an argument as to why this paper belonged in the program despite the shortcomings. Indeed, the TPC meeting itself is fundamental in determining the fate of each submission. This is why it is so important to capture the TPC summary in the comments returned to the authors.

- C3 has the most positive score distribution - a possible artifact due to the use of the smaller scale?
- All conferences accepted a paper with average score below 0.5!!
- The average score of the top paper at each conference was 0.76, 0.84, 0.95, 0.86, respectively.

**Lessons learnt**

So what did I actually learn from this exercise? I was surprised by a few findings: i) 50% of the submitted papers are recommended for acceptance by at least one reviewer! ii) the behavior observed across conferences is not that different, iii) 3 out of the 4 conferences finally accepted a paper that one of the reviewers scored with the lowest possible score, iv) all conferences accepted a paper with average score below 0.5.

My thinking is that the process we follow in the selection of our conference programs works rather well, despite its limitations. The difficult task is to balance the expectations of the average TPC member and what constitutes an appropriate paper for inclusion to the program. The position I have adopted over time is that as long as the paper is not flawed and it makes you think, to me that is an absolute valid contribution.

What does this mean for the newer members of our community? Do not despair. A conference review process is by default an objective, but also subjective process. As you see from the above, even accepted papers often have an “opponent” in a TPC. The norm in review scores is low across all four venues I mention. Proper positioning of your work, with appropriate motivation, and clear discussion on limitations is often what is needed to push a paper in the accepted list.

I hope this helps you obtain a little more perspective on how things work. I am looking forward to your comments on CCR Online.

*Dina Papagiannaki*

CCR Editor