Analyzing Cyber Security Research Practices through a Meta-Research Framework

Victor Le Pochat
victor.lepochat@kuleuven.be
imec-DistriNet, KU Leuven
Leuven, Belgium

Wouter Joosen
wouter.joosen@kuleuven.be
imec-DistriNet, KU Leuven
Leuven, Belgium

ABSTRACT

Sound research practices are the foundation of valid, reliable, and trustworthy research results. The discipline of *meta-research* critically evaluates research practices and proposes new methods to improve and refine the way in which research is conducted. In this paper, we apply the framework by Ioannidis et al. [59] for categorizing meta-research work that analyzes cyber security research practices, with the goal of gaining a better understanding on the research community's efforts to examine its own research practices. We use this framework to characterize which areas of meta-research are most commonly studied, and which areas receive the most attention in terms of developing and enforcing improved research practices. We also compare these meta-research findings with experiences from another academic community, in this case the Internet measurement community, to observe areas where academic communities can learn from each other. Our work is meant as an encouragement for the research community to continue its self-reflective practices, and we hope that it can contribute to these ongoing efforts to improve cyber security research.

CCS CONCEPTS

• **Security and privacy**; • **General and reference** → *Surveys and overviews*; *Cross-computing tools and techniques*;

KEYWORDS

meta-research, research practices, research methods, cyber security research

ACM Reference Format:
Victor Le Pochat and Wouter Joosen. 2023. Analyzing Cyber Security Research Practices through a Meta-Research Framework. In *2023 Cyber Security Experimentation and Test Workshop (CSET 2023), August 07–08, 2023, Marina del Rey, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3607505.3607523

1 INTRODUCTION

Cyber security is becoming an ever more important concern in society, with the number of cyber attacks being steadily on the rise, and these attacks having more and more impact on essential infrastructure such as power grids [66]. Coupled with this trend, the security research field is rapidly expanding in size and coverage of topics. This field seeks to examine the interactions between adversaries and the systems that they target, with attacks ranging from exploiting vulnerabilities in these systems to engaging in 'traditional' crimes in an online world. The goal of such research is to study these security issues, understand the modi operandi of the attackers and defenders, and design solutions that mitigate attacks and help to better protect users against current and emerging threats. With attackers deploying more sophisticated operations to attack their targets while seeking to remain hidden from defenders, the techniques necessary to study security also need to be more elaborate and well-designed in order to keep up with this increase in complexity.

This leads to a desire for cyber security research to be as valid and sound as possible, i.e., whether the research uses methods, data sources, etc., that are appropriate, and is well evaluated and communicated, therefore generating reliable data and ultimately making the claims derived from its results well-justified. This is not only to ensure that the research's analyses and findings reflect the actual security issues occurring in the real world, but also to ensure that the mitigations that are proposed are actually effective at preventing people from being harmed.

A critical reflection on the practices that are used to conduct (security) research can help in achieving the goals of valid and sound research. Such a reflection can be framed within the discipline of *meta-research* (or *metascience*). This discipline addresses the critical evaluation of the various methods and practices in scientific research in general [59], i.e., 'does research on research'. Its goal is to understand whether current research is sufficiently sound and reliable, and develop new best practices to improve and refine the way in which research is conducted. This supports enabling research findings and claims to be credible and trustworthy, allowing to build upon them for making informed decisions, or, e.g., within the scope of security, to use them for developing better countermeasures against attacks. One of the main underlying topics is understanding the range of biases that may emerge when doing research, as they negatively impact principles such as correctness or soundness, and searching mitigation strategies for these biases [25, 48, 59].

Ioannidis et al. [59] introduced a categorization of meta-research for the different phases of the research cycle. This categorization captures five areas of meta-research:

- the *methods* used when designing and conducting studies. This includes the development of sound data collection or the use of appropriate data sets. The analysis of methods seeks to account for a variety of biases from flawed methods

CSET 2023, August 07–08, 2023, Marina del Rey, CA, USA

ACM ISBN 979-8-4007-0788-9/23/08…$15.00
https://doi.org/10.1145/3607505.3607523

and instruments, selection (e.g., sampling) biases, to inappropriate statistical analyses.
- *reporting* or communicating research, also avoiding biases due to misinterpretation of results, with a risk of the wrong conclusions being drawn if the research results are not accurately and completely conveyed.
- *reproducibility* of research, allowing others to verify research, and avoiding biases from one-off observations and instead allowing to understand if the research captures a genuine trend.
- *evaluation* of research, primarily corresponding to the peer review process, which might suffer from biases such as favoring positive results or randomness in paper acceptance.
- *incentives* for research, or an understanding of what research is favored, including perceptions related to metrics for papers (e.g., citations) or funding criteria.

In this paper, we apply this categorization by Ioannidis et al. to cyber security-related meta-research work, to gain a better understanding of the cyber security research community's efforts to examine its own research practices. We use the framework to characterize which areas of meta-research are most commonly studied, and which areas receive the most attention in terms of developing and enforcing improved research practices. We also compare these meta-research findings with experiences from another academic community, in this case the Internet measurement community, to understand how these practices may differ between research communities and observe areas where the communities can learn from each other. With our overview, we seek to understand current practices, priorities, or open questions related to the way in which cyber security research is conducted – in a way, we will engage in 'meta-research on meta-research'. Our work is meant as an encouragement for the research community to continue its self-reflective practices, and we hope that it can contribute to these ongoing efforts to improve cyber security research.

2 BACKGROUND AND RELATED WORK

The concept of "Science of Security" covers efforts to define, analyze, and promote scientific practices in security research. Herley and van Oorschot [54] and Spring et al. [105] give a historical perspective on the development of this concept, mainly through US government initiatives. The proponents of this concept see the application of the scientific method as a way to move the security field forward, make it more rigorous, and therefore help in achieving more secure systems [105]. Next to a lack of concrete definition of 'Science of Security', critics see two major flaws, current and inherent. Herley and van Oorschot [54] find that current security research is not conducted scientifically, failing to adopt practices that the 'rest of science' has accepted and learned over the course of history, and suffering from crises pertaining to the disconnect between academic research and the real world. More fundamentally, critics find there to be inherent obstacles in applying science to security, that make the scientific method inapplicable to security. Spring et al. [105] summarized these perceived obstacles as impossible experiments (as they are unethical or too risky), impossible reproducibility, a lack of laws of nature to discover, a lack of a shared ontology, and security being not a science but 'just engineering'. Spring et al. [105] refute these obstacles, finding that they rely on outdated views on science. They even state that, based on a more modern view on science, the research community currently already engages in a good scientific practice – opposing the view of Herley and van Oorschot. Overall, security research is seen as a nascent field [54], going through the 'growing pains' of any scientific field. Spring et al. [105] see security as facing unique challenges that define it as a science, but Herley and van Oorschot [54] warn not to use these challenges as an excuse not to approach security scientifically.

Herley and van Oorschot [54] use their observations of flaws in current security research practices to summarize eleven suggestions for ways forward to make security research more scientific. Al-Zyoud et al. [7] compile guidelines for essential information that should be mentioned in security papers, covering that paper's evaluation subject (what is being evaluated), subject source (who generated the evaluated artifact), and approach (how is it evaluated). They find through a survey of two top security conferences that papers often fail to include this information. Carroll et al. [25] compare security to other scientific fields and survey these fields for desirable characteristics of (deductive) experimental methods to be applied to security: falsifiable hypotheses, reproducible results, controlling of variables, and understanding of biases. Peisert et al. [83] give an example of how to design security experiments that abide by those desirable characteristics. Kott [65] develops a semiformal model of 'the science of cyber security', and applies it to example case studies to finally classify the major problem groups.

In a broader reflection on security and privacy research, Baset and Denning [16] track the evolution of topics covered in 36 years of research, describing emerging and dying topics as well as the author communities that publish on these topics, through topic modeling on a paper text corpus. Katsikeas et al. [61] similarly identify communities within security and privacy research and describe the main topics they study, through community detection on the citation graph. The Cyber Security Body of Knowledge (CyBOK) [89] systematizes foundational cyber security knowledge into 21 knowledge areas, an endeavor that is already established in "mature scientific disciplines". These works focused on examining cyber security research on the *type* of research being done; in the remainder of this paper, we focus on works that examine the *way* in which research is done.

3 PAPER SELECTION

3.1 Method

To compile our overview of meta-research work related to cyber security, we start from the categories defined in Ioannidis et al.'s framework. We use the descriptions of these categories in their work as well as on their online repository[1] to compile an initial set of general topics and keywords for each category, e.g., "peer review" for the *evaluation* category. We extend this set with relevant topics and related keywords from our own domain knowledge, e.g., "large-scale web measurements" for the *methods* category.

