Challenges with Reproducibility

Vaibhav Bajpai
TU Munich

SIGCOMM Reproducibility Workshop
Los Angeles, USA

Joint work with

Mirja Kühlewind
ETH Zürich, Switzerland

Jörg Ott
TU Munich, Germany

Jürgen Schönwälder
Jacobs University Bremen, Germany

Anna Sperotto
University of Twente, Netherlands

Brian Trammell
ETH Zürich, Switzerland

August 2017

Introduction

- ∼15% of MobiHoc simulation papers (2000 - 2005) were repeatable[1] [2].
- ∼33% (out of 134 papers) ToIP papers release datasets while only 9% release code [3].
- ∼32% (out of 600) CS papers published in ACM events exhibit weak repeatability [4].

- We are less strict on reproducibility but tend to accept papers that appear *plausible*.
- This is a cultural issue and changing a culture is hard.
- Despite continued advice [5, 6, 7, 8, 9], reproducibility exists as an ongoing problem.

[1] ACM provides formal definitions [1] of repeatability, replicability and reproducibility.

Challenges

- Authors' perspective —
  - Lack of incentive to reproduce research
  - Double-blind review requires obfuscation
- Reviewers' perspective —
  - Fetching artifacts breaks review anonymity
  - Lack of appreciation for good review work

Challenges | Lack of incentive to reproduce research

- The CS networking discipline is extremely fast-paced –
  - Network measurement results become stale within a span of few years.
  - Race of putting together findings quickly to be first, tends to hurts reproducibility.
  - Ability to properly store, document, and organize data requires time.
  - Norm is to get the paper accepted, release artifacts later (after peer-review)
- Conferences[2] do not provide incentives for authors to release artifacts.
- Despite encouragement[3], few papers that reproduce results get published.
  - Papers with novel ideas tend to excite paper acceptance.

[2] unlike IMC that bestows best dataset awards
[3] IMC and TMA CFP solicit submissions that reproduce results

Challenges | Double-blind review requires obfuscation

- Reviewer cannot check for reproducibility of a submission with obfuscated artifacts.
- Datasets cannot be understood without the metadata [10] which breaks anonymity.
- Time invested in obfuscating paper can be used to prepare artifacts.
- Top venues need to setup a role model to initiate a cultural change.

Challenges | Fetching artifacts breaks review anonymity

- Paper submission systems do not allow authors[4] to upload artifacts with paper.
  - Artifacts are made available for review via external resources.
  - Reviewers are expected to fetch artifacts without leaving a trail.
- Authors rely on URL shortening services (another level of indirection) for artifacts.
- Artifacts made available on external resources may not remain permanently available.
  - Resources become hard to maintain over time.
  - Resources prone to garbage collection when authors switch jobs.

[4] SIGCOMM CCR now provides means to make artifacts available during the submission phase

Challenges | Lack of appreciation for good review work

- Limited pool of reviewers that provide good (substantial and constructive) reviews.
- Checking for reproducibility increases review expectations further.
- Conferences experimenting with automated review assignment systems [11, 12].

- Publicly releasing reviews[5] of an accepted paper helps with reproducibility.
  - Helps future readership to critically examine an accepted paper.

[5] IMC trailed making reviews publicly available for few years

Recommendations

- Discuss reproducibility considerations
- Allow authors to upload artifacts
- Ask review questions on reproducibility
- Highlight reproducible papers

Recommendations | Discuss reproducibility considerations

- A reproducibility considerations[6] section:
  - To ensure authors think about reproducibility.
  - Describes where code is available or how to get (or produce) datasets.

- Make measurement papers runnable [13, 14] (in the long run):
  - Play the process of consuming raw to data to produce results.
  - Helps see intermediate results; makes analytical errors visible.
  - Creates an incentives for carefulness.
  - Encourages application of analysis to an independent dataset.

[6] similar to an ethical considerations section

Recommendations | Allow authors to upload artifacts

- ACM SIGPLAN conferences employ an Artifacts Evaluation Committee (AEC) [15].
- SIGCOMM CCR allows authors to submit artifacts during submission phase.
- SIGCOMM CCR relaxes page limits for reproducible papers.
- Conferences can split paper and artifact (few weeks after) submission deadlines[7].
- Conferences can encourage authors to demo software to increase plausibility of results.
- Publishers (ACM *et al.*) should allow authors to upload artifacts with the paper.

[7] This involves a risk of releasing artifacts to anonymous reviewers before paper acceptance.

Recommendations | Ask review questions on reproducibility

- Accomodate questions in the review form concerning reproducibility:
  - Are artifacts available? Is advise on how results can be reproduced provided?
  - Can the released code be easily run on alternate datasets?
  - Is the methodology suitably explained to allow rewriting code?

Recommendations | Highlight reproducible papers

- Not practical to reject all non-reproducible papers.
- Good, working and reproducible papers should get attention they deserve.
  - Publishers can badge[8] and highlight reproducible papers.
  - Conferences can bestow best dataset awards.
  - AEC can be used to sample and evaluate papers on reproducibility.
  - Journals receiving extended conference papers can be strict on reproducibility.
  - SIGCOMM CCR can dedicate a column for papers that reproduce [16] results.
  - New venues [17] that solicit papers that reproduce research may help.

[8] This will require a mechanism to ensure badges do not become fake over time

Challenges with Reproducibility

- Despite challenges, state of reproducibility is not dismal, but improving –
  - Research is being reproduced [18, 19, 20], albeit rarely.
  - DatCat [21] & CRAWDAD [22] provide index of existing measurement data.

