Thoughts and Recommendations from the ACM SIGCOMM 2017 Reproducibility Workshop

Damien Saucez
Univresité Côte d'Azur, Inria, France
damien.saucez@inria.fr

Luigi Iannone
Telecom Paristech, France
luigi.iannone@telecom-paristech.fr

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 It has NOT been peer reviewed.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

Ensuring the reproducibility of results is an essential part of experimental sciences, including computer networking. Unfortunately, as highlighted recently, a large portion of research results are hardly, if not at all, reproducible, raising reasonable lack of conviction on the research carried out around the world.

Recent years have shown an increasing awareness about reproducibility of results as an essential part of research carried out by members of the ACM SIGCOMM community. To address this important issue, ACM has introduced a new policy on result and artifact review and badging. The policy defines the terminology to be used to assess results and artifacts but does not specify the review process or how to make research reproducible.

During SIGCOMM'17 a side workshop has been organized with the specific purpose to tackle the issue. The objective being to trigger discussion and activity in order to craft recommendations on how to introduce incentives for authors to share their artifacts, and the details on how to use them, as well as defining the process to be used.

This editorial overviews the workshop activity and summarizes the main discussions and outcomes.

CCS Concepts

•**General and reference** → *General conference proceedings; General literature; Computing standards, RFCs and guidelines;*

Keywords

Reproducibility, Artifacts

1. INTRODUCTION

The Reproducibility'17 workshop accepted 7 papers out of 11 submissions. The paper review process included an evaluation phase by program committee members, followed by an online discussion of the top ranked papers, out of which the best 7 papers were accepted to appear in the program.

The resulting program featured papers focusing mainly on two different and complementary aspects of reproducibility. The first aspect concerns the question of *"Why is reproducibility so hard?"*, aiming at triggering discussion around what can be done to create the right set of habits and incentives, may be supported by a clear set of policies, in order to make research more easily reproducible ([5, 12, 6]). The second aspect concerns the question *"How reproducible is our research?"*, aiming at providing a glimpse of what is the situation today and what kind of lessons we can learn looking at the current landscape ([8, 7, 10, 11]).

Because of the above mentioned types of accepted papers, the workshop was organized in two technical sessions, each one followed by a keynote. To discuss on the difficulty of making reproducible research, Christian Collberg, University of Arizona, presented his experience in maintaining *FindResearch.org* a website that lists information about the artifacts of thousands of papers in Computer Science [1]. For the second session, Lisa Yan, Stanford University, presented the experience of asking graduate students to reproduce papers for a student project in the *CS244: Advanced Topics in Networking* course at Stanford [13]. The last part of the workshop consisted in a final *brainstorm* session to discuss the major questions of how to make artifacts available to ensure reproducibility and how to introduce reproducibility in the habit of our community.

The remainder of the paper is organized as follows. As a background, Sec. 2 recalls the terminology introduced by ACM concerning reproducible research. Sec. 3 summarizes the content of the first technical session, while Sec. 4 the content of the second one. Then, Sec. 5 overviews the activity and discussion of the final brainstorm session, whose outcome is summarized in Sec. 6.

2. ACM TERMINOLOGY

Reproducibility covers a large scope in science in general and no uniform terminology exists. In this editorial, we consider the terminology defined by ACM [2] when introducing the badging system.

An *artifact* is *a digital object that was either created by the authors to be used as part of the study or generated by the experiment itself* [2]. In our community measured data, pre/post processing scripts, software or even computation results can then be considered as artifacts.

A work is defined as *repeatable* if the researchers that are at the origin of the work can obtain the same results on multiple trials by using the same experimental setup.

A work is defined as *replicable* if researchers independent of the original team can obtain the same results on multiple trials by using the same experimental setup as the original work, potentially on different locations.

Finally, a work is defined as *reproducible* if researchers

independent of the original team can obtain the same results on multiple trials by using different systems than the original work, on different locations.

