Sharing is caring — so where are your data?

Tristan Henderson
School of Computer Science, University of St Andrews,
St Andrews, Fife KY16 9SX, UK
tristan@cs.st-andrews.ac.uk

ABSTRACT

The networking research community lacks a tradition of sharing experimental data, or using such data for reproducing results. But are we really that bad? Are we worse than researchers in other fields? And if so, how can we do better?

Categories and Subject Descriptors

H.1.m [**Information Systems**]: Miscellaneous

General Terms

Experimentation

Keywords

network data, data archiving, measurement, experimental method

1. INTRODUCTION

Economics has long been referred to as the "dismal science" [1]. But in one aspect we, as computer "scientists", are even more dismal than our economist colleagues. I refer to our reticence to share experimental data: a fundamental part of the experimental method in that it aids the replicability and extensibility of experiments.

In 1985 the U.S. Committee on National Statistics published a report entitled *Sharing Research Data* [2]. This made sixteen recommendations, the first and most important of which was "Sharing data should be a regular practice." The U.S. National Science Foundation also "expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work" [3].

So why, over twenty years later, have we failed to follow many of these recommendations? Why aren't data archives and catalogues such as DatCat [4] and CRAWDAD [5] teeming with experimental data from the last twenty years of SIGCOMM, INFOCOM, MobiCom and ToN papers?

2. THE SITUATION ELSEWHERE

Let us take a quick tour of what our counterparts in other fields are up to. In the social sciences, the Inter-University Consortium for Political and Social Research [6] has archived and provided data to the research community over for forty years. Interestingly, the ICPSR offers access to public data as well as a formal mechanism for providing access to restricted sensitive data. The latter is relevant given the oft-heard cry from parts of our community that data cannot be shared because of their commercially-sensitive nature.

In the space sciences, it is somewhat expensive to run experiments. Luckily, the relevant space agencies, such as NASA [7] and the ESA [8], provide data for the benefit of all. The data collectors themselves often become authors on papers analysing the data, providing recognition and incentive to collect data in the first place.

Our dismal scientific colleagues, the economists, have studied the need to share data and the effects of this on research output [9]. While there are several economics data archives [10, 11, 12], it has been argued that these could be better [13]: contribution levels are low, and some of the data that are submitted are insufficient to enable replicable results.

The life sciences appear to have the most organised system for data-sharing. Journals have strict policies about sharing data before articles can be published [14, 15], funding bodies have rules about making data available [16], and so unsurprisingly there exist several large data archives, e.g., [17, 18].

There is also ongoing work in data-archiving in other fields, with the JISC Digital Repositories programme [19] funding new archives in environmental sciences [20] and chemistry [21], amongst others.

3. WHAT CAN YOU DO?

Hopefully by now you feel shamed by your failure to share your exciting experimental data, and you have acquired a paper bag to wear over your head to hide from your friends in other fields. So what next? I humbly present some advice for sharing:

Be prepared. Consider data-sharing at the outset of experiments. Document *everything* — if your work becomes as groundbreaking or as well-renowned as many of us hope, you may well be asked for obscure details about network topologies, kernel configurations or hardware versions several years after the experiments were conducted.

Be resourced. Allocate resources for data-sharing. Even though (as a random example) the CRAWDAD project can help you in creating metadata for your data, sanitising and making the data available, you as the collector of the data know the most about the data. Thus you will need to spend some time verifying the metadata and so forth. Of course if you are prepared from the outset then you will save time in the long run.

Be ethical and legal. The growing awareness of the legal issues in conducting networking research [22, 23] means that ignorance is not a defence. But as a networking researcher, it is likely that YANAL (You Are Not A Lawyer). So you may wish to speak to lawyers, or at the very least the data protection officers, ethics committee or institutional review board at your institution as appropriate. Remember that network users are human subjects, so treat them as such; make sure to get consent from users where possible.

Conversely, if you have no data to share, but instead would like to use data from other researchers, I have some tips for you as well:

Be grateful. Cite data sources as well as (or instead of) papers. Archives such as DatCat and CRAWDAD provide easy-to-use identifiers for citation, and the use of these can help justify the existence of the archives to funding bodies, and convince data contributors of the worth of doing so. This in turn will hopefully lead to more data being made available.

Be communicative. Inform data providers when you use their data. This will again be of use in funding said data provision. Equally, if there are problems in a set of data, or useful information that is missing, let the data providers know, as it may be possible to fix the problems, improve the metadata, or to collect additional information in future experiments.

Be nice. If the data providers have any conditions of use, please respect them. If providers participate in analysis or process data, consider making them co-authors. Don't use the data in ways that might get the data providers into trouble [24].

Be persistent (but not annoying). If there are particular data that you would find useful for your research, let people know. Bug data collectors or archivers for their data. We poll attendees of our CRAWDAD workshops to find out what data would be of interest; see the workshop reports [25, 26], and please get in touch if you have new areas of interest, or if you know of any data that belong in the archive. Persistence is required to encourage data contribution [27].

