Experiment Artifact Sharing: Challenges and Solutions

Clémentine Maurice, CNRS
@BloodyTangerine

Panel @ 2021 LASER workshop

Artifact Evaluation is awesome

→ A paper is **not just a paper**, it is also a lot of data, code, benchmarks...

→ **Improving science**: ideally everybody could **replicate** the results to have a higher **confidence** on the paper, **build on it**, and **compare it** with related (passed or future) work

→ Artifact Evaluation is relatively new in security (compared to, e.g., software engineering), but everybody agrees that it is awesome

People are very happy about it!

Vijay Chidambaram @vj_chidambaram · 15 janv.
Papers introducing tools, benchmarks, or solutions to known problems need to pass **Artifact Evaluation** to be accepted at @jsysresearch. Every paper should have an artifact we can run, and build on!

Dave Levin @DistributedDave · 13 août 2020
For the first time, the @ACMSIGCOMM conference did **artifact evaluation**! Very happy to see the community adopt this. The badges are listed in the program; I hope it encourages more authors to make their artifacts available.
conferences.sigcomm.org/sigcomm/2020/p...

Jack Kolokasis @JackKolokasis · 6 nov. 2020
I like very much the introduction of **artifact evaluation** in systems paper! Very helpful for the systems community! #osdi20

Christopher Patton @cjpatton_ · 12 janv.
#CHES is going to start doing **artifact evaluation**! Excellent! #realworldcrypto

Mathias Payer @gannimo · 22 nov. 2019
For HALucinator, our firmware analysis framework, we're working with the @USENIXSecurity **artifact** evaluation committee. Let me just say that those folks are doing an amazing job! 😍

# Artifact Evaluation process (for WOOT and USENIX Security)

"Does the artifact conform to the expectations set by the paper?"

➡ Authors can submit artifacts **after acceptance** of their paper -- **optional** process
  ○ They submit: the accepted paper, bidding instructions + sw/hw requirements, and the artifact itself

➡ AEC members bid on artifacts (so far nobody had more than 1 artifact each session)

➡ **Discussion phase** between AEC members and authors: ~12 days
  ○ AEC members are fantastic, this is quite short and makes for an intense phase, but it proved to be enough

➡ Review phase -- AEC members now have a good idea whether the artifact passed or not: ~ 2 days

➡ If the paper passed the Artifact Evaluation, the authors add a **badge** before camera ready

Artifact Evaluation is a lot of work

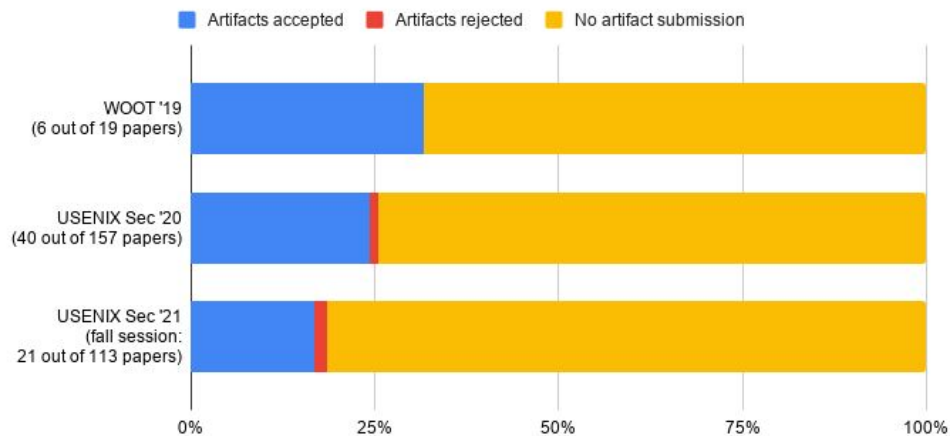Time spent evaluating artifacts (10 respondents)

