Data citation practices in the CRAWDAD wireless network data archive

Tristan Henderson[1] and David Kotz[2]

[1] School of Computer Science, University of St Andrews, St Andrews, UK
tnhh@st-andrews.ac.uk
[2] Department of Computer Science, Dartmouth College, Hanover, NH, USA
kotz@cs.dartmouth.edu

Abstract. CRAWDAD (Community Resource for Archiving Wireless Data At Dartmouth) is a popular research data archive for wireless network data, archiving over 100 datasets used by over 6,500 users. In this paper we examine citation behaviour amongst 1,281 papers that use CRAWDAD datasets. We find that (in general) paper authors tend to cite datasets in a manner that is sufficient for providing credit to dataset authors, but also providing access to the datasets that were used. Only 11.5% of papers did not do so; common problems included (1) citing the canonical papers rather than the dataset, (2) describing the dataset using unclear identifiers, or (3) not providing URLs or pointers to datasets.

1 Introduction

The archiving, sharing and reuse of research data is a fundamental part of the scientific process, and the benefits of doing so have been increasingly recognised [4,13]. Indeed such data sharing is now being encouraged or mandated by research funders [3,10].

Since 2005 we have run the CRAWDAD network-data archive as a resource for wireless-network researchers to deposit and share their data, and for other researchers to download and use data in their research. Wireless network data is crucial for conducting research into future wireless networks; much research is based on analytical or simulation models that might not reflect the real world. We have therefore encouraged researchers to share their data, such as mobility measurements of wireless network users, or radio measurements of networks. By many measures CRAWDAD has been a success, with over six thousand users using datasets in over a thousand papers, and in teaching and standards development. CRAWDAD datasets are not used solely by wireless-network researchers; we have observed researchers from other fields such as geography, epidemiology, and sociology using our datasets.

In this paper we investigate how these CRAWDAD users cite our datasets when they use them. Understanding how people cite data, and the problems that they have in citing data, is important if we are to maximise the usefulness of shared research data. If data are cited in a way that makes it difficult to find or interpret the data, then this might limit future research. These issues have recently been recognised by the Force 11 Data Citation Principles [5].

D-Lib Magazine, volume 21, number 1/2, January 2014. doi:10.1045/january2015-henderson
Corporation for National Research Initiatives (CNRI). ©Copyright the authors.

2 The CRAWDAD data archive

CRAWDAD[3] (Community Resource for Archiving Wireless Data at Dartmouth) was founded in 2005 [7], and initially funded for three years through a Community Resource award from the US National Science Foundation (NSF). The original NSF proposal summarised the archive as follows:

> The investigators propose to develop an archive of wireless network data and associated tools for collecting and processing the data, as a community resource for those involved in wireless network research and education. Today, this community is seriously starved for real data about real users on real networks. Most current research is based on analytical or simulation models; due to the complexity of radio propagation in the real world and a lack of understanding about the behavior of real wireless applications and users, these models are severely limited. On the other hand, the difficult logistical challenges involved in collecting detailed traces of wireless network activity preclude most people from working with experimental data.
>
> The investigators have years of experience collecting data from wireless networks, and have released this data to the research community. The community's hunger for this data clarified the need for a community-run facility with a larger capacity and staff to develop the necessary tools.

Starting with a single wireless network dataset collected by the investigators, CRAWDAD has grown to become what we believe is the largest data archive of its type.[4] As of October 2014 we host 116 datasets and tools. We require users to register and agree to a license before they can download datasets (viewing of metadata is free) and according to our registration records we now have over 6,200 users from 101 countries.

