Thoughts about Artifact Badging

Noa Zilberman
University of Oxford, UK
noa.zilberman@eng.ox.ac.uk

Andrew W Moore
University of Cambridge, UK
andrew.moore@cl.cam.ac.uk

This article is an editorial note submitted to CCR. It has NOT been peer reviewed.
The authors take full responsibility for this article's technical content. Comments can be posted through CCR Online.

ABSTRACT

Reproducibility: the extent to which consistent results are obtained when an experiment is repeated, is important as a means to validate experimental results, promote integrity of research, and accelerate follow up work. Commitment to artifact reviewing and badging seeks to promote reproducibility and rank the quality of submitted artifacts.

However, as illustrated in this issue [15], the current badging scheme, with its focus upon an artifact being reusable, may not identify limitations of architecture, implementation, or evaluation.

We propose that to improve the insight into artifact reproducibility, the depth and nature of artifact evaluation must move beyond simply considering *if* an artifact is reusable. Artifact evaluation should consider the methods of that evaluation alongside the varying of inputs to that evaluation. To achieve this, we suggest an extension to the scope of artifact badging, and describe both approaches and best practice arising in other communities. We seek to promote conversation and make a call to action intended to strengthen the scientific method within our domain.

CCS CONCEPTS

• **General and reference → Evaluation**;

KEYWORDS

Reproducibility, Artifact Evaluation, Robustness

1 INTRODUCTION

Science depends upon trust and verification; for experimental work this trust is derived from confidence that results are able to be verified given the same artifact (experiment) under the same conditions. Critically, this trust permits researchers to put their ideas among their peers — confident the peers will understand both artifact, and experiment — allowing interpretation and discussion of the results. Additionally, this trust permits a reader to verify the work and also to build upon it; our community has been tackling verification, this paper focusses upon reproducibility: the extent to which consistent results are obtained when an experiment is repeated. Reproducibility is a core requirement for a scientific discipline to be trusted. It enables us to validate that artifacts exist, that an evaluation is genuine, and that the obtained results and-or interpretation are reproducible. Reproducibility is recognised as of core importance to the ACM SIGCOMM community [2, 13]; to this end our community has adopted the ACM artifact evaluation and badging scheme [12].

ACM SIGCOMM currently supports three tiers of artifacts badging. The highest, "Reusable", means that the artifact is documented, consistent, complete, exercisable, includes appropriate evidence of verification and validation, and exhibits a quality that significantly exceeds minimal functionality [12]. ACM recognizes two further tiers of reproducibility that focus upon validation of reported results: *results replicated* and *results reproduced* [1]. ACM specifically considers robustness a goal for these badges; although that has not been without its own inconsistencies: some disagreement remains over the precise difference between replicate and reproduce.

Scope

We keenly recognise we will leave questions unanswered, such as how to provide strong incentives to publish reproducible artifacts, and how do we balance artifact evaluation needs and initiatives with maintaining engagement of the, already over-stretched, reviewer community. We must both balance and manage the expectations of author, reviewer, and reader alike, to this end we encourage the community to continue to explore these critical issues.

After all, a good evaluation should not become the metaphorical stick to reject a submission merely because the (previously unknown) limitations are described alongside the innovations.

In this work we ask *What should be the level of evaluation applied to SIGCOMM research artifacts?*. In this issue, Zilberman observes shortcomings that are not detectable under SIGCOMM's reusable badge [15]: limitations of the architecture, of the implementation, and of the evaluation. An evaluation of the performance of NDP [7], the Best Paper Award winner of SIGCOMM'17, is used to illustrate how sensitivity tests and robustness experiments can uncover limitations of a solution.

Based on these observations, we suggest that artifact evaluation should take place as part of the review process. Our recommendations are not novel [4, 6, 9], but rather a call to arms of the SIGCOMM community to embrace and adapt the good practices evolving across the related systems, measurement, performance and evaluation communities. We further suggest the following improvements to the SIGCOMM reproducibility initiative:

- Mandating an artifact description as part of papers submission process [11].
- Providing in calls-for-papers a checklist, to be used by authors and reviewers [4, 5].
- Allowing early career researchers to self-nominate to Artifact Evaluation committees.
- Publish the results of artifacts evaluation as part of a conference proceedings, or soon thereafter.

Our ambition is to start a conversation: how much trust can we, or others, place in our outcomes and how will we create and strengthen that trust?