We then search papers that match the keywords related to each topic across Google Scholar and the ACM Digital Library, adding "web/cyber", "security/privacy" or "Internet measurement" to search

[1] https://metrics.stanford.edu/research

Table 1: We structure the selected cyber security-related meta-research works according to the framework by Ioannidis et al. [59]. For each category, we list the topics that relate to it and for which we search relevant papers. These topics are annotated with a characterization of their prevalence and general interest, as apparent from the studied cyber security-related meta-research work.

Meta-research category	Topic	Prevalence
Methods (*performing*)	Best practices and pitfalls	High
	Data collection	High
	Data sets	High
	Qualitative methods	High
	Ethical considerations	High
Reporting (*communicating*)	Science communication	Low
	Publication bias	Medium
	Preregistration	Low
Reproducibility (*verifying*)	Artifacts	High
Evaluation (*evaluating*)	Peer review	Medium
Incentives (*rewarding*)	Rankings	Medium
	Citations	Medium
	Good scientific practices	Low

queries where appropriate, and using the default search settings for both repositories. We stop our search when we seemingly reach saturation in relevant works within the search results (i.e., until we only retrieve works that are irrelevant and are only listed because they have incidental collisions with the search query). We then skim each paper's title and abstract, and retain those papers whose main subject relates to the searched topic for an in-depth reading. We also iteratively process the references of discovered papers to find additional relevant papers, to then ultimately compile the final set of papers that we discuss in our overview. We do not set an explicit time range for our search query; we observe that some works already date back 20 years or more. For areas where the relevant literature is broad and extensive, we exemplify the work in the area with papers related to the field of web security and privacy, with which we are the most familiar, but seek to extend it to the broader computer security field in general. We also leverage our familiarity with web-related research by selecting the Internet measurement community to compare research practices, as this community addresses similar topics and issues. We conduct our search for relevant meta-research work in this community in the same way as for the cyber security community.

The remainder of this paper is structured as one section per category in Ioannidis et al.'s framework – *methods*, *reporting*, *reproducibility*, *evaluation*, and *incentives* –, ending with a section concluding our overview and examining the trends observed and the lessons we can learn going forward. Table 1 summarizes the topics that we cover for each category, alongside a characterization of their (relative) prevalence within the cyber security-related meta-research work that we selected.

3.2 Limitations

Our overview focuses on identifying meta-research works that relate to one of the five categories in the framework used, and we design our search process to discover relevant works in those five areas. We therefore do not consider meta-research works that more broadly analyze research trends in the cyber security field [16, 92]. In general, and in line with other work in this space of meta-research overviews [53, 59], our goal is not to be fully systematic or provide a quantitative evaluation of cyber security-related meta-research work, but rather be descriptive and highlight and illustrate trends within this research body. It is therefore also not meant to be prescriptive in which research practices are the most appropriate or should be applied.

Our choice of Google Scholar and the ACM Digital Library as paper repositories may cause us to miss certain relevant publications that are only listed in other repositories. Nevertheless, we believe these to be the most appropriate databases for our analysis, notably as they both provide full-text search of paper contents. While Google Scholar lacks a transparent policy for indexing, its coverage has steadily increased over time [51]. In contrast, controlled databases such as Web of Science or Scopus do not exhaustively cover works in, a.o., the discipline of computer science [51]. This further motivates our choice of paper repositories, as well as their usage in prior computer science meta-research work [44, 45].

4 METHODS

Crucial to the validity of research is *conducting* it using the best scientific methods and practices possible. Otherwise, there is a risk that the experiments and their results are not truly representative or accurate. Of the meta-research categories, methods tend to be the most specific to a given research field.

Best practices and pitfalls. Across the board, many domains of (cyber) security research have seen studies on best practices and pitfalls. Given the breadth of our field, we give a non-exhaustive selection of example studies that address these issues. Rossow et al. [95] studied issues in malware research, ranging from incorrect datasets, a lack of transparency on methods or results, unrealistic settings, to a lack of safety procedures for containing the malware. Botacin et al. [23] identified twenty pitfalls in malware research through a literature review, adding issues such as closed data sets. Arp et al. [9] identified ten common pitfalls in the application of machine learning in security research, at all stages of the machine learning workflow. Eberz et al. [37] found that evaluations of behavioral biometric authentication systems failed to report error distributions, which may have led to incorrect evaluations. Sugrim et al. [106] proposed robust metrics for the evaluation of authentication systems that use machine learning, as they found that existing commonly used metrics were incomplete or hard to compare. Das et al. [33] analyzed how studies use hardware performance counters and whether they acknowledged and/or addressed limitations in using them for security applications. Van der Kouwe et al. [111] analyzed how pitfalls may affect the validity of performance benchmarking in systems security papers, if they cause flaws such as an incomplete evaluation, irrelevant or unsound results, or a lack of reproducibility. Polakis et al. [87] described the various methods used in all phases of a measurement study on social networks, from ethical considerations to data collection and processing techniques.

Data collection. Data collection is a crucial phase of a research project, as all subsequent analyses and results depend on the accuracy and validity of the acquired data. A common denominator to many studies related to web security, web privacy, and Internet measurement in general is the use of *large-scale web measurements* for this data collection. Pour et al. [88] survey the use of various Internet measurement techniques in recent cyber security work, creating a taxonomy based on the type of security issue that was studied. Unsurprisingly, recent work has critically analyzed methods that are regularly used as part of large-scale web measurements, often formulating recommendations for how researchers should use them or proposing improved solutions. Already in 2004, Paxson [82] outlined strategies for sound Internet measurement, such as calibrating measurements, inspecting raw data, and designing for reproducibility. Collecting web data often involves 'crawlers' that scrape and store a web page's contents. Ahmad et al. [5] compared web crawlers with varying technologies and feature sets, finding that the choice of crawler may significantly impact measurements. Zeber et al. [118] compared crawlers with each other and with human-generated traffic, finding that crawling results can vary significantly over time as well as across platforms. Krumnow et al. [68] analyzed how the popular OpenWPM crawling framework is detectable and how its measurements can therefore be prevented or poisoned, which introduces errors into the obtained results. Szurdi et al. [107] found that cybercrime must be measured using multiple vantage points and profiles, with special attention to cloaking, in order to obtain reliable results. Jueckstock et al. [60] measured how the browser configuration and network vantage point cause significant biases for web privacy and security measurements. Demir et al. [35] measured how different experimental setups such as the browser, location, user interaction, and time may significantly influence web measurements. Roth et al. [96] measured how websites have inconsistent security policies between browsing profiles. This also has implications for measurements, as these may misreport findings if website behavior changes between profiles or page accesses. Wan et al. [115] found that Internet scan results depend on their origin, i.e., the location, network type, or protocol. Cassel et al. [26] found that frameworks for emulating mobile browsers on desktop may produce results that differ from real mobile browsing, causing incorrect findings about the mobile web specifically.

Data sets. Data sets form another subject of scrutiny, as there are often questions about reliability and validity, especially if these data sets are difficult to acquire or generated opaquely (e.g., by a commercial third party) [84]. For example, VirusTotal is a commonly used, but commercial source for labeling entities such as files and URLs as benign or malicious. Peng et al. [85] studied how reliable VirusTotal is for detecting phishing websites, finding varying and inadequate detection performance as well as inconsistent labeling. Zhu et al. [122] studied how researchers use VirusTotal to label malware, and analyzed how reliable the data set is in terms of accuracy, independence, and stability over time. More broadly, Feal et al. [40] found that blocklists are opaquely constructed, may be slow to update, may either label records differently or share labels and therefore have high overlap, and are not always well documented. Scheitle et al. [99] and Le Pochat et al. [70] found that commonly used rankings of top domains exhibit undesirable properties for research, such as opaque methods, volatility, disagreement, and vulnerability to manipulation. Vallina et al. [109] found similar shortcomings in terms of opaque methods and disagreement for third-party domain categorization services. Researcher-generated data sets may also suffer from a lack of coverage. For example, Cuevas et al. [32] found that scraping-based measurements 'by proxy' on online anonymous marketplaces systematically underestimate metrics such as revenue or the number of discovered listings.

Another example is the tension between using *real-world versus simulated data sets*. Real-world data has the perception of being more accurate and representative, but comes with substantial challenges for data collection and publication, not in the least due to the need to obtain permission to collect data and publish a (usable) anonymized version if the data pertains to human behavior [1]. Simulated data overcomes these issues and better allows for repeatable and comparable security experiments, but the community often questions its validity, as it is difficult to assess the quality and representativeness of generated data [1]. Indeed, problems with simulated data sets are known to exist and significantly affect research results. For example, the data collection strategy affects the perceived performance of website fingerprinting attacks [94], and standard data sets for evaluating intrusion detection systems contain significant noise or even errors that impact attack performance [38, 72].