■ 1/2 day ■ 1/2 to 1 day ■ 1 to 2 days ■ 2 to 4 days

Respondents
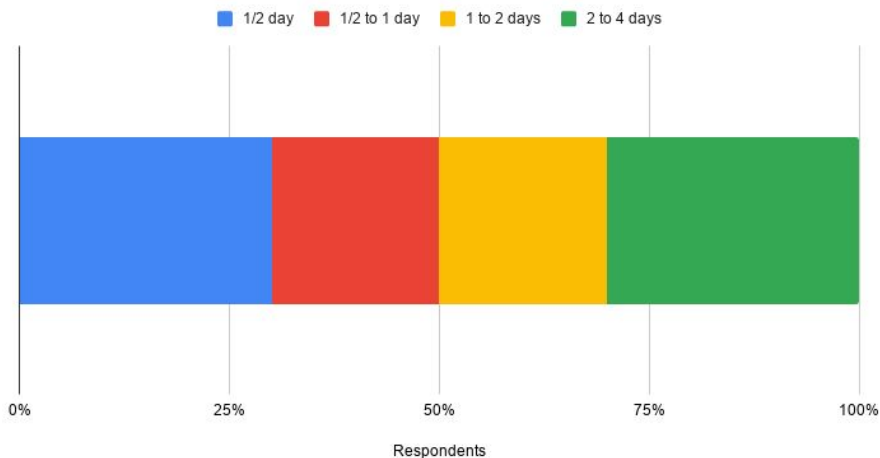
Feedback from WOOT '19 AEC

→ Median time: **1 day**, up to 4 days

→ Requires to be **very reactive**

→ Important point: the evaluation is **not adversarial**! AEC members want to make it work!

All the kudos to AEC members!

Artifact sharing

Artifacts accepted · Artifacts rejected · No artifact submission

WOOT '19
(6 out of 19 papers)

USENIX Sec '20
(40 out of 157 papers)

USENIX Sec '21
(fall session:
21 out of 113 papers)

0%   25%   50%   75%   100%

→ USENIX Sec '21 numbers are not definitive: some artifact submissions were **delayed** to the winter session

→ **25% to 30%** of accepted papers went to the Artifact Evaluation

→ No big trend in terms of artifact sharing between WOOT and Sec

→ Most submitted artifacts are accepted, most of them are **code**

Caveat of these numbers: only reflect papers gone through the formal evaluation process, not informal sharing

Motivators (1/3)

We collectively agree that Artifact Evaluation Is Awesome, yet less than 30% of papers have an artifact: **what can we do**?

Yanick Fratantonio 🌴 @reyammer · 4 oct. 2020
En réponse à @matteodellamico et @JethroGB
Again, I guess that's "no strong incentives" in doing that. Preparing code/dataset to be shared with referees takes time, but that has not been rewarded much. BUT: the **artifact** eval thing is a GREAT step forward, so I'm quite positive about this aspect for long term

→ We have **limited time** and there are **very little incentives**

Motivators: short term solutions (2/3)

A very prosaic answer: "**appealing to our inner first graders**"

Motivators: short term solutions (2/3)

A very prosaic answer: "**appealing to our inner first graders**"

STICKERS! Everybody loves stickers!

Konrad Rieck @mlsec · 17 juil. 2019
En réponse à @thorstenholz et @USENIXSecurity
Will we get a sticker? That would be great.

Motivators: long term solutions (3/3)

→ The immense majority of researchers want to do impactful work: **intrinsic motivation**

→ More powerful incentives would not hurt, but **we need to rethink how we evaluate research**

- Is "number of accepted papers" a good metric? (no, but we already knew that)

- Can Artifact Evaluations be taken into account in **hiring committees, tenure track committees**?