2 DEFINING ARTIFACT EVALUATION

The evaluation of NDP in [15] was not constrained by ACM badge definitions, it was when the authors of [15] sought clarity a gap among definitions became clear.

ACM defines Replicability as:
Replicability [1] *(Different team, same experimental setup):* "The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts."

While reproducibility is defined as:
Reproducibility [1] *(Different team, different experimental setup):* "The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently."

The artifact evaluation task in [15] is ill defined as it neither replicates nor reproduces the original artifact evaluation work. The artifact evaluation is conducted by a different team, and uses the same experimental setup. This implies replication rather than reproduction. However, the evaluation explores a change of parameters, such as seeds (in simulation), and packet sizes (in hardware testing), that do not fall under the replication category, nor are they clearly defined as reproduction. Such robustness or sensitivity testing is both good practice and has become increasingly common; for example, the evaluation of NDP arose through a study of Stardust's robustness [16]. The measurement community also expects a high level of scrutiny, e.g., as network measurements can easily be biased by the selection of vantage points.

SIGCOMM currently does not support either the "Results Replicated" nor the "Results Reproduced" badges. In the following sections we try to motivate the community not only to adopt these two badges, but also to apply and extend them to cover aspects of robustness.

3 MOTIVATION

Why care? The SIGCOMM community has a long track-record of research contributions that have had real-world impact: describing approaches that are quickly adopted in industry or arose in industry. As such, we should aspire to develop solid artifacts. Artifacts don't necessarily need to be mature or complete, but a good evaluation can expose important characteristics (limitations and advantages) of a solution. A good idea can too easily be tossed away by users if weaknesses are discovered, whereby if the same weaknesses are discovered by their originators, they can be quickly identified and perhaps fixed too.

Who cares? Teaching early career researcher proper evaluation practices provides a lesson for life. These researchers later lead development teams, or become senior researchers themselves, and the practices they have learned as students are carried on to their students. Many of the senior researchers reading this paper are probably thinking "but of course sensitivity testing is required", however unless we set an expectation from the community to maintain certain standards, such practices may simply be deemed unnecessary, a waste of paper, or irrelevant.

Should we separate papers from artifacts? Artifacts are the means to establish trust between papers' authors and readers. The source code tells a reader about the nature of a solution, the evaluation environment describes how the solution was evaluated, and results datasets provide a record of this evaluation. Papers sometimes describe a vision or an architecture, where an artifact does not exist. However, where a paper claims to cover aspects of implementation or evaluation, the artifact is our way of confirming such claims. When an artifact does not support claims made in a paper, a reviewer has the flexibility to decide whether the difference is core to the paper. However, separating the artifact from a paper leaves reviewers with no means to assess the trustworthiness of the paper.

4 RELATED INITIATIVES

How do other communities evaluate artifacts? Many communities: ACM and others, (e.g., SIGPLAN, SIGHPC) now run artifact evaluation as part of their conferences. However, the scope of evaluation is often limited. In most cases, only artifacts of accepted papers are evaluated (e.g., SOSP'19, PLDI'20, CGO'20, ASPLOS'20), and the artifact is not taken into consideration when reviewing the paper. An increasing number of conferences are seeking to badge to "Results Replicated" level. Examples include CGO, PPOPP, MLSys and workshops such as ASPLOS'18 ReQuEST. While the artifact evaluation committee is many times set up by the organizing committee, in some cases, e.g., OOPSLA'20, there is a call for self-nomination, targeted at post-docs and senior PhD students. Further community movement has recently seen USENIX ATC to direct authors to online checklists [8]. Yet, while the security performance communities also recognise the problems, e.g., [14], a coherent community strategy is still to coalesce.

SIGPLAN's Checklist. In the SIGPLAN community, empirical evaluation guidelines were developed, and the findings were used to produce a 1-page checklist [4, 5]. The checklist is comprised of seven categories, each with a few example violations. Despite the differences between SIGCOMM and SIGPLAN, all the categories and example violations are relevant to the SIGCOMM community. Only in the description of a few examples wording may require adoption, e.g., SIGPLAN checklist refers to "compiler optimization" whereas SIGCOMM may refer for example to "network stack optimization", yet the overarching goal of the example does not change. The SIGPLAN checklist is different to SIGCOMM's artifact evaluation guidelines, as it focuses on the quality of the evaluation, rather than on the reusability of the artifact.

