Challenges with reproducibility

Author(s):
Bajpai, Vaibhav; Kühlewind, Mirja; Ott, Jörg; Schönwälder, Jürgen; Sperotto, Anna; Trammell, Brian

Publication Date:
2017

Permanent Link:
https://doi.org/10.3929/ethz-b-000197662 →

Originally published in:
http://doi.org/10.1145/3097766.3097767 →

Challenges with Reproducibility

Vaibhav Bajpai
TU Munich

Mirja Kühlewind
ETH Zürich

Jörg Ott
TU Munich

Jürgen Schönwälder
Jacobs University Bremen

Anna Sperotto
University of Twente

Brian Trammell
ETH Zürich

ABSTRACT

The Computer Science (CS) culture is gentle to accepting papers that are non-reproducible as long as they appear plausible. In this paper, we discuss some of the challenges with reproducibility and a set of recommendations that we as a community can undertake to initiate a cultural change.

1 INTRODUCTION

Association for Computing Machinery (ACM) defines a research to be reproducible [1] when its results can be obtained by a group using an independently developed dataset. Kurkowski *et al.* in [11] have shown that less than 15% of MobiHoc papers (2000 - 2005) that used simulations (114 out of 151 papers) for MANET analysis were repeatable. We refer the reader to [1] for further definitions of repeatability and replicability which have less stringent goals. Vandewalle *et al.* in [22] checked 134 papers published in IEEE Transactions on Image Processing and found that 33% of papers release datasets, while only 9% of papers release code needed to reproduce the results. Recently, Collberg *et al.* in [5] examined ~600 CS papers from ACM conferences and journals and found weak repeatability in ~32% of papers. This shows that we are less strict on reproducibility but tend to accept papers that appear plausible. This is a cultural issue and changing a culture is generally hard and takes time. CS practitioners continue to do community service to help authors embrace reproducibility. For instance, Paxson in [15] provides guidance on how to develop a discipline for reproducible data analysis. Krishnamurthy *et al.* in [10] propose a socratic method to allow measurers and reusers of datasets to validate measurement-based research. Sandve *et al.* in [19] list down ten simple rules for reproducible research. Recent Dagstuhl seminars [2, 6] on global measurements also stress on the value of reproducibility. However, despite these continued efforts, reproducibility of research in CS and in networking in particular appears to exist as an ongoing problem.

2 CHALLENGES

In this section, we highlight some of the challenges that we as authors (see § 2.1 and § 2.2) and reviewers (see § 2.3 and § 2.4) face when handling papers from a reproducibility perspective.

2.1 Lack of incentive to reproduce research

Preparing a submission for a venue is usually last minute work. Getting the data into a shape that makes it easily accessible and understandable often requires even more work. This model of submitting papers for consideration under a strict deadline does not seem to fit well for reproducible research. The fast-paced nature of our discipline which involves a race of putting together findings quickly to be first, tends to hurts reproducibility. This is because networking research evolves quickly and results (especially network measurement results) tend to become stale within a span of few years. This is a tradeoff since the ability to properly store, document, and organize experimental data for reproduction requires time. Towards this effect, the norm generally is to get the paper accepted and then prepare artifacts for release by when the reviewers cannot help with curating the released artifacts. Furthermore, conferences (unlike Internet Measurement Conference (IMC) that bestows best dataset awards, see § 3.4) usually do not provide much incentives (additional points) for authors to make this extra effort to release artifacts to allow reproducibility.

Despite the encouragement from the community (IMC and Traffic Measurement and Analysis (TMA) call for papers explicitly solicit submissions that reproduce results), few papers that reproduce results get published. It is not easy to identify the root cause of this. It could be that papers with novel ideas tend to excite paper acceptance more than ideas that reproduce existing research. This may also affect research papers with negative results and studies that revisit known observations or provide incremental improvement on existing datasets.

2.2 Double-blind review requires obfuscation

Few top venues within our community require double-blinded paper submissions. This makes it impossible for a reviewer to check for reproducibility of a submitted work because the authors cannot reveal or may even have to obfuscate artifacts (see [1] for a formal definition) at review time to allow double-blinding reviewing. Furthermore, datasets cannot be properly understood and appreciated without the metadata [3] that describes them which often tends to break anonymity. The time invested in obfuscating the paper for a double-blinded submission can instead be used to prepare artifacts for reproducibility and to improve science. Indeed, authors who care about reproducibility can choose to submit papers to single-blind venues only, but top venues need to setup a role model to allow such a cultural change in our community.