To initially bootstrap and publicise the archive, we held a series of workshops at the largest wireless network research conferences [16,14,15]. We also approached premier publication venues in our research community (in computer science these are typically conferences rather than journals) with a view to encouraging or even mandating data sharing. These approaches were not particularly successful, with the exception of the Internet Measurement Conference,[5] which now requires data sharing for those papers that wish to be considered for a best paper award. Other mechanisms for encouraging researchers to contribute to the data archive include the all-important crawdad toys (Figure 1) and stickers that are sent out to contributors! We also attempt to make it as easy as possible for researchers to contribute their data, by helping them to create metadata for their datasets. Because wireless data might well contain sensitive data (e.g., locations, or application-usage information), we help researchers with santising and removing sensitive data. As there do not exist standard formats for much of the data collected in our field, however, we generally point to existing tools and algorithms and help data contributors, but do not carry out the sanitisation ourselves.

[3] A "crawdad" is a crustacean, also known as a crayfish, crawfish or yabby.

[4] This is perhaps not as impressive as it might sound, as in general, there are few network data archives and even fewer dedicated to wireless networks.

[5] `http://imc-conf.org/`

Fig. 1. To encourage deposits in the archive, data contributors are given a “crawdad” toy.

3 Encouraging data citation

To encourage users to cite CRAWDAD datasets, we provide BibTeX (e.g., Figure 2) that authors can use. Each of our datasets has a unique name, e.g., `dartmouth/campus`, which indicates the institution that collected the data (in this case, Dartmouth College), and a short descriptive identifier (in this case, the collection was from a campus-wide wireless network, hence `campus`). Each identifier is also used to provide a persistent URL of the form `crawdad.org/{identifier}`, e.g., `http://crawdad.org/dartmouth/campus/`.[6] While the main CRAWDAD site is hosted in the USA at `http://crawdad.org/`, we have mirrors in the UK and Australia at `http://uk.crawdad.org/` and `http://au.crawdad.org`, respectively. Our preference, as indicated in Figure 2, is that authors cite the main site.

4 Methodology

To study data citation practices amongst CRAWDAD users, we obtained a corpus of papers that use and cite CRAWDAD datasets. We urge CRAWDAD users to inform us when they publish a paper. This request has not been successful; only three authors have done this. As an alternative we further ask authors to add their papers directly to our group library on the CiteULike online bibliographical reference service;[7] similarly this request has been infrequently fulfilled, with only five authors adding their papers.

Therefore, to build a corpus of papers that use CRAWDAD datasets, we trawled various paper databases (ACM Digital Library, Google Scholar, IEEE Xplore, ScienceDirect and Scopus) to find any paper that mentioned the word “crawdad”. We then

[6] Our use of persistent identifiers predates the introduction of DOIs for datasets. We are currently investigating the use of DOIs, although thus far we have found most options to be of prohibitive cost for a data archive of our size.

[7] `http://www.citeulike.org/group/5303/library`

```
@MISC{dartmouth-campus-2007-02-08,
    author = {David Kotz and Tristan Henderson and Ilya Abyzov
        and Jihwang Yeo},
    title = {{CRAWDAD} data set dartmouth/campus
        (v. 2007-02-08)},
    howpublished = {Downloaded from
        http://crawdad.org/dartmouth/campus/},
    month = feb,
    year = 2007
}
```

Fig. 2. Example BibTeX

removed any false positives; that is, papers that discussed crayfish, the Crawdad text-mining tool, the Crawdad neurophysiology tool, and any other references that were not related to our data archive. This produced a list of 1,544 papers.

We manually examined each of these papers. 249 papers only mentioned the CRAWDAD archive (e.g., papers that discuss data archiving, or papers that considered using a dataset but did not). Of the remaining 1,295 papers that appeared to use CRAWDAD datasets, we were able to source PDF files for 1,281 of these. These were converted to text using pdftotext[8] to simplify some of the analysis.

Each paper was examined to determine whether it met the following criteria, based on a subset of the Force 11 Data Citation Principles:

1. **Credit and attribution**: do the data citations appropriately credit the creators of the dataset?
2. **Unique identification**: we provide unique names for each dataset; are these mentioned?
3. **Access**: do the data citations provide sufficient information for a reader to access the dataset?
4. **Persistence**: we provide persistent URLs for each dataset; are these used?

