"Flawed, but like democracy we don't have a better system": The Experts' Insights on the Peer Review Process of Evaluating Security Papers

Ananta Soneji*, Faris Bugra Kokulu*, Carlos Rubio-Medrano†,
Tiffany Bao*, Ruoyu Wang*, Yan Shoshitaishvili*, Adam Doupé*
*Arizona State University, †Texas A&M University - Corpus Christi
*{asoneji, fkokulu, tbao, fishw, yans, doupe}@asu.edu, †carlos.rubiomedrano@tamucc.edu

2022 IEEE Symposium on Security and Privacy (SP) | 978-1-6654-1316-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/SP46214.2022.9833581

***Abstract*—The academic computer security community has traditionally adopted peer review as an integral part of scientific publishing and dissemination, in a process that grows organically and nourishes itself by internal communications and intuitions, rather than repeatable experiments and investigations. Recently, key community members have shared a series of concerns regarding this process in public. To support or disprove some of these concerns, this paper presents the first qualitative study to examine the peer review process in the computer security field. Through semi-structured interviews (n=21) with Program Committee members, we systematically collect the reviewers' insights on how papers are evaluated in top-tier security conferences and investigate their concerns regarding the current security peer review system. Based on the collected data, we identify several issues in the security review system: whereas some have been previously observed by the community (e.g., the randomness in reviewers' decisions), others (e.g., reviewers have much more diverse and concrete opinions on the metrics of rejecting papers) have been observed for the first time in our study. Finally, through a series of recommendations, we aim to encourage the collaborative establishment of community norms that will significantly improve the security peer review process.**

I. Introduction

On January 5th, 1665, the first recognizable academic journal in history, the "Journal des sçavans," was published [61]. By the end of the 1700s, humanity had produced over 500 different journals. Since then, academics around the world have worked tirelessly to not just advance the state-of-the-art in their fields but also to communicate this advancement by entering it into the permanent scientific record.

As a fairly recent academic area, computer security is still learning to understand *itself* even as it strives to make advancements in the state-of-the-art. To advance the Science of Security in a fair, effective, and scalable way, the community has created various high-quality academic conferences and used them to select and present the latest scientific advancements. Researchers and practitioners submit their new findings to these conferences in the form of papers. Program Committees (PCs), composed of experts in different sub-areas from both academia and industry, select the best submissions for publication by adopting peer review as an integral part to help ensure the validity of scientific results [10], [14], [34], [51]. Although Herley et al. [22] discussed necessary properties of the *Science of Security*, there is no investigation to date on how the Science of Security is evaluated in practice, e.g., through the peer review process.

Moreover, all of the commonly-considered "top-tier" security conferences, namely IEEE Security & Privacy, USENIX Security, NDSS, and ACM CCS, have observed an exponential increase in the number of paper submissions in recent years, with a total of 3,039 papers submitted across all of the top four venues in 2020 [8]. Nevertheless, the computer security community has been facing new challenges associated with significant growth. As an example, in 2019, a panel at ACM CCS touched upon several challenges, including the increase in paper submissions, the workload on PC members, the quality and quantity of reviews, and the shift to a different submission model of rolling deadlines [5]. Moreover, in a recent article from July of 2021, a reviewer shared and compared experiences being on the PC of top-tier conferences from three distinct fields: software engineering, programming languages, and security [63].

To address these concerns, Program Chairs continuously introduce changes in the peer review process based on their knowledge, intuitions, conversations, and feedback from reviewers. Among other factors, the increase in paper submissions has led to changes such as rolling submission deadlines and major revision outcomes. However, to the best of our knowledge, in the span of its 40 years of existence, there has never been an investigation detailing how the peer review process is conducted within the computer security community. Overall, the interest in such a process via informal discussions in conference panels and on social media serves as a long-overdue wake-up call: now, more than ever, it is crucial to closely analyze the security peer review process for any systematic issues and obtain insights for potential improvements.

In this paper, we take the first step toward scientifically examining the peer review of top-tier computer security conferences by using qualitative data analysis techniques. Through in-depth, semi-structured interviews with 21 PC members who served on the PCs of top-tier security conferences between 2015 and 2019, we explore two main research questions:

1) How is the Science of Security currently served by the peer review process?
2) What are experts' opinions on current peer review mech-

DOI 10.1109/SP46214.2022.00131

anisms, especially with new challenges caused by the increasing number of paper submissions?

To this end, we explore reviewer responsibilities, reviewer interactions with the review system, delegation, and rolling submissions, reviewing approaches, evaluation metrics to assess security papers, and reviewer expectations for authors in terms of writing and conducting better security research.

Overall, our study resulted in the following key findings:

Evaluation metrics are subjective. 19 of our 21 participants mentioned that they check for novelty in top-tier security submissions (§ V). Our participants assessed novelty through various aspects such as advancement over the state-of-the-art, interesting research problem, novel solution, novel insight, and novel methodology, making it a subjective metric. We also find that only four evaluation metrics were commonly mentioned by more than six participants, while the rest of the evaluation metrics were relatively uncommon (Table II). Hence, we conclude that reviewers do not have a shared standard, and the metrics are rather subjective.

Comparing the metrics for evaluating papers, reviewers have much more diverse and concrete opinions on the metrics of rejecting papers. While the participants stated 16 different paper evaluation metrics, they pointed out 52 different "red flags" causing paper rejection. Furthermore, compared to the description of evaluation metrics, which was usually high-level and vague, the description of rejection metrics was much more specific and concrete. This observation may imply that reviewers have a more clear mind in rejecting a paper than accepting one. Additionally, two of our participants (P09 and P10) reported that reviewers of the security community might tend to look for reasons to reject because of the competitive nature of security conferences (§ V-B).

Program Chairs understand their responsibilities better than PC members. There was a high-level agreement that the primary responsibility of reviewers is to accept *high-quality* papers (§ VI). However, we observed that participants who had chaired PCs adopted a more fine-grained approach in checking for quality than non-chair participants. Our chair participants also believed that a reviewer's responsibility does not end at the submission of their reviews; they should also help Program Chairs shape the best possible program. Non-chair participants, generally, did not express this view.

"Randomness" of reviews is a serious and exploitable problem. Perceived "randomness" of reviews was the most frequent issue of all of the issues that were discussed by our participants in the reviewing system of top-tier security conferences (§ VIII). Strong accepts and rejects seem to be consistent, but reviewers' decisions are subjective and random for papers in the middle range. Our participants stated that authors could "game" the system, taking advantage of the randomness and lack of precision in the current reviewing process by resubmitting at other venues with next to no changes until they "get lucky," increasing the reviewing load and potential reviewer burnout.

Based on these and other findings, we develop several recommendations for the academic security community to improve and maintain its processes. We hope that the computer security community will show interest in its peer review process, move toward a more testing-based approach, and design further investigative studies to measure the quality of its processes.

II. Related Work

Qualitative human factors research in security. Qualitative research methods are used in studies involving humans to understand *human behaviors*, *emotions*, *needs*, *desires*, *routines*, and other *personality characteristics* [32] and provide scientific interpretation of qualitative data [37].

Because security researchers study and develop systems and solutions mainly for human use, qualitative human factors research in security is crucial. Researchers used interviews to study risk perceptions of SMEs (small and medium-sized enterprises) [23] and identified Security Operations Center issues [29]. Surveys were used to understood students' password knowledge [56] and users' perceptions of hardware security tokens [43]. In a mixed-method study, researchers investigated sex workers' unique security and privacy challenges [35].

Our study, similarly, uses qualitative methods to investigate the current state of the security reviewing process and to identify approaches that reviewers follow while selecting papers.

Peer review process. Peer review is designed to assess the validity, quality, and (often) the originality of articles for publication. Its ultimate purpose is to maintain the integrity of science by filtering out invalid or poor quality articles [4].

Peer review can be single-blind (where authors' identity is revealed but not reviewers'), double-blind (where both authors and reviewers are unknown), and open review (where both identities are revealed) [6], [36]. Single-blind can suffer from known or unknown reviewer biases (gender, race, nationality, reputation, seniority, affiliation, familiarity, etc.) [6]. Whereas a double-blind policy helps early career researchers not be penalized for their lack of seniority [25] and increases representation of female authors [11]. Double-blind may encourage reviewers' candor, impartiality, and freedom of expression, however reviewers can be disrespectful or harsh while hiding behind their anonymity [13], [60]. Open review may help to mitigate the lack of accountability, credibility, and transparency [48], [58], however it may increase the chances of reviews being less honest, critical, and rigorous [6].

A major challenge to maintaining double-blind is de-blinding of authors through various mechanisms such as social media, writing style, research field, thesis, multiple "salami-sliced" publications, self-citations, prolific publication history, funding agencies releasing a list of research proposals, and publishing on preprint servers [7], [9], [16], [20]. Two studies show that there is a strong correlation between well-known authors and the dissemination of working papers prior to submission [9], [20]. To mitigate the challenges to the double-blind peer review process, studies have suggested to weigh other alternatives [16], [40] such as open review as it would potentially promote shared responsibility between authors,

reviewers, and Program Chairs to maintain the integrity of the peer review process [7], [15].

Stelmakh et al. showed that novice reviewers tend to have a bias against resubmissions [54]. Cabanac and Preuss evaluated effects that lead to unconscious favoring of early-submitted papers to the detriment of later submitted papers [12].

There have been studies related to the peer evaluation process of specific fields. Ragone et al. analyzed peer reviews and reviewers' behavior in several review processes in computer science and reported on a theoretical model's development, definition, and rationale for those processes [47]. Olsen Jr. published a research paper on evaluating user systems research and pointed out that the reviewers of such papers should give more importance if the study made important progress and advanced the state-of-art [39]. In the biomedical field, Haffar et al. presented various possible mechanisms by which the peer review process can distort research results [21]. Squazzoni et al. aimed to promote peer review as an interdisciplinary research field and stimulate further quantitative research on the current peer review systems [53].

