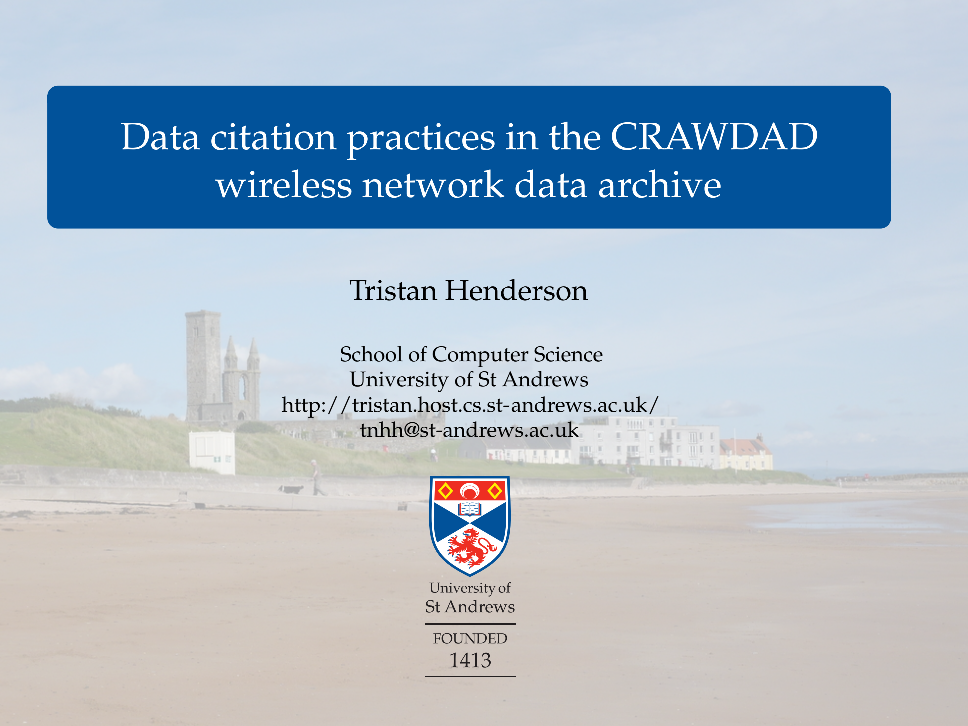
# Data citation practices in the CRAWDAD wireless network data archive

Tristan Henderson

School of Computer Science
University of St Andrews
http://tristan.host.cs.st-andrews.ac.uk/
tnhh@st-andrews.ac.uk



University of
St Andrews

FOUNDED
1413

*"If you can not measure it,
you can not improve it."*

*— Lord Kelvin*

# Me

- Computer systems researcher
- *Measure* and *improve* the behaviour of real-world computer systems and their users
- e.g.,
  - networked games
  - wireless networks
  - pervasive computing
  - opportunistic networks
  - online social networks
- Archive and share the data

# CRAWDAD

CRAWDAD
Community Resource for Archiving Wireless Data at Dartmouth
http://crawdad.org

- World's largest (!!) wireless network data archive
  - Funded by NSF, ACM SIGCOMM, ACM SIGMOBILE, Intel, Aruba
  - (always looking for more!)
- Archives:
  - wireless network traces
  - tools for collecting traces
  - (always looking for more!)
- Community support:
  - event calendar
  - bibliography
  - HOWTO documents / wiki
  - workshops

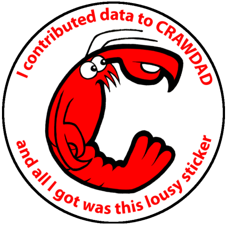
# CRAWDAD is popular!

As of September 2014:

- 6,501 users from 104 countries
- 116 datasets and tools used in over 1,500 papers (that we know of), including all of our top venues
- Some popular datasets:
  - Cambridge Bluetooth encounters: 328 papers
  - Dartmouth WLAN data: 268 papers
  - MIT Reality Mining: 149 papers
  - EPFL taxi cabs: 127 papers
- Definition of "wireless" is broad
  - have recently started archiving mobile/social datasets
  - datasets have been used for security, network management, geography, epidemiology, animal sociology, . . .
  - i.e., ESRC, BBSRC, MRC, NERC as well as EPSRC remit

# How did we get popular?

- Data sharing is good for science[1]
  - Indeed it is now required by RCUK[2]
- So everyone should be lining up to give us their datasets!
  - Not quite…
- Useful carrots:
  - citation/download tracking
  - literature on increased citations [3]
  - letters for department chairs/heads
  - toys! stickers!