- A good start: in our regular evaluations, my employer (CNRS) asks about software production

Artifact quality

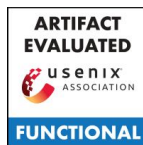**ARTIFACT EVALUATED usenix ASSOCIATION PASSED**: the artifact conforms to the expectations set by the paper

➜ says more about the paper than the artifact, **very variable artifact quality**

More complete badges by USENIX (ACM has equivalent badges), used at OSDI

ARTIFACT EVALUATED usenix ASSOCIATION AVAILABLE available for retrieval, permanently and publicly

ARTIFACT EVALUATED usenix ASSOCIATION FUNCTIONAL documented, completeness, successfully executed

**ARTIFACT EVALUATED usenix ASSOCIATION REPRODUCED** independently repeatable experiments

Improving artifact quality

Feedback from WOOT '19 AEC members from **what helped or would have helped them**:

1. Good **documentation**

2. Providing a step-by-step **running example** or automated test cases

3. **Packaging**: VM, docker… anything that avoids Dependency Hell

4. (Providing access to a remote machine)
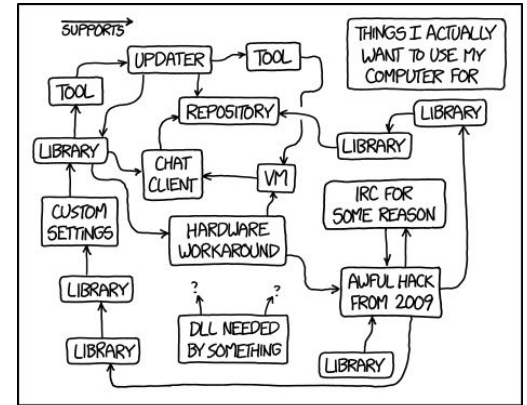
https://xkcd.com/1579/

A few hurdles we experienced

→ **Tight timeline** that has been retrofitted to fit AE, e.g., shepherding and AE at the same time

→ Complicated to **fix hard and fast rules** for all artifacts due to the **diversity**
   ○ I feel like we run into one or more unexpected questions each AE session

→ Sometimes **only a part** of the paper has a corresponding artifact (for various reasons)
   ○ Not ideal, but we asked the authors to clarify this in their paper for camera ready

Challenges (1/n)

What about **hardware**?

Brendan Dolan-Gavitt
@moyix

Slightly frustrating thing about embedded research is the hardware platforms used in past evaluations become completely unobtainable. Good luck finding an Econotag in 2021 :\

Traduire le Tweet

5:57 PM · 15 févr. 2021 · Twitter Web App

→ Hardware **requirements** can be problematic for the evaluation

→ Hardware **availability** will be an issue in a few years

→ Nice suggestion from AEC member: having AWS accounts for AEC

Challenges (2/n)

Actually… what about **software**?

David Brumley
@thedavidbrumley

En réponse à @thorstenholz et @USENIXSecurity

Artifacts in theory are great. I do have an issue with maintaining them. Getting asked 10 years later about code you barely remember written by a grad student long gone is hard. And funding doesn't cover sysadmin work needed for backups and access. Please set an expiry date.

Traduire le Tweet

6:11 PM · 17 juil. 2019 · Twitter Web Client

→ Authors can package beautifully their artifacts to help with software requirements

→ But code probably won't be **maintained** for years

→ Artifact Evaluation probably has a **timestamp**

Challenges (3/n)

Licensing can get in the way of the evaluation

Brendan Dolan-Gavitt
@moyix

Artifact eval question: is it kosher to include SPEC2006 in your artifact package?

Traduire le Tweet

6:30 PM · 23 août 2020 · Twitter Web App

→ Some artifacts may include **proprietary code**, e.g., SPEC CPU benchmarks are only available for purchase

Challenges (4/n)

It would be great for Artifact Evaluation to **happen during reviews** instead of after acceptance

Hernan Ponce De Leon
@h_poncedeleon

⋯

En réponse à @vj_chidambaram @jsysresearch et @eeide

That's the way to go! I hope conferences follow the lead and make use of the Artifact Evaluation as an input for acceptance decision

Traduire le Tweet

5:15 PM · 15 janv. 2021 · Twitter for Android

→ Where to find the **workforce**?

Let's make more people happy
Let's share more artifacts

———