Artifact Description. An Artifact Description form (AD) includes information on the experiments reported in the paper, associated datasets, hardware and software, baseline experimental setup and more. An Artifact Evaluation form (AE) also extends on steps taken to ensure that results are trustworthy. SuperComputing introduced AD/AE as part of its reproducibility initiative in 2016 [10], and mandated submitting an Artifact Description form since 2019 [3, 11], while submitting an Artifact Evaluation form remains optional. The AD/AE are online forms and part of the submission process, comprised of checkboxes and a few open questions, reducing the load on both authors and reviewers, as the forms assist in the assessment. Similar AD/AE requirements in SIGCOMM conferences would improve both submission and review quality, without adding significant overheads.

5 RECOMMENDATIONS

Starting Early. Artifact evaluation practices should be interleaved throughout a research project. Testing an artifact's robustness just before a paper deadline is too late. Building an artifact evaluation environment that supports parameter variation and enables repeatable executions simplifies debugging during the development process, increases portability between platforms, and reduces the burden when artifact submission is due. An evaluation environment can often be carried from previous projects, thus properly setting the infrastructure once alleviates future efforts. As a community, we need to emphasize the need for such practices.

Using Checklists. The empirical evaluation guidelines developed by the SIGPLAN community [4, 5] are relevant also to SIGCOMM researchers. Still, conferences within SIGCOMM differ significantly from each other: an artifact submitted to IMC will be evaluated (by its authors) differently than artifacts submitted to ANCS or e-energy. We propose that SIGCOMM conferences publish a checklist with a conference's call for papers, setting the expectations from submissions and providing clarity for early career researchers. For instance, a difference between conferences will be in "Fails to measure all important Effects" under "Relevant Metrics", where ANCS will give as an example "Failing to measure the effect of packet size on throughput", while IMC will use "Failing to evaluate the effect of using different vantage points".

Including Artifact Description. Including an artifact description in paper submissions is a low hanging fruit. The artifact description can be part of the submission form on hotcrp, and some information, e.g., identifying platforms names, can be blinded from the reviewers. Proprietary corporate platforms can be catered for, as done by the supercomputing community. An artifact description form can assist reviewers in many different ways, including common questions such as "is this a result of a simulation, or was it running in hardware?".

We do believe that robustness should be a goal for artifact evaluation, but we believe that the way to get there is by educating the community to submit artifacts that were tested for robustness, rather by putting the burden on artifact evaluation committees.

Artifact Evaluation Committee. Enabling early career researcher participation in artifact evaluation committees, rather than established researchers alone, will benefit all parties: evaluating the artifacts early on, making more researchers familiar with other work and evaluation practices, providing a development opportunity to early stage researchers and not adding load to more senior ones. The key to the success of such an evaluation will be, through the existence of pre-determined evaluation check lists, the setting of expectations. Some communities publish reports of their results, which can serve as an incentive to participate in artifact evaluation committees. While already today CCR publishes the results of artifact evaluation, we propose to do so as part of the conference proceedings, as one or two pages summaries.

Badging During the Review Process. It is easy under the pressure of a deadline to evaluate only to "the minimal level required for acceptance" rather to the "adequate level of quality", without anyone (possibly even the authors) being able to notice. As shown in [15], the disparity between results can be significant. We therefore believe that artifact evaluation results need to inform conferences: from eligibility and awards, to publishing artifact evaluation reviews (in addition to badges) along with the papers. The important part is doing so in a timely manner, so researchers can be made aware of strengths and weaknesses of projects before they begin follow up works. With the artifact evaluation committee extended, and with more supporting materials (checklists, AD/AE forms) provided to artifact reviewers, evaluating the quality of submitted artifacts beyond current 'reusable' badge becomes easier. In order to do so, artifact evaluation committees should become part of conferences, just like poster or shadow PC committees. As described earlier, this is already implemented in other SIGs' conferences.

Making a Difference. Each of the recommendations above provides a layer of improvement to artifacts submissions. Education, Checklists and Artifact Evaluation forms help improve the *quality* of artifacts. An Artifact Description form improves the completeness of an artifact, and does not mean that the quality is better, but it is the fist step toward mandating Artifact Evaluation and it encourages researchers to improve their artifacts. Increasing participation in artifact evaluation committees not only reduces the load on heavily-loaded reviewers, but also improves the quality of future artifacts, as new reviewers learn from their experience and adapt practices. Badging during the review process provides the trust in a published paper, which is, eventually, our goal.