I contributed data to CRAWDAD
and all I got was this lousy sticker

[1] T. Henderson. Sharing is caring: so where are your data? *ACM SIGCOMM Computer Communication Review*, 38(1):43–44, Jan. 2008. doi:10.1145/1341431.1341439

[2] www.rcuk.ac.uk/research/DataPolicy/

[3] 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

# Tracking usage

- We provide canonical URLs, e.g., `crawdad.org/dartmouth/campus`
  - indexed by Google Scholar (and Thomson Reuters when we get around to it)
  - no DOIs yet, although DataCite etc may help
- We provide BibTeX etc for authors, e.g., `G. Bigwood, D. Rehunathan, M. Bateman, T. Henderson, and S. Bhatti. CRAWDAD data set st_andrews/sassy (v. 2011-06-03). Downloaded from http://crawdad.org/st_andrews/sassy, June 2011`
- We request that authors tell us when they publish, or add to our CiteULike group[4]

[4] citeulike.org/groupfunc/5303/home

# Tracking usage

How many people have told us when they have published a paper using CRAWDAD datasets?

# Tracking usage

How many people have told us when they have published a paper using CRAWDAD datasets?

3

# Tracking usage

How many people have told us when they have published a paper using CRAWDAD datasets?

3

How many people (other than ourselves) have added papers to the CiteULike group?

# Tracking usage

How many people have told us when they have published a paper using CRAWDAD datasets?

3

How many people (other than ourselves) have added papers to the CiteULike group?

5

# Tracking usage in practice

1. Google Scholar/ScienceDirect/IEEExplore/…searches for "CRAWDAD"
2. filter out all the references to shellfish, CRAWDAD text analysis tool, CRAWDAD neurophysiology tool
3. check paper manually to determine which (if any) datasets were used

# Tracking usage in practice

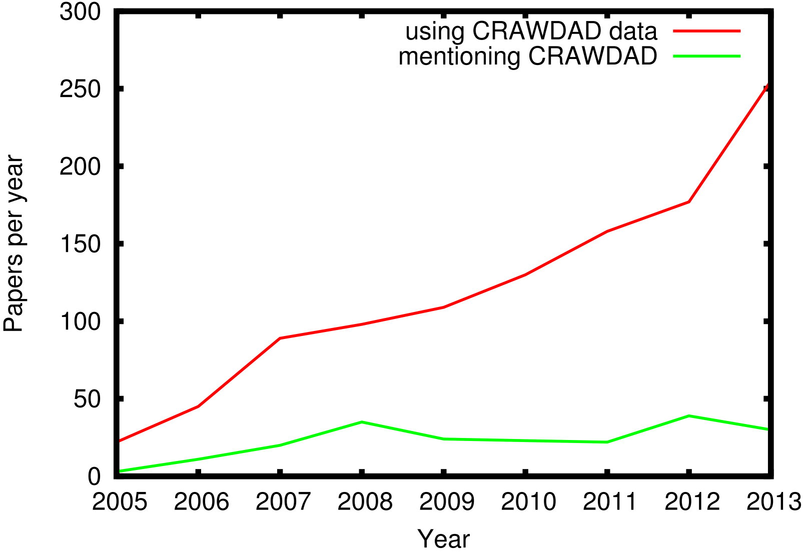
1. Google Scholar/ScienceDirect/IEEExplore/…searches for “CRAWDAD”
2. filter out all the references to shellfish, CRAWDAD text analysis tool, CRAWDAD neurophysiology tool
3. check paper manually to determine which (if any) datasets were used

- There *must* be a better way!

# CRAWDAD usage: healthy?

- ≈3,800 papers matching “CRAWDAD” full-text search
- 1,219 papers appear to use CRAWDAD datasets
    - able to find PDF files for 1,206 of them

# CRAWDAD usage: healthy



# Data Citation Principles

- force11.org/datacitation

1. Importance
2. Credit and Attribution
3. Evidence
4. Unique Identification
5. Access
6. Persistence
7. Specificity and Verifiability
8. Interoperability and Flexibility

# Data Citation Principles

- force11.org/datacitation

1. Importance
2. Credit and A
3. Evidence
4. Unique Ider
5. Access
6. Persistence
7. Specificity and Verifiability
8. Interoperability and Flexibility

In short:

- data are important
- cite them accurately
- different fields have different conventions

# Data Citation Principles

- force11.org/datacitation

1. Importance
2. Credit and
3. Evidence
4. Unique Ide
5. Access
6. Persistence
7. Specificity and Verifiability
8. Interoperability and Flexibility

