Report on The Eurocrypt 2021 Review Process

Anne Canteaut[*] François-Xavier Standaert[†]

June 14, 2021

Abstract

We report on the Eurocrypt 2021 review process. We start with a list of high-level goals we pursued and follow with a description of the strategies we implemented to try ensuring them, for the different steps of the review process. We finally discuss the advantages of these strategies and the challenges they raise (as we perceived them), with suggestions for future PC chairs. We hope this document can also help authors understanding how their paper was evaluated.

1 High-level goals

As program chairs, our primary goals were to build a balanced program based on the best papers submitted to the conference and to ensure equal standards for all the submitted papers. Yet, in view of the continuously increasing competition in the main IACR venues, a secondary goal was to try making the review process less adversarial. Informally, we wanted to avoid "*strong reject recommendations due to subjective (non-technical) reasons*". More precisely, we wanted to:

1. Decorrelate the more objective (e.g., technical correctness, editorial) parts of the reviews from the more subjective (e.g., scientific interest or scientific quality) parts of the reviews;
2. Connect the recommendations made by the reviewers to the parts of the review that motivate them, leading to an easier interpretation of the final decisions to the authors.

2 Assembling the Program Committee

The first task of program chairs is to assemble a Program Committee (PC), which turns out to be challenging. A first reason is that a good PC should be able to review submissions on many different topics with confidence. For this purpose, we relied on the IACR main areas (CHES, FSE, PKC, RWC and TCC) and tried having PC members coming these different fields. But the distribution of the submissions per topic is typically unbalanced. For example, we plot such a distribution for Eurocrypt 2021 in Figure 1. We give it for all submissions (which can span multiple topics) and for single-topic submissions. It illustrates that more applied/implementation topics like CHES and RWC usually have some multi-disciplinary nature (reflected by their reduced proportions when moving to the single-topic plot). The same seemed to hold for PKC papers, while an opposite trend held for FSE and TCC. Without such data available a priori, we had to trade the need to deal with an expected higher number of theoretical submissions (which we could infer from past Eurocrypt programs) and our goals to build a balanced program and to have all papers treated with the same

[*] Project-team COSMIQ, Inria, Paris, France.
[†] Crypto Group, ICTEAM Institute, UCLouvain, Belgium.

standards (e.g., regarding the need to rely on subreviewers). For this purpose, we chose 10 PC members per topic and 20 for the TCC topic, trying to prioritize PC members with expertise on multiple topics. We further tried to avoid multiple PC members with the same affiliation, to select PC members mixing junior and senior researchers and to favor gender and geographical balance.

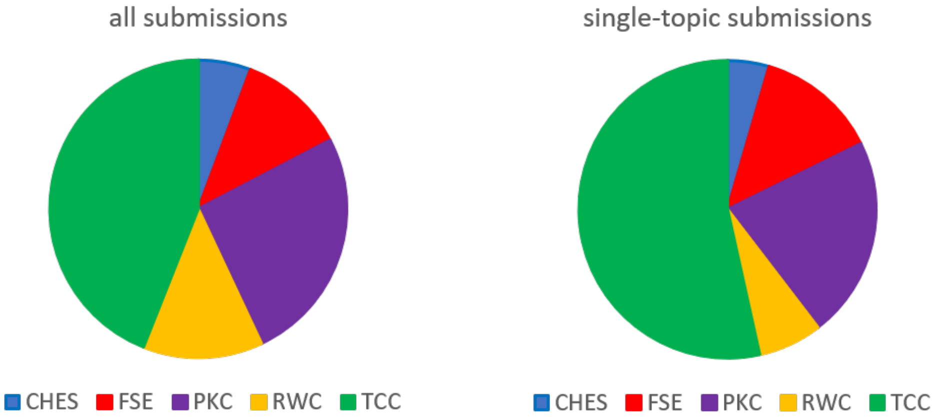

Figure 1: Topic distribution of the Eurocrypt 2021 submissions.