3. WHY IS REPRODUCIBILITY SO HARD?

Our role is not to stigmatize the work with limited reproducibility properties, rather to try to understand the reason for the lack of reproducibility in our community. To this end, the session dedicated to understanding why reproducibility is hard to achieve was a gold mine.

3.1 Reproducibility: a sociological problem?

The main conclusion of the session is that the lack of scientific reproducibility is mostly sociological, not technological. Actually, in our community there is a general lack of incentives to be reproducible. All things being equal, being reproducible does not influence much the chances of being accepted or to be referenced. It just represents more work to be done, with no foreseeable return of investment. In addition, given that our community emphasis on novelty there is little if no consideration for studies reproducing prior work, e.g., of a measurement work.

At a first glance, authors may think that reproducibility incurs extra cost without benefits. First, publishing reproducible work takes time and must be thought at the earliest stage as it requires to be rigorous. For instance, after every change in the system (e.g., change of code, upgrade of a library, model modification) all results have to be re-validated. In addition, time is required to precisely understand which factors impact the results and to document the experimental environment. Additionally, once documented, it is hard to integrate such meta-information in the manuscript as it would consume space and break the reading flow with overly technical terms. A possible solution could be then to write a technical report referenced in the paper but it also consumes time and challenges the anonymity required by the double-blind reviewing process. And last but not least, providing artifacts and their meta-information can cause intellectual properties and privacy issues.

While privacy and anonymity may be seen as real hard problems that call for a deep reflexion on the reviewing process, the rest should be seen as a chance for everyone rather than a cost. To start with, it simplifies the life of the authors themselves on the long run, since usually automation is required. This is also useful for the authors' origin team as usually people of the same team tend to work on similar topics. Therefore, if the work of their direct colleagues is more structured and made to be usable easily by many, they increase their possibility of re-using those artifacts, hence improving work effectiveness, instead of re-doing the same things over and over again. In a team it also helps to have a well documented and re-usable artifacts as it simplifies sharing the knowledge and avoids such knowledge to fade out when the researcher leaves. This allows, new students coming to work on an extension of a previous work, they don't have to spend an awful amount of time to rebuild all the research environment and knowledge. Finally, it simplifies and may even enable collaborations as the learning and integration costs are reduced, for both sides.

3.2 Reviewing Reproducibility

Reproducibility has intrinsic incentives for the authors, as described above, yet, unfortunately incentives and benefits are less obvious for reviewers. Indeed, assessing the reproducibility of papers requires an important amount of work as it requires much more than simply reading the paper. It requires, to dig into the details of the artifacts, potentially retrieve them, maybe read some code, or even execute it (assuming the reviewer has the resources to do it). We also have to keep in mind that research is a collaborative activity, hence papers have usually multiple authors, each having some specific competences, for instance one author can be more oriented toward the technology, one towards the modeling parts, another on coding. On the contrary, the reviewer is alone and has to assess the quality of all of the different facets of the research work. Considering that finding good reviewers, experts of a domain, is already hard, finding reviewers that are also able to look at the reproducibility of the work is even harder, particularly as the reviewing process is poorly rewarded and is seen as a duty.

3.3 Artifacts Availability

While the sociological aspects are very important in limiting reproducibility of research, it would be unfair blame it as the only issue. It sure is an important problem, but some technical issues have been highlighted during the workshop. The most important technical problem relates to the sharing of artifacts both during the reviewing process and after publication. During the reviewing process, the problem is mostly related to intellectual property and anonymity. As of today, no reviewing platform guarantees that an artifact will not be seen by a competitor or that a reviewer will not use it even if the paper is rejected. In addition, when the work requires to access a specific platforms (e.g., a local testbed or some hardware prototype) their is no real mean to guarantee the anonymity of both authors and reviewers.

