Report – Improving the Quality and Reuse of Cybersecurity Datasets, Software, and Other Artifacts

2022 SaTC PI Meeting Breakout Group Report
Co-Leads: Eric Eide (University of Utah) and S. Jay Yang (RIT)
Scribes: David Balenson (SRI International), Laura Tinnel (SRI International), and Trevor Dunlap (NC State)

1. Topic and its importance to society, SaTC, and beyond

The breakout group focused on best practices and future approaches to increase and improve the quality and reuse of cybersecurity research datasets, software, and other artifacts. It explored issues, challenges, and opportunities to:

- better package and describe research artifacts, including their intended uses and limitations, and
- better share and find relevant cybersecurity research artifacts, as well as knowledge and experience about their use,

in order to increase the effectiveness of artifacts and broaden their reuse. The group discussed current initiatives, infrastructure, and incentives for producing, sharing, and reusing high-quality artifacts. And, it explored and identified other potential solutions and opportunities to increase and improve the quality and reuse of high-quality artifacts as a solid practice in the cybersecurity research as well as practitioner communities.

Motivation. There is a need for more and high quality shared cybersecurity research datasets, software, and other artifacts. Datasets can include curated data from real operations or synthetically generated ones. Software can include code, scripts, configuration files, etc. Other artifacts include paper, experiment designs/methodologies, etc. The group noted that artifacts can also include special hardware, such as Arduino boards or Raspberry Pis.

Researchers are increasingly producing and sharing their artifacts. Many conferences, workshops, journals are encouraging and recognizing research artifacts and practitioners are appreciating the value of research artifacts. However, there are many challenges and issues with artifacts, including the value and quality of artifacts is not always well understood and artifacts can be difficult to package, share, find, and use.

Approach. The community needs a “sustainable” process to “produce and/or curate” cybersecurity datasets, software, and other artifacts that include a clear description of their intended uses and limitations. This will provide researchers with a better understanding of the quality (and how to measure it) and limitations of artifacts as well as an improved capability to

share and find relevant artifacts as well as knowledge and experience about their use. This, in turn, will help the community reproduce experimental results and accumulate and advance knowledge, instead of repeating past mistakes and misunderstanding the implications of previous research results.

2. Existing body of research and/or practice and highlights/pointers?

NSF RFI. The NSF 2021 Computer and Network Systems Research Dataset Needs Request for Information (RFI) requested input from the research community on its specific needs for datasets to conduct research on computer and network systems (NSF 21-056). The RFI sought input from the community on the specific needs related to collecting, sharing, and utilizing public or private datasets for networking and computer systems research, and any challenges associated with each. NSF was interested in assessing where research progress is slowed due to the lack of datasets, especially when such data may either already exist or can be generated using existing infrastructure (including NSF-funded infrastructure). NSF received 33 responses on the specific needs for datasets to conduct research on computer and network systems, comprising contributions from 75 named contributors from at least 39 research institutions and other organizations. Additional information about the RFI and the response can be found at https://www.nsf.gov/cise/cns/research_datasets/rfi_responses.jsp.

General Sharing Infrastructure. A number of platforms and resources seek to facilitate the sharing of cybersecurity artifacts as well as computer science artifacts more broadly:

- The NSF-funded SEARCCH project (Sharing Expertise and Artifacts for Reuse through Cybersecurity Community Hub) is developing a community-driven platform intended to lower barriers to sharing and reusing research artifacts. It uses a rich metadata representation that enables researchers to better describe and find relevant artifacts and provides import, curation, and search functions to enable greater scientific quality of cybersecurity research. https://searcch.cyberexperimentation.org/
- Github is a common software development and version control platform used by researchers to share software and other artifacts. https://github.com/
- Many researchers also make their research artifacts available via personal or project websites.
- DVC is an open-source version control system for machine learning projects. It is built to make ML models shareable and reproducible and is designed to handle large files, data sets, machine learning models, and metrics as well as code. https://dvc.org/
- Papers with Code is a free and open community-based resource with machine learning papers, code, datasets, methods and evaluation tables. https://paperswithcode.com/
- FindResearch.org is a catalog of research artifacts for computer science developed by Christian Collberg and Todd Proebsting from the Department of Computer Science at the University of Arizona. It aims to be an authoritative and complete catalog of research artifacts (e.g., code and data) related to recent computer science publications to support repeating, reproducing, and extending published research. http://www.findresearch.org/