The second reason is that in view of the increasing number of papers submitted to the main IACR venues, the review workload also increases and so does the amount of researchers who decline the invitation to the PC (for admittedly good reasons: most of them declined because they were just done with another PC). For illustration, we contacted 93 researchers for a 59-member PC.

With 400 submissions, Eurocrypt 2021 was no exception. So on the one hand, we can only insist on how much we are indebted to our PC for their professionalism and the time they devoted to reviewing papers. On the other hand, *the (average) 20-paper per PC workload also questions the need to further incentivize or reward the PC work*, which we leave as a question to the IACR.

3 Review form and guidelines

We modified the review structure as follows in order to try reaching our high-level goals:

A. Paper summary. Give a succinct and positive description of the paper's main contributions.

B. Suitability. Does the paper belong to the conference (yes/no)?

C. Novelty, methodology and technical correctness:
Question 1. Does the paper ignore related works that overlap with the claimed results?
Question 2. Is the methodology used in order to answer the research question and to demonstrate improvements over previously published state-of-the art solutions appropriate?
Question 3. Are there technical flaws that affect the correctness of the claims?
Try to precisely identify the overlapping related works and the methodological/technical flaws.

D. Editorial quality and technical details:
Question 4. Is the editorial quality of the paper sufficient to understand the contribution?
Question 5. Are there technical details missing to verify the contribution?
Use this section to add any technical or editorial comment you want to send to the authors.

E. Scientific quality. (In case answers to questions Q1 to Q5 are sufficiently positive).
Question 6. How do you rate the scientific importance of the research question?
Question 7. How do you rate the scientific contribution to this research question?
Note that there are no universal rules for evaluating scientific quality, and each reviewer is entitled to her or his own view. Try to motivate your opinions based on your specific field of research and whether you would be interested to listen to a talk on the paper content during the conference.

F. Confidence level:
1. Weak: an educated guess.
2. Medium: quite confident but I could not check many of the details.
3. Good: I know the area and I studied the paper in sufficient detail.

G. Recommendation:
1. Strong reject (novelty, methodology or correctness issues).
2. Weak reject (editorial quality requires improvement or technical details are missing).
3. Borderline (the research question/result is deemed of limited interest by the reviewer).
4. Accept (the paper is found to improve the state-of-the-art on an important research question).
5. Strong accept (breakthrough paper, best paper award candidate).

H. Comments to the PC. This part will always remain hidden to the authors.

The rationale behind this review form is twofold. On the one hand, it aims at *encouraging the reviewer to focus first (and as long as possible) on the technical/editorial aspects of the submission*, since this is the part of the review which usually contains constructive comments for improving the paper. In other words, our goal was to postpone the more subjective discussion about whether the topic of the paper or its result are deemed important. On the other hand, it aims at *connecting the recommendations with a part of the review.* Typically, a strong reject means Part C is negative, a weak reject means Part D is negative and a borderline means Part E is negative.

As for more general guidelines, we asked reviewers to write reviews that they would like to receive as authors, remembering that these reviews can fall in the hands of a first-year PhD student; to be concrete in the parts of the review related to technical and editorial issues; to be respectful and inclusive in the parts of the review related to scientific quality (Eurocrypt is a general crypto conference covering all theoretical and practical aspects of cryptology). We finally suggested to pose explicit questions whenever in doubt, so that they can be clarified during rebuttal.

4 Steps of the review process

We now detail the steps of the review process and the challenges we encountered in their application.

4.1 Paper bidding and Conflicts of Interest

Assigning 400 papers to 59 reviewers inevitably requires some automation. We relied on the HotCRP review system to assist us with this task, which comes with two limitations:

1. The default COI management by HotCRP is approximate for PC members (we removed many fake COIs manually) and missing for subreviewers (each PC member had to contact us in case of doubts and some conflicts were only spot close to the deadline for reviews).

2. HotCRP does not provide any visualization of assignments. As a result, it is difficult (time consuming) to check whether each PC member has received a majority of papers he is interested in, and that each paper will be reviewed by at least one confident PC member.