Alas, artifact sharing problems do not stop at the publication of the work and a major issue is to guarantee perennial reproducibility. This translates on how to guarantee the availability and the meaning of the artifacts for long period of time, knowing that our work environment changes constantly. The agility of our environment causes even issues in the way to identify the artifacts in the papers. The artifact may have to move to different sharing platforms or some platforms may not be accessible in some locations because of their policies. In addition, artifact often rely on complex engineering systems composed of software and hardware and with time these systems change. A piece of software perfectly working today, may stop working on new versions of an operating system, or a new release of a library, or an upgrade of a piece of hardware. This raises the highly technical question of knowing how to extract the information that is really needed to be able to reproduce the results. Indeed even very well documented papers can be hardly reproducible because of some dependencies that not even authors knew about, impairing future reproducibility of the work.

4. REPRODUCING OUR RESEARCH: YES, BUT HOW?

There is a general consensus that we can (and should) improve reproducibility in our community. However, before trying to find ways to improve the situation, it is necessary to understand how reproducible are our current papers. At the workshop we had a session presenting efforts to reproduce papers of various propositions with artifacts.

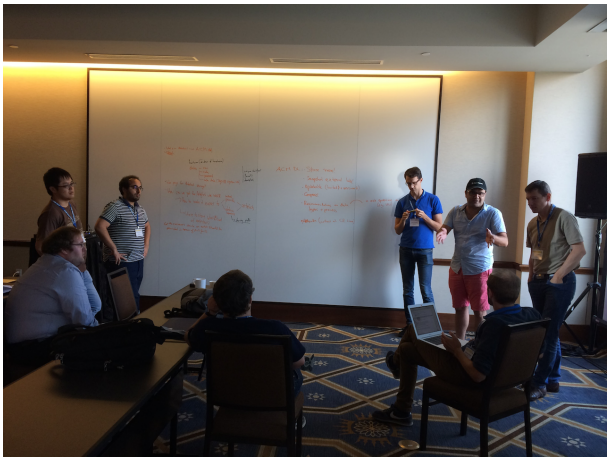

Figure 1: During last session attendees were split into two groups brainstorming in front of whiteboards.

4.1 Some Lessons Learned

Keynote speaker Lisa Yan presented her experience over the last five years in training graduate students with a reproducibility project [13]. During the project, graduate students have to select a paper and reproduce one of their key results and then communicate about their experience. After 5 years and more than 40 papers studied, the experience appears to be all in all positive. Most students were able to reproduce the papers, even if in general it required extra efforts to make it work, i.e., often students had to write their own code or generate the workload by themselves, based on what was described in the paper. It also appeared that while some papers are well documented, they are sometimes just impossible to reproduce years later because the environment changed too much, e.g., studies of web pages. The main conclusion of this long term experience is that by reproducing anterior work, students are rapidly exposed to tools that will help them in their career. It sparks fruitful scientific and critical discussions, and ease technology transfer.

4.2 Documentation is Key

From the presentations in the second technical session of the workshop, it can be clearly identified a general inconsistent/fragmented way of documenting research. Most of the paper in the session did not try to reproduce directly other research papers, but rather studied them from a more abstract point of view, aiming at extracting the information necessary to their reproducibility.

A first observation is that even when papers are treating similar questions, the evaluation environments may differ largely because of the plethora of simulators, emulators, traffic generators, or testbeds. This lack of unification also appears in the terms that are used. In addition to this absence of standard, it also appears that authors do not really know what really needs to be documented. Typically in performance studies, authors provide OS version and CPU frequency but do not specify other hardware specifications such as caches or memory that can have a significant impact on performances. Finally, because of the intrinsic complexity of the topics we explore in our community, multiple approaches are followed with complex setups, yet, the majority of the papers do not provide any link to the artifacts.