Qualitative methods. A particular body of research focuses on correctly applying qualitative methods, usually for studying usable security and privacy. This body of research usually entails collecting data from humans through specific methods (e.g., interviews) and analyzing that data qualitatively as opposed to quantitatively [67]. Fujs et al. [47] surveyed the use of such qualitative methods in security research, finding that interviews are most common. Since the rest of the security community may be unfamiliar with these methods, as their research tends to be quantitative, special care is taken to show the validity of research results that originate from these qualitative methods. Schechter [98] summarized pitfalls and good practices for describing security and privacy experiments that involve human subjects, including the experiment design and setup but also the reporting on statistical tests. Redmiles et al. [90] compiled guidelines for conducting surveys in security and privacy studies, including how to design the questions, achieve a representative sample of participants, and test the questions upfront. Ortloff et al. [80] examined the process of coding (or labeling) data qualitatively for usable security and privacy studies, recommending that the number of coders should be adapted to the data type. Unfortunately, these best practices do not appear to always be followed. Groß [48] analyzed the reliability of statistical analyses in security user studies, finding systemic issues such as low statistical power that put the validity of the results into questions. They use their findings to provide recommendations for supporting and requiring more reliable studies. Kaur et al. [62] surveyed human factors security research over ten years, finding, a.o., biases in population sampling, and a lack of theorization that should be the result from inductive methods such as grounded theory.

Ethical considerations. Ethical considerations for conducting research are meant to ensure that no harm is done while studying a

security or privacy system. Existing frameworks for ethical review may not be adapted to the needs of the cyber security field. Van der Ham and van Rijswijk-Deij [110] describe the shortcomings of processes involving ethical review boards such as an Institutional Review Board for Internet measurements as these often fall out of those boards' scope, and design an alternative framework with guidelines for ethical measurements. Macnish and van der Ham [74] continue this line for security research ethics, using two case studies of controversial studies to motivate how current methods and guidance are inadequate, as review boards provide insufficient guidance and ethical oversight for practitioners is lacking. It is then often up to the community itself to set their own ethical standards and provide guidelines to researchers. The Menlo Report [10], which outlines the principles of respect for persons, beneficence, justice, and respect for law and public interest, is commonly seen as the main framework for ethical computer science research. Reidsma et al. [91] propose a practical framework for addressing the specificities of cybersecurity research when passing through ethical review boards or designing relevant university policies. Allman and Paxson [8] provide guidelines for ethically sharing data from network measurements, preventing risks such as privacy leaks and setting acceptable use policies including appropriate acknowledgments. Conducting research ethically is increasingly enforced at top-tier security conferences, with measures ranging from mandatory descriptions of the ethical considerations made, to research ethics committees reviewing potentially contentious cases [56]. Zhang et al. [119] surveyed ethical considerations in computer security research, including what ethical requirements conferences impose, how papers discuss ethics, and whether researchers apply ethical practices. They also give recommendations on how to learn about ethical requirements, apply them in practice, and describe them appropriately. Feitelson [41] uses the 2021 controversy on the "Hypocrite Commits" paper, which analyzed developer reaction to intentionally introduced bugs, as a starting point for surveying developers and researchers on what they consider ethically acceptable research practices, formulating recommendations based on the insight that developers are willing to contribute to research if it is conducted transparently and in good faith. Pauley and McDaniel [81] describe the ethical considerations seen recently in practice in Internet measurement research, finding that this community still lacks a cohesive approach.

5 REPORTING

Communicating research well is essential for ensuring that it reaches the intended audience(s) without being misinterpreted or misrepresented.

Science communication. A research study and its results can be of interest to multiple stakeholders. Fellow researchers can build upon prior work, relate the findings of prior work to their work, or learn about methods and data sets used. Policymakers can use research results as a foundation for new regulations that seek to improve security and privacy, e.g., by prohibiting privacy-invasive practices that research has found in the wild. Industry companies can integrate state-of-the-art research solutions into their tools or processes to improve their security posture. Finally, researchers can communicate the real-world impact of their findings to the public at large, e.g., directly or through the media, and give actionable guidance such that the public can improve their own security and privacy practices. However, it appears there is little research into how these different forms of science communication are used in security research. As one example, Narayanan and Lee [78] reflected on the success of their engagement with policymakers, carriers, journalists, and users for their security policy audit of SIM swapping attacks. Pennekamp et al. [86] proposed a framework for conducting cybersecurity research for industrial applications, and collaboration with companies to enable such interdisciplinary research.

Publication bias. Next to studying research that *is* communicated, there is a concern for research that *is not* being made public, either because its results are negative, deemed insignificant, or deemed undesirable, or because it is kept proprietary. Publication bias broadly refers to any bias that may cause specific research to be overrepresented or underrepresented in what is actually published, based on the outcomes of that research [36, 48]. The most commonly regarded form is the omission of negative results, where a hypothesis could not be confirmed nor falsified, or an expected phenomenon was not observed, because researchers are less inclined to submit them for publication, and reviewers and other research gatekeepers (e.g., editors, funders) are less inclined to appreciate them. This causes positive results to be overrepresented, extending to an incentive to always find (statistically significant) results. This may trigger questionable practices such as performing many analyses on data until significant results are found ("p hacking"). Not publishing negative results may also mean that other researchers waste time and resources retrying those experiments, only to find (and discard) the negative results. This bias also forms a threat for meta-analyses through literature surveys, as these may erroneously conclude only positive findings, as the negative results that run counter to those findings have simply not been published.

Groß [48] showed empirically that the cyber security user study field suffers from a publication bias, with smaller studies without significant results going unpublished. Such user studies might be among the type of study that is most vulnerable to publication bias, as they heavily rely on statistical inferences across relatively small populations, where there is a higher risk of selectively executing analyses and reporting results that support a hypothesis as well as reporting results with small effect sizes and low statistical power. In security, publication bias may also be due to potential underreporting of vulnerabilities, where papers are not submitted or published in the first place, for example if the vulnerable entity requests that the publication is delayed or stopped altogether, leading to unreliable aggregate vulnerability statistics [27]. Afterwards, there is also a belief that within the research community, papers presenting attacks are more readily accepted than papers proposing defenses [104], potentially giving an appearance that attacks are more prevalent (if they are allowed to be published, as mentioned above). Boucher and Anderson [24] discuss one example of the difficulties that may emerge in academically publishing a discovered vulnerability, as their public disclosure was used as grounds for paper rejection.

Preregistration. One proposed solution to alleviate some publication bias is preregistration, where the intended aim, research

questions, hypotheses, methods, data sets, analyses, etc. are established in a document before the actual experiments take place [79]. However, it seems that this practice is very uncommon in security and privacy research, possibly also due to the exploratory or vulnerability-oriented nature of many studies, which does not always allow for a detailed experimental design upfront.

6 REPRODUCIBILITY

Verifying research can be achieved by seeking to reproduce it. Successfully repeating a study serves as a confirmation of its results, and increases the likelihood that the studied hypothesis is correct [77]. Conversely, failing to repeat a study puts the validity of its results into question, in particular when this failure is due to flawed methods. The challenges in reproducing past work has given rise to a perceived 'replication crisis' [58], although this notion is also being challenged [39].

Artifacts. The ability to reproduce studies hinges on the availability and quality of (descriptions of) the artifacts used, comprising data sets, methods and tools. One set of high-level guiding principles are the FAIR principles [117]: artifacts should be findable, accessible, interoperable, and reusable. Within the field of cyber security, efforts to support scientific reproducibility focus on sharing data sets and tools to allow for repeating studies and building upon prior work. Benzel [17] describes how associations such as ACM [2] and USENIX [108] have an artifact evaluation process where papers can receive badges based on the extent to which artifacts are available, functional, and able to be used for reproducing results. However, these badges may give a false sense of research validity, as the fact that, e.g., methods are reproducible does not mean that they are appropriate or complete [86]. Balenson et al. [14] introduced SEARCCH, an online catalog supporting better discovery of security research artifacts. Hamm et al. [52] found that security papers with user studies generally publish their questionnaires or interview guides, but not the actual participant data that was used in the analysis. More broadly in systems research, Frachtenberg [44] found that the availability of artifacts quickly decays over time. In Web measurement research, Demir et al. [35] evaluated recent work on 18 criteria that enable replicability and reproducibility, finding that they often fail to meet these criteria and omit crucial information that would allow reproduction.

The Internet measurement community has recently made reproducibility a topic of community debate and academic work. Reproducibility was the focus of a workshop at the 2017 SIGCOMM conference. Based on this workshop, Bajpai et al. [13], Saucez and Iannone [97] and Scheitle et al. [100] identified challenges for reproducibility, including ambiguous definitions, unavailability of authors or artifacts, and a lack of incentives. They formulate recommendations to improve reproducibility such as artifact review and badges. Notably, the IMC conference has not implemented such a review and badging process, unlike the security community. Bonaventure [22] and Flittner et al. [43] surveyed authors at computer networking conferences on the composition and availability of paper artifacts. Among their findings, they discuss obstacles such as insufficient descriptions of software and data sets, incomplete tools or broken links, and the influence of research cultures on the type of tools and data sets used, which impacts artifact availability. In 2018, reproducibility was the subject of a Dagstuhl Seminar [11], which resulted in a set of recommendations and best practices for documenting the research process to allow for reproduction [12]. Zilberman and Moore [123] describe experiences with and recommendations for the artifact evaluation process at networking conferences. IMC 2019 featured a 'reproducibility track' [3], inviting short papers replicating prior work, but these were only presented as posters, i.e., not featured at the main conference track.