We consider a paper that cites data in such a way that fulfils all four of these criteria to cite data *sufficiently*. A paper that uses the BibTeX that we provide will be sufficient.

We are additionally interested in how other researchers cite papers that cite our datasets. To determine citation counts for papers, we used a publicly available command-line tool for querying the Google Scholar database.[9]

5 Results

Figure 3 shows the number of papers that have cited CRAWDAD datasets each year since the archive was launched. As the archive has grown in visibility and the number

[8] http://www.foolabs.com/xpdf/home.html

[9] https://github.com/ckreibich/scholar.py

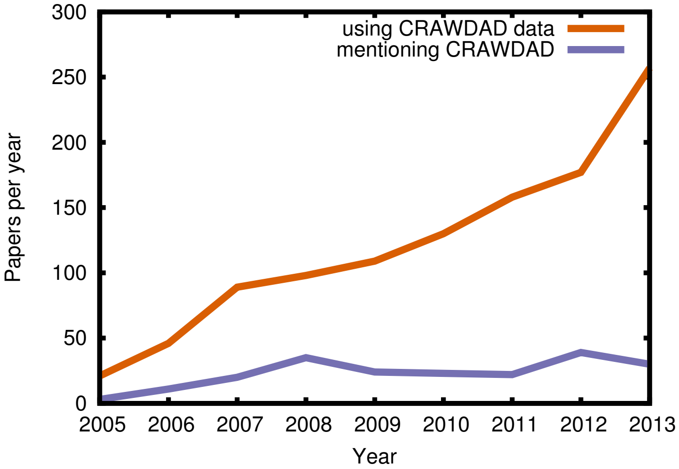

Fig. 3. Papers that cite CRAWDAD datasets, by year of publication (2014 is omitted as the year is not complete).

of datasets available has also increased (Figure 4), we see that the number of papers published each year using our datasets has risen.

Of the 1,281 papers that we studied, 1,134 papers (88.5%) fulfilled all four of our sufficiency criteria. This result is reassuring as it indicates that most authors are either using our provided citation information, or choosing to cite CRAWDAD datasets in a way that is useful to other researchers. But 147 papers exist that used our data in a way that is unclear. Some of the most common citation errors that we found include:[10]

- 48 papers cited the canonical papers for datasets, i.e., the original papers that described the studies that led to the collection of the data. While this is useful for providing credit and attribution, it neither helps readers to identify nor to access a dataset. In some cases authors have contacted us and asked to deposit datasets and be issued with a URL before the publication of a paper, but this is not the norm. Thus, most of the canonical papers do not have information on how to access the dataset used in these papers.
- 49 papers chose to describe the dataset, rather than use the identifier that we provide. For instance, “we used a dataset of campus Wi-Fi users from Dartmouth College” rather than “we used the CRAWDAD dartmouth/campus dataset”. Conversely, 19 papers chose to reference papers by a name of their own design, such as a truncated version of our identifier, e.g., “we used the Infocom06 dataset” instead of the “cambridge/haggle” dataset which includes a trace called “Infocom06”.
- 83 papers cited the main CRAWDAD site (e.g., “we obtained data from CRAWDAD” or simply mentioned `http://crawdad.org/`). In some cases this was in conjunction with a description of the dataset, or a citation to the canonical dataset

[10] Note that these numbers do not add up to 148 since papers may have made more than one error.