Additionally, there have been studies focused on specific venues together with their unique review processes [18], [26], [33], [42], [46], [49], [52], [59], [62]. ACM Conference on Human Factors in Computing Systems (CHI), Neural Information Processing Systems (NeurIPS), and International Conference on Software Engineering (ICSE) are considered top-tier venues in their respective research fields. Mackay, in their book, wrote about the review process that takes place in CHI [33]. Wilson discussed the different expectations of various venues and submission types, reviewers' processes to make decisions, and good techniques to produce reviews [62]. Shah et al. discussed the vast number of submissions that NeurIPS receives and calls for an analysis of the peer review system for improvement [49]. A survey of 241 authors and PC members of ICSE 2014, 2015, and 2016 was conducted to uncover that faulty reviews, insufficient time spent on reviews, and unfamiliarity with the topic areas are the main issues in the ICSE peer review system [45]. Unfortunately, the computer security field lacks research on its peer review system. In this study, we make an attempt to fill that gap and to be a spark that can ignite further research. We interviewed 21 PC members to investigate the double-blind peer review system of top-tier security conferences.

Research paper quality. Besides the peer review process, many studies have focused on improving submitted paper quality and guidance on how to write research papers in computer science [28], [31], [38], [50], [57]. In a peer review course for CHI, Nacke provided hands-on advice on writing papers with clarity, substance, and style [38]. Shaw analyzed the abstracts of research papers submitted to ICSE 2002 and observed the Program Committee discussions to help researchers design research projects and write papers in a way that they would have a higher chance of acceptance [50]. Levin and Redell pointed out the common problems that appear in technical papers to help future authors avoid them [31].

In this study, we highlight the evaluation metrics used by security reviewers while evaluating security papers and some red flags for authors to avoid.

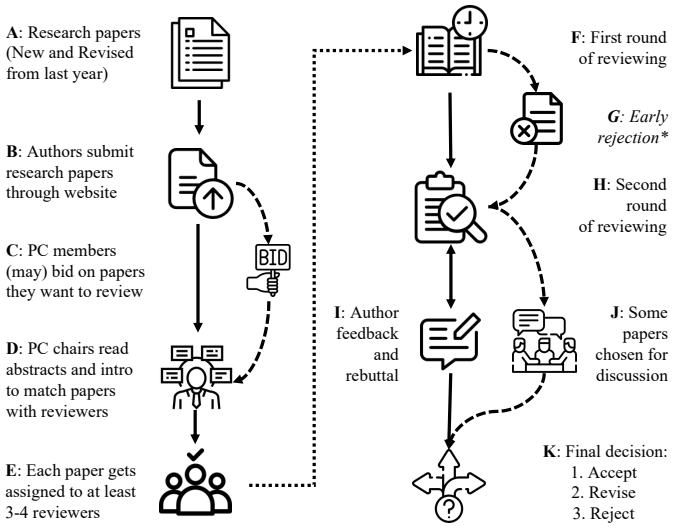

Figure 1. The reviewing process of a typical top-tier security conference. *Some conferences might not have an early rejection phase.

III. Background and Overview

In this paper, we study the peer evaluation process in top-tier computer security conferences. Figure 1 illustrates how such a process is typically conducted in top-tier security conferences. A more detailed description of how security peer review works can be found in Appendix A.

A. Background

For the readers' convenience, we highlight two key topics in our study: Rolling submissions and Delegation process.

Rolling submissions. Recently, top-tier security conferences adopted a new model called *rolling submissions*. In this model, a conference has two to four submission deadlines in a year's time. Authors can submit their research papers before any of these deadlines. As a middle-ground between journal submission and traditional annual conference submission, rolling submissions aims to help reviewers mitigate overloaded reviewing tasks crammed in a short period (while allowing authors more chances for submission).

Delegation process. Reviewers may delegate papers to external reviewers due to reviewing overload, time constraints, or topic unfamiliarity. The reviewer decides whom to delegate and is responsible for the review quality. Top-tier conferences allow delegation, although it remains controversial for various reasons, which we will detail in the rest of this paper.

B. Overview

Our study is based on the following research questions:

RQ1. How does the community enforce and advance the Science of Security through peer review? The peer review process is crucial for any scientific community to evaluate merits in scientific manuscripts. However, there is little documented understanding of *how* the security community conducts its peer

review. We show the internals of the peer review process in security conferences from three perspectives: the criteria under which reviewers evaluate papers (§ V), what responsibilities reviewers have (§ VI), and the characteristics of high-quality reviews (§ VII). Our answers will serve as the basis of RQ2 and will help the community, especially paper authors, understand metrics used for evaluating science and establish reasonable expectations for reviews.

RQ2. What do experts think of the current security peer review process? In recent years, top-tier security conferences adopted some significant changes to cope with new challenges (e.g., record-breaking numbers of submissions) and long-standing ones (e.g., low-quality reviews).

The perceptions of Program Chairs on such problems, along with the feedback from PC members, drive changes in the publishing processes of security conferences. Therefore, interviewing experts (both Program Chairs and PC members) from the computer security community will provide a holistic view and add more nuance to the current peer review of top-tier security conferences. Moreover, this will help the Program Committee and security conference organizers identify areas of improvement in the security peer review process.

We present and analyze the experts' concerns with the current peer review system in § VIII. Further, we focus on a critical change that all top-tier security conferences adopted, rolling submissions, in § IX. Finally, we discuss the experts' opinions on a controversial topic: delegated reviews, in § X.

IV. Methodology

We conducted 21 semi-structured interviews with PC members from top security conferences between November 14, 2019 and March 17, 2020. To obtain rigorous and exhaustive qualitative results, we conducted interviews until new themes stopped emerging [19]. Despite receiving an exemption from our Institutional Review Board (IRB), we followed policies and procedures designed for human research studies that are specific to our institution. We anonymized all personally identifiable information (PII) when transcribing interviews.

Participant recruitment. We first extracted all 585 PC members who had participated in top security conferences between 2015 and 2019. We then built a pool of 313 PC members who had served at least twice and randomly sampled 70 PC members from the pool. We sent invitation emails in small batches to not immediately exhaust our recruitment list[1]. Our invitation emails contained details about the motivation behind our study, the interview procedure, ethical considerations, and potential benefits of the study. We sent consent forms detailing the study once our participants signed up for the interview. In total, we contacted 70 PC members and interviewed 21.

Data collection. Each interview lasted between 23 and 73 minutes, with an average interview length of 45 minutes. Apart from 8 initial interviewees, all other participants received two interview options: 25 minutes (short) or 60 minutes (regular). We began providing the short interview option when interviewees started sharing their concerns about the interview length. We saw more responses from PC members to our email invitations with explicit interview lengths. Overall, one interview was conducted in-person while the rest were over Zoom, Skype, or phone calls.

[1]Inviting everyone on our recruitment list would unnecessarily inflate the number of conflicts of interest when we submitted this paper.

Table I
Participant Demographics. We do not provide reviewing experience of industry professionals to avoid deanonymization. (†No survey data for this participant)

ID	Current Role	Ever been a chair?	Estd. Reviewing Exp. (Years)	Self-reported Research Interests
$P01$	Assistant Professor	No	4	Security and Privacy
$P02$	Associate Professor	No	4	Cryptography
$P03^{\dagger}$	Industry Professional	No	-	Security, Distributed Systems
$P04$	Associate Professor	Yes	2	Hardware Security
$P05$	Associate Professor	Yes	3	System Security, Hardware Security
$P06$	Assistant Professor	Yes	7	Security and Privacy, Machine Learning
$P07$	Associate Professor	Yes	6	Security and Privacy, Internet Measurement, Data Science
$P08$	Professor	Yes	11	System Security, Malware Analysis, Web Security, Social Network, Mobile Security, Privacy
$P09$	Associate Professor	No	2	Cryptography, Provable Security, Anonymity, Privacy Enhancing Technology
$P10$	Assistant Professor	Yes	7	System Security
$P11$	Assistant Professor	No	2	Cryptography, Cloud Computing
$P12$	Professor	Yes	4	Cybersecurity
$P13$	Industry Professional	No	-	Malware, Cyber threats, Predictive analytics, Risk, Privacy
$P14$	Assistant Professor	No	1	Data-driven Security, AI Security, Fuzzing, Data Analytics
$P15$	Associate Professor	No	4	Cryptography
$P16$	Associate Professor	Yes	9	Computer Security, Theoretical and Applied Cryptography, Human Factor Issues
$P17$	Professor	Yes	11	Software Security, Network Security, Privacy
$P18$	Professor	No	3	Cryptography
$P19$	Associate Professor	Yes	10	System Security, Network Security
$P20$	Associate Professor	Yes	9	Software Security, System Security, Network Security, Social Sciences Studies
$P21$	Professor	Yes	3	Cybersecurity, Privacy, Network Security, Mobile Computing

The primary researcher of this project conducted all the interviews in a semi-structured style with the option to skip or ask follow-up questions. The interviewer might skip a question if they felt that the interviewee had already answered that question. Our interview (see Appendix C for interview questions) covered three topics: interactions of PC members with the review system, evaluation metrics used by PC members to review security papers, and recommendations from PC members to write better security papers. Participants were free not to answer any questions. Before conducting interviews, we pre-tested interview questions with three pilot interviews (whose data is not included in our results) and updated the questionnaire accordingly. Besides the interview, our participants also completed a short demographic survey (details in Appendix B). Table I provides demographic information of our participants, including current roles, estimated years of reviewing experience, and research interests. In total, we had 21 participants, where 9 participants had no prior chairing experience and 12 participants had chairing experience.

Data analysis. We analyzed the data using an iterative open-

coding process [55]. The primary coder crafted a codebook in MaxQDA (software for qualitative data analysis) [3] and identified themes by coding the interview data. The secondary coder used this codebook to code all interviews, provided feedback on the codebook, and worked with the primary coder to adjust code definitions wherever required. After the coding process, we compared both coders' codes to determine inter-coder reliability and calculated the Cohen's Kappa coefficient ($\kappa < 0.7$). We believe that a large amount of disagreement was because of the varying answers to open-ended questions and varying interpretations of the codebook. The final codebook had 15 principal codes. Both coders re-coded the ten codes where they disagreed to increase the reliability score and achieve more conclusive results. After every five transcripts, both coders would compare their codes, calculate Cohen's Kappa, and refine the codebook. Ongoing inter-coder reliability score calculations helped to ensure that the coding process did not unintentionally drift, mainly because the two coders collaborated remotely (due to COVID-19). The Cohen's Kappa score for this study is 0.96, which is considered an almost perfect inter-coder reliability score [30]. We also decided to include manual sentiment analysis for questions where we asked the participants about their thoughts on different topics. The coders were instructed to weigh positive and negative sentiments and assign an overall sentiment to each response. We provide the complete codebook in Appendix D.