In short:

- find a data archive that helps others cite your data
- find the conventions for your particular field

# CRAWDAD usage: healthy?

- ≈3,800 papers matching "CRAWDAD" full-text search
- 1,219 papers appear to use CRAWDAD datasets
    - able to find PDF files for 1,206 of them

# CRAWDAD usage: healthy?

- ≈3,800 papers matching “CRAWDAD” full-text search
- 1,219 papers appear to use CRAWDAD datasets
  - able to find PDF files for 1,206 of them
- 1,091 (90%) cited CRAWDAD data in a “reproducible” way

# CRAWDAD usage: healthy?

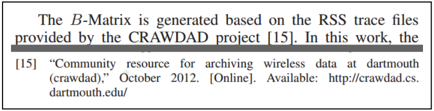
- ≈3,800 papers matching “CRAWDAD” full-text search
- 1,219 papers appear to use CRAWDAD datasets
  - able to find PDF files for 1,206 of them
- 1,091 (90%) cited CRAWDAD data in a “reproducible” way:
  - **credit and attribution**: do the data citations appropriately credit the creators of the dataset?
  - **unique identification**: we provide unique names for each dataset; are these mentioned?
  - **access**: do the data citations provide sufficient information for a reader to access the dataset?
  - **persistence**: we provide persistent URLs for each dataset; are these used?

# CRAWDAD usage: healthy?

- ≈3,800 papers matching "CRAWDAD" full-text search
- 1,219 papers appear to use CRAWDAD datasets
  - able to find PDF files for 1,206 of them
- 1,091 (90%) cited CRAWDAD data in a "reproducible" way:
  - **credit and attribution**: do the data citations appropriately credit th
  - **unique identification**: we provide unique names for each dataset; are these mentioned?
  - **access**: do the data citations provide sufficient information for a reader to access the dataset?
  - **persistence**: we provide persistent URLs for each dataset; are these used?

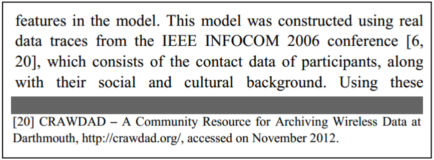
i.e., used our BibTeX

# Data citation; not always as intended



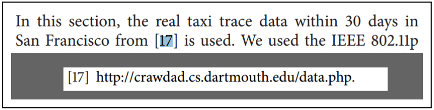
The *B*-Matrix is generated based on the RSS trace files provided by the CRAWDAD project [15]. In this work, the

[15] "Community resource for archiving wireless data at dartmouth (crawdad)," October 2012. [Online]. Available: http://crawdad.cs.dartmouth.edu/



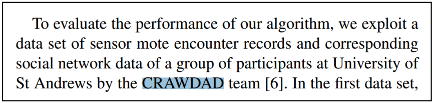
features in the model. This model was constructed using real data traces from the IEEE INFOCOM 2006 conference [6, 20], which consists of the contact data of participants, along with their social and cultural background. Using these

[20] CRAWDAD – A Community Resource for Archiving Wireless Data at Darthmouth, http://crawdad.org/, accessed on November 2012.

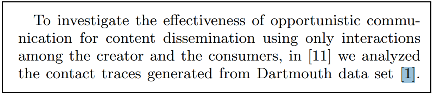


In this section, the real taxi trace data within 30 days in San Francisco from [17] is used. We used the IEEE 802.11p

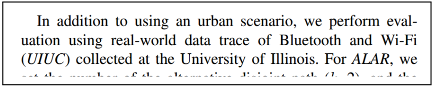
[17] http://crawdad.cs.dartmouth.edu/data.php.



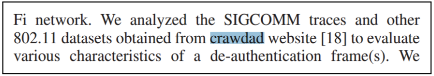
To evaluate the performance of our algorithm, we exploit a data set of sensor mote encounter records and corresponding social network data of a group of participants at University of St Andrews by the CRAWDAD team [6]. In the first data set,



To investigate the effectiveness of opportunistic communication for content dissemination using only interactions among the creator and the consumers, in [11] we analyzed the contact traces generated from Dartmouth data set [1].



