Author Feedback Experiment at PAM 2007

Konstantina Papagiannaki
Intel Research Pittsburgh
dina.papagiannaki@intel.com

ABSTRACT

This editorial article is put together to disseminate the experience gained through the author feedback experiment, performed at the 2007 Passive and Active Measurement (PAM) conference.

Categories and Subject Descriptors

A.m [General Literature]: Miscellaneous

General Terms

Documentation

Keywords

Author feedback, review quality.

1. INTRODUCTION

After the formation of the technical program committee (TPC) for PAM 2007 the TPC chair circulated the idea of performing an experiment in order to assess how authors perceive the feedback generated through the reviewing process. Such an idea was motivated by regular discussions in the TCCC mailing list regarding poor quality in the reviews returned by Tier-1 conferences in the area of computer networking and communication. Such an effect is typically attributed to the fact that TPC members are rarely judged on the quality of the work they perform[1].

The goal of the experiment was twofold. First, the TPC members, that agreed to participate in the experiment, wanted to find out how authors perceive the feedback provided. Second, we wanted to understand how useful author feedback could be in a conference feedback mechanism; a solution that has been previously proposed in the community and is currently supported by EDAS.

2. EXPERIMENT DESIGN

In designing the author feedback experiment we wanted to ensure two things: (i) honest participation by the authors, (ii) confidentiality of the outcome.

In order to ensure honest participation by the authors, we aimed for a mechanism that can ensure author anonymity, i.e. the TPC member would not know which author provided the feedback. Our solution to this requirement was the introduction of a single, trusted entity, that would collect and analyze all received feedback. We thought that given that the TPC chair has complete visibility over the entire procedure, he/she would be a natural choice for the collection and analysis of the author-provided feedback. In addition, feedback was not going to be sent back to the TPC member as received, but aggregated across all the papers the TPC member reviewed.

Upon the decision notification authors were requested to provide feedback on *each* review received. In order for authors to provide honest feedback, the mapping between review score and the paper it corresponds to was visible to the TPC chair alone. The authors did not only provide an overall score but also answered a series of questions that aimed at identifying the reason behind such a score. All scores were studied by the TPC chair. In addition, each TPC member received all scores for all the reviews he/she authored without having the ability to map a specific score to a specific review.

The whole experiment was voluntary for authors as well as TPC members and did not affect the decision made on the paper. TPC members opted out of the experiment by default. Only those that asked to participate received feedback (there was no warning when they were invited).

The individual TPC scores remain confidential and will not be propagated to the next year's TPC chair. The intent of the experiment is to understand author perception of the reviews provided by TPC members and not to stigmatize the TPC members.

2.1 Questionnaire

The contact author for each paper was asked to fill in the following questionnaire for each review they received:

Reviewer N:

1. Please rank the review: (Best) 1 2 3 4 5 (Worst)
2. Was the review feedback useful in terms of helping you identify problems?
3. Did the reviewer understand the work?
4. Did the review have any factual errors?
5. Was there sufficient justification behind the reviewers' decision?
6. Did the reviewer simply use "easy" comments in order to reject the paper (please answer if paper was rejected)?
7. Please add any other comments here. These comments will be seen by the TPC chair and will NOT be forwarded to the reviewer.

[1] This problem was also the focus of [3].

In retrospect, the aim at anonymity may have limited the usefulness of such an experiment. A final question asking whether the author agrees with the feedback being forwarded to the TPC member as is, would have been useful. It is not clear at this point how much exposure of the author's identity would affect the participation in the scheme or the honesty of the review.

2.2 Potential issues with selected design

- The authors may decide not to participate in the experiment at all.

  We believe that such a danger was minimized given that author anonymity was preserved.

- The authors may be biased in their assessment depending on whether the paper was accepted or rejected.

  This question is studied based on the feedback received later in this report. Each author was asked to provide a score for each individual review he/she receives, so perhaps biases are not going to be the dominant factor in the authors feedback. Nonetheless, we intend to study the correlation of author score and the outcome itself.

- Authors of accepted papers may not participate in the scheme at all.

  Overall we received feedback from 47 (out of the 80) authors, thus providing a sufficient sample for further analysis. In contrast to what we originally expected, the majority of the responses came from authors of accepted papers (19 out of the 21 accepted full papers, and 10 out of the 11 posters). Participation was not so high for authors with rejected papers; only 18 out of 48 responded. Therefore the author feedback mechanism is hardly a way for authors to "take it out on" the reviewers. In most cases (almost 60%), authors of rejected papers did not even respond.