Enhancing the COI management so that it becomes less error-prone and can automatically spot conflicts with subreviewers would be of great help. For now, removing manually the fake COIs for 400 papers is obviously infeasible for the PC chairs, and this cannot be handled by the PC members since the submissions are anonymous. Therefore, the COI management mainly relies on the information given by the authors. *It is then of utmost importance that the authors carefully check the default COIs proposed by HotCRP and remove any fake COI from the list.*

We noticed that a few (sub)reviewers were not completely aware of COIs for closely related technical works (`https://www.iacr.org/docs/conflicts.pdf`). In this respect, we note that reviewing papers on topics too close from ones' submission is not only ethically disputable, it can also prevent having comparative discussions that all reviewers can join, which is detrimental to all authors. Based on our experience, we believe it is beneficial to be quite conservative regarding this type of COI, especially if the PC is broad enough to completely avoid them.

Besides, enhancing the assignment procedure in HotCRP and integrating visualization features would be helpful. For now, we relied on external programs shared by Gaëtan Leurent, which were useful to guarantee that each PC member was assigned (a sufficient number of) papers she/he could review with confidence, and each paper was assigned to at least one confident PC member.

4.2 Pre-rebuttal discussions

We used a two-week period to prepare the rebuttals. Our main goal during these weeks was to identify the papers for which rebuttal was critically needed in order to clarify technical or editorial points typically related to the novelty or the correctness of the results. The papers for which reviews were positive excepted for those technical or editorial doubts were of particular interest here. For the other papers, we mentioned that "*rebuttal was not identified as critically needed for this paper but authors are welcome to react if they want to*" to make explicit that the rebuttal is optional.

In view of the massive amount of papers in the loop at this stage of the process, we used area chairs to assist us with it. It enabled us to tag the papers needing a rebuttal within a few days, so that PC members could easily spot them and help determining the issues to address in the rebuttal. We assigned a discussion leader to each paper in order to summarize the rebuttals' requirements, trying to ensure that each discussion was led by an expert of the papers' topics.

We also took advantage of the two weeks before the rebuttal to identify papers for which we lacked confidence. For most of them, we looked for an additional opinion after the rebuttal and, in case this new opinion changed the reviewers' views, allowed the authors to react. The rebuttal was the only communication channel between the authors and the reviewers: we did not allow other interactions in order to ensure equal standards for all the submitted papers.[1]

We then managed the post-rebuttal discussions and the selection of accepted papers in two phases:

1. Rebuttal updates and reject decisions due to objective reasons.

2. Iterative acceptance of the best papers by piles of 25.

4.3 Rebuttal updates and reject decisions due to objective reasons

As a first phase, we asked the reviewers to read the rebuttals of the papers they were assigned, and to check whether these rebuttals and discussions with other reviewers modified their judgment. We then asked the PC to try reaching consensus on one of the following two options:

[1] For a similar reason, we did not communicate revised versions to the reviewers at any stage of the process.

a. The paper remains with technical or editorial issues and reviewers do not think it is ready for this years' Eurocrypt (i.e., it requires more than a few days of work to be in a publishable shape).

b. The paper is correct and can reach a sufficient editorial quality in its final version.

Papers (a) were then tagged "maybe reject" for a buffer period, before being tagged "reject". They correspond to papers rejected for mostly objective (e.g., technical or editorial) reasons. Papers (b) remained in the loop during the next phase of the discussion. They correspond to papers for which the (sub)reviewers did not find strong (i.e., objective) reasons to reject.

It is important to note that Papers (a) contain submissions rejected for different objective reasons. Papers with a majority of strong reject recommendations indicate fundamental errors so that the submission should not be re-submitted without substantial corrections and revisions. Papers with a majority of weak reject recommendations rather indicate that the submission lacks a minimum level of editorial quality or technical details so that the reviewers could not be convinced by the result. Hence, a re-submission clarifying these aspects may lead to a different outcome.