In addition to using an urban scenario, we perform evaluation using real-world data trace of Bluetooth and Wi-Fi (*UIUC*) collected at the University of Illinois. For *ALAR*, we



Fi network. We analyzed the SIGCOMM traces and other 802.11 datasets obtained from crawdad website [18] to evaluate various characteristics of a de-authentication frame(s). We

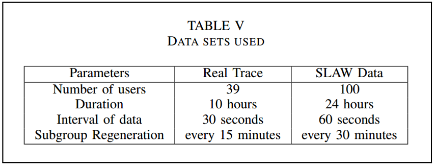


TABLE V
DATA SETS USED

| Parameters | Real Trace | SLAW Data |
|---|---|---|
| Number of users | 39 | 100 |
| Duration | 10 hours | 24 hours |
| Interval of data | 30 seconds | 60 seconds |
| Subgroup Regeneration | every 15 minutes | every 30 minutes |

# 90% isn't bad

**115** papers that use data but we don't know which data or how to find them…

# 90% isn't bad

**115** papers that use data but we don't know which data or how to find them…

- **36** papers cite the original papers that created the data
  - good, but papers often published before data are released and don't foreshadow location

# 90% isn’t bad

**115** papers that use data but we don’t know which data or how to find them…

- **36** papers cite the original papers that created the data
  - good, but papers often published before data are released and don’t foreshadow location
- **45** papers describe the dataset rather than use our identifiers
  - good, but makes it hard to track usage

# 90% isn’t bad

**115** papers that use data but we don’t know which data or how to find them…

- **36** papers cite the original papers that created the data
  - good, but papers often published before data are released and don’t foreshadow location
- **45** papers describe the dataset rather than use our identifiers
  - good, but makes it hard to track usage
- **72** cited CRAWDAD website without specifying dataset
  - 23 cited the website and original papers (so not space issue)

# 90% isn’t bad

**115** papers that use data but we don’t know which data or how to find them...

- **36** papers cite the original papers that created the data
  - good, but papers often published before data are released and don’t foreshadow location
- **45** papers describe the dataset rather than use our identifiers
  - good, but makes it hard to track usage
- **72** cited CRAWDAD website without specifying dataset
  - 23 cited the website and original papers (so not space issue)
- **21** papers provided no means to find the used data at all
  - 1 paper provided a non-existent URL

# 90% isn't bad

**115** papers that use data but we don't know which data or how to find them…

- **36** papers cite the original papers that created the data
  - good, but papers often published before data are released and don't foreshadow location
- **45** papers describe the dataset rather than use our identifiers
  - good, but makes it hard to track usage
- **72** cited CRAWDAD website without specifying dataset
  - 23 cited the website and original papers (so not space issue)
- **21** papers provided no means to find the used data at all
  - 1 paper provided a non-existent URL
- **6** papers cited me (yay h-index!) or Dartmouth as authors of data when they were not our data
  - Does the subject-specific database hinder rather than help?

# 90% isn't bad

**115** papers that use data but we don't know which data or how to find them...

- **36** papers cite the original papers that created the data
  - good, but papers often published before data are released and don't foreshadow location
- **45** papers describe the dataset rather than use our identifiers
  - good, but makes it hard to track usage
- **72** cited CRAWDAD website without specifying dataset
  - 23 cited the website and original papers (so not space issue)
- **21** papers provided no means to find the used data at all
  - 1 paper provided a non-existent URL
- **6** papers cited me (yay h-index!) or Dartmouth as authors of data when they were not our data
  - Does the subject-specific database hinder rather than help?
- **31** papers were so vague that I could not work out which datasets were used!
  - 3 were so vague that I couldn't work out if they used any data at all

# And another problem…

- Redirects…
- Old version of website used PHP
  - http://crawdad.org/dartmouth/campus/ would redirect to http://crawdad.cs.dartmouth.edu/meta.php?name=dartmouth/campus
- 74 papers used these `meta.php` URLs
- New version of website has to redirect these `meta.php` URLs back to original persistent URL
- DOIs would help
  - (but only if people cite the DOIs and not the resolved URL)

# 90% isn't great

- This sample is only the papers that mention CRAWDAD or that we were told about
- What about all the papers that don't even do this?
- ≈6,500 users, but only ≈1,200 papers?
- Are we better than other fields?
  - other people have looked at data contribution rather than citation, and rates are poor unless pressure is applied (e.g., can't publish until data are deposited) [5]
  - "evaluation research" is highlighted as a future topic of research [6]

[5] B. D. McCullough, K. A. McGeary, and T. D. Harrison. Do economics journal archives promote replicable research? *Canadian Journal of Economics/Revue canadienne d'économique*, 41(4):1406–1420, 30 Sept. 2008. doi:10.1111/j.1540-5982.2008.00509.x