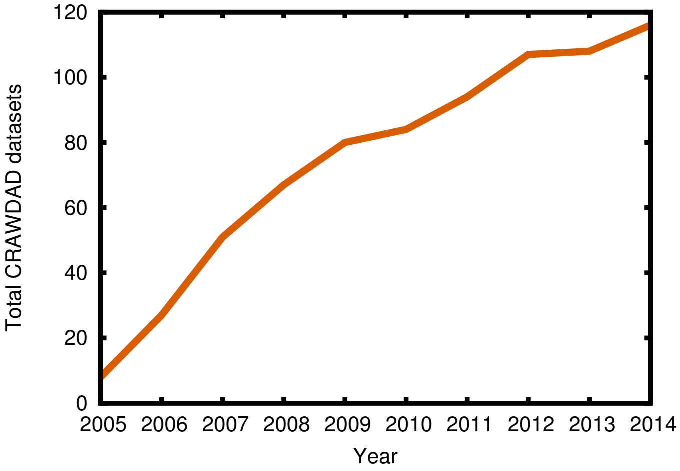

Fig. 4. Total datasets in the CRAWDAD archive over time. Growth appears to be relatively constant. The recent slight slowdown in growth can be attributed to changes in personnel and funding.

paper. But this does not provide sufficient credit. It also does not necessarily provide access, since a reader would have to visit the main site and then attempt to determine which of the datasets had been used. One paper provided a direct pointer to a dataset URL with no other description, which enables easy access but insufficient credit.

- 28 papers did not provide any URL or pointer to where data might be obtained, thereby failing to provide sufficient access. One paper provided a URL that did not exist.
- Eight papers attributed authorship of a dataset to “CRAWDAD” or “Dartmouth College” rather than the actual dataset authors. Thus even if the access criteria were met, the authors were not correctly attributed.
- 32 papers described the datasets used in such a vague way that it was impossible for us to determine what particular data had been used. This included cases where a dataset with multiple traces was used, but the particular traces were not denoted. Or cases where authors simply referred to “real scenarios available in the CRAWDAD data repository” or even more vaguely “a packet trace” with a pointer to the main CRAWDAD website.

We also observed one further issue to do with persistent URLs. The first version of our website used PHP and would redirect from the persistent URL to a dynamic page that was automatically generated from the dataset’s metadata. This page had a different URL, such that `http://crawdad.org/dartmouth/campus/` would redirect to `http://crawdad.cs.dartmouth.edu/meta.php?name=dartmouth/campus`. 80 papers used URLs of the latter form. In 2014 we redesigned our website, and the new website no longer uses PHP nor these redirected URLs. Having noticed this citation practice, we have since reconfigured the new website such that these `meta.php` links now redirect back to the original persistent URL. As we maintain these links, we consider papers that use these links to meet our sufficiency criterion for persistence, but we note this as a problem that needs to be addressed. The use of DOIs should help with this problem (assuming that authors do indeed cite the DOIs rather than URLs).

We were also interested in how many datasets were cited by papers. Table 1 shows that most papers used a single dataset, but a quarter of papers use more than one. One paper, which compares several of the datasets in the archive, used 11 datasets.

Number of datasets used in a paper	*Number of papers*
0	6
1	958
2	219
3	74
4	29
5	7
6	1
11	1

Table 1. The number of CRAWDAD datasets used by research papers. 331 (25.8%) of the papers where we were able to identify the used datasets use more than one dataset. The six papers with 0 datasets are those where we could not determine if any data had actually been used, despite referencing the CRAWDAD archive.

It has been observed that papers that share data tend to be cited more [12]. We further investigated whether papers that use shared data and cite them sufficiently are also cited more. Figure 5 shows a boxplot of Google Scholar citations for the two sets of papers that cite sufficiently and insufficiently. It does appear that papers that cite sufficiently are cited more, but further work is needed to investigate this more deeply.

6 Related work

We are not aware of any study that looks at citation practices in wireless network research. Most studies of research practice in this field have focused on other issues such as reproducibility [8]. Data citation has, however, been studied in the fields of environmental science [2] and gene expression and phylogenetics [11]. Similar to our study, these papers report on preliminary and ongoing work.

More broadly, various working groups have been examining data citation and proposing ways to improve citation practices [1,9]. The aforementioned data citation principles (Section 4) are another example of this trend.