7 EVALUATION

Peer review. The primary way of *evaluating* research is through the peer review process, where fellow scientists judge the quality of a research paper, such as the soundness of its methods or the originality of its findings, and decide whether it is acceptable for formal publication. This process is meant to maintain the integrity of science [102]. However, as peer review remains a human endeavor, concerns prevail about subjectivity in the review process leading to subpar papers with fundamental flaws being published while papers that advance the state of the art are rejected. Ultimately, this could lead to spreading false scientific beliefs and hindering scientific progress, respectively.

In 2022, Soneji et al. [104] studied the peer review process in computer security through interviews with PC[2] members for top-tier conferences. Among their key findings, they found that reviewers did not share common evaluation metrics. Only novelty was a metric considered by most reviewers, although they acknowledged that this was a subjective metric. In contrast, 'red flags' that give reason to reject a paper are more diverse and concrete. This suggests that reviewers may have a mindset of looking for reasons to reject rather than accept papers. While reviewers felt the responsibility to provide high-quality reviews, high workloads, a lack of accountability, and a PC that has insufficient expertise or experience to review a paper run counter to this goal. These yield a risk of subjective reviews and contributes to a sense of 'randomness' as to whether a paper is deemed scientifically worthy. One 'countermovement' to the focus on novelty is the increased appreciation for Systemization of Knowledge papers, which evaluate and systematize existing knowledge on a specific research topic [18]. Specifically for usable security and privacy, Ortloff et al. [80] surveyed reviewers on their criteria for qualitative studies. Overall, the reviewers expected detailed methods descriptions and the use of some method for reaching agreement among coders. There was more disagreement on acceptable task division and agreement levels across coders.

The top-tier security conferences have recently moved to a more journal-style model, with multiple submission deadlines and the possibility of revisions. As one possible word of encouragement, Vardi [112] posits that the time and workload pressure brought about by the preference in computer science for conferences over journals reduces review quality. The trend ostensibly started with IEEE S&P adopting rolling deadlines in 2018 [55]. Interestingly, IEEE S&P has since started to backtrack, scrapping revisions for its 2024 edition, due to a concern that papers were no longer being immediately accepted, but instead (unnecessarily) put through a

[2]The collective of reviewers for one scientific conference is also known as the 'program committee' or PC.

revision process to cater to reviewer interests [57]. The top conferences also start to give more attention to encouraging good reviewing practices, including adding public meta-reviews, avoiding re-reviewing by the same reviewers of a resubmitted paper [57] or recognition through awards. Frachtenberg and Koster [46] surveyed authors of papers at systems conferences, including top security venues. Among their findings, they conclude that authors find review rebuttals and longer reviews very valuable. Sion [103] discusses the shortcomings of the peer review process for computer science conferences from his viewpoint as a PC chair, and proposes to request reviewers to rate more papers more favorably to then increase agreement on whether a paper should be accepted. Lee [71] laments a "toxic culture of rejection" with computer science conferences chasing low acceptance rates, with rejections of otherwise high-quality papers on the basis of lack of novelty or obviousness causing "detrimental effects" to the community.

The Internet measurement and computer networking community has had a longer (academic) experience and experimentation regarding the peer review process. In 2005, Feldmann reported on her experience organizing a 'shadow PC' (also called 'student PC') for SIGCOMM 2005 [42], a parallel PC of mostly junior researchers that runs similarly to a real PC but does not actually decide on the papers that are accepted to the conference. The goal is to give novice researchers an opportunity to experience the review process first hand. Among the findings, Feldmann discussed the differences in paper decisions between the actual and shadow PC, observed a more varied review depth and breadth for the shadow PC, and noted that the experience was well received. The concept of a shadow PC also made it to some editions of security conferences, e.g., USENIX Security in 2014 and 2015, and IEEE S&P from 2016 to 2021. In 2008, Mogul and Anderson [76] summarized prior and future work on best practices for organizing the conference review process. Schulzrinne [101] opines that double-blind reviewing, where authors are anonymous to reviewers, improves perceived fairness but must be implemented judiciously to account for its unintended side effects and limitations such as properly addressing submitted papers that build upon prior publications. Beverly and Allman introspectively measured the IMC 2010 review process [21], with the goal of improving transparency and the process itself. They focused in particular on whether review biases can be measured empirically. The 2011 through 2013 editions of the IMC conference published (meta-)reviews openly, but a community survey led to this practice being discontinued as there were no apparent benefits [6]. Keshav [64] commented on the "spirit of harsh criticism" that led to an attitude in measurement conferences and the computer science field at large of finding reasons to reject rather than accept a paper. Mogul [75] provided advice on how to reduce 'hypercriticality' and negativity in the reviewing process.

8 INCENTIVES

Rewarding research involves evaluating the quality, value, and impact of research, and providing the right incentives and support for research, including appropriate funding.

Rankings. Research is often compared by compiling rankings. Based on the conference acceptance rate and community input, several conference rankings are used as indicators for quality, both specific to security and privacy venues [49, 121] and for all of computer science [29, 63], with the 'top-tier' conferences being the most attractive and easiest to identify [69]. There is a connection to the peer review process, as the restrictiveness of selecting papers there leads to a division of conferences into tiers of prestige and selectivity. For example, Ortloff et al. [80] commented that replication of qualitative usable security and privacy studies is worthwhile for improving insights, but that such papers may struggle to be accepted to highly valued conferences, therefore disincentivizing researchers from taking the risk of doing such "underappreciated" replication work given a "publish or perish" culture. Publication counts at the most reputable conferences are also used to compile rankings of researchers (e.g., Balzarotti's 'System Security Circus' [15]) and/or institutions (e.g., CSRankings [19]), next to survey-based approaches for the latter [113]. Such rankings are not considered reliable or useful by all, with criticism ranging from questionable methods for survey-based rankings [20, 113] to biases towards established, US-based, 'traditional' institutions and conferences [50]. More fundamentally, such rankings and the data they are based on may say very little about actual quality or other aspects that are harder to measure.

Citations. Next to assigning value to a research work based on where it is published, citations by other papers are usually used to quantitatively measure the subsequent impact of individual works on the academic field. Rieck [93] maintains a list of highly cited security papers, again only at 'tier 1' and 'tier 2' conferences. Frachtenberg [45] analyzed trends within citations of computer systems papers, with security being one of the most cited subfields. Wendzel et al. [116] measured potential factors influencing the citation count of information security papers, using bibliometrics to draw conclusions that, a.o., papers with longer abstracts and more references are cited more often, as well as journal papers, although they also suggest this may be due to a higher number of low-tier conferences with many papers with few citations skewing the data. Vrhovec et al. [114] expanded this analysis, with a contrasting finding that top conference papers are cited more often than journal papers, and described how paper title lengths and references may impact citation counts. While these findings may be statistically validated, there is however no proposed theory that would clearly explain these trends. Overall, the creators of these rankings and counts are often quick to stress that they are merely informal metrics [15, 49, 93, 121] and "are insufficient to characterize all aspects contributing to the relevance of scientific work" [93]. Citations, venue reputation, and quality may also have little relation to each other [34]. 'Altmetrics' are designed to measure research impact online beyond only citations, comprising metrics such as read counts, social media mentions, or media coverage [4]. However, these may not (yet) be a viable alternative [30].

Good scientific practices. Particular attention also goes to incentivizing good scientific practices beyond pure publications. For example, for reproducibility, Collberg and Proebsting [28] proposed additional research funding tied to enforceable 'sharing contracts' in systems research. As another potential incentive, Zheng et al. [120] found that security papers that create and share data sets are likely to be cited more often. Frachtenberg [44] found that systems papers with shared artifacts were cited around 75% more often than those

without. On the front of evaluation and peer reviews, Crowcroft et al. [31] proposed mechanisms to incentivize authors, reviewers, and the community to submit higher-quality papers and reviews as well as reward reviewing, and therefore improve the review process. Longstaff et al. [73] found that the time pressure to publish ('breakthrough') results reduces the quality of research experiments and a worse application of a scientific approach. They suggested funding agencies could incentivize security research work that is more based in the scientific method.

9 DISCUSSION AND CONCLUSION

From our overview of cyber security meta-research work, we can see that gradually more work is being published that critically examines cyber security research practices, with varying emphasis on the different categories of meta-research. A strong focus is put on improving methods, especially from an observation that significant pitfalls can present themselves and may be prevalent due to a lack of awareness or critical study. This has also led to the identification of best practices for several domains of (cyber) security research. We see a growing trend of critical analyses of state-of-the-art and commonly used data collection tools or data sets, both identifying flaws within them and iteratively proposing improvements. However, given a lack of a central repository for best practices or another clear way to discover them, it may depend on researchers themselves to be aware of the latest developments, especially when prior work still relies on outdated practices. Adoption of best practices may therefore be slow. The enforcement of using such state-of-the-art methods appears to become a task left for the peer review process, where (individual) reviewers are expected to equally be aware of current best practices and require that submitted work applies them, which may not always be tenable. A collective, up-to-date, and easily referenceable resource of best practices may therefore be helpful to support researchers in selecting the most appropriate methods and data sets for their study.