2.3 Fetching artifacts breaks review anonymity

Authors that try to provide artifacts [1] that are necessary to completely comprehend the paper (assuming that there are no obstacles to making artifacts available such as ownership or anonymization issues) usually do this with good intentions. However, these artifacts are made available in an ad-hoc way that may break review anonymity. This is because paper submission systems usually do not allow authors (barring one or more venues, see § 3.2) to upload these artifacts with the paper submission. Consequently, reviewers are expected to fetch this information from external resources (likely from the university infrastructure of the author's affiliation)

which leaves a trail. As a result, it is left to the reviewer to make an effort to fetch things using mechanisms that immediately do not reveal the identity of the reviewer. Authors also tend to sometimes rely on URL shortening services (such as TinyURL *et al.*) to save space which creates another level of indirection for fetching these artifacts. Papers using URL shortening services may become disassociated with their artifacts in unfortunate situations where the used URL shortening service ceases to exist in the future. Furthermore, artifacts released by authors on university resources also may not remain permanently available online. It becomes hard to maintain resources that are prone to garbage collection in situations where authors switch universities. As such, providing artifacts in such an ad-hoc fashion does not scale with time.

2.4 Lack of appreciation for good review work

Good reviews take time and the community usually has a limited pool of people providing good (substantial and constructive) reviews. Matching reviewers with submitted papers is also becoming a challenge, to such a degreee that conferences are now experimenting [12, 17] with automated review assignment systems. Checking for reproducibility increases review expectations even further thereby shrinking the pool of good reviewers. The limited number of good reviews is not due to lack of expertise, but generally due to lack of appreciation for doing good review work. This is a major structural problem whereby checking whether work is reproducible is just one facet of doing good reviewing work.

Publicly releasing reviews written by experts in the field for an accepted paper also helps with reproducibility. This allows future readership to critically examine an accepted paper. IMC trialed making reviews publicly available for few years, but doing this repeatedly every year is an overhead that the conference organization committee has to factor in.

3 RECOMMENDATIONS

In this section, we attempt to provide some recommendations on how we as a community can improve the state of reproducibility in networking research.

3.1 Discuss reproducibility considerations

Measurement papers should have (similar to an ethical considerations [14] section) a reproducability considerations section that forces authors to think about reproducibility. The description of where code is available or how to get (or produce) necessary datasets should go into such a section. The benefit of a dedicated section is to ensure authors think about reproducibility and provide input on how their work can be reproduced. In the long term, we should strive to make measurement papers runnable [7] so that a reader can *play* the process of consuming raw data to produce results described in the paper. This will allow one to see intermediate steps, which makes errors due to analysis (particular cherry picking of outliers) obvious. Furthermore, knowing that the reviewers will see these intermediate steps is a nice incentive for carefulness. This also allows the methodology described in the paper to be applied on an independent raw dataset encouraging further investigation of the same phenomenon by the larger research community.

3.2 Allow authors to upload artifacts

Paper submission systems should allow authors to upload artifacts for review purposes. The authors should be encouraged to make use of this feature. Several ACM SIGPLAN and closely related conferences have embraced an artifact evaluation process [9] that allows authors to submit artifacts to back up their results. An Artifact Evaluation Committee (AEC) in addition to the regular Programmme Committee (PC) is installed to facilitate this process. Within the networking community as well, SIGCOMM Computer Communication Review (CCR) now provides means to make artifacts available during the submission phase and it also relaxes restrictions on page limit for reproducible papers that require space to adequately describe the artifacts needed to reproduce results. Traditional conferences can also split deadlines into a paper submission and an artifact submission deadline (with a few weeks time window) to allow authors to prepare artifacts for review. Although, this involves certain risk of releasing artifacts to anonymous reviewers before paper acceptance. Traditional conferences can also encourage authors to demo the code used in the paper to increase plausibility of produced results. Finally, publishers should allow authors of accepted papers to upload artifacts on the publishers website with the premise that both the paper and artifacts remain available online together, as one entity, at one hosting location.