7 Conclusions and Future Work

It appears that, in general, our datasets are cited in a useful fashion, with 88% of the papers in our study citing datasets in what we deem to be a sufficient fashion. But the

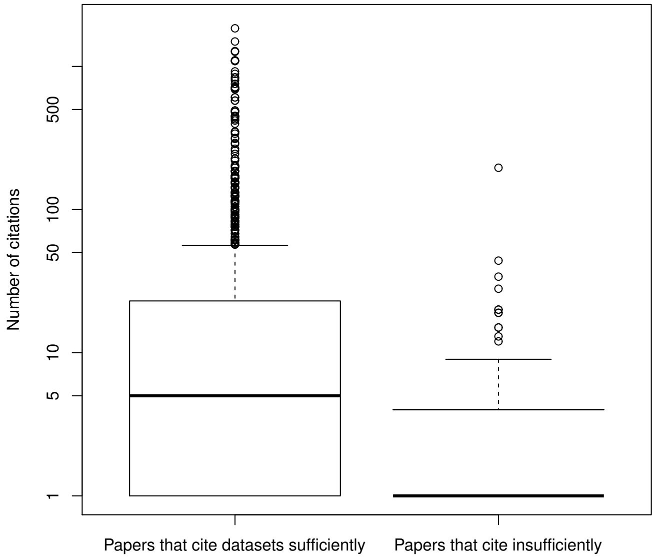

Fig. 5. Boxplot showing the number of citations for papers published before 2014 that cite CRAWDAD datasets sufficiently versus those that do not. Note the logarithmic y-axis.

biggest drawback of this corpus is that it is biased towards those papers that have actually mentioned “CRAWDAD” somewhere in the text. There are no doubt a number of papers that have used our datasets without mentioning the name of our archive. Finding these papers is a challenge. But we note that several papers in our existing corpus cite the canonical papers for datasets rather than the dataset themselves. So one search strategy would be to extend our search to all of the papers that reference any of the canonical papers for any of our datasets, and then search these for mentions of data use.

Of the papers that cited data insufficiently, we note that a common problem is citing papers rather than datasets. The issue here is that often a paper is published before the datasets are made available, and so by reading a paper, one has no way of finding the related datasets. Indeed in many cases the decision to share data can be made sometime after the paper is published (e.g., after prodding from data archive curators such as ourselves!) and so there might not even be an indication in the paper that the data are available. Clearly some mechanism for linking datasets and publications, and for not treating publications as static immutable objects, is needed. Being able to update the links between publications and datasets would also help with some of our other observed problems, such as the citing of redirected URLs rather than persistent links, or papers that cited incorrect URLs. The most common venues for computer science publications, such as the ACM (Association for Computing Machinery) and IEEE (Institute of Electrical and Electronics Engineers), do not currently permit the updating of the PDF papers that have been uploaded to the digital libraries. Versioning of papers, or more widespread provision of additional online information or resources, might help with this.

In addition to observing citation behaviour, we are also trying to better understand the motivations for data citation. For instance, do people insufficiently cite datasets because they are pressed for space? Or because they assume that readers will know where to source the relevant datasets? To this end we are currently conducting a survey of the CRAWDAD userbase.

Finally, having collected this corpus we believe that it might be of use to other researchers in understanding data citation practices. To this end we have made the data available (as a BibTeX bibliography with keywords indicating our classifications and an additional field for citations) on FigShare [6].

8 Acknowlegements

We are thankful for the generous support of our current funders ACM SIGCOMM and ACM SIGMOBILE, and our past funders Aruba Networks, Intel and the National Science Foundation.

Authors

Tristan Henderson is a Senior Lecturer in Computer Science at the University of St Andrews. His research aims to better understand user behaviour and use this to build improved systems; an approach which has involved measurements and testbeds for networked games, wireless networks, mobile sensors, smartphones, online social networks and opportunistic networks. Together with David Kotz he runs the CRAWDAD wireless network data archive described in this paper. Dr Henderson holds an MA in Economics from Cambridge University and an MSc and PhD in Computer Science from University College London. For more information, see `http://tristan.host.cs.st-andrews.ac.uk/`.