Compared to these methods, the other categories in our framework are less commonly addressed in cyber security meta-research literature, but in contrast are enforced or encouraged more strictly or explicitly, providing more clarity as to what the research community expects. We also see a noticeable evolution in this space, with ongoing changes to various research processes. For example, ethical considerations are rapidly gaining prominence and have become a required element of cyber security research projects and papers. Similarly, the introduction of artifact evaluation processes and their encouragement through badging supports reproducibility. The peer review process also evolves to incorporate revisions or increase accountability through practices such as public reviews. These fit a trend towards aspiring higher scientific rigor and objectivity, and should therefore be welcomed, although these improvement efforts are mostly applied to the processes around the publication of research papers, and less the actual scientific content itself.

Improving the soundness and validity of research should be a collective community effort, and there should be venues where the processes and practices that form research can be discussed. For example, in computer security, the *Cyber Security Experimentation and Test* (CSET) and *Learning from Authoritative Security Experiment Results* (LASER) workshops are of interest. Simultaneously, the community can learn from the experiences of other research communities, as was illustrated throughout with examples from the Internet measurement community – observe for example how a top Internet measurement conference stopped publishing meta-reviews in 2013 due to an apparent lack of benefits, yet a top-tier security conference introduced them ten years later. Introspectively, we find that the framework initially proposed by Ioannidis et al. within the biology community generalizes to and is reusable for analyzing research practices within the cyber security community, further encouraging knowledge sharing across disciplines. However, we find that the topics that populate the categories within the framework are community-specific, with their own accents and prevalence, and sufficient domain knowledge is therefore required to fully apply this framework to the body of meta-research work within our community. For example, we see that preregistration is nearly nonexistent in our community, although it is gaining traction in other scientific disciplines. Nevertheless, certain themes are more common to all scientific fields, such as concerns on publication bias or the proper incentivization of scientific research.

This iterative process of reflecting about the way in which cyber security research is conducted, implementing improvements, and evaluating how effective they are – i.e., applying the scientific process to study our research –, can help to make cyber security research become more reliable and trustworthy. By proxy, this further contributes to making cyber security itself a (more) scientific practice, and to helping ensure that the research done within the field proves to be beneficial for improving the state of cyber security and helping to protect against current and emerging threats.

ACKNOWLEDGMENTS

We thank the anonymous reviewers as well as Lieven Desmet, Michel van Eeten, Maciej Korczyński, Frank Piessens, and Katrien Verbert for their valuable feedback. This research is partially funded by the Research Fund KU Leuven, and by the Flemish Research Programme Cybersecurity.

REFERENCES

[1] Sebastian Abt and Harald Baier. 2014. Are We Missing Labels? A Study of the Availability of Ground-Truth in Network Security Research. In *3rd International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*. 40–55. https://doi.org/10.1109/badgers.2014.11

[2] Association for Computing Machinery 2020. *Artifact Review and Badging*. Association for Computing Machinery. https://www.acm.org/publications/policies/artifact-review-and-badging-current

[3] Association for Computing Machinery 2019. *Call For Posters, ACM IMC 2019, Amsterdam, The Netherlands*. Association for Computing Machinery. https://conferences.sigcomm.org/imc/2019/call-for-posters/

[4] Euan Adie and William Roe. 2013. Altmetric: enriching scholarly content with article-level discussion and metrics. *Learned Publishing* 26, 1 (2013), 11–17. https://doi.org/10.1087/20130103

[5] Syed Suleman Ahmad, Muhammad Daniyal Dar, Muhammad Fareed Zaffar, Narseo Vallina-Rodriguez, and Rishab Nithyanand. 2020. Apophanies or Epiphanies? How Crawlers Impact Our Understanding of the Web. In *The Web Conference 2020*. 271–280. https://doi.org/10.1145/3366423.3380113

[6] Aditya Akella and Nina Taft. 2014. *IMC 2014 Decision on Public Reviews*. https://conferences.sigcomm.org/imc/2014/news3.html

[7] Mahran Al-Zyoud, Laurie Williams, and Jeffrey C. Carver. 2017. Step One Towards Science of Security. In *2017 Workshop on Automated Decision Making for Active Cyber Defense*. 31–35. https://doi.org/10.1145/3140368.3140374

[8] Mark Allman and Vern Paxson. 2007. Issues and Etiquette Concerning Use of Shared Measurement Data. In *7th ACM SIGCOMM Conference on Internet Measurement*. 135–140. https://doi.org/10.1145/1298306.1298327

[9] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. 2022.

Dos and Don'ts of Machine Learning in Computer Security. In *31st USENIX Security Symposium*. 3971–3988.

[10] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. 2012. The Menlo Report. *IEEE Security & Privacy Magazine* 10, 2 (2012), 71–75. https://doi.org/10.1109/msp.2012.52

[11] Vaibhav Bajpai, Olivier Bonaventure, Kimberly Claffy, and Daniel Karrenberg. 2019. Encouraging Reproducibility in Scientific Research of the Internet (Dagstuhl Seminar 18412). *Dagstuhl Reports* 8, 10 (2019), 41–62. https://doi.org/10.4230/DagRep.8.10.41

[12] Vaibhav Bajpai, Anna Brunstrom, Anja Feldmann, Wolfgang Kellerer, Aiko Pras, Henning Schulzrinne, Georgios Smaragdakis, Matthias Wählisch, and Klaus Wehrle. 2019. The Dagstuhl Beginners Guide to Reproducibility for Experimental Networking Research. *ACM SIGCOMM Computer Communication Review* 49, 1 (February 2019), 24–30. https://doi.org/10.1145/3314212.3314217

[13] Vaibhav Bajpai, Mirja Kühlewind, Jörg Ott, Jürgen Schönwälder, Anna Sperotto, and Brian Trammell. 2017. Challenges with Reproducibility. In *Reproducibility Workshop*. 1–4. https://doi.org/10.1145/3097766.3097767

[14] David Balenson, Terry Benzel, Eric Eide, David Emmerich, David Johnson, Jelena Mirkovic, and Laura Tinnel. 2022. Toward Findable, Accessible, Interoperable, and Reusable Cybersecurity Artifacts. In *15th Workshop on Cyber Security Experimentation and Test*. 65–70. https://doi.org/10.1145/3546096.3546104

[15] Davide Balzarotti. 2023. *System Security Circus*. https://www.s3.eurecom.fr/~balzarot/security-circus/

[16] Aniqua Baset and Tamara Denning. 2019. A Data-Driven Reflection on 36 Years of Security and Privacy Research. In *12th USENIX Conference on Cyber Security Experimentation and Test*. https://www.usenix.org/system/files/cset19-paper_baset.pdf

[17] Terry Benzel. 2023. Security and Privacy Research Artifacts: Are We Making Progress? *IEEE Security & Privacy* 21, 1 (January 2023), 4–6. https://doi.org/10.1109/MSEC.2022.3222887

[18] Terry Benzel and Frank Stajano. 2020. IEEE Euro S&P: The Younger Sibling Across the Pond Following in Oakland's Footsteps. *IEEE Security & Privacy* 18, 3 (2020), 6–7. https://doi.org/10.1109/MSEC.2020.2980180

[19] Emery Berger. 2022. *CSRankings: Computer Science Rankings*. https://csrankings.org/

[20] Emery Berger, Stephen M. Blackburn, Carla Brodley, H. V. Jagadish, Kathryn S. McKinley, Mario A. Nascimento, Minjeong Shin, Kuansan Wang, and Lexing Xie. 2019. GOTO Rankings Considered Helpful. *Commun. ACM* 62, 7 (June 2019), 29–30. https://doi.org/10.1145/3332803

[21] Robert Beverly and Mark Allman. 2012. Findings and Implications from Data Mining the IMC Review Process. *ACM SIGCOMM Computer Communication Review* 43, 1 (January 2012), 22–29. https://doi.org/10.1145/2427036.2427040

[22] Olivier Bonaventure. 2017. The January 2017 issue. *ACM SIGCOMM Computer Communication Review* 47, 1 (January 2017), 1–3. https://doi.org/10.1145/3041027.3041028

[23] Marcus Botacin, Fabricio Ceschin, Ruimin Sun, Daniela Oliveira, and André Grégio. 2021. Challenges and pitfalls in malware research. *Computers & Security* 106, Article 102287 (July 2021). https://doi.org/10.1016/j.cose.2021.102287

[24] Nicholas Boucher and Ross Anderson. 2022. Talking Trojan: Analyzing an Industry-Wide Disclosure. In *2022 ACM Workshop on Software Supply Chain Offensive Research and Ecosystem Defenses*. 83–92. https://doi.org/10.1145/3560835.3564555

[25] Thomas E. Carroll, David Manz, Thomas Edgar, and Frank L. Greitzer. 2012. Realizing Scientific Methods for Cyber Security. In *2012 Workshop on Learning from Authoritative Security Experiment Results*. 19–24. https://doi.org/10.1145/2379616.2379619

[26] Darion Cassel, Su-Chin Lin, Alessio Buraggina, William Wang, Andrew Zhang, Lujo Bauer, Hsu-Chun Hsiao, Limin Jia, and Timothy Libert. 2021. OmniCrawl: Comprehensive Measurement of Web Tracking With Real Desktop and Mobile Browsers. *Proceedings on Privacy Enhancing Technologies* 2022, 1 (September 2021), 227–252. https://doi.org/10.2478/popets-2022-0012

[27] Steve Christey and Brian Martin. 2013. *Buying Into the Bias: Why Vulnerability Statistics Suck*. Technical Report.