V. Evaluation Metrics

To understand how reviewers evaluate the Science of Security through peer review, we studied the reviewers' evaluation metrics and their priority. Here, we present evaluation metrics and red flags that may negatively impact their decision on a paper. Interested readers may find reviewers' recommendations on writing high-quality security papers in Appendix D.

A. Common Evaluation Metrics

During the interview, we asked our participants "What evaluation metrics do you use to evaluate security papers?" All participants answered this question. We created a list of all the evaluation metrics with their prevalence in Table II.

We noticed that none of the evaluation metrics are common to all of our participants. After novelty, mentioned by 90.47% of all participants, the following common metric was only acknowledged by 42.85% of our participants. This observation aligns with our interviews, where P19 mentioned that:

> *So, security is an area where there are not any kind of hardened established metrics for evaluating security itself.*

Next, we discuss metrics that were mentioned by more than 20% of participants. These metrics are sorted by their prevalence in descending order.

Is it novel? Here, our discussions are based on the participants who check for novelty (17), advancement of the state-of-the-art (7), and the research problem (7). Altogether, 19 participants mentioned any of these three evaluation criteria.

Table II
Evaluation Metrics (with their prevalence) used by our participants to review papers at top-tier security conferences.

Evaluation Metric	Count	Percentage
Is it novel?	19	90.47
Is it correct?	9	42.85
Was the evaluation self-contained?	7	33.33
Is it well-written?	6	28.57
Is the problem relevant to the audience and the conference?	4	19.05
Is it practical?	3	14.28
Is it reusable or has any utility for the community?	3	14.28
Does it have a discussion around related work?	3	14.28
Is it impactful?	3	14.28
Whether contributions are valid and support the claims made? (internal consistency)	2	9.52
Is the paper well-executed?	2	9.52
Is it inspiring future research?	2	9.52
Does the paper answer the research questions?	1	4.76
To what extent are the questions answered?	1	4.76
Whether the paper provides deeper and generalizable insights?	1	4.76
What is the proposed methodology?	1	4.76

We observed that our participants assessed a paper for novelty in various aspects, such as having a novel problem, solution, insight, method, technique, awareness, and comparison with the state-of-the-art, or just something surprising, making it a subjective metric. P19 and P08, supporting our observation, mentioned that the assessment of novelty is in some sense a judgment call. P19 continued:

> *Novelty is definitely subjective. This is something where different reviewers will see different values out of a paper. Novelty is possibly multi-dimensional in itself in terms of, what are we learning from this, and what information from this is valuable?*

Fourteen of our participants reported checking for novelty when evaluating security papers from top-tier conferences. P01 does not expect the papers to be the *firsts* in their subarea, but the papers should provide a novel insight either by addressing a novel problem or by providing a novel solution to a known problem. P05, P06, P09, P10, P11, P14, and P21 emphasize the idea and research problem. P10 checks the relevance of the problem, then looks for novelty in the solution or key idea. P17 strives for novelty in terms of a new method, attack, or an introduction of something unknown. Additionally, P08 mentioned that they evaluate novelty in terms of the *interestingness* level of the paper. For P09, the paper should be exciting and must advance science.

Seven of our participants evaluated security papers for their comparisons with and advancements over previous work, and six of them also mentioned that they evaluate a paper for novelty, excluding P04. According to P04, frequently, papers ignore the state-of-the-art and only present their attacks or countermeasures. Evaluating a paper against the state-of-the-art helps a reviewer understand if the authors show unbiased

views and benefits of their research work. P19 shared their approach in checking for novelty:

> *What's the related work, who solves similar problems in the same or possibly different domains, have the authors talked about those other papers sufficiently? Is the nuance and difference between them of sufficient delta?*

Is it correct? Correctness was among the list of evaluation metrics for nine of our participants. P01 placed novelty and correctness at the top of their list to evaluate security papers. According to P01, P08, and P18, an incorrect paper is a weakness, and P08 said:

> *If it is wrong then it does not really matter what else is in the paper.*

P03, P08, and P11 also mentioned that they evaluated security papers for technical soundness, correctness, and depth.

Was the evaluation self-contained? Seven of our participants, P05, P08, P10, P12, P13, P14, and P19, emphasized the evaluations presented in the paper. According to P05, a security paper should show that they have evaluated their work from all angles:

> *From a performance point of view, from a security point of view, and you have to be self-contained. So, you have to use the evaluation to convince the reviewer to say that hey, this work is really complete and self-contained.*

P19 mentioned that papers should use right benchmarks to evaluate their work's performance and P10 believed that evaluations and analysis should support the conclusions drawn in the paper.

Is it well-written? Six of our participants, P02, P03, P11, P12, P16, and P21, considered writing of the paper as an evaluation metric. P02 reported that well-written papers are easier to be disseminated and understood. P21 requests authors to spend enough time on the writing and presentation of their papers. Moreover, P11 mentioned that a well-written paper ties everything together and added that:

> *Something that is well-written, well-structured, has good flow, can prepare the reader for what is coming next, can help the reader ask the right questions or provide the answers. This is always very welcome.*

B. Red Flags of Paper Rejection

We asked our participants "What do you believe are some serious misfits or red flags in a security paper that can get it rejected?" Compared to evaluation metrics (§ V-A) where the total number of metrics is only 16, there are 52 different red flags mentioned by the participants, which are *much more diverse*.

As security conferences are competitive, P09 stated:

> *We know that the acceptance rate is so low (at these conferences) that sometimes there can be a tendency from the reviewer side to look for reasons to reject instead of reasons for accepting a paper.*

Specifically, P09 reported that when one looks for reasons to reject, they focus on the paper's execution and not the content. They stated that it is easy to find reasons to reject a paper when it forgets to cite something, does not perform experiments that reviewers like, or does not have the writing of reviewers' choice.

As the acceptance rate of top security conferences is very low, most papers are rejected. Hence, P10 asks the reviewers to fight the instinct of *finding reasons to reject* and frame constructive and positive reviews. Also, the major/minor revision outcomes in the current reviewing model are pushing reviewers to write constructive reviews as the reviewers are now looking for improvements.

We grouped 52 different red flags mentioned by our participants in three main categories: content-related, argument-related, and writing-related. Content-related criticisms include the red flags mentioned by our participants about the objective information contained in different sections of the paper and the problem that the paper is attempting to solve, for example, selecting the wrong benchmark for evaluations (Table III). Argument-related criticisms include red flags concerning the authors' claims, such as contributions (Table IV). In Table V, we present writing-related criticisms, including mistakes in the writing and overall presentation of the paper.

Table III
CONTENT-RELATED CRITICISMS FROM OUR PARTICIPANTS.

Not novel or insignificant	
reinventing a known problem	Not novel
incremental papers	
trivial advancement	
resubmitting without making changes	
uninteresting papers	Insignificant
lacking real world applicability	
Mistakes in different sections of the paper	
misleading title	Title
not clearly outlining contributions	Introduction
not pointing out conceptual ideas	
not explaining the methodology	Methodology
mistakes in the methodology	
mistakes in formulas/algorithms	
technical mistakes	
not mentioning the attack model	
not doing or describing experiments thoroughly	Experiments
improper or insufficient or shoddy experiments	
lacking proper execution	
picking the wrong benchmark	Evaluation
evaluating on suitable conditions	
not comparing against competing solutions	
using the wrong dataset	
incomprehensible results	Results
not mentioning a takeaway message	
not mentioning or unclear limitations	
plagiarism	Unethical
lacking ethical considerations in human studies	
Incompatible venue	
out of scope for a security venue	Wrong fit

Table IV
ARGUMENT-RELATED CRITICISMS FROM OUR PARTICIPANTS.

Poorly argued	
over-claiming or incorrect claims	Inaccurate
incorrectly building the expectation	
not backing up the claims	Unsupported
not linking results with claims	
not explaining or discussing the results	Unexplained
using raw data without explanation	
unnecessary obfuscation	Obfuscation
hiding details of reproducibility	
Relevance	
unclear problem statement or motivation	Motivation
not motivating certain choices	
non-thorough literature review	Related Work
not having a solid comparison with the state-of-the-art	
not showing competency in the topic	
not being aware of the related work	
not treating literature fairly or not objective comparison with literature	
not having a convincing security argument	Application
not having clear security application	

Table V
WRITING-RELATED CRITICISMS FROM OUR PARTICIPANTS.

Poor writing	
bad writing	Poor Writing
incomprehensible writing	
rushing papers	
ineffective communication	
writing inconsistencies with multiple authors	
having bad grammar	Poor English
having bad or colloquial language	Jargons
incomprehensible graphs, tables, figures	Graphs, tables, figures

C. Evaluation Metrics Consistency

Reviewers are not always assigned papers from their core area of research. In such cases, participants P01, P06, P07, and P16 would often provide their judgment to the best of their abilities in the form of high-level reviews as it is difficult for them to properly review the state-of-the-art in a short amount of time. However, P10 and P17 mentioned that if they receive a non-core paper to review, it would most certainly mean that the paper is not fit for the venue, and it was assigned because someone *had* to review it. Frequently, reviewers would reject such papers with the suggestion to submit to a different venue.

We asked 16 of our participants if their evaluation metrics stay consistent when evaluating security papers. Nine of them reported *trying* to be consistent with evaluation metrics for top-tier security papers, whereas six did not. Moreover, P03 reported that they try to be consistent such that each paper receives a fair review. Five participants reported that they lower their expectations, sometimes for novelty, when reviewing security papers from non-top-tier venues, but the correctness and other evaluation metrics they use stay consistent.