David Kotz is the Champion International Professor in the Department of Computer Science, and Associate Dean of the Faculty for the Sciences, at Dartmouth College. His research interests include security and privacy, pervasive computing for healthcare, and wireless networks. He is an IEEE Fellow, a Senior Member of the ACM, a 2008 Fulbright Fellow to India, and an elected member of Phi Beta Kappa. After receiving his A.B. in Computer Science and Physics from Dartmouth in 1986, he completed his Ph.D in Computer Science from Duke University in 1991 and returned to Dartmouth to join the faculty. See `http://www.cs.dartmouth.edu/~dfk/`.

References

1. CODATA-ICSTI Task Group on Data Citation Standards and Practices. Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal*, 12:CIDCR1–CIDCR75, 13 Sept. 2013. doi:`10.2481/dsj.osom13-043`.
2. V. Enriquez, S. W. Judson, N. M. Walker, S. Allard, R. B. Cook, H. A. Piwowar, R. J. Sandusky, T. J. Vision, and B. E. Wilson. Data citation in the wild. Poster paper, Dec. 2010. doi:`10.1038/npre.2010.5452.1`.
3. EPSRC policy framework on research data. Accessed 14 October 2014, Online at `http://www.epsrc.ac.uk/index.cfm/about/standards/researchdata/`.
4. B. A. Fischer and M. J. Zigmond. The essential nature of sharing in science. *Science and Engineering Ethics*, 16(4):783–799, Dec. 2010. doi:`10.1007/s11948-010-9239-x`.
5. Force11. Joint declaration of data citation principles, 2014. Accessed 14 October 2014, Online at `https://www.force11.org/datacitation`.

6. T. Henderson and D. Kotz. CRAWDAD wireless network data citation bibliography, 14 Oct. 2014. doi:10.6084/m9.figshare.1203646.
7. D. Kotz and T. Henderson. CRAWDAD: A Community Resource for Archiving Wireless Data at Dartmouth. *IEEE Pervasive Computing*, 4(4):12–14, Oct. 2005. doi:10.1109/mprv.2005.75.
8. S. Kurkowski, T. Camp, and M. Colagrosso. MANET simulation studies: the incredibles. *ACM SIGMOBILE Mobile Computing and Communications Review*, 9(4):50–61, Oct. 2005. doi:10.1145/1096166.1096174.
9. M. S. Mayernik. Bridging data lifecycles: Tracking data use via data citations workshop report. Technical Report NCAR/TN-494+PROC, National Center for Atmospheric Research, Jan. 2013. doi:10.5065/D6PZ56TX.
10. NSF policy on dissemination and sharing of research results. Accessed 14 October 2014, Online at http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/aag_6.jsp#VID4.
11. H. A. Piwowar, J. D. Carlson, and T. J. Vision. Beginning to track 1000 datasets from public repositories into the published literature. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–4, 2011. doi:10.1002/meet.2011.14504801337.
12. H. A. Piwowar, R. S. Day, and D. B. Fridsma. Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, 2(3):e308+, 21 Mar. 2007. doi:10.1371/journal.pone.0000308.
13. J. C. Wallis, E. Rolando, and C. L. Borgman. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE*, 8(7):e67332+, 23 July 2013. doi:10.1371/journal.pone.0067332.
14. J. Yeo, T. Henderson, and D. Kotz. Workshop report - CRAWDAD workshop 2006. *Mobile Computing and Communications Review*, 11(1):67–69, Jan. 2007. Online at http://tristan.host.cs.st-andrews.ac.uk/pubs/mc2r07.pdf.
15. J. Yeo, T. Henderson, and D. Kotz. CRAWDAD workshop 2007. *ACM SIGCOMM Computer Communication Review*, 38(3):79–82, July 2008. doi:10.1145/1384609.1384619.
16. J. Yeo, D. Kotz, and T. Henderson. CRAWDAD: a community resource for archiving wireless data at Dartmouth. *ACM SIGCOMM Computer Communication Review*, 36(2):21–22, Apr. 2006. doi:10.1145/1129582.1129588.