4.3 Reproducibility vs. Time

Still during the second technical session, Mahfoudi et al. presented their attempt to reproduce the beamforming feature proposed in the OpenRF paper [9]. In a nutshell, beamforming compatible wireless network cards can adapt their signal to focus it towards their target. To enable such feature, network cards must support the capability. While trying to reproduce the OpenRF experiment in their own lab, Mahfoudi et al. faced two issues. The first issue came from the numerous operating system's updates since the original paper, and the fact that the OS used in OpenRF was not compatible with the infrastructure used for the their experiment. Fortunately the software problem could be fixed but the authors tried with multiple compatible wireless network cards but never managed to obtain the same results. This lack of success did not come from a lack of documentation or the absence of artifacts, but rather from the impossibility for the authors to obtain the exact same network cards than the ones used in the original paper. Actually, the authors had the same type of network cards but with newer hardware release, changing the behavior of the cards. As a conclusion, we can see that because of the time span between the original paper and the trial (4 years), the software and hardware was simply not available anymore, which prevented the paper to be replicated.

5. IMPROVING REPRODUCIBILITY

Reproducibility is a broad topic and one workshop cannot answer all questions related to it. During the last "brainstorming" session of the workshop, discussion led to estimate that the first two points to be addressed by our community were *i*) how to provide incentives for reproducible papers and *ii*) how to share artifacts. To tackle the two questions, the attendees have been split into two groups and worked in front of a whiteboard (cf., Fig. 1). In the following we provide outcome these activities.

5.1 Encouraging Reproducibility

The vast majority of researchers are convinced by the need of reproducibility, yet, as we have seen earlier, sociological factors make it hard to reach that goal in practice.

Reproducibility is a fundamental part of science but it

is still legitimate that some papers are not directly reproducible and even if we have to encourage reproducibility it cannot be at the expense of papers that are not reproducible. For that reason reproducibility can only be *encouraged* but *not forced* in the publication process. To go in that direction the *badging* system proposed by the ACM [3] is a great tool that we recommend to generalize. Nevertheless, it should only be used on a voluntary basis, where authors decide if they want to be explicitly evaluated on that point. The voluntary basis is important as it may require extra work for the authors and interactions with the reviewers.

Regarding the review procedure, it appears that a *committee* independent of the technical program committee should be formed, with the specific task of assessing reproducibility. The reason is that assessing the reproducibility level of papers is complex and time consuming and not all reviewer can do it (they have other competencies essential to review a paper). As this task is potentially complex, the group proposes to only review *accepted papers* that produced artifacts in order to limit the amount of paper to be evaluated and not bias reviews. To build a strong and motivated group of reviewers, an *independent and open reproducibility reviewing committee* is proposed. If the committee is open and public and if it is involved only after acceptance of papers, it simplifies the problems linked to intellectual property, double-blind reviewing and incentives in general. As the paper is already accepted interactions can be simplified as anonymity can be broken. However, if anonymity is broken there is a risk of collusion, that's why a *public badging explanation* should be provided. It is important to remark that to avoid negative social impact, this justification must strictly explain in what the paper fulfills the requirements of the badge, not saying why it does not have another badge. With this we can see clear incentives for the authors as their papers gain extra visibility and the label can be used to support career plans.

While the committee solution looks a great way to encourage authors to share artifacts and improve reproducibility, the incentives for reviewers (members of the committee) are not that clear. However, if a committee is open and on a voluntary basis, we can imagine that members of the committee used it as a tool for their career but also a mean to speed-up their research and open new collaborations. As a matter of fact, the committee has to review novel papers potentially not yet public, having the chance to directly interact with the authors which should increase the creation of collaborations and extensions to the work, particularly if the work is reproducible, since efforts of the reproducers and the authors can be put in common. Experience shows that interactions between authors and reproducers spark collaborations and ease and accelerate work extensions [13].