[28] Christian Collberg and Todd A. Proebsting. 2016. Repeatability in Computer Systems Research. *Commun. ACM* 59, 3 (February 2016), 62–69. https://doi.org/10.1145/2812803

[29] Computing Research and Education Association of Australasia 2022. *CORE Rankings Portal*. Computing Research and Education Association of Australasia. https://www.core.edu.au/conference-portal

[30] Rodrigo Costas, Zohreh Zahedi, and Paul Wouters. 2015. Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology* 66, 10 (2015), 2003–2019. https://doi.org/10.1002/asi.23309

[31] Jon Crowcroft, S. Keshav, and Nick McKeown. 2009. Viewpoint Scaling the Academic Publication Process to Internet Scale. *Commun. ACM* 52, 1 (January 2009), 27–30. https://doi.org/10.1145/1435417.1435430

[32] Alejandro Cuevas, Fieke Miedema, Kyle Soska, Nicolas Christin, and Rolf van Wegberg. 2022. Measurement by Proxy: On the Accuracy of Online Marketplace Measurements. In *31st USENIX Security Symposium*. 2153–2170.

[33] Sanjeev Das, Jan Werner, Manos Antonakakis, Michalis Polychronakis, and Fabian Monrose. 2019. SoK: The Challenges, Pitfalls, and Perils of Using Hardware Performance Counters for Security. In *2019 IEEE Symposium on Security and Privacy*. 20–38. https://doi.org/10.1109/SP.2019.00021

[34] James Davis. 2019. Do top conferences contain well cited papers or junk? https://doi.org/10.48550/arxiv.1911.09197 arXiv:1911.09197

[35] Nurullah Demir, Matteo Große-Kampmann, Tobias Urban, Christian Wressnegger, Thorsten Holz, and Norbert Pohlmann. 2022. Reproducibility and Replicability of Web Measurement Studies. In *ACM Web Conference 2022*. 533–544. https://doi.org/10.1145/3485447.3512214

[36] Kay Dickersin. 1990. The Existence of Publication Bias and Risk Factors for Its Occurrence. *JAMA* 263, 10 (March 1990), 1385–1389. https://doi.org/10.1001/jama.1990.03440100097014

[37] Simon Eberz, Kasper B. Rasmussen, Vincent Lenders, and Ivan Martinovic. 2017. Evaluating Behavioral Biometrics for Continuous Authentication: Challenges and Metrics. In *2017 ACM on Asia Conference on Computer and Communications Security*. 386–399. https://doi.org/10.1145/3052973.3053032

[38] Gints Engelen, Vera Rimmer, and Wouter Joosen. 2021. Troubleshooting an Intrusion Detection Dataset: the CICIDS2017 Case Study. In *2021 IEEE Security and Privacy Workshops*. 7–12. https://doi.org/10.1109/SPW53761.2021.00009

[39] Daniele Fanelli. 2018. Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences* 115, 11 (2018), 2628–2631. https://doi.org/10.1073/pnas.1708272114

[40] Álvaro Feal, Pelayo Vallina, Julien Gamba, Sergio Pastrana, Antonio Nappa, Oliver Hohlfeld, Narseo Vallina-Rodriguez, and Juan Tapiador. 2021. Blocklist Babel: On the Transparency and Dynamics of Open Source Blocklisting. *IEEE Transactions on Network and Service Management* 18, 2 (2021), 1334–1349. https://doi.org/10.1109/TNSM.2021.3075552

[41] Dror G. Feitelson. 2021. "We do not appreciate being experimented on": Developer and Researcher Views on the Ethics of Experiments on Open-Source Projects. https://doi.org/10.48550/arxiv.2112.13217 arXiv:2112.13217

[42] Anja Feldmann. 2005. Experiences from the Sigcomm 2005 European Shadow PC Experiment. *ACM SIGCOMM Computer Communication Review* 35, 3 (July 2005), 97–7–102. https://doi.org/10.1145/1070873.1070889

[43] Matthias Flittner, Mohamed Naoufal Mahfoudi, Damien Saucez, Matthias Wählisch, Luigi Iannone, Vaibhav Bajpai, and Alex Afanasyev. 2018. A Survey on Artifacts from CoNEXT, ICN, IMC, and SIGCOMM Conferences in 2017. *ACM SIGCOMM Computer Communication Review* 48, 1 (April 2018), 75–80. https://doi.org/10.1145/3211852.3211864

[44] Eitan Frachtenberg. 2022. Research artifacts and citations in computer systems papers. *PeerJ Computer Science* 8 (February 2022), e887. https://doi.org/10.7717/peerj-cs.887

[45] Eitan Frachtenberg. 2023. Citation analysis of computer systems papers. *PeerJ Computer Science* 9 (May 2023), e1389. https://doi.org/10.7717/peerj-cs.1389

[46] Eitan Frachtenberg and Noah Koster. 2020. A survey of accepted authors in computer systems conferences. *PeerJ Computer Science* 6 (September 2020), e299. https://doi.org/10.7717/peerj-cs.299

[47] Damjan Fujs, Anže Mihelič, and Simon L. R. Vrhovec. 2019. The Power of Interpretation: Qualitative Methods in Cybersecurity Research. In *14th International Conference on Availability, Reliability and Security*. Article 92, 10 pages. https://doi.org/10.1145/3339252.3341479

[48] Thomas Groß. 2020. Statistical Reliability of 10 Years of Cyber Security User Studies. In *10th International Workshop on Socio-Technical Aspects in Security and Trust*. 171–190. https://doi.org/10.1007/978-3-030-79318-0_10

[49] Guofei Gu. 2022. *Computer Security Conference Ranking and Statistic*. https://people.engr.tamu.edu/guofei/sec_conf_stat.htm

[50] Mark Guzdial. 2020. *Why I Don't Recommend CSRankings.org: Know the Values You are Ranking On*. BLOG@CACM. https://cacm.acm.org/blogs/blog-cacm/248078-why-i-dont-recommend-csrankingsorg-know-the-values-you-are-ranking-on/fulltext

[51] Gali Halevi, Henk Moed, and Judit Bar-Ilan. 2017. Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—Review of the Literature. *Journal of Informetrics* 11, 3 (2017), 823–834. https://doi.org/10.1016/j.joi.2017.06.005

[52] Peter Hamm, David Harborth, and Sebastian Pape. 2019. A Systematic Analysis of User Evaluations in Security Research. In *14th International Conference on Availability, Reliability and Security*. Article 91, 7 pages. https://doi.org/10.1145/3339252.3340339

[53] Tom E. Hardwicke, Stylianos Serghiou, Perrine Janiaud, Valentin Danchev, Sophia Crüwell, Steven N. Goodman, and John P.A. Ioannidis. 2020. Calibrating the Scientific Ecosystem Through Meta-Research. *Annual Review of Statistics and Its Application* 7, 1 (2020), 11–37. https://doi.org/10.1146/annurev-statistics-031219-041104

[54] Cormac Herley and P.C. Van Oorschot. 2017. SoK: Science, Security and the Elusive Goal of Security as a Scientific Pursuit. In *2017 IEEE Symposium on Security and Privacy*. 99–120. https://doi.org/10.1109/SP.2017.38

[55] IEEE Symposium on Security and Privacy 2017. *Call For Papers*. IEEE Symposium on Security and Privacy. https://www.ieee-security.org/TC/SP2018/cfpapers.html

[56] IEEE Symposium on Security and Privacy 2023. *Call For Papers*. IEEE Symposium on Security and Privacy. https://sp2024.ieee-security.org/cfpapers.html

[57] IEEE Symposium on Security and Privacy 2023. *CFP Changes for the 2024 Conference*. IEEE Symposium on Security and Privacy. https://sp2024.ieee-security.org/changes-cfp.html