Cybersecurity Datasets. A number of sites provide various datasets or listings of datasets. Examples include:

- Canadian Institute for Cybersecurity Datasets identifies datasets used around the world by universities, private industry, and independent researchers. It maintains an interactive map indicating datasets downloaded by country. https://www.unb.ca/cic/datasets/index.html
- VizSec.org maintains a list of potentially useful data sets for the computer security visualization and data mining research and development community. https://vizsec.org/data/
- The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. The repository includes the Kitsune Network Attack Dataset Data Set which contains nine different network attacks on a commercial IP-based surveillance system and an IoT network. https://archive.ics.uci.edu/ml/index.php, https://archive.ics.uci.edu/ml/datasets/Kitsune+Network+Attack+Dataset
- The Avast AIC laboratory created and shares Aposemat IoT-23, a labeled dataset with malicious and benign IoT network traffic. https://www.stratosphereips.org/datasets-iot23
- The Intelligent Security Group at UNSW Canberra maintains the UNSW-NB 15 dataset, which contains a hybrid of real modern normal activities and synthetic contemporary attack behaviors and the BoT-IoT dataset, which contains a contains a combination of normal and botnet behavior collected from a realistic network environment in their Cyber Range Lab. https://research.unsw.edu.au/projects/unsw-nb15-dataset, https://research.unsw.edu.au/projects/bot-iot-dataset
- Many cyber competitions and challenges, such as the National Collegiate Cyber Defence Challenge (NCCDC) and Collegiate Penetration Testing Competition (CPTC) generate and share datasets. https://www.nationalccdc.org/, https://cp.tc/
- The Department of Homeland Security Science and Technology Directorate (DHS S&T) Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT) project supports the global cyber-risk research community by coordinating and developing real-world data and information-sharing capabilities—tools, models and methodologies. https://www.dhs.gov/science-and-technology/cybersecurity-impact, https://www.impactcybertrust.org/
- The UCSD Center for Applied Internet Data Analysis (CAIDA) conducts network research and builds research infrastructure to support large-scale data collection, curation, and data distribution to the scientific research community. https://www.caida.org/

Conferences. A growing number of cybersecurity conferences as well as conferences in other areas encourage the sharing of research artifacts:

- The Annual Computer Security Applications Conference (ACSAC) Artifact Initiative encourages authors of accepted papers to submit software and data artifacts and make them publicly available to the entire community. Artifacts are evaluated and authors are

rewarded with a special mention during the conference and on the ACSAC webpage, an ACM Artifacts Evaluated badge on their papers, and a Distinguished Paper Award reserved for the group. https://www.acsac.org/2022/submissions/papers/artifacts/
- The USENIX Security Symposium Call for Artifacts encourages authors of accepted papers to submit artifacts for evaluation. https://www.usenix.org/conference/usenixsecurity22/call-for-artifacts
- ACM SIGSAC Conference on Computer and Communications Security (CCS) just started requiring submissions whose claimed contributions rely on artifacts (e.g., code, models, data sets) to make these accessible to the reviewers, unless there are good reasons not to. https://www.sigsac.org/ccs/CCS2022/call-for/call-for-papers.html
- ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI). https://pldi22.sigplan.org/track/pldi-2022-PLDI-Research-Artifacts
- USENIX Symposium on Operating Systems Design and Implementation (OSDI). https://www.usenix.org/conference/osdi22/call-for-artifacts
- The Learning from Authoritative Security Experiment Results (LASER) Workshop, which is collocated with ISOC NDSS and ACSAC invites authors of accepted papers at the respective venues to present and discuss the experimental aspects of their work. While the workshop does not collect artifacts, per se, it encourages deeper discussions of experimental work including software and data artifacts used and/or produced. https://www.acsac.org/2021/program/final/p43.html, https://www.ndss-symposium.org/ndss2022/call-for-participation-laser-workshop/

Many of these venues leverage and follow ACM's Reviewing and Badging process. https://www.acm.org/publications/policies/artifact-review-badging.