Furthermore, P04, P11, P14, P15, and P19 mentioned that they change their evaluation metrics based on the type of paper. P15 wished to be consistent with their evaluation metrics, but they review papers from a broad spectrum, making it challenging to apply the same rubric for every paper. P11 added that every paper has its *own story to tell*, and factors, such as stress and emotion, take effect when reviewing a paper. Although P01 does not customize their evaluation metrics for top security venues, they mentioned that reviewers have to customize their evaluation metrics because security papers are unique, and established benchmarks do not exist.

VI. RESPONSIBILITIES OF PC

Reviewers of security conferences evaluate the submitted papers in a peer review, double-blind fashion. Once accepted, these papers become part of the knowledge base that other researchers refer to while conducting their research. Hence, it becomes crucial to understand reviewers' responsibilities as they filter and accept papers to be published at security conferences. In this section, we report the primary responsibilities discussed by our participants. We also report findings specifically from chair participants and non-chair participants to highlight perspective differences, wherever possible.

Accept papers of quality. Nine of our participants, including five chairs and four non-chairs, ensured that the accepted papers are of quality. According to our non-chair participants, P01, P02, P03, and P11, a high-quality paper can have one or more characteristics: advancing the field, novelty, having a significant breakthrough, being well-written, or having some unseen technique. However, some of the characteristics of a high-quality paper mentioned by our chair participants, P05, P06, P08, P16, and P19, included: novelty, correctness, well-written, significant contributions, interest to the community, helpfulness to society, advancing the field, meeting the bar set by the conference, standing out in the area, having substantial evaluations, or having scientific implementation.

From our analysis, we observe that Program Chairs have a very fine-grained approach to evaluate papers for high quality.

Provide constructive feedback. Six of our chair participants considered providing constructive feedback to authors as one of their primary responsibilities, whereas only two of our non-chair participants mentioned this. Our participants believed that reviewers must show how a paper can be improved and explaining this concept, P17 stated:

> *What kind of extra experiments are needed, where some of the numbers need additional support, or where the explanation could be improved, or even simple things like typos or whether the paper structure needs to be changed.*

P10 also believed that constructive feedback could be extremely helpful to the authors at any stage and said:

> *...helping the authors with feedback to improve their papers in the best possible way. Be it for the final version, for the next iteration, for the revision of the paper, and so on.*

Moreover, P04, P19, and P20 shared that reviewers often forget to provide valuable, constructive feedback, especially those who are new to the reviewing process.

Evaluate correctness, novelty, and validity. Among the responsibilities of PC, nine participants mentioned that the correctness of security papers is critical to evaluate. P12 stated that their responsibility as a reviewer also includes validating the correctness of the paper's approach. Our chair participants, P04 and P07, considered checking for validity as one of their primary responsibilities. P07 stated that they evaluate the validity of papers by checking if the claims made in the paper follow the evidence shown.

Eight participants mentioned checking for novelty in security papers. P20 defines novelty as:

> *Traditionally speaking, something that has not been published in a peer-reviewed setting such as a journal or a conference.*

Review and advocate papers fairly. Two of our non-chair participants, P11 and P13, considered fair assessment of the security papers as one of their primary responsibilities as reviewers. Another non-chair participant, P15, believed that reviewers should also be fair in their representation and advocation of papers during the discussion phase after the reviewing phase. None of our chair participants reported fair assessment as their responsibility.

Help shape the best program. Only our chair participants, P06, P10, and P17, firmly believed that the responsibility of reviewers does not end at inserting their reviews in the system, but they should also support Program Chairs in assembling the program by formulating recommendations and coming up with a set of papers that can be presented at the conference.

VII. Characteristics of High-quality Reviews

Reviews are crucial for the peer review process as they are the only interface where reviewers present their assessment and suggestions of a submission. In our study, we observed the following review characteristics acknowledged by reviewers.

Providing constructive and actionable feedback. Twenty participants mentioned that a good review should provide constructive and actionable feedback. P02 reported that it is important to have a review that suggests improvements, inspires other ideas, or even offers different points of view on the results. They added:

> *What I would not like to have is only an opinion that says, I don't think this is an interesting result or technique. I think that is the most useless one.*

Moreover, P18 mentioned that a constructive review can help improve even the bad papers for their next submission.

Being detailed and informative. Twelve participants stated that a good review should be detailed with explanations and informative to not just the authors but to the entire PC. P09 mentioned that based on a review other PC members should be able to provide their opinions on the paper. Participant P01 compared review writing with paper writing and said:

> *I feel that writing a review is just like writing a paper. In that, you establish a position, you make some claims regarding that position, and then you provide evidence to support that position. If any of these things are missing, it is not a good review or a good paper.*

Being comprehensive and well-structured. P10 stated that a good review should evaluate a paper against *all* dimensions—the problem, the idea, the contribution, the impact, the execution, the evaluation, and the comparison against prior work.

P20 provided a guide to write well-structured reviews: First, the review should demonstrate that the reviewer understood the paper and summarize the decision. Then, the review should list the strengths and weaknesses of the paper along with necessary pieces of evidence. The review should have a thorough explanation for each comment and provide suggestions on how to fix issues. Finally, the review should end with comments on the language and typographical issues of the paper.

Being clear and carefully written. Six participants considered that a good review should be clearly written. A review should not be riddled with incorrect statements and typographical errors. A review is a report that should be self-contained and understandable to the readers.

P11 reported that they re-read their reviews after they are drafted to ensure that they do not come across incorrectly. P16 mentioned that they spellcheck and proofread their reviews to ensure that they do not sound "grumpy." P15 stated that even if they are going to reject a paper, they will try to be as humble as possible and make constructive comments.

P18 and P20, our chair participants, shared a tip for writing positive and constructive reviews: Stay in a good mood. P20 added that reviewers should show empathy in reviews and imagine that they are reading reviews to someone in person.

Being objective. Nine participants stated that objectivity is essential with review writing. Participants P06 and P13 mentioned that the reviews should *avoid* subjective remarks such as "*I do not think you are advancing the field or this application is not relevant*" and "*I do not like the results*."

Including a paper summary. Eight participants believed that having a summary statement is crucial to portray the understanding of the paper. P04, P07, P15, P19, and P20 consider that the paper summary helps other reviewers in the discussion phase and the authors when they see that reviewers have understood their paper. P15 further added:

> *... and some people do not really do that well (write a summary statement about the paper), and I think it affects the quality of the review.*

Being anonymous. P08 and P17, our chair participants, take certain precautions to not de-anonymize themselves because they believe that if one always writes their review in the same style, it is easier to identify them. To maintain anonymity, P08 refrains from suggesting authors to include citations to their work, and P17 varies their review writing style.

VIII. Systemic Issues with the Review Process

As reviewers regularly interact with the review system, they may have complaints or expectations about it, which can highlight to conference organizers specific issues that need

attention. In this section, we report our participants' opinions on the current state of the review process.

Randomness. Participants P02, P06, P07, P09, P11, and P20 complained about the randomness in reviews for papers that reviewers do not unanimously accept or reject. For papers in the gray area, reviewers' decisions are subjective and seemingly random. Randomness exists in the review system because PC members may draw very different conclusions on the same paper. P07 mentioned that randomness is the result of priorities set by reviewers and added:

> *My main concern is you would not be able to use papers that are accepted in selective conferences the same way you have been using them in the past to signal quality and academic excellence.*

P11 mentioned that authors might be motivated to *game the system* because of randomness in reviews, even at top-tier conferences. Authors can keep submitting until they get lucky with reviews such that their paper gets accepted. P07 reported that there is a considerable amount of randomness in who gets to review, what they think, how the discussion goes, or who is advocating the paper. P07 added that reviewers could not be angry at authors who try to game the system until the review system is sufficiently deterministic. They further commented:

> *If we can be more accurate in our reviews, then yeah, it (gaming the system) is a horrible thing to do. But, we are not; it works. And so, somebody whose job depends on getting these papers in, why would you blame them for doing something that works.*

Usefulness of reviewing history. Chair participants P06 and P17 find it useful to carry over reviews to other conferences. P06 believed that reviewing history can serve two purposes: (a) reviewers can hold authors accountable when they do not make any changes to their papers during re-submission, and (b) reviewers can identify if previous reviews were unfair. According to P17, reviewing history ensures that the same reviews are not repeated because most top-tier security conferences have an acceptance rate of about 20%. If 80% of papers are getting rejected and re-submitted to somewhere else (without improvements), reviewers' reviewing effort is lost.

Re-submission with unfair reviews is acceptable. P01, a non-chair participant, had different views on reviewing history and reported that they do not like conferences asking for prior reviews. Sometimes reviews are of bad quality and without any constructive feedback or with incorrect criticisms. Criticizing prior reviews is difficult for authors as it would "send a negative connotation to reviewers." Authors may mention that prior reviews were constructive, but when new reviewers find out that the authors did not make any changes to their re-submission, the reviewers would feel that the authors did not meaningfully address prior criticisms. This is a dilemma for authors, and P01 believed that in such scenarios, re-submission without providing prior reviews is the only option.

P06, P11, and P15 also mentioned that re-submission of papers without changes is acceptable as long as the authors justify that they received wrong, unactionable feedback or their paper was judged unfairly.

Need for accountability. Participants P10 and P11 reported that the current reviewing system lacks accountability of reviews and some reviews are opinionated or incorrect. According to P11, regardless of the review being good or negative, the reviewer must be responsible for their review, and an ideal solution is open reviews. P17 suggested having open reviews to increase reviewers' responsibility to write good reviews.

Huge reviewing load due to being on multiple PCs. Four of our participants, P04, P11, P12, and P19, mentioned that reviewing load drastically increases if a reviewer serves on multiple PCs. P12 explained that when a reviewer reviews too many papers, they can not put enough effort into understanding everything, which is why review quality may vary a lot from person to person. They added that everyone has a capacity and suggested that top-tier conferences should limit the number of papers a reviewer reviews in a year.

Scalability challenges with rolling submissions. P10, a chair participant, said that there is still much redundancy with rolling submissions because many papers that get rejected from one conference get reviewed *verbatim* by other PC members at other venues. Re-submitted papers increase reviewers' workload and pose huge scalability problems to conferences.

