

# Reliable, reproducible and responsible data collection from online social networks

---

Tristan Henderson

School of Computer Science  
University of St Andrews

<http://tristan.host.cs.st-andrews.ac.uk/>  
[tnhh@st-andrews.ac.uk](mailto:tnhh@st-andrews.ac.uk)



University of  
St Andrews

---

FOUNDED  
1413

---

# Who am I?

- Data collector
- Data archiver
- Data analyser
  
- for various things:
  - networked games
  - wireless networks
  - pervasive computing
  - opportunistic networks
  - online social networks



# Who am I?

- Data collector
- Data archiver
- Data analyser
  
- for various things:
  - networked games
  - wireless networks
  - pervasive computing
  - opportunistic networks
  - **online social networks**



# Who am I?

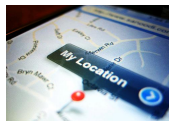
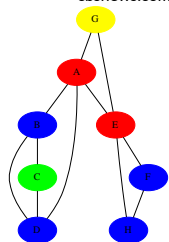
- Data collector
- Data archiver
- Data analyser
  
- for various things:
  - networked games
  - wireless networks
  - pervasive computing
  - opportunistic networks
  - **online social networks**
  
- *NOT* a statistician!



# Online social network research



cbsnews.com

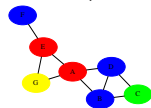


commnexus.org

- Online social network (OSNs) are an important part of today's Internet
  - hundreds of millions of users
  - and correspondingly large valuations (and profits?)
- OSNs have become an important source of “big” data and an avenue for research in many disciplines
  - healthcare
  - urban planning
  - epidemiology
  - politics
  - location-based services
  - mobile networks



help-desk.org



# How does this research study sound?

- *Goal:* collect social graph data
- Ask users for informed consent
- Ask users before they give any data to researchers
- Remove any identifiable data (user names, content, etc)



# How about this study?

- *Goal*: measure students' privacy preferences
- Do not ask users for informed consent
- Pay students' friends to use their credentials to collect data from students' accounts
- Remove some identifiable data (name, institution) but not others (age, gender, content)



# How about this study?

- *Goal:* understand interactions in mobile social applications
- Create innocuous mobile application (e.g., “Really Angry Birds”) that surreptitiously records all mobile activities and uploads to server
- Distribute application on ‘app store’ without any informed consent





# How about this study?

- *Goal*: understand disagreements on social network sites
- Create application to encourage “dislikes” of “enemies”
- Complain publicly when experiment does not lead to the desired cyber-bullying



# How about this study?

- *Goal:* understand social network sharing behaviour
- Ask users for informed consent
- Collect data from both users and friends of users
- Do not ask friends for informed consent (as they are not “participants” in the experiment)



# How about this study?

- *Goal:* understand spread of emotions through social networks
- Present different information to different OSN users
- Do not ask users for consent



# Ethics and social network research

- Ethics is a charged term...
- Let's talk about *responsible* research instead:
  - “Responsible Research and Innovation is a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view on the (ethical) acceptability, sustainability and societal desirability of the innovation process” [1]
- Lots and lots of key actors
  - Who owns data?
- Lots of issues
  - Are “public” data fair game for research?
  - Are OSN users human subjects?
  - Does informed consent make sense?
  - Do we need IRB/ethics approval?

---

[1] European Commission Directorate-General for Research and Innovation. *Towards responsible research and innovation in the information and communication technologies and security technologies fields*. EUR-OP, 2011.  
doi:10.2139/ssrn.2436399

# Key actors in OSN research

- Researchers
- OSN user (participants)
- Friends of users
- Other users
- Other researchers
- OSN operator
- Institutions



# Key actors in OSN research

- Researchers
- OSN user (participants)
- Friends of users
- Other users
- Other researchers
- OSN operator
- Institutions
- Anyone else?