[58] John P. A. Ioannidis. 2005. Why Most Published Research Findings Are False. *PLoS Medicine* 2, 8, Article e124 (August 2005). https://doi.org/10.1371/journal.pmed.0020124

[59] John P. A. Ioannidis, Daniele Fanelli, Debbie Drake Dunne, and Steven N. Goodman. 2015. Meta-research: Evaluation and Improvement of Research Methods and Practices. *PLOS Biology* 13, 10, Article e1002264 (October 2015). https://doi.org/10.1371/journal.pbio.1002264

[60] Jordan Jueckstock, Shaown Sarker, Peter Snyder, Aidan Beggs, Panagiotis Papadopoulos, Matteo Varvello, Benjamin Livshits, and Alexandros Kapravelos. 2021. Towards Realistic and Reproducible Web Crawl Measurements. In *The Web Conference 2021*. 80–91. https://doi.org/10.1145/3442381.3450050

[61] Sotirios Katsikeas, Pontus Johnson, Mathias Ekstedt, and Robert Lagerström. 2021. Research communities in cyber security: A comprehensive literature review. *Computer Science Review* 42, Article 100431 (2021). https://doi.org/10.1016/j.cosrev.2021.100431

[62] Mannat Kaur, Michel van Eeten, Marijn Janssen, Kevin Borgolte, and Tobias Fiebig. 2021. Human Factors in Security Research: Lessons Learned from 2008-2018. https://doi.org/10.48550/arxiv.2103.13287 arXiv:2103.13287

[63] Leonid Keselman. 2019. Venue Analytics: A Simple Alternative to Citation-Based Metrics. In *2019 ACM/IEEE Joint Conference on Digital Libraries*. 315–324. https://doi.org/10.1109/JCDL.2019.00052

[64] Srinivasan Keshav. 2011. Editor's Message. *ACM SIGCOMM Computer Communication Review* 41, 3 (July 2011). https://doi.org/10.1145/2002250

[65] Alexander Kott. 2014. *Towards Fundamental Science of Cyber Security*. 1–13. https://doi.org/10.1007/978-1-4614-7597-2_1

[66] Tim Krause, Raphael Ernst, Benedikt Klaer, Immanuel Hacker, and Martin Henze. 2021. Cybersecurity in Power Grids: Challenges and Opportunities. *Sensors* 21, 18, Article 6225 (2021). https://doi.org/10.3390/s21186225

[67] Kat Krol, Jonathan M. Spring, Simon Parkin, and M. Angela Sasse. 2016. Towards Robust Experimental Design for User Studies in Security and Privacy. In *2016 Learning from Authoritative Security Experiment Results Workshop*. 21–31. https://www.usenix.org/conference/laser2016/program/presentation/krol

[68] Benjamin Krumnow, Hugo Jonker, and Stefan Karsch. 2022. How Gullible Are Web Measurement Tools? A Case Study Analysing and Strengthening OpenWPM's Reliability. In *18th International Conference on Emerging Networking EXperiments and Technologies*. 171–186. https://doi.org/10.1145/3555050.3569131

[69] Peep Küngas, Siim Karus, Svitlana Vakulenko, Marlon Dumas, Cristhian Parra, and Fabio Casati. 2013. Reverse-engineering conference rankings: what does it take to make a reputable conference? *Scientometrics* 96, 2 (January 2013), 651–665. https://doi.org/10.1007/s11192-012-0938-8

[70] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *26th Annual Network and Distributed System Security Symposium (NDSS '19)*. https://doi.org/10.14722/ndss.2019.23386

[71] Edward Lee. 2022. *The Toxic Culture of Rejection in Computer Science*. ACM SIGBED. https://sigbed.org/2022/08/22/the-toxic-culture-of-rejection-in-computer-science/

[72] Lisa Liu, Gints Engelen, Timothy Lynar, Daryl Essam, and Wouter Joosen. 2022. Error Prevalence in NIDS datasets: A Case Study on CIC-IDS-2017 and CSE-CIC-IDS-2018. In *2022 IEEE Conference on Communications and Network Security*. https://doi.org/10.1109/CNS56114.2022.9947235

[73] Tom Longstaff, David Balenson, and Mark Matties. 2010. Barriers to Science in Security. In *26th Annual Computer Security Applications Conference*. 127–129. https://doi.org/10.1145/1920261.1920281

[74] Kevin Macnish and Jeroen van der Ham. 2020. Ethics in cybersecurity research and practice. *Technology in Society* 63, Article 101382 (2020). https://doi.org/10.1016/j.techsoc.2020.101382

[75] Jeffrey C. Mogul. 2013. Towards More Constructive Reviewing of SIGCOMM Papers. *ACM SIGCOMM Computer Communication Review* 43, 3 (July 2013), 90–94. https://doi.org/10.1145/2500098.2500112

[76] Jeffrey C. Mogul and Tom Anderson. 2008. Open issues in organizing computer systems conferences. *ACM SIGCOMM Computer Communication Review* 38, 3 (July 2008), 93–102. https://doi.org/10.1145/1384609.1384623

[77] Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1, 1 (January 2017). https://doi.org/10.1038/s41562-016-0021

[78] Arvind Narayanan and Kevin Lee. 2023. Security Policy Audits: Why and How. *IEEE Security & Privacy* 21, 2 (2023), 77–81. https://doi.org/10.1109/MSEC.2023.3236540

[79] Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115, 11 (2018), 2600–2606. https://doi.org/10.1073/pnas.1708274114

[80] Anna-Marie Ortloff, Matthias Fassl, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz, and Matthew Smith. 2023. Different Researchers, Different Results? Analyzing the Influence of Researcher Experience and Data Type During Qualitative Analysis of an Interview and Survey Study on Security Advice. In *2023 CHI Conference on Human Factors in Computing Systems*. Article 864, 21 pages. https://doi.org/10.1145/3544548.3580766

[81] Eric Pauley and Patrick McDaniel. 2023. Understanding the Ethical Frameworks of Internet Measurement Studies. In *2nd International Workshop on Ethics in Computer Security*. 8 pages. https://doi.org/10.14722/ethics.2023.239547

[82] Vern Paxson. 2004. Strategies for Sound Internet Measurement. In *4th ACM SIGCOMM Conference on Internet Measurement*. 263–271. https://doi.org/10.1145/1028788.1028824

[83] Sean Peisert and Matt Bishop. 2007. How to Design Computer Security Experiments. In *5th IFIP World Conference on Information Security Education*. 141–148. https://doi.org/10.1007/978-0-387-73269-5_19

[84] Sean Peisert and Matt Bishop. 2007. I Am a Scientist, Not a Philosopher! *IEEE Security & Privacy* 5, 4 (2007), 48–51. https://doi.org/10.1109/MSP.2007.84

[85] Peng Peng, Limin Yang, Linhai Song, and Gang Wang. 2019. Opening the Blackbox of VirusTotal: Analyzing Online Phishing Scan Engines. In *2019 Internet Measurement Conference*. 478–485. https://doi.org/10.1145/3355369.3355585

[86] Jan Pennekamp, Erik Buchholz, Markus Dahlmanns, Ike Kunze, Stefan Braun, Eric Wagner, Matthias Brockmann, Klaus Wehrle, and Martin Henze. 2020. Collaboration is not Evil: A Systematic Look at Security Research for Industrial Use. In *2020 Learning from Authoritative Security Experiment Results Workshop*. 16 pages. https://doi.org/10.14722/laser-acsac.2020.23088

[87] Iasonas Polakis, Federico Maggi, Stefano Zanero, and Angelos D. Keromytis. 2014. Security and Privacy Measurements in Social Networks: Experiences and Lessons Learned. In *3rd International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*. 18–29. https://doi.org/10.1109/BADGERS.2014.9

[88] Morteza Safaei Pour, Christelle Nader, Kurt Friday, and Elias Bou-Harb. 2023. A Comprehensive Survey of Recent Internet Measurement Techniques for Cyber Security. *Computers & Security* 128, Article 103123 (May 2023), 35 pages. https://doi.org/10.1016/j.cose.2023.103123

[89] Awais Rashid, George Danezis, Howard Chivers, Emil Lupu, Andrew Martin, Makayla Lewis, and Claudia Peersman. 2018. Scoping the Cyber Security Body of Knowledge. *IEEE Security & Privacy* 16, 3 (2018), 96–102. https://doi.org/10.1109/MSP.2018.2701150

[90] Elissa M. Redmiles, Ziyun Zhu, Sean Kross, Dhruv Kuchhal, Tudor Dumitras, and Michelle L. Mazurek. 2018. Asking for a Friend: Evaluating Response Biases in Security User Studies. In *2018 ACM SIGSAC Conference on Computer and Communications Security*. 1238–1255. https://doi.org/10.1145/3243734.3243740

[91] Dennis Reidsma, Jeroen van der Ham, and Andrea Continella. 2023. Operationalizing Cybersecurity Research Ethics Review: From Principles and Guidelines to Practice. In *2nd International Workshop on Ethics in Computer Security*. 7 pages. https://doi.org/10.14722/ethics.2023.237352