Need for a balanced PC. P12 mentioned that sometimes there is a lack of suitable matches for papers within a PC and added that reviewers should not review papers that are outside their areas of expertise. P03 wished to have more qualified reviewers for intelligible feedback.

On the contrary, another chair participant, P17 believed:

> *There is also a time when a paper is not a good fit for the conference because if there is no PC member to review this paper in a meaningful way, then this probably is not the right audience for this type of paper.*

Such papers are typically not in the core focus of Call for Papers and hence, out of scope for the program attendees.

Need more objectivity from reviewers. Participant P10 mentioned that the security community is impact- and hype-oriented, and the reviewers favor such papers. P12 stated that some reviewers have technical and research biases, which usually reflect when reviewing new ideas. P20 mentioned that the security community has reviewers with philosophical biases against certain areas of research and added:

> *I think it's very hard to divorce a reviewing system if the community is not huge and everyone knows each other, even if it's double-blind. I mean, it's difficult to fix social problems with cabals that are accepting each other's papers.*

P12 mentioned that *favoritism* is the most common problem in security conferences. P17 acknowledged that both double-blindness and peer review have problems and added:

> *It might not always be really objective. There are of course humans involved so, they could also give*

sometimes wrong decisions, but I haven't seen any proposal on how to improve the process significantly.

A shift in PC discussions. PC discussions are an essential part of the peer review in security conferences and can considerably impact a paper's outcome. Although our questionnaire did not cover the topic of PC discussions, eight participants shared their concerns and opinions.

PC discussions are more important than review writing for P06. Additionally, P20 mentioned that borderline papers are challenging to be reviewed in isolation such that outcomes vary among reviewers. Hence, articulating one's case and advocating for or against the paper is critical for the selection process. Frequently, a strong champion or detractor can make or break a paper. P05 shared that if junior PC members are intimidated by seniors, they might not speak freely or fight against them. Such an issue defeats the purpose of PC discussions—to decide the outcome collectively.

P10 stated that with major/minor revision outcomes in the revised model, reviewers could avoid prioritizing one dimension over the other and communicate the improvements in each dimension to the authors. However, P13 mentioned that reviewers are much less invested in online discussions. In the previous model with physical meetings, a reviewer would go through other reviews to judge their review, see if they missed something, and afterward discuss the paper in front of other PC members. P17 reported that reviewers are confined to the assigned papers in the online discussions and do not have a broad overview of other submitted papers. They appreciated the active discussions around submitted papers and the bar on expectations in the previous model. P17 added that PC discussions help bring new faculty into the reviewing community by discussing the reviewing structure and their responsibilities.

Negative sentiments. When asked about the current state of the review system in security conferences, thirteen of our participants shared negative sentiment, including five non-chair participants (P03, P11, P13, P15, P18) and eight chair participants (P05, P06, P07, P10, P12, P16, P17, P20). When P12 was asked about the review system, they said:

> *The review system that we have currently is broken, and it is not systematic in the way that papers are reviewed.*

P20 mentioned some issues in the review system and said:

> *It is a flawed system, but like democracy, we do not have a better system to replace it.*

Positive sentiments. Only four participants, P08, P09, P14, and P21, thought positively of the current review system. P08 stated that even though security is growing as a community and there are more submissions, reviewers are still spending much of their time writing great, extensive, and meaningful reviews. P21 appreciated that reviewers provided constructive feedback, even for papers they did not like. P09 expressed their happiness for attempts made by our community in trying to innovate itself and moving to a rolling submission system. They added that the community is learning how to use it better, but as researchers, we should be willing to experiment a little.

IX. Rolling Submissions

We asked our participants about their experiences with this new shift to multiple rolling deadlines. Here, we share our participants' opinions on rolling submissions from authors' and reviewers' perspectives except from two participants (P03, P12) who did not experience rolling submissions.

Blessing to authors. Ten participants considered rolling submissions good for authors, including four non-chair participants (P01, P02, P11, P18) and six chair participants (P04, P07, P08, P10, P16, P17). P01, P04, P07, P10, P16, and P17 reported that authors have more flexibility in what they want to submit, when, and where with rolling submissions. P10's research group's submission quality has improved as they have no pressure to submit papers on a given deadline. P16 strongly believed that the previous system of the single deadline was the least good from the authors' perspective.

However, P02, P04, P17, and P21 recognized that multiple deadlines can allow for procrastination and authors can get more relaxed. P17 and P21 added that authors could miss several deadlines because of postponing their submissions.

Negative sentiments from reviewers' perspective. Participants P07, P14, P17, P18, P19, and P20 reported that workload distribution has increased with rolling submissions. They felt an increase in workload even though the number of papers to review per day remained the same because with rolling deadlines review turnaround times have been reduced, and our participants find it difficult to manage.

Non-chair participant P15 reported that they accepted papers with more confidence having a longer time to review and believed that putting the committee under time pressure would eventually decrease review quality. P07, a chair participant, mentioned that it has become difficult to write satisfactory reviews with more papers and less time to commit.

P01, P08, P09, P10, P13, P17, P18, and P19 complained that they keep getting papers to review. P01, P08, and P10 also mentioned how strenuous it must have been to be on the PC when Oakland had monthly deadlines (2018–2020).

Participants P04, P18, and P19 expected that reviewers could get exhausted with rolling deadlines and P18 believed that self-motivation was essential to be on the PC. P04 added that it could be challenging for Program Chairs to keep the reviewers engaged throughout the review process. Exhaustion may also affect the review quality, and there is a high chance that experts do not participate if the workload is too heavy, resulting in gaps in expertise.

P09 complained that the rolling deadlines model forgets how significant the paper is and only focuses on how ready the paper is for publication. P09 is worried that later submission rounds might receive a positive boost because they might have accepted fewer papers in earlier rounds.

Overall, according to P07, it has been a growing pain in switching from single shot to rolling submissions. They stated:

Just getting used to the details of how do you assign PC members? How do you return reviews and talk? How do you motivate authors to submit to any of them as opposed to the last deadline? Once everybody has figured it out, it is not a hard problem.

Positive sentiments from reviewers' perspective. Non-chair participants P01 and P02 mentioned that they could review papers more fairly as they would not exactly know what they compete against. Five participants believed that their review quality increased because of workload distribution throughout the year, and they could focus more on the details of the paper.

According to our participants (P01, P10, P17, P19, and P20), a benefit of having rolling submissions is the possibility of dialogue with the authors and conveying reviewer expectations to revise the paper instead of having the vicious cycle of rejection and resubmission. According to P10, the revised model with major/minor revision is pushing reviewers to write more constructive reviews and added:

This revised model is pushing more people to write reviews where you (reviewers) are by construction more constructive and positive towards papers because they're fighting perhaps the common instinct to find reasons to reject.

To add to this, P20 and P10 mentioned that overall paper quality also increases because now reviewers can ask for improvements when they see *potential* in a paper.

X. Review Delegation

Reviewers may delegate their reviews to Ph.D. students, postdocs, and other external experts. With the increase in delegation and the opaqueness of the process, delegation has become one of the most controversial topics in the community.

In our study, all nine non-chair participants and nine out of twelve chair participants mentioned that they have participated in delegation in some form. In this section, we will unveil the delegation process and present the reviewers' opinions on delegation.

A. Delegation Process

The purpose of delegation. The participants mentioned two following purposes:

- **Training students.** Seven participants believed that delegating reviews is essential for students' personal development. Participants P01, P06, P18, and P19, delegate reviews even to younger Ph.D. students because they believe students should learn how to review early on. P18 stated that if students read other papers, they focus on the results and techniques and wonder how to use them in their research. However, if they have to write a paper review, they might "*focus much more on using different ways to make that paper more understandable*."
- **Leveraging external expertise.** Participants P04, P09, P10, P13, P16, P18, and P20 stated that they seek external help in case of low confidence. P04, P13, and P20 consult with external reviewers after getting permission from the chair or only if they were allowed to assign external reviewers from within the reviewing system. P10 mentioned that reviewers need to be accountable for their reviews. They added that full delegation would be a great idea only when the reviewers know an expert in that area and Program Chairs can open a slot for additional review.

Notably, although participants delegate reviews to external reviewers, they object to complete delegation. P12 reported that top conferences have papers with controversial and novel ideas, so reviewers should "*never opt for complete delegation*."

Selection of the delegatee. Ten participants mentioned that they select the delegates based on their expertise on the subject matter. P14, P16, P18, and P21 mentioned that they only delegate to senior Ph.D. students as they have higher confidence in reviewing papers. Three participants described their approach in deciding whom to delegate. P12 would read the paper first, beginning to end, and then decide whom they would want to delegate. P09 would delegate to other members in their group, Ph.D. students, post-doctoral researchers who are *"very willing"* to do the reviews. P07 would show their students paper titles and abstracts and ask them to pick the papers they would like to review.

Handling the delegation process. Ten participants stated that they do not submit their delegated reviews without supervision. P02 and P09 reported that PC members should try and understand the paper, because, as reviewers, they have to understand how that paper relates to other submitted papers and the state-of-the-art. There is a chance that not many people are active in PC discussions if a paper only receives delegated reviews. When P19 assigns papers to their students, they also explain confidentiality and other reviewing-related concepts.

Other participants delegating reviews to students wrote reviews in parallel. They gave their students an earlier deadline and discussed with them paper strengths and weaknesses. They provided their feedback on students' reviews. P19 mentioned they would anonymize other reviewers' reviews and send them to the students to explore different perspectives.

B. Opinions on Delegation

Negative sentiments. While seven out of twelve chair participants (P04, P08, P10, P12, P16, P17, P20) showed negative sentiment on review delegation, only two of our non-chair participants (P13, P14) shared the same sentiment.

P20 stated that delegation is *abused a lot* within the security community and is considered a *tradition*. Senior members may not agree to be on the PC if they could not delegate. Such pushback from senior members can hurt the program, especially when the Program Chairs are trying to balance the PC with a range of seniority. Program Chairs may not receive reviews on time if they forbid delegation. P20 mentioned that Program Chairs could decide not to invite reviewers when they ignore the rules and concluded with:

But in practice, it doesn't really hurt people as black lists are kind of ad hoc and irregularly enforced.