- The anonymized feedback provided back to the TPC member will be of little value.

  In order to preserve author anonymity TPC members were forwarded the answers to questions 1 to 6 for their reviews in an anonymized fashion, i.e. all feedback simultaneously without an indicator to the paper in question. On one hand, this kind of information is still more than the information currently available to TPC members. On the other hand, unless all feedback is positive or negative, it is hard to tell how much a TPC member can learn from it. Indeed, what we found out is that preserving author anonymity definitely limited the usefulness of such feedback. A non-negligible number of authors provided comments in response to Question #7 that would definitely help the TPC member understand the reasons behind particular scores.

2.3 TPC reaction

The reaction throughout the TPC after the dissemination of the author feedback was lukewarm. Some TPC members tried to make some sense out of the scores correlating the different responses to Questions 1 to 6. Some others thought that the provided feedback did not help them at all. They thought that such feedback could only be meaningful if they could map scores with reviews and see the responses to question #7. Finally, some of the TPC members from academia paralleled the collected feedback to faculty evaluations within Universities. Previous studies of such feedback showed that they are only useful to identify the exceptional and problematic cases, but cannot offer an objective, quantitative measure of quality (in teaching in this particular case). The analysis of Student Evaluations of Faculty (SEF) has been the topic of [1, 2].

Finally, because such feedback remains confidential and is not propagated to the next TPC chair, some TPC members also doubted the potential of such a scheme to motivate TPC members to put more effort into their reviewing. The fundamental problem with propagating such information is that it needs to somehow be validated beforehand. And in this particular case it is hard to even derive the ground truth. Other disciplines use schemes by which each TPC member needs to grade each review submitted for the papers he/she reviews, without being aware of the identity of the TPC member behind them. This happens anonymously and the feedback is propagated to the TPC chair for dissemination to next year's chair. Even though such a mechanism does not capture author feedback, it may be beneficial in identifying TPC members that may be underperforming. Interestinly enough, though, even expert evaluations have been shown to be problematic in faculty evaluations [2]. According to [2], *"Other methods of evaluating teaching effectiveness do not appear to be valid. Ratings by colleagues and trained observers are not even reliable (a necessary condition for validity) – that is, colleagues and observers do not even substantially agree with each other in instructor ratings"*.

3. ANALYSIS

3.1 Participation

Participation in the experiment was voluntary both for authors and TPC members. All but 2 TPC members participated in the experiment (19 out of 21). We received feedback for 47 out of the 80 submitted papers. In what follows we will process the feedback provided by 43 authors. The remaining 4 papers were reviewed out-of-band due to a conflict of interest of the TPC chair. As a result, the TPC chair had no information on who reviewed those papers.

3.2 Processing of feedback

There are two different types of analysis that need to be performed on the feedback received by the authors:

- We need to assess the validity of the collected feedback. Do the results actually make sense? Potential worrisome feedback would rank a reviewer very high but state that the reviewer mis-understood the work.

- We compute statistics that could quantify the quality of the individual TPC members (anonymized) and the TPC as a whole.

More specifically, we tested the following:

1. Distribution of review scores for accepts and rejects. Is the review feedback highly correlated with the outcome of the reviewing process?

2. Correlate the review scores with the number of words per review. Does the length of a review influence the decision of the authors?

3. Study the deviation in the scores provided by the same author on the same paper. Such an analysis could help us understand whether authors grade each review on its own merit and whether they are significantly influenced by the decision on the paper.

4. Average and median score per reviewer.

5. Drill down into the justification questions and see whether the scores make sense.

4. RESULTS

During the reviewing phase PAM TPC members were asked to rank a paper as A, B, C, or D. The interpretation of these scores are:

- A: I will champion this paper at the PC meeting (Advocate/Accept).
- B: I can accept this paper, but I will not champion it (accept, but could reject).
- C: This paper should be rejected, though I will not fight strongly against it (reject, but could accept).
- D: Serious problems. I will argue to reject this paper (Detractor).

For ease of labeling, in what follows we denote the different categories as "Strong Accept", "Weak Accept", "Weak Reject", and "Strong Reject" respectively.

4.1 TPC score and author review score

Figure 1 shows the correlation of the author-provided TPC score and the TPC member's recommendation (A to D).