To ease the process of writing reproducible papers, the group suggests to make a *sharing contract*: at publication, the authors provide a moral contract where they engage themselves in making their paper reproducible and list what they did to ensure reproducibility. A way to provide such contract could be to fill a form at submission or to provide a *"reproducibility"* section in the paper. The moral contract has two advantages. On the one hand, it encourages the authors to think about reproducibility while working on the paper. This will help authors as we have noticed that while many authors were willing to be reproducible, and thought they were reproducible, when queried about explaining what actions they made to ensure reproducibility they noticed they missed some important points. On the other hand, clarifying the actions taken by authors will ease the reviewers or researchers willing to reproduce a paper to understand the points that deserve attention.

Precisely describing experimental setup or reproducibility matters takes space and authors often neglect the description of their setup because of the lack of room (due to page number limitation). The group proposes to authorise free *unlimited size appendixes* for matters related to reproducibility. The intuition is that if authors have no restriction for describing experimental setup, they will provide important details, which will avoid the work to be open to a number of unfunded interpretations.

5.2 Sharing Artifacts

A particularity of our community is to produce a lot of artifacts as the systems we work on change rapidly and also because producing such artifacts is relatively simple and cheap compared to some other fields. As a results, we publish at a rapid pace and continuously make incremental changes in the artifacts themselves.

By nature artifacts in our community depend on complex and evolving systems and it is needed to attach *meta-information* to every of them. The meta-information must provide the technical description of the artifact environment, typically obtained using automatic tools, but often it is not enough and additional text by the authors can clarify obscure points (e.g., the reason of a particular configuration setting). How to understand the level of details to provide for describing artifacts is still an open question [7, 8, 1].

To help providing good meta-information, the group recommends to list *good sharing practices* with tools used by the community. For example, the amount of meta-information to provide is very different between a paper doing simulations in $ns-3$ or doing hardware-specific beamforming. The sharing practices of a tool would simplify the work of authors and reviewers to asses the level of reproducibility of a work.

As artifacts are part of the scientific work, they must be part of the reviewing process (in the form previously suggested). Submission systems must provide a way to share artifacts *without breaking anonymity and privacy* when required. The system should also provide a way for the reviewers to *interact with the authors* if more information is required to assess the reproducibility level of the paper. Moreover, as artifact description require precision and thus space, submission should not limit the amount of pages that can be provided to describe the artifacts and how to use them.

As artifacts change with time and rely on complex systems, once a paper is published, it is essential to provide a way to be able to *version* the artifact to pinpoint the exact one that was used to produce the results shown in a paper when *archives* are not possible. It is then needed to *identify* precisely this artifact and version as over its life time, it may have to change location or ownership. The group recommends the use of *Digital Object Identifier System* (DOI) [4] for identification of artifacts once published. To foster the reuse of the artifacts, the group recommends to use an *open access* policy whenever possible. Finally, the artifact sharing system should provide a way to associate *feedback* from the community in order to clarify points that may have been unclear at the publication time, but that can only be seen when other teams work on reproducing the paper or reusing its artifacts. In the long term, by looking at the feedback

that artifacts receive, it will be possible to calibrate the process by progressively understand what meta-information is really needed.

To start with the process, it is probably easier to remain at small scale, e.g., the special interest group, and the ACM digital library seem to be the right place to experiment the proposals.

6. CONCLUSION

Reproducibility'17 has been an atypical workshop, closer to the original meaning the word *workshop* itself. Attendees did not only listen to presentations concerning the accepted papers, or to the invited keynotes. Rather, they were asked to engage in intense discussions in a brainstorm session. The workshop pointed out that there are several hurdles concerning reproducibility, namely the absence of incentives and the bad habit that our community has grown accustomed to. This is evident in the current typical review process which is not adapted to handle reproducibility. Furthermore, there is no general way to share and preserve artifacts (and related documentation), every author does it in their own way.

The discussions in the brainstorm session focused on the two most important points to be tackled, namely, *i*) how to provide incentives for reproducible papers and *ii*) how to share artifacts.