Journals. Some research journals specifically document data curation (not specifically in cyber though):
- Nature's Scientific Data journal. https://www.nature.com/sdata/
- Elsevier's Data in Brief journal. https://www.journals.elsevier.com/data-in-brief

3. Remaining challenges / new challenges?

How to foster generation/curation of high quality artifacts with documentation?
- We need to develop and apply the appropriate incentive structures including both carrots and sticks.
- As noted earlier, a growing number of cybersecurity conferences as well as conferences encourage the sharing of research artifacts along with papers and reward authors through various forms of recognition. Is such recognition sufficient? What else can we do to reward or recognize researchers for producing and sharing high quality artifacts?
- NSF supplemental funding?

How to measure/evaluate artifacts properly?

- Do we have the right criteria for packaging and sharing datasets, software, and other artifacts, including (specialized) hardware?
- How to leverage the evaluations as a way to accumulate knowledge about the artifacts? The knowledge and experience gained through artifact evaluation initiatives (such as the ones mentioned earlier) is not currently captured and shared with the community.
- How to encourage self-reported artifact metadata, such as intended uses, limitations, and attributes, to help users learn about and appropriately reuse artifacts in their own research?

How to foster the sharing/use of high quality artifacts?

- What are the appropriate incentives for encouraging sharing/use of high quality artifacts?
- Today's students seem more inclined to share their work. Many students will share artifacts on github and/or their own personal webpages. Some professors are increasingly encouraging their students to produce and share their software and data. How can we learn more about, tap into, and better encourage this mindset?
- We need to develop and encourage the use of community-based infrastructure (such as the NSF SEARCCH hub) for cataloging/indexing artifacts in support of broader adoption.
- How can we better facilitate active and timely communication between producers and consumers of artifacts?
- How can we expand beyond cybersecurity artifacts and interoperate with artifacts from other experimental computer science disciplines?
- Should we align with publishers such as ACM, Elsevier, IEEE, ISOC, Springer, and USENIX) who already publish papers? Some publishers are even starting to share artifacts along with papers.
- Can we align artifact sharing initiatives with cataloging/indexing services such as DBLP, Google Scholar, and Scopus to both incentivize sharing and broaden reach? Could we develop the artifact equivalent of a "citation count"?

4. Promising directions?

NSF funding? Researchers should request and/or NSF should provide funding to support the software development and data curation activities necessary to provide and maintain high-quality research artifacts. This can include building funding into a new proposal or requesting supplemental funding to support an existing proposal. Ideally, such funding would require researchers to propose and track measurable outcomes.

Cyber Competitions and Challenges? Work with cyber competitions and challenges produce, curate, review, and/or rate artifacts? There should be clear goals to address specific intended uses of data as well as documentation describing limitations of and experiences with datasets.

Conference Artifact Initiatives? We should encourage more conferences, including all of the "big four" (i.e., ISOC NDSS, IEEE S&P, USENIX Security, and ACM CCS) to support and maintain artifact initiatives. Conferences should consider instituting "artifact tracks" separated from the traditional main conference in which researchers can present and/or publish papers describing work producing and/or reusing research artifacts. Conference artifact evaluation committees should explore the development criteria and best practices for artifact evaluation as well as processes for capturing and sharing the results of evaluations.

Education? Professors and other educators should teach and mentor students to produce and share high quality software, datasets, and other artifacts. The community should develop and publish guidelines for developing and sharing reusable/replicable experiments, including required software, datasets, and other artifacts. Guidelines should include privacy and other ethical considerations, including identifying appropriate uses and limitations of datasets. Ideally, the production and sharing of high quality artifacts should be taught as part of fundamental research methods in computer science.

5. Additional questions/issues?

The breakout group did not have time to discuss any additional questions/issues. Other questions/issues envisioned by the organizers included:

- What resources are available for researchers to refer to when learning how to produce and share artifacts? Are there any open source references? What role can software containers, such as Docker containers, play?
- How can we evaluate/ensure the integrity/resilience of cybersecurity artifacts in light of real-world adversaries? Could artifacts be the subject of attacks and how can we defend against them?