We also mention that we made the reviews accessible to the subreviewers, while the decision to make comments visible to the subreviewers was left at the discretion of the PC members.

4.4 Iterative acceptance of the best papers by piles of 25

Iterative Process. Assuming that the first phase of the post-rebuttal discussions led us to detect the papers that had to be rejected for objective reasons, we reversed the priorities in the second phase of the discussions. This time, we asked the PC to look at the papers from the top of the submissions, and to iteratively identify the best papers in their assignments. We repeated this operation three times, with the goal to select the top 5 to 10% papers, the top 10 to 15% papers and the top 15 to 20% papers. Each time, we let two weeks to the reviewers to discuss these papers and proposed a pile of around 25 submissions to be accepted. We then let another week to further discuss these decisions and any PC member could object by checking a "do-not-accept" box.

Among the strategies we that we used for processing the submissions, we rate the selection of papers by piles (rather than individually) positively. In several cases, it allowed the PC to spot inconsistencies (i.e., papers that were tagged "accept" despite seemingly weaker than other papers still in discussion). Piles of approximately 25 submissions appeared well dimensioned for this purpose (since easily browsable by the PC members). We also used comparative discussion threads in several cases, and ensured that similar submissions were reviewed by on common reviewer.

On our side, we managed the whole process by examining together all submissions in the same area. Our rationale was that it is much easier to have a clear view of the respective merits of around 100 submissions on similar topics than to try directly comparing all submissions. We used the IACR main areas for this purpose which appeared well dimensioned and are also easy to handle in practice since HotCRP allows the PC to easily monitor the submissions by topic.[2]

As an indication, the top 5 to 15 % papers were easy to identify: they correspond to papers for which there was a clear consensus that they should be in the program (directly reflected by a majority of accept and strong accept grades). Selecting the last third of the program was more challenging, since based on the subjective parts of the evaluation of different (sub)reviewers.

[2] This requires that the topics to be chosen by the authors is defined by the chairs before the submission deadline.

Selecting the last 25 papers. For this last part of the program, we re-insisted to the PC that there are no universal rules for evaluating scientific quality, and that each reviewer is entitled to her or his own view. We asked reviewers to try motivating opinions based on their specific field of research, and whether they would be interested to listen to a talk on the paper content during the conference, with the hope to build a program that each PC member would be willing to attend.

We additionally made the following general suggestions to the PC members:

(i) To have a regular look at the papers that are already tagged as “accept” or “maybe accept” to get an idea of how the preliminary program looks like – since in case several papers end up with similar quality, ensuring a certain diversity of topics could serve as a tie breaker;

(ii) To remember that the final acceptance rate will be below 20 (so despite the quality of the assignments may differ, to consider that a significantly higher or lower acceptance rate for one reviewer may indicate that she/he was more or less generous than other PC members);

(iii) That we should evaluate the submissions, not potential updates that appeared on ePrint.

We finally mentioned shepherding as an option for strong papers with minor technical/editorial issues. But given the large amount of high-quality papers submitted, we chose to use it scarcely.

For this part of the process, we gave more weight to the discussions (compared to the grades) as we were moving to the selection of papers based on subjective criteria. We also gave priority to papers with strong support without consensus over papers with weak support and consensus.

We additionally monitored the program balance in an indicative-only way. Precisely, we checked that our acceptance rates did not significantly deviate for the different research areas (in the same way as we asked PC members to check that their personal acceptance rate was not significantly deviating from the global one). It turned out the recommendations of the PC expressed in the discussions were quite balanced and we only used the topics as a tie-breaker for a couple of papers. Figure 2 shows the proportion of accepted papers per topic and Figure 3 shows the success rates per topic, confirming that the final program reasonably reflects the received submissions.

We hope these figures can strengthen the motivation of authors from all the IACR research areas to submit their best works to future venues, in order to further improve program balance.