3.3 Ask review questions on reproducibility

Conference review forms should also accommodate specific questions concerning reproducibility. This will remind reviewers to pay attention to reproducibility when reviewing papers. For instance: *a) Are the artifacts [1] made available? In situations where the artifacts cannot be released, do authors provide advise on how the results can be reproduced?* The idea here is not to ask authors why they cannot release the artifacts (which gives authors an escape channel [18] to put an obligatory disclaimer) but instead encourage them to provide constructive ways to help their work get reproduced or validated. This will allow reviewers to give bonus points to authors that think about reproducibility, *b) Can the released code be easily run to allow reproduction of results using alternate datasets?* The idea here is that released code is a necessary but not sufficient condition unless independent groups can run it without inordinate effort, *c) In situations where the code cannot be released, is the methodology suitably explained to allow independent groups to rewrite code that produces same results?* The idea here is that since papers are usually limited by space (some venues relax page limits in favor of reproducibility, see § 3.2) to sufficiently explain every detail to allow complete reproduction of results, authors must think about releasing code since this is a feasible way to ensure that papers become reproducible.

3.4 Highlight reproducible papers

There may also be limits to the lengths a reviewer can go to assess the reproducibility of a paper during the conference review phase. As such, it may not be practical to reject all non-reproducible research, but it is important to ensure that good, working and reproducible ideas get the attention they deserve. Conferences (such as IMC initiated this effort) can bestow awards to papers with best datasets. Publishers can be encouraged to badge [1] and highlight

reproducible papers on their webpage. An AEC (see § 3.2) can be used to sample and evaluate papers based on reproducibility on a regular basis. Extended conference papers that get submitted for consideration to a journal can be more strictly judged from the reproducibility perspective. This will help bubble up reproducible papers from the lot. However, such an initiative will also require installing processes in place to ensure that the badges do not become fake over time. SIGCOMM CCR can (in addition to relaxing page limits, see § 3.2) dedicate a column for papers that reproduce results. Summary outcomes of graduate seminars that encourage networking students [21] to reproduce existing research can be published in such a column. Furthermore, new venues that specifically solicit papers that reproduce previous work may be needed. Recent reproducibility initiatives [13] already attempt to provide new formal publication venues to specifically solicit papers that reproduce previous work.

4 CONCLUSION

Despite these challenges, research is being reproduced [4, 8, 16], albeit rarely. Institutions are also making efforts to make data easily accessible. CAIDA [20] provides a searchable index of existing measurement data and invites the community to reproduce results. As such, the state of reproducibility is not dismal but improving with time. The recommendations provided in this paper may not be concluding wisdom, but we hope these ideas transfer and eventually help the cause to incentivise the community to reproducibility.

ACKNOWLEDGEMENTS

Georg Groh, Ljubica Kärkkäinen, Aaron Yi Ding and Leonardo Tonetto provided useful input to this manuscript. This work received funding from the European Union's Horizon 2020 research and innovation programme 2014-2018 under grant agreement No. 644866. This work was also partially supported by the European Commission under Horizon 2020 grant agreement no. 688421 Measurement and Architecture for a Middleboxed Internet (MAMI), and by the Swiss State Secretariat for Education, Research, and Innovation under contract no. 15.0268. This support does not imply endorsement. This work was also partially funded by the D3 (NWO 628001018) project.

REFERENCES

[1] ACM. 2016. Result and Artifact Review and Badging. (2016). Retrieved March 15, 2017 from https://www.acm.org/publications/policies/artifact-review-badging

[2] Vaibhav Bajpai, Arthur W. Berger, Philip Eardley, Jörg Ott, and Jürgen Schönwälder. 2016. Global Measurements: Practice and Experience (Report on Dagstuhl Seminar #16012). *Computer Communication Review* 46, 2 (2016), 32–39. DOI: http://dx.doi.org/10.1145/2935634.2935641

[3] Vaibhav Bajpai, Steffie Jacob Eravuchira, and Jürgen Schönwälder. 2015. Lessons Learned From Using the RIPE Atlas Platform for Measurement Research. *Computer Communication Review* 45, 3 (2015), 35–42. DOI: http://dx.doi.org/10.1145/2805789.2805796

[4] Bryan Clark, Todd Deshane, Eli M. Dow, Stephen Evanchik, Matthew Finlayson, Jason Herne, and Jeanna Neefe Matthews. 2004. Xen and the Art of Repeated Research. In *Proceedings of the FREENIX Track: 2004 USENIX Annual Technical Conference, June 27 - July 2, 2004, Boston, MA, USA*. 135–144. http://www.usenix.org/publications/library/proceedings/usenix04/tech/freenix/clark.html

[5] Christian S. Collberg and Todd A. Proebsting. 2016. Repeatability in Computer Systems Research. *Communications of the ACM* 59, 3 (2016), 62–69. DOI: http://dx.doi.org/10.1145/2812803