P04 and P17 strongly believe that PC members are invited to the committee for their expertise and not somebody else's. P17, P08, and P20 mentioned that PC members need to

offer valuable comments on the papers during online PC discussions. If any reviewer had not read the paper or had not written their review, they would not be adding anything constructive toward making a value judgment for the paper. P20 continued and added that PC members should write reviews on their own to ensure quality and their reputation.

P12 and P13 do not like when PC members, especially senior reviewers, completely delegate to students who are not experts in the papers' topic area. According to P12:

> *It will be unfair for the authors, and it will be unfair for the students because they will not learn how to evaluate. So, it is not good for either.*

P16 and P18 reported that the community is tired of PC members who delegate everything to students and postdocs and still get rewarded for being on the PC. P18 believed that it is terrible when reviewers outsource the reviews and add PC memberships to their CV.

Positive sentiments. While seven out of nine non-chair participants (P01, P02, P03, P09, P11, P15, P18) shared positive sentiments, only four out of twelve chair participants (P06, P07, P19, P21) shared the same sentiment. P15 reported that reviews should be delegated to experts to ensure high-quality. They added that PC members should participate in the delegation process to increase the review quality and not as an excuse to do no work. They continued advocating delegation:

> *We (reviewers) try to be as expert as we can and cover as many areas as we can, but we cannot be an expert for everything that lands on our desks.*

P06, P19, and P21 reported that delegation is essential for the education and training of future reviewers. If students do not have reviewing experience before being on a PC, they would not know what to do and may affect the entire system. P06 also believed that students should know how the review process works, how experienced reviewers write their reviews, and how reviewers interact during PC discussions.

XI. Recommendations

In this study, 21 security reviewers shared their insights and concerns regarding the current reviewing system of top-tier security conferences and described a number of concrete suggestions which are presented in Table VI. In this section, we discuss several recommendations—inspired by our participants' suggestions, but further expanded based on our own deliberations and analyses of different options—to the security community. These recommendations should not be taken as mandates, but rather as starting points for deeper discussions within the security community.

Focus on review quality when mentoring novice reviewers. Our participants suggested expanding the Program Committee for various reasons: overwhelmingly increased reviewing load (P01, P07, P21), discouraging delegation (P02), and balancing expertise between reviewing cycles (P11). Moreover, seven of our participants delegate reviews to students for training purposes and P19 provides anonymized reviews from other reviewers to students to explore different perspectives (§ X). Notably, reviewers might introduce bias when teaching students to review from their experience and not universally accepted best practices.

IEEE S&P organizes student/shadow PC intending to educate Ph.D. students and post-docs about the reviewing process. Although the 2017 IEEE S&P report suggests that students were more negative than seniors, they do not evaluate students' performance in terms of their review quality, perhaps the most critical metric for success of the conference peer review process [41]. Proceedings on Privacy Enhancing Technologies (PoPETs) provides opportunities for recent graduates and senior Ph.D. to contribute as external reviewers [1]. They believe direct participation in the review process is the most effective way to train novice reviewers.

To address the problem of reviewing overload with a scarcity of qualified reviewers (P08), we recommend that conferences recruit reviewers through a vetting process inspired by the shadow/student PC processes [17], [24], [44]. Early career reviewers may start as Probationary PC members, with senior PC members (e.g., as part of a Reviewer Mentorship Committee) mentoring them through the peer review process and evaluating the quality of their reviews. After a satisfactory performance as a probationary member, reviewers may become full PC members.

Assist reviewers in performing timely reviews. Authors expect constructive, actionable, detailed, and timely feedback from their reviewers. However, we observed that reviewers could not start reviewing early on because of various commitments (e.g. *P02: But, let's say I don't postpone everything to the end...* and *P17: Of course, I want to review early but quite often I did not*). P09 suggested that reviewers could decline conference invitations to reduce workload but younger faculty are ambitious and accept all invitations because of the pressure of building their CV while in tenure-track positions.

To assist security reviewers with timely reviews such that they do not have to engage in last minute delegation or provide last minute high-level reviews, we make certain recommendations. First, to ensure adequate resources of individual reviewers, Program Chairs can request that reviewers limit the number of PCs that they serve on concurrently (P12) in addition to ensuring that the reviewers do not have scheduling conflicts and are well aware of the format before accepting the invitation. Second, to avoid overwhelming the reviewer, conferences can assign papers in more batches with a shorter turnaround time (e.g., through more reviewing rounds)[2]. Third, to help individual PC members *start* their reviews, Program Chairs can ask reviewers to submit paper *summaries* early, ensuring that reviewers do not postpone starting to read the paper until the last minute. Lastly, P08 stated that the frustration of last-minute, late, and bad reviews could be solved by leveraging more automation in managing the reviewing process (e.g., better and automated paper-reviewer matching, timely and automated reminders to submit reviews on time).

[2]Additional rounds may also mitigate problems caused with reviewers not returning reviews during early rounds of reviewing.

Reward and recognize good reviewer behavior. Reviewers provide their volunteer effort to select good science from the submitted papers. P13 complained that some reviewers are consistently negative in their reviews and mentioned that conferences could be stricter with regulations if reviewing were to be compensated. We make two recommendations to Program Chairs to recognize good reviewers: (1) Leverage characteristics of high-quality reviews (§ VII) to design a list of quality indicators. Program Chairs could assess if reviewers are consistently negative and not invite them for future reviewing cycles. (2) Monitor reviewers' performance by tracking certain variables such as the number of papers reviewed, time since the last paper was assigned, average review turnaround time, review length, inter-reviewer agreement, participation in rebuttals and discussions, and review quality. Better the reviewer's performance, better the likelihood of them receiving more invitations to review, not forgetting that there should be continuous monitoring.

Program Chairs can create a systematic reviewer recognition process by using quality indicators. PC memberships are crucial for reviewers' academic success: if bad reviewer behavior were to reliably impact a reviewer's PC invitations, they would be more disciplined toward it. Moreover, good reviewing can be recognized by Program Chairs with a letter describing the reviewer's valuable contribution, which reviewers may use when being considered for promotion. Even outside of the context of promotion, a visible recognition such as a Good Reviewer award may motivate reviewers to strive for quality in their reviews.

Make authors accountable for their submissions. P06 and P13 strongly believe that authors who resubmit without changes, without a justification, and try to game the system, should be penalized. P06 shared that at the Conference on Neural Information Processing Systems (NeurIPS), those who submit commit to being external reviewers, making authors accountable for their submissions.

As some authors might be hesitant to submit their previous reviews (§ VIII), we believe that organizations of top-tier security conferences could come together and build a shared database to keep track of each paper in the reviewing pipeline. This system could detect and flag if a submitted paper crosses a certain match threshold with another paper. To handle false negatives, more than one reviewer would be assigned to review flagged papers. Program Chairs would have to ensure that new reviewers consider review history as additional information without any bias.

Social media makes it difficult to enforce double-blind. P10 and P12 both mentioned that double-blind does not stop people from exchanging their papers for "comments and reviews", including on social media platforms such as Twitter. P12 and P20 additionally believe that this can impact reviewer impartiality. Un-blinding of authors has been explored in prior work [7], [9], [16], [20], [27]. In double-blind peer review, authors' identity and affiliation are not revealed until after the paper is accepted but social media seems to be a threat to double-blind [9], [20].

To preserve this, Program Committees need to consider the dissemination of information on social media and sharing among non-collaborators while designing their Call for Papers. Alternatively, the community should honestly discuss if double-blind can stand in the face of social media, and how to ensure that double-blind remains.

Meet with the community to listen, identify, and reflect. The security community should create spaces to foster communication *beyond* those participants who happen to be in the conferences. For example, the IEEE S&P change to rolling submissions was accompanied by an extended community discussion period carried out on GitHub [2].

Having a vast audience for such meetings, including newcomers, authors, reviewers, and conference organizers, will help illuminate different perspectives on the reviewing process of security conferences. Such meetings might help reach a consensus on what modifications are needed to improve the review process.

XII. Limitations and Future Work

Our study has certain limitations, and here, we discuss those in detail, along with the specific methods that we employed to mitigate them. We also discuss opportunities for future work.

Generalizability. We study the peer review process specifically in security; however, our results may not be generalized to *all* other fields of research. Moreover, there are many sub-areas of security, and it is challenging to recruit participants from all sub-areas of security. Researchers could repeat our study focusing on a specific area to understand and provide security sub-area-specific criticisms and recommendations.

Social desirability bias. Participants can share their thoughts partially, present themselves more favorably, or withhold information in self-report studies. Though none of our participants stated that they would be disrespectful in their reviews, the reality might differ. Future research can observe PC discussions to strengthen our results and investigate reviews of rejected papers to improve the understanding of red flags.

Non-response bias. We contacted 70 potential participants, and only 21 accepted our invitation. 49 out of 70 refused or did not respond to our recruitment email. Information from non-participants could have affected the outcome of this study.

Recall bias. We asked our participants questions on overall reviewing experience, evaluation metrics, unique perspective toward the review system, and typical features they look for in security research papers. Participants might have had trouble recollecting every detail from the time when they last reviewed top-tier security papers.

Interview time limitation. We initially designed an hour-long interview; however, it was challenging to recruit some of the potential participants. After getting rejections only because of interview time, we shortened our interview to 20 minutes, prioritizing the data collection. Even though we observed an increase in participation, our participants might not have had sufficient time to respond to later questions.

XIII. Conclusion

In this work, we presented reviewers' opinions and concerns on the current reviewing system adopted by top-tier security conferences. We conducted an exploratory qualitative study with 21 experts of the security community to understand the peer review process, discover issues, and explore potential improvements. Our findings reveal that security Program Committee members have genuine complaints about the security reviewing system, and the security community should address them. We hope that this paper encourages the computer security community to bring more focus on its peer review process.