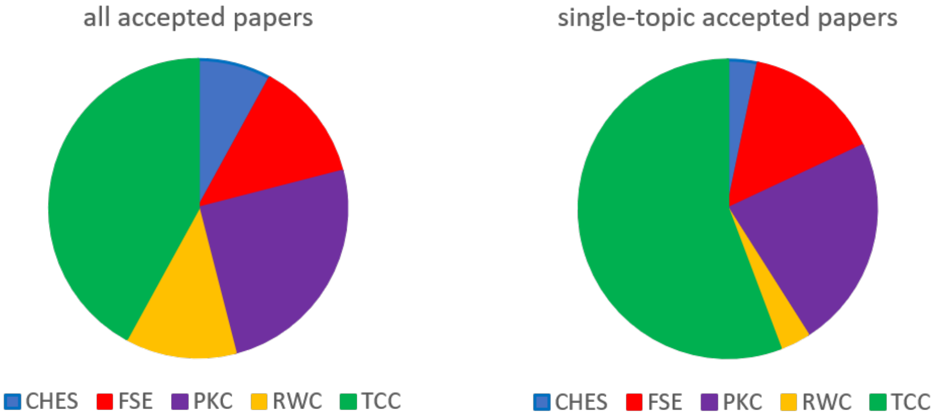

Figure 2: Topic distribution of the Eurocrypt 2021 accepted papers.

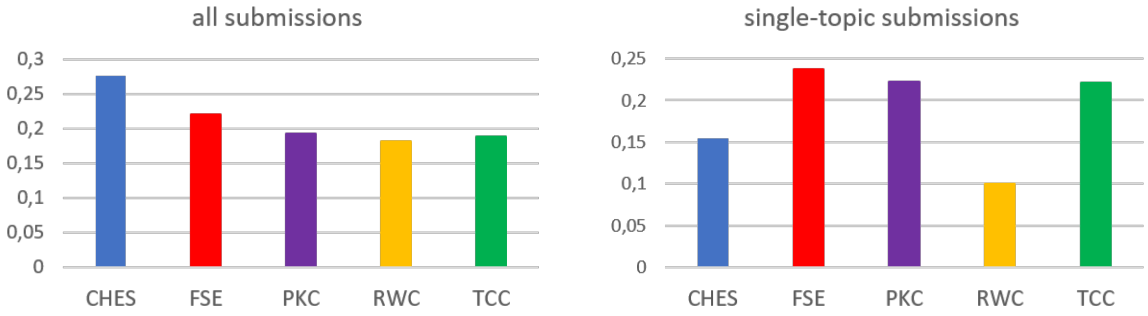

Figure 3: Success rates of the Eurocrypt 2021 accepted papers.

We finally add a few words on the interpretation of the submissions rejected with mostly "accept" or "borderline" recommendations. In the first case, it indicates that the paper could appear in the conference and the only reason it did not is the limited number of slots. So re-submission addressing the reviews is encouraged. The second case (i.e., a majority of "borderline" recommendations) rather indicates a (subjective) lack of interest of the PC members who reviewed the paper. In case it is a first submission, re-submitting to a different PC may be worth trying. If the paper has already been submitted multiple times, it is advisable to try a less competitive venue. We note that it may be that that papers with a majority of "weak reject" recommendations (with a clear list of improvements to implement) have a better chance to be accepted to a next IACR venue than papers with a majority of "borderline" recommendations (since a subjective lack of interest of the PC members who reviewed the paper may be harder to fix). This happened at Eurocrypt 2021 where some submissions were rejected due to editorial reasons despite a very positive appreciation of the scientific result (which would have received a revision recommendation in a journal).

4.5 Reviews updates and interpretation

The last step of the review process is to update the reviews based on the discussions that motivated the decisions. This final step is important for authors of rejected papers, so that they can decide about their next plans. It is even more important given the structure of the review form and the phases of the discussions we used, which had a tendency to make the reviews more positive and to focus the discussions on the top papers (i.e., selecting by the top rather than from the bottom). Both effects may lead more authors to believe their paper was close to acceptance, which therefore makes the final explanations more critical (which we discuss in the conclusions of this report).