[6] 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

# 90% isn't great

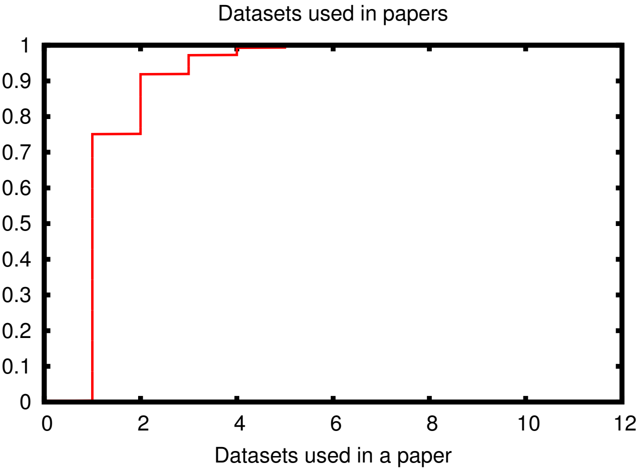
- This sample is only the papers that mention CRAWDAD or that we were told about
- What about all the papers that don't even do this?
- ≈6,500 users, but only ≈1,200 papers?
- Are we better than other fields?
  - other people have looked at data contribution rather than citation, and rates are poor unless pressure is applied (e.g., can't publish until data are deposited) [5]
  - "evaluation research" is highlighted as a future topic of research [6]
  - help?

[5] B. D. McCullough, K. A. McGeary, and T. D. Harrison. Do economics journal archives promote replicable research? *Canadian Journal of Economics/Revue canadienne d'économique*, 41(4):1406–1420, 30 Sept. 2008. doi:10.1111/j.1540-5982.2008.00509.x

[6] 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

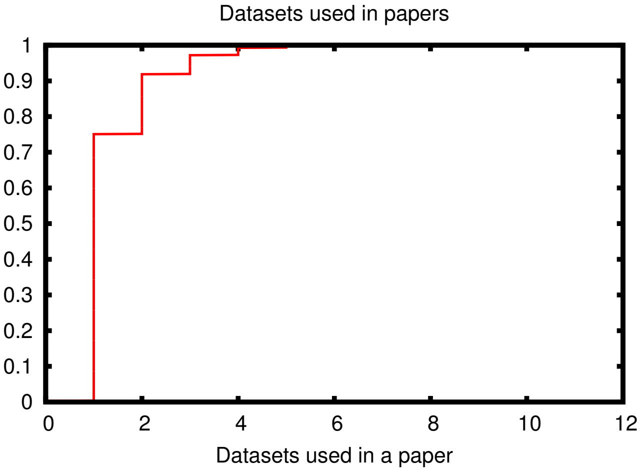
# What datasets do people use?

35% use more than one dataset



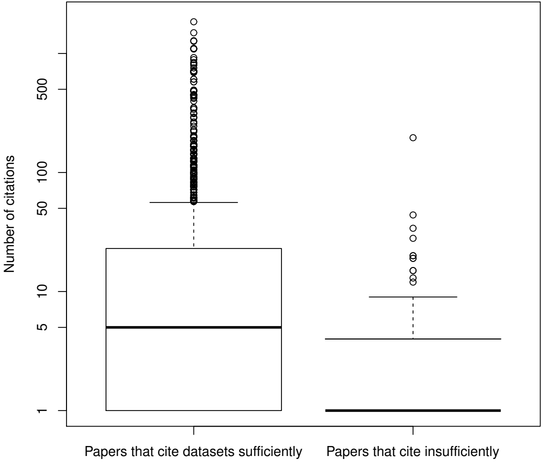
# What datasets do people use?

35% use more than one dataset



- Does the subject-specific database help rather than hinder?

# Are better-citing papers cited better?



# What can we do?

- Education
  - authors/reviewers/publishers
- Research
  - why don't people cite?
- DOIs (surprisingly complicated)
  - DataCite not really cost-effective
  - looking at EZID through Dartmouth

# Linking papers and datasets

- Many cite papers rather than datasets
- Papers don't link to datasets
  - authors might not even have thought of sharing data at point of writing/publishing paper
- Why are papers immutable?
- Updating links between papers and datasets would also be useful (cf. our `meta.php` links)

# What next?

- Find more missing papers
- Understand motivations for data citation
- Wrapping this all up into reproducible papers: i.e., not just datasets!

# Thank you

CRAWDAD

Community Resource for Archiving Wireless Data at Dartmouth

http://crawdad.org

| | | |
|---|---|---|
| 🏠 | tnhh.org | crawdad.org |
| ✉ | tnhh@st-andrews.ac.uk | crawdad@crawdad.org |
| 🐦 | @tnhh | @CRAWDADdata |