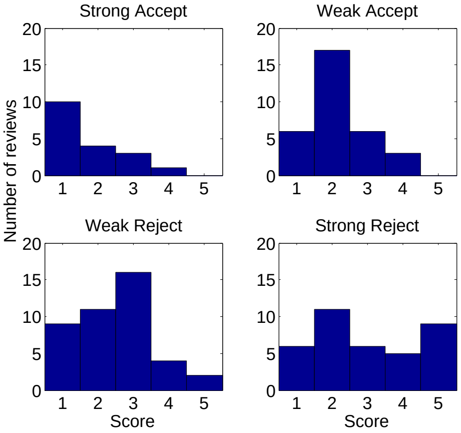

Figure 1: Correlation of TPC score and author review score

Interpretation: The score of the review is loosely correlated with the score given by the TPC member. One notices that if the TPC feedback is positive then authors are more likely to offer positive feedback, but there are cases where authors have ranked a review low while the TPC score was high and they have ranked a review high while the TPC member recommended strong rejection. Interestingly, the feedback received on papers that were recommended for "strong rejection" appear to cover the entire range of scores.

5. AUTHOR FEEDBACK AND REVIEW WORD COUNT

Figure 2 shows the correlation between the author-provided feedback and the number of words in the TPC member's review.

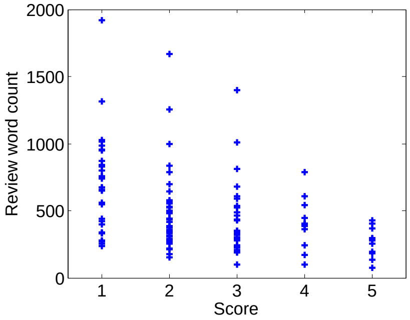

Figure 2: Correlation of author feedback and review word count

Interpretation: Author feedback scores and review word count appear to be correlated. However, content appears to also be important since reviews between 300 and 700 words can be ranked anything between 1 (best) and 4.

6. AUTHOR FEEDBACK AND FINAL OUTCOME

Figure 3 presents the author provided feedback categorized according to the final decision on the paper. Notice that each paper is contributing 3 different author scores (1 for each review).

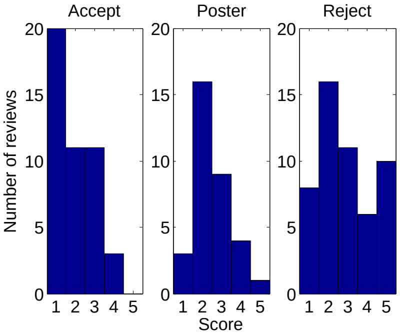

Figure 3: Author feedback scores for papers accepted as full papers, posters, and rejected

Interpretation: Accepted papers tend to receive high author feedback scores. Rejected papers receive the full range of author scores. This implies that author feedback is slightly affected by the outcome of the paper. However, interestingly, authors of rejected papers do recognize the reasons behind the reviewer's recommendation. Poster papers which faced more criticism during the reviewing process received feedback scores in the center of the scale.

7. REVIEW SCORES PER PAPER

Figure 4 shows the different scores provided by each author on the three different reviews.

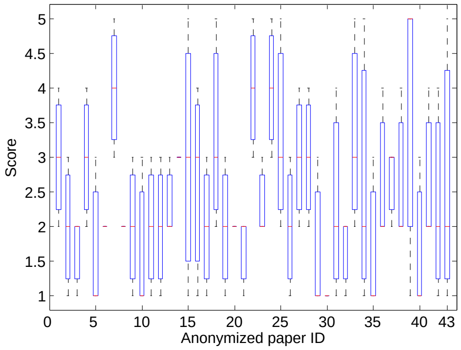

Figure 4: Whisker plot of review scores per paper (anonymized ID)

Interpretation: Authors appear to be ranking each review on its own merit leading to great diversity in the scores provided for individual reviews. Very few papers show small deviation in the review scores provided.

8. STATISTICS PER TPC MEMBER

Figure 5 shows the average and median score received by each TPC member.

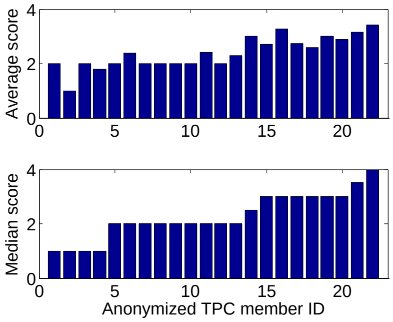

Figure 5: Average and median review score per TPC member (anonymized)

Interpretation: 70% of the committee received excellent(!) author scores (1 to 3). This particular metric would be interesting for comparisons between conferences. In addition, we observe that median score may actually be a little more meaningful in this context. The large majority of TPC members had one outlier score in the feedback they received.