4. CONCLUSION

Sharing data is good. We should do more of it. Obviously I have glossed over all of the difficulties of sharing data, but I hope that most of these can be overcome. If not, then I welcome dissenting opinions in future CCR columns or in CCR Online; a constructive discourse on how to encourage more data-sharing would be great. If we as a community recognise the efforts of those who do share data, or who reproduce experiments, then perhaps that will encourage others? Or perhaps we should just mandate data-sharing? The community should decide. But given that other fields are able to this, and that we are supposed to know something about ICT, we really should be able to do better.

About the author

Tristan has degrees in both of the dismal sciences, and helps to run the CRAWDAD wireless network data archive. Whether that means you should listen to him is up to you.

Acknowledgements

Thanks to Denise Anthony and Victoria Cowling for pointers to data archives in other fields, to Saleem Bhatti, Jon Crowcroft, David Kotz and Richard Mortier for feedback, and to Christophe Diot for persistently reminding me to write this. And of course a big thanks to all of the CRAWDAD data contributors for doing their part.

5. REFERENCES

[1] T. Carlyle. Occasional Discourse on the Negro Question. *Fraser's Magazine*, 40:670–679, December 1849.

[2] S. E. Fienberg, M. E. Martin, and M. L. Straf. *Sharing Research Data*. National Academy Press, 1985. http://www.nap.edu/openbook.php?isbn=030903499X.

[3] National Science Foundation. General grant conditions (GC-1), June 01 2007. http://www.nsf.gov/publications/pub_summ.jsp?ods_key=gc160107.

[4] DatCat: Internet measurement data catalog. http://www.datcat.org/.

[5] CRAWDAD: A community resource for archiving wireless data at Dartmouth. http://crawdad.cs.dartmouth.edu/.

[6] Inter-University Consortium for Political and Social Research. http://www.icpsr.umich.edu/.

[7] NASA space science data archives. http://science.hq.nasa.gov/research/space_science_data.html.

[8] ESA Earthnet. http://earth.esa.int/.

[9] R. Anderson, W. H. Greene, B. D. McCullough, and H. D. Vinod. The role of data/code archives in the future of economic research. *Journal of Economic Methodology*, 2007. forthcoming.

[10] Economic and Social Data Service. http://www.esds.ac.uk/.

[11] The UK data archive. http://www.data-archive.ac.uk/.

[12] Council of European Social Science Data Archives. http://extweb3.nsd.uib.no/cessda/.

[13] B. D. McCullough. Got replicability? The Journal of Money, Credit and Banking archive. *Econ Journal Watch*, 4(1):326–337, September 2007.

[14] Nature Publishing Group. Availability of data & materials. http://www.nature.com/authors/editorial_policies/availability.html.

[15] Science Magazine. Database deposition policy. http://www.sciencemag.org/about/authors/prep/gen_info.dtl#datadep.

[16] NIH data sharing policy. http://grants1.nih.gov/grants/policy/data_sharing/.

[17] National Center for Biotechnology Information. Entrez nucleotide sequence database. http://www.ncbi.nlm.nih.gov/sites/entrez?db=Nucleotide.

[18] UNC microarray database. https://genome.unc.edu/.

[19] JISC digital repositories programme. http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories.aspx.

[20] CLADDIER (Citation, Location, and Deposition in Discipline & Institutional Repositories). http://claddier.badc.ac.uk/.

[21] Project SPECTRa (Submission, Preservation and Exposure of Chemistry Teaching and Research Data). http://www.lib.cam.ac.uk/spectra/.

[22] JANET activities: Network monitoring and investigation. http://www.ja.net/development/legislation/activities/monitoring.html.

[23] P. Ohm, D. Sicker, and D. Grunwald. Legal issues surrounding monitoring during network research. In *Proceedings of the Internet Measurement Conference*, pages 141–148, San Diego, CA, USA, October 2007.

[24] M. Allman and V. Paxson. Issues and etiquette concerning use of shared measurement data. In *Proceedings of the Internet Measurement Conference*, pages 135–140, San Diego, CA, USA, October 2007.

[25] J. Yeo, D. Kotz, and T. Henderson. CRAWDAD: A Community Resource for Archiving Wireless Data at Dartmouth. *ACM Computer Communication Review*, 36(2):21–22, April 2006.

[26] J. Yeo, T. Henderson, and D. Kotz. Workshop report — CRAWDAD workshop 2006. *Mobile Computing and Communications Review*, 11(1):67–69, January 2007.

[27] B. D. McCullough, K. A. McGeary, and T. D. Harrison. Do economics journal archives promote replicable research?, September 2006. Available at SSRN: http://ssrn.com/abstract=931231.