Acknowledgments

We thank the anonymous reviewers as well as our major revision shepherd for their valuable feedback. We extend our sincere gratitude to the participants of this study for taking out the time and providing valuable insights into the review process of security conferences. This material is based upon work supported in part by the National Science Foundation (NSF) under Grant No. 2131263, the Defense Advanced Research Projects Agency (DARPA) under Grants No. FA875019C0003, N6600120C4020, and HR00112190093, and by a Startup Funds Grant from Texas A&M University - Corpus Christi.

References

[1] About PoPETS reviews. https://petsymposium.org/reviews.php. Accessed: 2022-06-04.

[2] IEEE S&P ongoing submission plan. https://github.com/ieee-security/ongoing-submission-plan. Accessed: 2021-12-02.

[3] MAXQDA the art of data analysis. https://www.maxqda.com/. Accessed: 2022-06-04.

[4] What is peer review? https://authorservices.wiley.com/Reviewers/journal-reviewers/what-is-peer-review/index.html. Accessed: 2021-08-15.

[5] CCS 2019. Scaling the academic security community. https://sigsac.org/ccs/CCS2019/index.php/ccs-2019-panel/. Accessed: 2021-08-11.

[6] Parveen Azam Ali and Roger Watson. Peer review and the publication process. *Nursing Open*, 3(4):193–202, 2016.

[7] Ashokan Arumugam, Poonam Mehta, and G David Baxter. Double-blind peer review of manuscripts: Opportunities, challenges, and way forward. *Physical Therapy Reviews*, 25(1):1–6, 2020.

[8] Davide Balzarotti. System security circus 2020. http://s3.eurecom.fr/~balzarot/notes/top4_2020/. Accessed: 2021-06-03.

[9] Homanga Bharadhwaj, Dylan Turpin, Animesh Garg, and Ashton Anderson. De-anonymization of authors through arXiv submissions during double-blind review. *arXiv preprint arXiv:2007.00177*, 2020.

[10] Raymond Bowers. Views: The peer review system on trial: Despite its imperfections, peer review is an indispensable means of allocating limited resources and maintaining the high quality of scientific research. *American Scientist*, 63(6):624–626, 1975.

[11] Amber E Budden, Tom Tregenza, Lonnie W Aarssen, Julia Koricheva, Roosa Leimu, and Christopher J Lortie. Double-blind review favours increased representation of female authors. *Trends in Ecology & Evolution*, 23(1):4–6, 2008.

[12] Guillaume Cabanac and Thomas Preuss. Capitalizing on order effects in the bids of peer-reviewed conferences to secure reviews by expert referees. *Journal of the American Society for Information Science and Technology*, 64(2):405–415, 2013.

[13] Bernie Carter. Peer review: A good but flawed system? *Journal of Child Health Care*, 21(3):233–235, 2017.

[14] Susan Crawford and Loretta Stucki. Peer review and the changing research record. *Journal of the American Society for Information Science*, 41(3):223–228, 1990.

[15] Daniel J Dunleavy. The cultivation of social work knowledge: Toward a more robust system of peer review. *Families in Society*, 102(4):556–568, 2021.

[16] Stefania Fatone, Michael P Dillon, Brian J Hafner, and Nerrolyn Ramstrand. The challenges of double-blind peer review in an era of increasing research transparency. *Prosthetics and orthotics international*, 44(4):189–191, 2020.

[17] Anja Feldmann. Experiences from the SIGCOMM 2005 European shadow PC experiment. *ACM SIGCOMM Computer Communication Review*, 35(3):97–102, 2005.

[18] Peter A Flach, Sebastian Spiegler, Bruno Golénia, Simon Price, John Guiver, Ralf Herbrich, Thore Graepel, and Mohammed J Zaki. Novel tools to streamline the conference review process: Experiences from SIGKDD'09. *ACM SIGKDD Explorations Newsletter*, 11(2):63–67, 2010.

[19] Patricia I Fusch and Lawrence R Ness. Are we there yet? Data saturation in qualitative research. *The Qualitative Report*, 20(9):1408, 2015.

[20] Yue Guo, Fu Xin, and Stuart J Barnes. The fiction of double-blind reviewing: Evidence from the social science research network. *International Journal of Business Communication*, 59(1):48–55, 2022.

[21] Samir Haffar, Fateh Bazerbachi, and M Hassan Murad. Peer review bias: A critical review. *Mayo Clinic Proceedings*, 94(4):670–676, 2019.

[22] Cormac Herley and Paul C Van Oorschot. SoK: Science, security and the elusive goal of security as a scientific pursuit. In *IEEE Symposium on Security and Privacy (SP)*, 2017.

[23] Nicolas Huaman, Bennet von Skarczinski, Christian Stransky, Dominik Wermke, Yasemin Acar, Arne Dreißigacker, and Sascha Fahl. A large-scale interview study on information security in and attacks against small and medium-sized enterprises. In *USENIX Security Symposium*, 2021.

[24] Rebecca Isaacs. Report on the 2007 SOSP shadow program committee. *ACM SIGOPS Operating Systems Review*, 42(3):127–131, 2008.

[25] Hamid R Jamali, David Nicholas, Anthony Watkinson, Abdullah Abrizah, Blanca Rodríguez-Bravo, Chérifa Boukacem-Zeghmouri, Jie Xu, Tatiana Polezhaeva, Eti Herman, and Marzena Świgon. Early career researchers and their authorship and peer review beliefs and practices: An international study. *Learned Publishing*, 33(2):142–152, 2020.

[26] Ralph E Johnson, Kent Beck, Grady Booch, William Cook, Richard Gabriel, and Rebecca Wirfs-Brock. How to get a paper accepted at OOPSLA (panel). In *ACM Conference on Object-oriented Programming Systems, Languages, and Applications (OOPSLA)*, 1993.

[27] Douglas S Katz, Anthony V Proto, and William W Olmsted. Incidence and nature of unblinding by authors: Our experience at two radiology journals with double-blinded peer review policies. *American Journal of Roentgenology*, 179(6):1415–1417, 2002.

[28] Donald Ervin Knuth, Tracy Larrabee, Paul M Roberts, and Paul M Roberts. *Mathematical writing*, volume 14. Mathematical Association of America, 1989.

[29] Faris Bugra Kokulu, Ananta Soneji, Tiffany Bao, Yan Shoshitaishvili, Ziming Zhao, Adam Doupé, and Gail-Joon Ahn. Matched and mismatched SOCs: A qualitative study on security operations center issues. In *ACM Conference on Computer and Communications Security (CCS)*, 2019.

[30] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

[31] Roy Levin and David D Redell. How (and how not) to write a good systems paper. *ACM SIGOPS Operating Systems Review*, 17(3):35–40, 1983.

[32] Howard Lune and Bruce L Berg. *Qualitative Research Methods for the Social Sciences*. Pearson, 2017.

[33] Wendy E Mackay. The CHI conference review process: Writing and interpreting paper reviews. In *Conference Summary on Human Factors in Computing Systems (CHI)*, 1998.

[34] Saunders MacLane. Peer review and the structure of science. *Science*, 190(4215):617–617, 1975.

[35] Allison McDonald, Catherine Barwulor, Michelle L Mazurek, Florian Schaub, and Elissa M Redmiles. It's stressful having all these phones: Investigating sex workers' safety goals, risks, and practices online. In *USENIX Security Symposium*, 2021.

[36] Durga Prasanna Misra and Vinod Ravindran. Peer review in academic publishing: Threats and challenges. *The Journal of the Royal College of Physicians of Edinburgh*, 49:99–100, 2019.

[37] Haradhan Kumar Mohajan et al. Qualitative research methodology in social sciences and related subjects. *Journal of Economic Development, Environment and People*, 7(1):23–48, 2018.

[38] Lennart E Nacke. How to write CHI papers. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2018.

[39] Dan R Olsen Jr. Evaluating user interface systems research. In *ACM Symposium on User Interface Software and Technology*, 2007.

[40] Nikolaos Papanas and Dimitri P Mikhailidis. Alice through the looking-glass: Can we improve peer review? *The International Journal of Lower Extremity Wounds*, page 15347346221084784, 2022.

[41] Bryan Parno, Ulfar Erlingsson, and Will Enck. Report on the IEEE S&P 2017 submission and review process and its experiments. https://www.ieee-security.org/TC/Reports/2017/SP2017-PCChairReport.pdf. Accessed: 2022-06-04.

[42] Craig Partridge. How to increase the chances your paper is accepted at ACM SIGCOMM. *Computer Communication Review*, 28:70–74, 1998.

[43] Katharina Pfeffer, Alexandra Mai, Adrian Dabrowski, Matthias Gusenbauer, Philipp Schindler, Edgar Weippl, Michael Franz, and Katharina Krombholz. On the usability of authenticity checks for hardware security tokens. In *USENIX Security Symposium*, 2021.

[44] Marina Picciotto. New reviewer mentoring program. *Journal of Neuroscience*, 38(3):511–511, 2018.

[45] Lutz Prechelt, Daniel Graziotin, and Daniel Méndez Fernández. A community's perspective on the status and future of peer review in software engineering. *Information and Software Technology*, 95:75–85, 2018.

[46] William Pugh, PDLI 1991 Program Committee, et al. Advice to authors of extended abstracts. In *ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*, 1991.

[47] Azzurra Ragone, Katsiaryna Mirylenka, Fabio Casati, and Maurizio Marchese. On peer review in computer science: Analysis of its effectiveness and suggestions for improvement. *Scientometrics*, 97(2):317–356, 2013.

[48] Birgit Schmidt, Tony Ross-Hellauer, Xenia van Edig, and Elizabeth C Moylan. Ten considerations for open peer review. *F1000Research*, 7, 2018.

[49] Nihar B Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. Design and analysis of the NIPS 2016 review process. *Journal of Machine Learning Research*, 2018.

[50] Mary Shaw. Writing good software engineering research papers. In *IEEE International Conference on Software Engineering (ICSE)*, 2003.

[51] Richard Smith. Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99(4):178–182, 2006.

[52] Alan Snyder. How to get your paper accepted at OOPSLA. In *ACM Conference on Object-oriented Programming Systems, Languages, and Applications (OOPSLA)*, 1991.

[53] Flaminio Squazzoni, Elise Brezis, and Ana Marušić. Scientometrics of peer review. *Scientometrics*, 113(1):501–502, 2017.