For the first, a promising approach is to put in place a Reproducibility Committee, which will run in parallel with the normal Technical Program Committee of conferences and workshops, which will assess the level of reproducibility of papers accepted for publication by the TPC. Such approach will solve some of the privacy and anonymity issues while reducing the volume of work for the reviewers that volunteer in assessing the reproducibility level.

For the second, a gradual approach has been suggested. The ACM digital library has been suggested as place to start sharing artifacts, which will be also identified via a DOI number. Beside the artifact itself it is important to share all of the meta-information necessary to actually reproduce prior work, as well as a way to provide feedback in order to make the community learn which meta-information is actually important and build guidelines on how to provide such information.

Acknowledgments. This work was partly funded by the French National Research Agency (ANR) through the "Investments for the Future" Program (#ANR-11-LABX-0031-01) and the REFLEXION Project (ANR-14-CE28-0019). It also benefited support from NewNet@Paris, Cisco's Chair "NETWORKS FOR THE FUTURE" at Telecom ParisTech (http://newnet.telecom-paristech.fr). Any opinions, findings or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of partners of the Chair.

7. REFERENCES

[1] A Catalog of Research Artifacts for Computer Science. http://www.findresearch.org. Accessed: 2017-09-11.

[2] Artifact Review and Badging. https://www.acm.org/publications/policies/artifact-review-badging. Accessed: 2017-09-11.

[3] Artifact Review and Badging. https://www.acm.org/publications/policies/artifact-review-badging. Accessed: 2017-09-14.

[4] ISO 26324:2012, Information and documentation – Digital object identifier system. https://www.iso.org/standard/43506.html. Accessed: 2017-09-13.

[5] Bajpai, V., Kühlewind, M., Ott, J., Schönwälder, J., Sperotto, A., and Trammell, B. Challenges with reproducibility. In *Proceedings of the Reproducibility Workshop* (New York, NY, USA, 2017), Reproducibility '17, ACM, pp. 1–4.

[6] Canini, M., and Crowcroft, J. Learning reproducibility with a yearly networking contest. In *Proceedings of the Reproducibility Workshop* (New York, NY, USA, 2017), Reproducibility '17, ACM, pp. 9–13.

[7] Ferreira, D. C., Vázquez, F. I., Vormayr, G., Bachl, M., and Zseby, T. A meta-analysis approach for feature selection in network traffic research. In *Proceedings of the Reproducibility Workshop* (New York, NY, USA, 2017), Reproducibility '17, ACM, pp. 17–20.

[8] Flittner, M., Bauer, R., Rizk, A., Geissler, S., Zinner, T., and Zitterbart, M. Taming the complexity of artifact reproducibility. In *Proceedings of the Reproducibility Workshop* (New York, NY, USA, 2017), Reproducibility '17, ACM, pp. 14–16.

[9] Kumar, S., Cifuentes, D., Gollakota, S., and Katabi, D. Bringing cross-layer mimo to today's wireless lans. In *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM* (New York, NY, USA, 2013), SIGCOMM '13, ACM, pp. 387–398.

[10] Mahfoudi, M. N., Turletti, T., Parmentelat, T., and Dabbous, W. Lessons learned while trying to reproduce the openrf experiment. In *Proceedings of the Reproducibility Workshop* (New York, NY, USA, 2017), Reproducibility '17, ACM, pp. 21–23.

[11] Nussbaum, L. Testbeds support for reproducible research. In *Proceedings of the Reproducibility Workshop* (New York, NY, USA, 2017), Reproducibility '17, ACM, pp. 24–26.

[12] Scheitle, Q., Wählisch, M., Gasser, O., Schmidt, T. C., and Carle, G. Towards an ecosystem for reproducible research in computer networking. In *Proceedings of the Reproducibility Workshop* (New York, NY, USA, 2017), Reproducibility '17, ACM, pp. 5–8.

[13] Yan, L., and McKeown, N. Learning networking by reproducing research results. *SIGCOMM Comput. Commun. Rev. 47*, 2 (May 2017), 19–26.