- Recommendations –
  - Discuss reproducibility considerations
  - Allow authors to upload artifacts
  - Ask review questions on reproducibility
  - Highlight reproducible papers

…may not be concluding wisdom, but maybe an incentive to reproducibility.

www.vaibhavbajpai.com
bajpaiv@in.tum.de | @bajpaivaibhav

References

[1] ACM. (2016) Artifact review and badging. [Online]. Available: https://www.acm.org/publications/policies/artifact-review-badging

[2] S. Kurkowski, T. Camp, and M. Colagrosso, "MANET simulation studies: The incredibles," *Mobile Computing and Communications Review*, vol. 9, no. 4, pp. 50–61, 2005. [Online]. Available: http://doi.acm.org/10.1145/1096166.1096174

[3] P. Vandewalle, J. Kovacevic, and M. Vetterli, "Reproducible Research in Signal Processing," *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 37–47, May 2009.

[4] C. S. Collberg and T. A. Proebsting, "Repeatability in computer systems research," *Communications of the ACM*, vol. 59, no. 3, pp. 62–69, 2016. [Online]. Available: http://doi.acm.org/10.1145/2812803

[5] V. Paxson, "Strategies for sound internet measurement," in *ACM SIGCOMM Internet Measurement Conference, IMC 2004, Sicily, Italy, October 25-27, 2004*, 2004, pp. 263–271. [Online]. Available: http://doi.acm.org/10.1145/1028788.1028824

[6] B. Krishnamurthy, W. Willinger, P. Gill, and M. F. Arlitt, "A socratic method for validation of measurement-based networking research," *Computer Communications*, vol. 34, no. 1, pp. 43–53, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.comcom.2010.09.014

[7] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig, "Ten simple rules for reproducible computational research," *PLoS Computational Biology*, vol. 9, no. 10, 2013. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.1003285

[8] V. Bajpai, A. W. Berger, P. Eardley, J. Ott, and J. Schönwälder, "Global measurements: Practice and experience (report on dagstuhl seminar #16012)," *Computer Communication Review*, vol. 46, no. 2, pp. 32–39, 2016. [Online]. Available: http://doi.acm.org/10.1145/2935634.2935641

[9] P. Eardley, M. Mellia, J. Ott, J. Schönwälder, and H. Schulzrinne, "Global measurement framework (dagstuhl seminar 13472)," *Dagstuhl Reports*, vol. 3, no. 11, 2013. [Online]. Available: http://dx.doi.org/10.4230/DagRep.3.11.144

[10] V. Bajpai, S. J. Eravuchira, and J. Schönwälder, "Lessons learned from using the RIPE atlas platform for measurement research," *CCR*, vol. 45, no. 3, pp. 35–42, 2015. [Online]. Available: http://doi.acm.org/10.1145/2805789.2805796

[11] B. Li and Y. T. Hou, "The new automated IEEE INFOCOM review assignment system," *IEEE Network*, vol. 30, no. 5, pp. 18–24, 2016. [Online]. Available: http://dx.doi.org/10.1109/MNET.2016.7579022

[12] S. Price and P. A. Flach, "Computational support for academic peer review: A perspective from artificial intelligence," *Communications of the ACM*, vol. 60, no. 3, pp. 70–79, 2017. [Online]. Available: http://doi.acm.org/10.1145/2979672

[13] C. Boettiger, "An introduction to docker for reproducible research," *Operating Systems Review*, 2015. [Online]. Available: http://doi.acm.org/10.1145/2723872.2723882

[14] N. Handigol, B. Heller, V. Jeyakumar, B. Lantz, and N. McKeown, "Reproducible network experiments using container-based

emulation," in *CoNEXT '12*, 2012. [Online]. Available: http://doi.acm.org/10.1145/2413176.2413206

[15] S. Krishnamurthi and J. Vitek, "The real software crisis: Repeatability as a core value," *Communications of the ACM*, vol. 58, no. 3, pp. 34–36, 2015. [Online]. Available: http://doi.acm.org/10.1145/2658987

[16] L. Yan and N. McKeown, "Learning networking by reproducing research results," *Computer Communication Review*, vol. 47, no. 2, pp. 19–26, May 2017. [Online]. Available: http://doi.acm.org/10.1145/3089262.3089266

[17] P. ONE. (2012) Reproducibility initiative. [Online]. Available: https://validation.scienceexchange.com

[18] B. Clark, T. Deshane, E. M. Dow, S. Evanchik, M. Finlayson, J. Herne, and J. N. Matthews, "Xen and the art of repeated research," in *USENIX Annual Technical Conference*, 2004, pp. 135–144.

[19] H. Howard, M. Schwarzkopf, A. Madhavapeddy, and J. Crowcroft, "Raft refloated: Do we have consensus?" *Operating Systems Review*, vol. 49, no. 1, pp. 12–21, 2015. [Online]. Available: http://doi.acm.org/10.1145/2723872.2723876

[20] D. A. Popescu and A. W. Moore, "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.*, 2016. [Online]. Available: http://tma.ifip.org/2016/papers/tma2016-final10.pdf

[21] C. Shannon, D. Moore, K. Keys, M. Fomenkov, B. Huffaker, and K. Claffy, "The internet measurement data catalog," *Computer Communication Review*, 2005. [Online]. Available: http://doi.acm.org/10.1145/1096536.1096552

[22] J. Yeo, D. Kotz, and T. Henderson, "CRAWDAD: a community resource for archiving wireless data at dartmouth," *Computer Communication Review*, vol. 36, no. 2, pp. 21–22, 2006. [Online]. Available: http://doi.acm.org/10.1145/1129582.1129588