[92] Christian Reuter, Luigi Lo Iacono, and Alexander Benlian. 2022. A quarter century of usable security and privacy research: transparency, tailorability, and the road ahead. *Behaviour & Information Technology* 41, 10 (2022), 2035–2048. https://doi.org/10.1080/0144929X.2022.2080908

[93] Konrad Rieck. 2023. *Influential Security Papers*. https://mlsec.org/topnotch/

[94] Vera Rimmer, Theodor Schnitzler, Tom Van Goethem, Abel Rodríguez Romero, Wouter Joosen, and Katharina Kohls. 2022. Trace Oddity: Methodologies for Data-Driven Traffic Analysis on Tor. *Proceedings on Privacy Enhancing Technologies* 2022, 3 (July 2022), 314–335. https://doi.org/10.56553/popets-2022-0074

[95] Christian Rossow, Christian J. Dietrich, Chris Grier, Christian Kreibich, Vern Paxson, Norbert Pohlmann, Herbert Bos, and Maarten van Steen. 2012. Prudent Practices for Designing Malware Experiments: Status Quo and Outlook. In *2012 IEEE Symposium on Security and Privacy*. 65–79. https://doi.org/10.1109/SP.2012.14

[96] Sebastian Roth, Stefano Calzavara, Moritz Wilhelm, Alvise Rabitti, and Ben Stock. 2022. The Security Lottery: Measuring Client-Side Web Security Inconsistencies. In *31st USENIX Security Symposium*. 2047–2064.

[97] Damien Saucez and Luigi Iannone. 2018. Thoughts and Recommendations from the ACM SIGCOMM 2017 Reproducibility Workshop. *ACM SIGCOMM Computer Communication Review* 48, 1 (April 2018), 70–74. https://doi.org/10.1145/3211852.3211863

[98] Stuart Schechter. 2013. *Common Pitfalls in Writing about Security and Privacy Human Subjects Experiments, and How to Avoid Them*. Technical Report MSR-TR-2013-5. https://www.microsoft.com/en-us/research/publication/common-pitfalls-in-writing-about-security-and-privacy-human-subjects-experiments-and-how-to-avoid-them/

[99] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D. Strowes, and Narseo Vallina-Rodriguez. 2018. A Long Way to

the Top: Significance, Structure, and Stability of Internet Top Lists. In *18th Internet Measurement Conference* (2018). 478–493. https://doi.org/10.1145/3278532.3278574

[100] Quirin Scheitle, Matthias Wählisch, Oliver Gasser, Thomas C. Schmidt, and Georg Carle. 2017. Towards an Ecosystem for Reproducible Research in Computer Networking. In *Reproducibility Workshop*. 5–8. https://doi.org/10.1145/3097766.3097768

[101] Henning Schulzrinne. 2009. Double-blind reviewing. *ACM SIGCOMM Computer Communication Review* 39, 2 (March 2009), 56–59. https://doi.org/10.1145/1517480.1517492

[102] Nihar B. Shah. 2022. Challenges, Experiments, and Computational Solutions in Peer Review. *Commun. ACM* 65, 6 (May 2022), 76–87. https://doi.org/10.1145/3528086

[103] Radu Sion. 2011. *Democracy in Peer Reviewing*. https://zxr.io/theseriousacademic/

[104] Ananta Soneji, Faris Bugra Kokulu, Carlos Rubio-Medrano, Tiffany Bao, Ruoyu Wang, Yan Shoshitaishvili, and Adam Doupé. 2022. "Flawed, but like democracy we don't have a better system": The Experts' Insights on the Peer Review Process of Evaluating Security Papers. In *2022 IEEE Symposium on Security and Privacy*. 1845–1862. https://doi.org/10.1109/SP46214.2022.9833581

[105] Jonathan M. Spring, Tyler Moore, and David Pym. 2017. Practicing a Science of Security: A Philosophy of Science Perspective. In *2017 New Security Paradigms Workshop*. 1–18. https://doi.org/10.1145/3171533.3171540

[106] Shridatt Sugrim, Can Liu, Meghan McLean, and Janne Lindqvist. 2019. Robust Performance Metrics for Authentication Systems. In *2019 Network and Distributed System Security Symposium*. 15 pages. https://doi.org/10.14722/ndss.2019.23351

[107] Janos Szurdi, Meng Luo, Brian Kondracki, Nick Nikiforakis, and Nicolas Christin. 2021. Where are you taking me? Understanding Abusive Traffic Distribution Systems. In *The Web Conference 2021*. 3613–3624. https://doi.org/10.1145/3442381.3450071

[108] USENIX Association 2022. *USENIX Security '22 Call for Artifacts*. USENIX Association. https://www.usenix.org/conference/usenixsecurity22/call-for-artifacts

[109] Pelayo Vallina, Victor Le Pochat, Álvaro Feal, Marius Paraschiv, Julien Gamba, Tim Burke, Oliver Hohlfeld, Juan Tapiador, and Narseo Vallina-Rodriguez. 2020. Mis-shapes, Mistakes, Misfits: An Analysis of Domain Classification Services. In *20th Internet Measurement Conference*. 598–618. https://doi.org/10.1145/3419394.3423660

[110] Jeroen van der Ham and Roland van Rijswijk-Deij. 2017. Ethics and Internet Measurements. *Journal of Cyber Security and Mobility* 5, 4 (2017), 287–308. https://doi.org/10.13052/jcsm2245-1439.543

[111] Erik van der Kouwe, Gernot Heiser, Dennis Andriesse, Herbert Bos, and Cristiano Giuffrida. 2019. SoK: Benchmarking Flaws in Systems Security. In *2019 IEEE European Symposium on Security and Privacy*. 310–325. https://doi.org/10.1109/EuroSP.2019.00031

[112] Moshe Y. Vardi. 2009. Conferences vs. Journals in Computing Research. *Commun. ACM* 52, 5 (May 2009), 5. https://doi.org/10.1145/1506409.1506410

[113] Moshe Y. Vardi. 2016. Academic Rankings Considered Harmful! *Commun. ACM* 59, 9 (August 2016), 5. https://doi.org/10.1145/2980760

[114] Simon Vrhovec, Luca Caviglione, and Steffen Wendzel. 2021. Crème de La Crème: Lessons from Papers in Security Publications. In *16th International Conference on Availability, Reliability and Security*. Article 92, 9 pages. https://doi.org/10.1145/3465481.3470027

[115] Gerry Wan, Liz Izhikevich, David Adrian, Katsunari Yoshioka, Ralph Holz, Christian Rossow, and Zakir Durumeric. 2020. On the Origin of Scanning: The Impact of Location on Internet-Wide Scans. In *20th Internet Measurement Conference*. 662–679. https://doi.org/10.1145/3419394.3424214

[116] Steffen Wendzel, Cédric Lévy-Bencheton, and Luca Caviglione. 2020. Not All Areas Are Equal: Analysis of Citations in Information Security Research. *Scientometrics* 122, 1 (January 2020), 267–286. https://doi.org/10.1007/s11192-019-03279-6

[117] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1, Article 160018 (March 2016). https://doi.org/10.1038/sdata.2016.18

[118] David Zeber, Sarah Bird, Camila Oliveira, Walter Rudametkin, Ilana Segall, Fredrik Wollsén, and Martin Lopatka. 2020. The Representativeness of Automated Web Crawls as a Surrogate for Human Browsing. In *The Web Conference 2020*. 167–178. https://doi.org/10.1145/3366423.3380104

[119] Yiming Zhang, Mingxuan Liu, Mingming Zhang, Chaoyi Lu, and Haixin Duan. 2022. Ethics in Security Research: Visions, Reality, and Paths Forward. In *2022 IEEE European Symposium on Security and Privacy Workshops*. 538–545. https://doi.org/10.1109/EuroSPW55150.2022.00064

[120] Muwei Zheng, Hannah Robbins, Zimo Chai, Prakash Thapa, and Tyler Moore. 2018. Cybersecurity Research Datasets: Taxonomy and Empirical Analysis. In *11th USENIX Workshop on Cyber Security Experimentation and Test*. 8 pages. https://www.usenix.org/conference/cset18/presentation/zheng

[121] Jianying Zhou. 2022. *Top Cyber Security Conferences Ranking*. http://jianying.space/conference-ranking.html

[122] Shuofei Zhu, Jianjun Shi, Limin Yang, Boqin Qin, Ziyi Zhang, Linhai Song, and Gang Wang. 2020. Measuring and Modeling the Label Dynamics of Online Anti-Malware Engines. In *29th USENIX Security Symposium*. 2361–2378. https://www.usenix.org/conference/usenixsecurity20/presentation/zhu

[123] Noa Zilberman and Andrew W. Moore. 2020. Thoughts about Artifact Badging. *ACM SIGCOMM Computer Communication Review* 50, 2 (May 2020), 60–63. https://doi.org/10.1145/3402413.3402422