[6] Philip Eardley, Marco Mellia, Jörg Ott, Jürgen Schönwälder, and Henning Schulzrinne. 2013. Global Measurement Framework (Dagstuhl Seminar 13472). *Dagstuhl Reports* 3, 11 (2013). DOI: http://dx.doi.org/10.4230/DagRep.3.11.144

[7] Nikhil Handigol, Brandon Heller, Vimalkumar Jeyakumar, Bob Lantz, and Nick McKeown. 2012. Reproducible Network Experiments Using Container-Based Emulation. In *Conference on emerging Networking Experiments and Technologies, CoNEXT '12*. 253–264. DOI: http://dx.doi.org/10.1145/2413176.2413206

[8] Heidi Howard, Malte Schwarzkopf, Anil Madhavapeddy, and Jon Crowcroft. 2015. Raft Refloated: Do We Have Consensus? *Operating Systems Review* 49, 1 (2015), 12–21. DOI: http://dx.doi.org/10.1145/2723872.2723876

[9] Shriram Krishnamurthi and Jan Vitek. 2015. The Real Software Crisis: Repeatability as a Core Value. *Communications of the ACM* 58, 3 (2015), 34–36. DOI: http://dx.doi.org/10.1145/2658987

[10] Balachander Krishnamurthy, Walter Willinger, Phillipa Gill, and Martin F. Arlitt. 2011. A Socratic method for validation of measurement-based networking research. *Computer Communications* 34, 1 (2011), 43–53. DOI: http://dx.doi.org/10.1016/j.comcom.2010.09.014

[11] Stuart Kurkowski, Tracy Camp, and Michael Colagrosso. 2005. MANET Simulation Studies: The Incredibles. *Mobile Computing and Communications Review* 9, 4 (2005), 50–61. DOI: http://dx.doi.org/10.1145/1096166.1096174

[12] Baochun Li and Y. Thomas Hou. 2016. The new automated IEEE INFOCOM review assignment system. *IEEE Network* 30, 5 (2016), 18–24. DOI: http://dx.doi.org/10.1109/MNET.2016.7579022

[13] PLOS ONE. 2012. Reproducibility Initiative. (2012). Retrieved March 15, 2017 from https://validation.scienceexchange.com

[14] Craig Partridge and Mark Allman. 2016. Ethical Considerations in Network Measurement Papers. *Communications of the ACM* 59, 10 (2016), 58–64. DOI: http://dx.doi.org/10.1145/2896816

[15] Vern Paxson. 2004. Strategies for Sound Internet Measurement. In *Proceedings of the 4th ACM SIGCOMM Internet Measurement Conference, IMC 2004, Sicily, Italy, October 25-27, 2004*. 263–271. DOI: http://dx.doi.org/10.1145/1028788.1028824

[16] Diana Andreea Popescu and Andrew W. Moore. 2016. Reproducing Network Experiments in a Time-controlled Emulation Environment. In *Traffic Monitoring and Analysis - 8th International Workshop, TMA 2016, Louvain La Neuve, Belgium, April 07-08, 2016*. http://tma.ifip.org/2016/papers/tma2016-final10.pdf

[17] Simon Price and Peter A. Flach. 2017. Computational Support for Academic Peer Review: A Perspective from Artificial Intelligence. *Communications of the ACM* 60, 3 (2017), 70–79. DOI: http://dx.doi.org/10.1145/2979672

[18] Michael Rabinovich. 2014. The Reproducibility versus Debuggability of Research. *Internet Computing* 18, 6 (2014), 4–6. DOI: http://dx.doi.org/10.1109/MIC.2014.123

[19] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. 2013. Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology* 9, 10 (2013). DOI: http://dx.doi.org/10.1371/journal.pcbi.1003285

[20] Colleen Shannon, David Moore, Ken Keys, Marina Fomenkov, Bradley Huffaker, and Kimberly C. Claffy. 2005. The Internet Measurement Data Catalog. *Computer Communication Review* 35, 5 (2005), 97–100. DOI: http://dx.doi.org/10.1145/1096536.1096552

[21] Stanford. 2012. Reproducing Network Research. (2012). Retrieved March 15, 2017 from https://reproducingnetworkresearch.wordpress.com

[22] Patrick Vandewalle, Jelena Kovacevic, and Martin Vetterli. 2009. Reproducible Research in Signal Processing. *IEEE Signal Processing Magazine* 26, 3 (May 2009), 37–47. DOI: http://dx.doi.org/10.1109/MSP.2009.932122