9. LOOKING THROUGH THE REST OF THE FEEDBACK

9.1 Was the feedback useful?

Figure 6 presents the responses received to Question #2, and their correlation to overall review score.

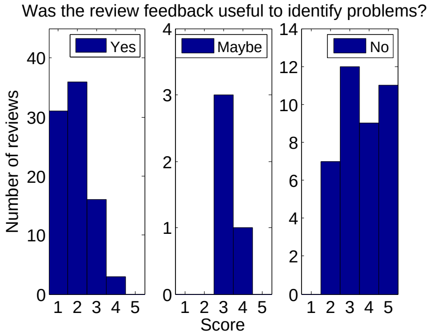

Figure 6: Was the feedback useful?

Interpretation: When the feedback is said to be useful, review scores concentrate to 1s and 2s. However, there is a small number of authors that found the feedback useful and still ranked the review as a 4. Negative (3-5) review scores tend to correlate with negative answers as to whether the review helped the authors identify problems. Authors tend to rank reviews low when they do not see ways in which the review helps them identify issues with their work. A number of authors answered “maybe”, associating a 3 or 4 score.

9.2 Did the reviewer understand the work?

Figure 7 shows the responses received to Question #3, and their correlation to overall review score.

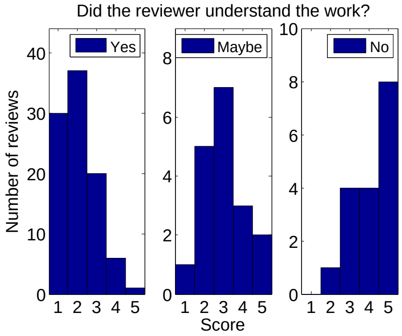

Figure 7: Did the reviewer understand?

Interpretation: Positive reviews are accompanied by answers that state that the TPC member appears to have understood the paper. Negative scores are accompanied by complaints about the TPC member not having understood the work. There is one author that ranked a review with a “2” even though he thought that the TPC member did not understand the work. There are authors that did rank a review low even though they thought the TPC member understood the work: this is to be expected because quality of feedback should not necessarily correlate with whether one understood the work. There was a number of authors that responded to this "Yes/No" question with a "Maybe". Scores associated with such reviews tend to concentrate in the more neutral score region, peaking at 3.

9.3 Did the review have factual errors?

Figure 8 shows the responses received to Question #4, and their correlation to overall review score.

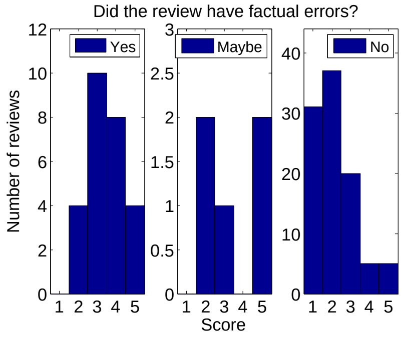

Figure 8: Did the review have factual errors?

Interpretation: When an author thought the review had factual errors, he/she tended to offer poor scores to the TPC member; even though we saw 4 cases where the review was ranked with a 2! (Apparently some authors are not as well calibrated as others.) When reviews were assessed to not have factual errors they were highly ranked. Notice that the majority of the reviews provided by the committee were thought to be accurate. Again a small number of authors responded with a maybe.

9.4 Was there enough justification behind the reviewer's decision

Figure 9 shows the responses received to Question #5, and their correlation to overall review score.

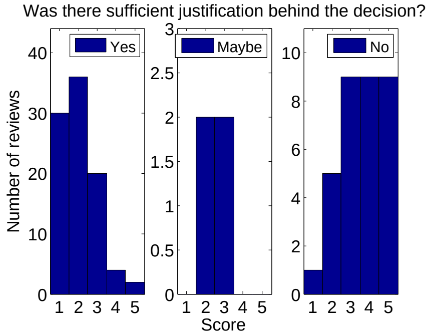

Figure 9: Was there enough justification behind the reviewer's decision?

Interpretation: Here we see a rather consistent trend. Low author provided scores tend to correlate with the fact that the authors believe that the TPC member did not offer enough justification behind his/her A/D ranking. However, one can still observe outlier behavior where authors ranked a review with a "5" even though they thought that the review contained enough justification.

9.5 Did the reviewer provide easy comments to reject the paper?

Figure 10 shows the responses received to Question #6, and their correlation to overall review score.