[54] Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–17, 2021.

[55] Anselm Strauss and Juliet Corbin. Basics of qualitative research techniques, 1998.

[56] Mary Theofanos, Yee-Yin Choong, and Olivia Murphy. Passwords keep me safe – Understanding what children think about passwords. In *USENIX Security Symposium*, 2021.

[57] Mary-Claire van Leunen and Richard Lipton. How to have your abstract rejected. *ACM SIGACT News*, 8(3):21–24, 1976.

[58] Susan Van Rooyen, Fiona Godlee, Stephen Evans, Nick Black, and Richard Smith. Effect of open peer review on quality of reviews and on reviewers' recommendations: A randomised trial. *Bmj*, 318(7175):23–27, 1999.

[59] Mark N Wegman. What it's like to be a POPL referee; or how to write an extended abstract so that it is more likely to be accepted. *ACM SIGPLAN Notices*, 21(5):91–95, 1986.

[60] Maureen Weicher. Peer review and secrecy in the "information age". *Proceedings of the American Society for Information Science and Technology*, 45(1):1–12, 2008.

[61] Ellen B Wells. A history of scientific & technical periodicals: The origins and development of the scientific and technical press, 1665-1790. *Bulletin of the Medical Library Association*, 64(4):441, 1976.

[62] Max L Wilson. How to: Peer review for CHI (and beyond). In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2020.

[63] Andreas Zeller. How do different fields review papers? Experiences from ICSE, PLDI, and CCS. https://andreas-zeller.info/2021/07/27/Reviewing-across-fields-ICSE-PLDI-CCS.html, 2021. Accessed: 2021-08-11.

APPENDIX

A. Background

a) Paper-Reviewer matching process: Researchers submit their unpublished or revised works in any area of security and privacy field to security conferences. A typical top-tier security conference has two to four reviewing cycles, and authors can submit their research papers to any of those reviewing cycles (A,B in Figure 1). Once the submission deadline has passed, initial filtering is done to check if submitted papers comply with conference submission guidelines and if there is any violation of the call for papers. All the conference submissions should meet the submission policies mentioned on the conference website, such as paper formatting, conflict of interest, anonymous submissions, ethical considerations, and concurrent submissions. Also, papers that fail to show a clear application in security or privacy are considered out of scope and not a good fit for the conference. Remaining papers are then assigned to PC members by the Program Chairs using a specific mechanism where they match the paper with an expert PC member in that area, mainly by reading the abstracts and introductions of the submitted papers (C in Figure 1). Program Chairs try their best to match the submitted papers' area with the expertise of the program committee, which in turn helps reviewers provide detailed feedback on the technical aspects of the paper. PC members may also be interested in reviewing certain papers and, therefore, bid on them (D in Figure 1). The top portion of Figure 1 depicts the paper-reviewer matching process in a graphical form.

b) Decision-making process for submitted papers: Three to five PC members typically review a submitted paper (E,F in Figure 1) in a double-blinded fashion: author names are unknown to the reviewers and vice versa. Security conferences typically have two rounds of reviewing, and some conferences may send out early reject notifications (G in Figure 1) after the first round of reviews. Papers that are not rejected in the first round advance to the second round of reviewing, where reviewers review and discuss them (H in Figure 1). Authors of papers that pass the first round may have the option of submitting a rebuttal on the reviewers' initial decision (I in Figure 1). A rebuttal is an opportunity for authors to highlight the importance and significance of their work during the peer review process. Then, the PC, at last, may select a few papers to be discussed extensively via virtual or in-person meetings (J in Figure 1). There is a deadline by which the Program Chairs need to receive the reviews for all the papers. A submitted paper can get a decision in three ways after the second round: accept, reject, or revise (K in Figure 1). An accept means a paper is accepted for publication. A rejected paper cannot be resubmitted to that conference, and the authors must wait for another reviewing year to resubmit. The reviewers may consider some papers as promising but with additional revision work. In that case, the reviewers convey their expectations to the authors, and authors may choose to revise and resubmit their paper by a certain deadline. If the authors satisfactorily fulfill the revision tasks, the revised paper will be accepted and published at that conference.

B. Demographic Survey

1) What is your current job title?
2) Are you a member of a university-affiliated research laboratory?
   a) Are you the director of this research laboratory?
   b) How long have you been a member of this laboratory (in months)?
   c) How many researchers (students, post doctoral) do you supervise?
3) What are your research interests?
4) How many times have you been a Program Committee member in the following security conferences?
   a) IEEE Security and Privacy?
   b) USENIX Security?
   c) ACM CCS?
   d) NDSS?
5) To date, approximately, how many papers have you accepted in the above mentioned security conferences?
   OR
   How has been your acceptance rate in the above mentioned security conferences?

C. Interview Questions

Most of the interview questions were the same for both short and regular interview types. Here, we will list all the interview questions and specify which questions were specific to the regular interview type.

Review system

1) What, according to you, is your primary responsibility as a reviewer?
2) What is your approach in reviewing a security paper? Is there a disciplined approach? *(regular)*
3) Do you review papers from your field of expertise?
   a) How do you establish a confidence level for the papers? *(regular)*
   b) What do you do when you are in low confidence?
4) How do you balance the time of evaluation with the depth/quality of evaluation?
5) Do you delegate?
   a) How do you decide to delegate?
   b) Do you supervise the delegated reviews?
   c) What are your thoughts about the delegation system?
6) What are your thoughts on the rolling submission model?
   a) Which type of submissions do you prefer: single shot or multiple? *(regular)*
   b) Has rolling submissions affected your evaluations? If so, how? *(regular)*
7) What are your thoughts on the current review system? Would you like to change anything?

Evaluation metrics

1) What evaluation metrics do you have for reviewing security papers?
   a) How would you rank the importance of the metrics you mentioned?
   b) Do these metrics stay constant?
2) In your opinion, what are the characteristics of a good review?
   a) Are there any precautions that you take while writing those reviews?
3) Do you judge the paper for its fit to the venue? *(regular)*

Criticisms and Recommendations

1) What kind of papers do you like to see? *(regular)*
2) What, according to you, are some important features for a security paper to get accepted? *(regular)*
3) What do you believe are some serious misfits/red flags in a security paper that can get it rejected?
4) What are your thoughts about papers that with great research but poor presentation?
5) Do you give importance to the title of the paper?
6) What, according to you, is the importance of properly articulating a problem statement?
7) Do you find that the experiments are complete? Are the authors exploring different possibilities?
8) What are your expectations from the results and validation section of the paper?
9) What are your thoughts about the pressure of publishing quickly? *(regular)*
   OR
   Is the pressure of publishing quickly bad? *(short)*
   a) Does it affect the research? *(regular)*
   b) Does it affect the writing? *(regular)*
10) What do you think is the end goal of security research papers?

D. Codebook and Recommendations

The codebook and recommendations to write high-quality security papers can be accessed from https://github.com/sonejiananta/Security-Review-Process.

The codebook contains category names, category descriptions, and associated codes. This codebook does not contain the category evaluation metrics as Table II includes all the codes associated with the category.

Regarding recommendations, we provide general suggestions from our interviews that authors could leverage to write high-quality security research papers.

Table VI

SUGGESTIONS AND EXPECTATIONS THAT PARTICIPANTS MENTIONED IN OUR INTERVIEW TO IMPROVE THE SECURITY REVIEWING PROCESS.

Participant ID	Participant Response
P01, P02, P07, P11, P21	Scale up the Program Committee given that the community is growing.
P01, P06, P18, P21	Important to educate future reviewers about the reviewing process.
P04	- Reviews and the reviewing process might suffer when reviewers are exhausted and program chairs are not engaged. Need to have a careful and delicate balance between having rolling submissions and doing a thorough process.
	- Conferences should invite papers that discuss what things failed and reviewers should be open to that.
P05	Two phased PC discussions: anonymized and deanonymized. Instead of having completely anonymized or de-anonymized discussions, having a mix of both would solve the cons: reviewers being unnecessarily harsh in anonymized discussions, and reviewers getting intimated by senior PC in de-anonymized discussions.
P06	- Perform more experiments on the PC to reduce randomness.
	- Move away from this peer review system that hinders publication and just have arXiv.
	- Reviewers should respond to the rebuttal to avoid the other side feeling that they wasted their time.
P09	Need to understand and use rolling submissions better.
P10	- For most people in the security community, it does not matter where they publish in the top four conferences. Hence, in an ideal world, there should be a *centralized reviewing system* with single or multiple PC and a single queue for paper submissions.
	- We need papers talking about best practices to bring the community in sync. often times if a paper is not properly evaluated and draws conclusions that are too strong then follow-on work struggles to improve on top of that baseline.
	- Have a new track on papers that do reproducibility studies with an emphasis on open sourcing.
	- Community should stop doing *follow-the-trend* research to avoid fatigue effect and emphasize fundamental problems.
	- Reviewers should reward authors for stating the limitations instead of using those for rejection.
P11	Have open reviews to bring accountability in reviewers.
P12	- Limit PC to review only certain papers in a year. Conferences should not invite those who have reached their limit.
	- In the top four, attack papers get more importance and acceptance than defense (P17). Have a conference for attacks, called *cyber-attack*, where all attack papers are also reviewed by professional hackers who know how to drive those attacks.
	- Make scientific innovation a metric for evaluating security papers.
P14	Authors should read and discuss more papers from top conferences to write a paper that finally gets accepted by the big four.
P15	People who submit to crypto conferences may resubmit to CCS. Reviewing can be more efficient if *cross-community communication* existed.
P16, P18, P19	Move to journal style where all the papers that meet certain criteria are accepted. Decouple conference presentations from paper publishing. This will reduce a lot of publishing pressure. Unpolished papers that increase reviewing load will not be submitted to get initial feedback.
P17	Conferences should have better coordination in terms of submission deadlines.
P19	- Need to encourage reviewers to be constructive. Bring in the *review task force*.
	- Reviewers are on multiple PC and they might not always read the instructions. Need to concretely establish major/minor revision norms across top venues so that reviewers do not have to change their reviewing style based on the venue.