6 CONCLUSION

Trust is hard won, and easily lost; evaluation within artefact badging defines an expected standard of a projects' quality. Badging up to Reusable level may not expose flaws in architecture, implementation and evaluation, and Results Replicated and Results Reproduced badges may not reveal such flaws either [15]. To attend to these limitations we propose pathways to improve artifact evaluation, from the publishing of evaluation check lists to the mandating of Artifact Description submissions. Embedding artifact evaluation within the papers' review process will provide important quality assurance both to authors and the community: not only our own, but to the many communities who must build their work upon ours.

This paper left some unanswered questions, such as how to provide strong incentives to publish reproducible artifacts. We encourage the community to continue to explore these critical issues.

ACKNOWLEDGMENTS

We thank Marcin Wójcik, Gianni Antichi and Costin Raiciu for their assistance in the reproduction of NDP's evaluation and the discussion of the results. We thank Jan Vitek, Vijay Chidambaram and Grigori Fursin for providing information of OOPSLA, SOSP and Supercomputing Artifact Evaluation (correspondingly). We thank Steve Blackburn, Matthias Hauswirth, Michael Hicks and Emery Berger for an insightful discussion of SIGPLAN's checklist. This work was partly funded by the Leverhulme Trust (ECF-2016-289) and the Isaac Newton Trust.

REFERENCES

[1] ACM. 2018. Artifact Review and Badging. https://www.acm.org/publications/policies/artifact-review-badging.

[2] 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.

[3] Lorena Barba and Grigori Fursin. 2019. Reproducibility Initiative. https://sc19.supercomputing.org/submit/reproducibility-initiative/.

[4] E. D. Berger, S. M. Blackburn, M. Hauswirth, and M. W. Hicks. 2019. A Checklist Manifesto for Empirical Evaluation: A Preemptive Strike Against a Replication Crisis in Computer Science.

[5] E. D. Berger, S. M. Blackburn, M. Hauswirth, and M. W. Hicks. 2019. Empirical Evaluation Guidelines.

[6] Ronald F Boisvert. 2016. Incentivizing reproducibility. *Commun. ACM* 59, 10 (2016), 5–5.

[7] Mark Handley, Costin Raiciu, Alexandru Agache, Andrei Voinescu, Andrew W Moore, Gianni Antichi, and Marcin Wójcik. 2017. Re-architecting datacenter networks and stacks for low latency and high performance. In *SIGCOMM*. ACM, Los Angeles, CA, USA, 29–42.

[8] Gernot Heiser. 2020. Systems Benchmarking Crimes. https://www.cse.unsw.edu.au/~gernot/benchmarking-crimes.html [accessed Jan. 2020].

[9] M Heroux. 2015. The TOMS initiative and policies for replicated computational results (RCR). *ACM Trans. Math. Softw.* 41, 3, Article Article 13 (June 2015), 5 pages. https://doi.org/10.1145/2743015

[10] Michael A. Heroux. 2018. SC Reproducibility Initiative. https://sc18.supercomputing.org/submit/sc-reproducibility-initiative/index.html.

[11] Beth Plale. 2020. Transparency and Reproducibility Initiative. https://sc20.supercomputing.org/submit/transparency-reproducibility-initiative/.

[12] Damien Saucez, Luigi Iannone, and Olivier Bonaventure. 2019. Evaluating the artifacts of SIGCOMM papers. *ACM SIGCOMM Computer Communication Review* 49, 2 (2019), 44–47.

[13] Quirin Scheitle, Matthias Wählisch, Oliver Gasser, Thomas C Schmidt, and Georg Carle. 2017. Towards an ecosystem for reproducible research in computer networking. In *Proceedings of the Reproducibility Workshop (Reproducibility '17)*. ACM, Los Angeles, CA, USA, 5–8. https://doi.org/10.1145/3097766.3097768

[14] Erik van der Kouwe, Dennis Andriesse, Herbert Bos, Cristiano Giuffrida, and Gernot Heiser. 2018. Benchmarking Crimes: An Emerging Threat in Systems Security. arXiv:1801.02381 http://arxiv.org/abs/1801.02381

[15] Noa Zilberman. 2020. An artifact evaluation of NDP. *ACM SIGCOMM Computer Communication Review* 50, 2 (2020).

[16] Noa Zilberman, Gabi Bracha, and Golan Schzukin. 2019. Stardust: Divide and Conquer in the Data Center Network. In *NSDI*. Boston, MA, USA, 141–160.