Concretely, this final phase is also quite difficult to implement. It comes directly after the last (exhausting) discussion phases and it is therefore harder to motivate discussion leaders to spend time on this additional effort (after, we insist again, tremendous work for several months already). As PC chairs, this is also a moment where we are quite constrained in time, since authors usually like to receive the feedback on their paper quite soon after the decisions (e.g., in case a re-submission is foreseen). We did not find an ideal solution to deal with these conflicting goals.

5 Conclusions

As far as our high-level goals are concerned, and despite further refinements are always necessary, we feel encouraged by the feedback we received about the changes we made to the review form, and the overall review process. We think the review structure contributed to write positive/constructive feedback (whenever this structure was actually followed, which was not always the case). As already mentioned, we believe organizing the selection of the papers from the top, by multiple (small enough) piles, also helped us (and the PC) to spot when comparative discussions were needed. We insist again that being conservative with respect to COIs for closely related technical works is quite helpful in this respect: it allows more expert reviewers to contribute to the discussions.

Overall, selecting papers with the diversity of topics of Eurocrypt end ensuring equal standards remains an important challenge. Beyond the organizational aspects, our feeling is that the opportunity to rely on a PC with broad expertise is the most important ingredient for this purpose. It is in particular critical to ensure that the discussion of each paper is lead by an expert PC member and, ideally, that it is not only based on subreviews that may be less actively followed. Not knowing the distribution of the submissions' topics in advance makes this goal difficult to optimize, especially if aiming to keep a similar workload for each PC member. Working with even larger PCs may become necessary if the number of submissions to Eurocrypt keeps on growing.

From the feedback we received from some PC members, we conclude that contributing to the selection of the Eurocrypt 2021 program was quite time consuming: first because of the heavy review workload, but also because of the quite demanding discussion process. Another less positive outcome of this process is that despite our goal to make the interpretation of the final decisions easier to the authors, the more positive comments that sometimes came without a final update of the reviews to explain the decisions have lead some authors to be even more surprised (and disappointed) by this decision. We hope this document can mitigate their surprise a posteriori.

We illustrate this last concern with Figure 4 which represents the proportion of papers that were accepted, rejected for objective reasons (i.e., with a majority of "strong reject" and "weak reject" recommendations) and rejected for subjective reasons (with a majority of "borderline" or even "accept" recommendations). Clearly, it would have been possible to double the program size with papers having sufficient technical and editorial quality that were either positively appreciated or rated borderline by the PC. We believe this last figure makes a case for ongoing discussions within the IACR about the need of a journal where the acceptance would be less driven by competition (i.e., a limited number of slots). This is especially true for papers in the borderline category, which deserve acceptance based on the comments and may lack the final twist to make reviewers fully enthusiast in one of the main IACR venues, therefore increasing the number of re-submissions.

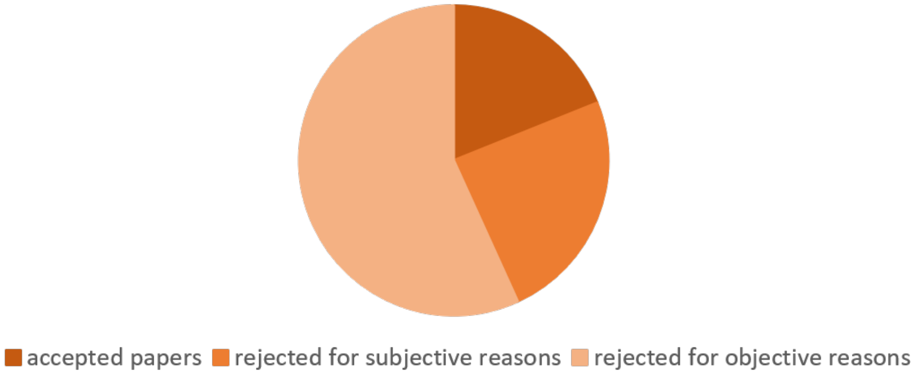

Figure 4: Fraction of papers rejected for subjective reasons.