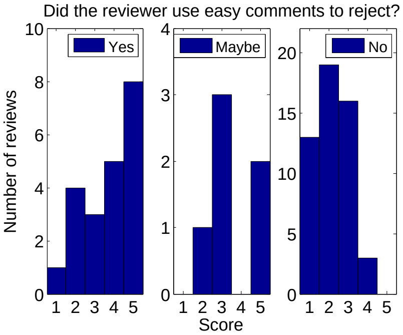

Figure 10: Did the reviewer provide easy comments to reject?

Interpretation: This question was conditional on the paper being rejected. A "Yes" answer to this question correlated well with poor feedback scores; even though a number of authors graded the review high despite providing easy comments to reject. When the answer was "No" feedback was primarily positive.

9.6 Additional comments by authors

One of the nicest features of the author feedback experiment was the unsolicited comments that authors left in justification of their scores and responses (either to Question #7 but also throughout). Reading those comments, one realizes that some authors actually put significant thought into their responses. Some sample comments from papers that received a reject decision follow:

> "Yes. The reviewer was very clear and explicit in his thoughts and his reasoning. He detailed pros and cons clearly. Such kind of reviews are very well appreciated."

> "Excellent comments which touched on the real problems. I would very much appreciate if you can kindly ask the reviewer for the permission to give me his/her name and contact. I sincerely wish to have more conversations with him/her."

On the opposite end of the spectrum. We also received comments such as:

> "It seems that he did not read the paper. His comments are totally untrue and impertinent (some are ridiculous)."

> "This first review is very short. The reviewer does not argue or detail its point of view. This is not helping me to improve the quality of my works."

Such comments are even more interesting to read when they come from the same author in response to the three individual reviews. In such cases, one can see that the authors did see a difference in the feedback provided and acknowledged it in their comments.

10. SUMMARY

In summary, we believe that one can draw the following conclusions from such an experiment.

- Authors are not necessarily calibrated! A TPC member can calibrate his/her scores by looking at tens of different papers. The perception of what constitutes a good review, after a sample of 3, is a much harder task.
- Authors do not appear to be strongly biased by the TPC decision on their paper. In particular, while one would assume that poor author scores would correlate strongly with a reject decision, we found that not to be true. However, one can notice that when the paper is accepted, then the probability for positive author feedback tends to be higher.
- In terms of the usefulness of this experiment to the TPC member themselves, we found its value significantly limited due to the requirement for double anonymity.
- Finally, while author feedback may be useful in pinpointing extreme cases, such as exceptional or problematic reviewers, it is not quite clear how such feedback could become an integral part of the process behind the organization of a conference. Probably, borrowing techniques from other fields, where such an assessment is part of the TPC member's duties, would be a better idea. Under such alternative schemes, TPC member performance would be ranked by the rest of the committee, hopefully offering a less biased outcome (one could argue both ways here).

We sincerely hope that our findings will be interesting to the community and we would be very interested in hearing any feedback regarding this report.

Disclaimer: This editorial was edited by the PAM 2007 TPC chair, Dr. Konstantina Papagiannaki, and may not represent the views of each member in the PAM 2007 Technical Program Committee.

The PAM 2007 Technical Program Committee

- Mark Allman, ICSI
- Suman Banerjee, University of Wisconsin, Madison
- Ethan Blanton, Purdue University
- Nevil Brownlee, University of Auckland
- Mark Claypool, WPI
- Christophe Diot, Thomson Research
- Christos Gkantsidis, Microsoft Research, Cambridge
- Gianluca Iannaccone, Intel Research Berkeley
- Balachander Krishnamurthy, AT&T Research
- Simon Leinen, SWITCH
- Bruce Maggs, CMU/Akamai Technologies
- Ratul Mahajan, Microsoft Research
- Alberto Medina, BBN Technologies
- Konstantina Papagiannaki, Intel Research Pittsburgh
- Lili Qiu, University of Texas, Austin
- Coleen Shannon, CAIDA
- Peter Steenkiste, CMU
- Steve Uhlig, Delft University of Technology
- Jia Wang, AT&T Research
- David Wetherall, University of Washington
- Tilman Wolf, University of Massachusetts, Amherst

11. REFERENCES

[1] A. G. Greenwald and G. M. Gillmore. Grading leniency is a removable contaminant of student ratings. *American Psychologist*, 52:1209–1217, 1997.

[2] M. Huemer. Student evaluations: A critical review. http://home.sprynet.com/∼owl1/sef.htm.

[3] S. Keshav. Reviewing the reviewers, September 2006. http://keshav-essays.blogspot.com/2006/09/reviewing-reviewers.html.