Terry Benzel
Associate Editor in Chief

Security and Privacy Research Artifacts: Are We Making Progress?

According to the Annual Computer Security Applications Conference (ACSAC):[1]

> *Security research is often criticized for the poor reproducibility of its results. Unfortunately, authors seldom release the software they develop and the datasets they use to perform their experiments. This makes it difficult to compare different solutions and forces, other researchers, to undergo the tedious and error-prone task of re-implementing previous approaches and comparing solutions on different datasets, which may not be a fair comparison.*

Increasingly, the security research community has embraced the formalization of artifact submission and evaluation to improve the quality of research and as a service to the community by increasing reproducibility. The intent behind artifact evaluation is to reach a place where all published research is supported by independently evaluated evidence. Leading conferences and some publications now include a “Call for Artifacts” as part of the Call for Papers and include an artifact evaluation committee that awards badges. Artifact evaluation is independent of the paper review process, and artifacts are generally evaluated after papers are (conditionally) accepted. In 2022, ACSAC initiated an early ad hoc artifact submission in parallel with second-round paper evaluations so that the submission of an artifact can help reviewers clarify minor concerns and make decisions for those papers that are borderline. These artifact evaluations are very time consuming, error prone, and often inconsistent in application and results. While the goals behind artifact evaluation are important, the question is: Are we making progress in increasing reproducibility?

Conference artifact submission began in the security research community in 2017 as part of ACSAC. The conference organizers offered an option for authors of accepted papers to submit the software and/or datasets used in their research and make them publicly available to the community. ACSAC instituted a formal evaluation process and awarded Association for Computing Machinery (ACM) Artifacts Evaluated—Functional badges to papers with successfully evaluated artifacts and reserved the Distinguished Paper Award for a paper with artifacts. The goal was focused on making the research artifacts available as one way to improve research and encourage the community to build on the work of one another.

During this period, ACM’s Reproducibility Task Force worked with ACM conferences and journals to define the common best practices and definitions of levels for assigning badges to artifacts. These badges range from Functional to Results Replicated.[2]

ACSAC should be commended for formalizing the review process and employing the ACM artifact badging definitions, initially using the Functional badging level and then, in 2022, adding the Reusable level. In fact, ACSAC has been quite successful in this regard with nearly 50% of the submitted papers also submitting artifacts that passed an evaluation as either Functional or Reusable.

In 2020, USENIX Security added artifact submission and evaluation with a custom Artifact Evaluated–Passed badge. In 2022, USENIX Security switched to ACM badges, using Artifacts Evaluated–Functional or, in some cases, Artifacts Evaluated–Reusable levels. Several other workshops and conferences, including the USENIX Security Workshop on Offensive Technologies

Digital Object Identifier 10.1109/MSEC.2022.3222887
Date of current version: 19 January 2023

(WOOT), began including artifacts in 2019. The most recent security conference to include artifact submission is the 2022 ACM Conference on Computer and Communications Security. It is notable to this associate editor that the IEEE Computer Society's leading security conference, the Symposium on Security and Privacy, does not (yet) include artifact evaluation.

What Is an Artifact?

Largely, artifacts are defined as the code and/or data used in a research project as reported on in a publication. Artifacts may also include software; experiment scripts; input datasets; data collected through an experiment; and curated or analyzed results. However, experience quickly showed that this definition of an artifact significantly underspecifies and underestimates what is required for the evaluation of an artifact. The evaluation of an artifact rapidly involved understanding a myriad of dependencies from algorithms to runtime environments, not to mention often overlooked underlying assumptions.

Each security conference Call for Artifacts attempts to navigate the complexity of packaging an artifact with a range of requirements.

How Is an Artifact Submitted?

Each security conference Call for Artifacts attempts to navigate the complexity of packaging an artifact with a range of requirements. USENIX Security 2022, for instance, includes 23 detailed questions to answer and a set of complex descriptions of hardware and software dependencies; third-party datasets; customizations; experimentation workflow; and ethical considerations.[3]

The call further indicates that artifacts should be:

- *Documented:* At a minimum, an inventory of artifacts is included, and sufficient description is provided to enable the artifacts to be exercised.
- *Consistent*: The artifacts are relevant to the associated paper and contribute in some inherent way to the generation of its main results.
- *Complete:* To the extent possible, all components relevant to the paper in question are included. (Proprietary artifacts need not be included. If they are required to exercise the package, then this should be documented, along with instructions on how to obtain them. Proxies for proprietary data should be included to demonstrate the analysis.)
- *Exercisable*: Included scripts and/or software used to generate the results in the associated paper can be successfully executed, and included data can be accessed and appropriately manipulated.

The level of detail and the sheer number of items required for an artifact submission create a significant workload. Of course, in an ideal research lab, researchers and students should practice good engineering and development discipline and easily have on hand the artifacts and their dependencies, but this is not the reality in most research labs. Perhaps one important outcome of artifact evaluation is increasing rigor and discipline in leading research labs. However, these requirements may unfairly burden smaller underprivileged research labs due to the sheer volume of work required to not only package an artifact but to maintain

Digital Object Identifier 10.1109/MSEC.2022.3223601

it for some period of time beyond submission and evaluation. This is a related issue with its own challenges!

How Are Artifacts Evaluated?

Artifacts are evaluated by a committee that is distinct from the program review committee. Most artifact evaluation committees include senior graduate students, postdocs, and researchers who are generally closer to the hands-on work required. It is expected that artifact evaluation will require communication between reviewers and authors (while still preserving reviewer anonymity). This adds requirements for artifact authors to be available and respond promptly over a specified period (on the order of weeks). Evaluators are expected to understand the connections between the artifact and the paper through numerous "artifacts;" access GitHub repositories; and install and run code, which may include adding patches, libraries, and even third-party datasets or licensed repositories. There are many reports of the challenges in artifact evaluation and reports of evaluators spending from 30 plus hours to multiple weeks to achieve some level of reproducibility.[4,5]

It should be clear at this point that artifact submission and evaluation are detailed and complicated. These complexities in artifact evaluation have given rise to new research topics evaluating the evaluations and creating an audience for best practices for submissions and evaluators.[6,7]

Is It Having the Desired Impact?

As Eric Eide[8] (University of Utah) discussed in his keynote address, "Reflections on Artifact Evaluation," an invited talk from LASER @ NDSS 2022, there remain many issues with achieving any form of reproducibility. Several case studies reviewed in his talk indicate that perhaps reproducibility is not the right metric as unexpected and biased results are still common.

It is clear that artifact evaluation is not straightforward and is very time consuming with the effort largely on the backs of graduate students. Hopefully, the work improves the research practices of these students. The best result from all of this artifact submission and evaluation effort might be educational. At the same time, the field is growing, and we are seeing better practices and tools; better evaluation platforms; sharing of runnable software artifacts; artifact evaluation indexes; and community artifact hubs.[4]

While there has been improvement in practices and education, we still owe it to the community to ask: What have we accomplished? Do we know more about the research reported on in the published papers? Can we use the results to study a more complex or compound problem by building on those results? The cynic in me wonders if the availability and access to software and/or data simply lay the groundwork for the next researcher to create a paper refuting some aspect of the publication, resulting in an ever-increasing class of security papers at the minimally publishable unit while failing to truly improve security.

Is it time to revisit the artifact submission and evaluation process and culture? Is there some other aspect of security research reproducibility that can yield more impact for our community? ■

Acknowledgments

Thanks to Dave Balenson for consultation and advice and for his role in the Applied Computer Security Associates (ACSA) organization and their leadership in artifacts submission and evaluation in the computer security field.

References

1. "Paper artifacts," Applied Computer Security Associates, Olney, MD, USA, 2019. [Online]. Available: https://www.acsac.org/2019/submissions/papers/artifacts/
2. "Artifact review and badging—Current," Association for Computing Machinery, New York, NY, USA, 2020. [Online]. Available: https://www.acm.org/publications/policies/artifact-review-and-badging-current
3. "USENIX security '22 artifact appendix guidelines (V20220119)," in *Proc. 31st USENIX Secur. Symp.*, 2022. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/artifact-appendix-guidelines
4. D. Balenson et al., "Toward findable, accessible, interoperable, and reusable cybersecurity artifacts," in *Proc. 15th Workshop Cyber Secur. Experimentation Test (CSET)*, New York, NY, USA: Association for Computing Machinery, 2022, pp. 65–70, doi: 10.1145/3546096.3546104.
5. M. Beller. "Why I will never join an Artifacts Evaluation Committee Again," Inventitech. Accessed: Oct. 26, 2022. [Online]. Available: https://inventitech.com/blog/why-i-will-never-review-artifacts-again/
6. "HOWTO for AEC submitters," Google Docs. Accessed: Oct. 26, 2022. [Online]. Available: https://docs.google.com/document/d/1pqzPtLVIvwLwJsZwCb2r7yzWMaifudHe1Xvn42T4CcA/edit
7. R. Padhye. "Artifact evaluation: Tips for authors," Primordial Loop. Accessed: Oct. 26, 2022. [Online]. Available: https://blog.padhye.org/Artifact-Evaluation-Tips-for-Authors/
8. E. Eide. (Feb. 25, 2021). "Reflections on artifact evaluation." Presented at 2021 NDSS LASER Workshop. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/reflections-on-artifact-evaluation/