“Just because data is accessible doesn’t mean that using it is ethical.”<sup>[2]</sup>

---

[2] D. Boyd. Privacy and publicity in the context of big data. Keynote at WWW '10: the 19th International Conference on the World Wide Web, Apr. 2010. Online at <http://www.danah.org/papers/talks/2010/WWW2010.html>

“conducting a social network study without truly informed consent is deceptive and wrong.”<sup>[3]</sup>

---

<sup>[3]</sup>S. P. Borgatti and J.-L. Molina. Toward ethical guidelines for network research in organizations. *Social Networks*, 27(2):107–117, May 2005. doi:10.1016/j.socnet.2005.01.004



## Alternatively...

- Does OSN research require ethics approval?[4]
- Is ethics approval relevant?[5]

---

[4] L. Solberg. Data mining on Facebook: A free space for researchers or an IRB nightmare? *University of Illinois Journal of Law, Technology & Policy*, 2010(2), 2010. Online at <http://www.jltp.uiuc.edu/works/Solberg.htm>

[5] E. Buchanan, J. Aycok, S. Dexter, D. Dittrich, and E. Hvizdak. Computer science security research and human subjects: Emerging considerations for research ethics boards. *Journal of Empirical Research on Human Research Ethics*, 6(2):71–83, June 2011. doi:10.1525/jer.2011.6.2.71



# Problems with using OSN data #1: reliability

- OSNs are an attractive and accessible source of “big data”
- But “big” data might be *inappropriate* data
  - Publicly-available data are *public*
  - But we might need private data
- Data might be *collected* inappropriately
  - Ethics?
  - DPA?
  - Science?
- Relevant key actors: OSN users; friends of users; other users; researchers



# Collecting private OSN data

- Our interest:
  - understanding privacy conceptions in OSNs
  - understanding methodologies for measuring users
- So can't merely use publicly-available data
  - and don't want to since we are interested in methodologies



<http://www.pvnets.org/>

# Experience Sampling Method

- Commonly-used method in psychology for diary studies<sup>[6]</sup>
- Ask participants to stop during their everyday activities and record their experiences
  - signal-contingent or event-contingent times
- Participants record *in situ* – less recall error
- Short, but numerous and repetitive, data points

---

[6] R. Larson and M. Csikszentmihalyi. The experience sampling method. In H. T. Reis, editor, *Naturalistic Approaches to Studying Social Interaction*, volume 15 of *New Directions for Methodology of Social and Behavioral Science*, pages 41–56. Jossey-Bass, San Francisco, CA, USA, 1983



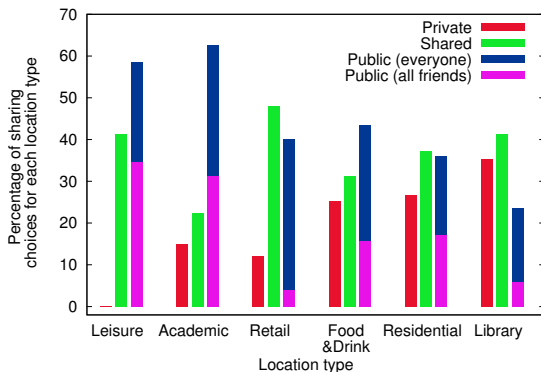
# ESM and mobile Facebook



- Give students (in St Andrews and London) smartphones (Nokia N95) with Wi-Fi/GPS/Bluetooth/accelerometer/...
- Track them (after obtaining informed consent)
- Periodically ask them questions about their current activities and social network sharing behaviour
- Let them share information on Facebook (or not?)



# Where do people share?



- More willing to share in Leisure and Academic areas, less willing in Library or Residential
  - “I don’t want friends to join”
  - “I don’t want friends to know I am staying home”
  - “I share my location when it is interesting”



## So what's wrong with crawling? Or surveys?

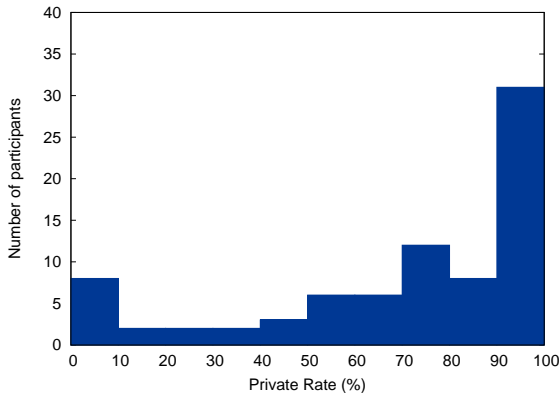
- Crawling: miss the unshared locations
- Surveys: self-reported data are unreliable

<i>Self-reported group</i>	<i>Responses to location-sharing requests</i>	<i>Locations that were shared</i>
Never share location on Facebook	431	77.5%
Share location on Facebook	95	78.9%



# What is the effect of poor data?

*Private rate*: proportion of sharing activities that were private (i.e., not shared with anyone)

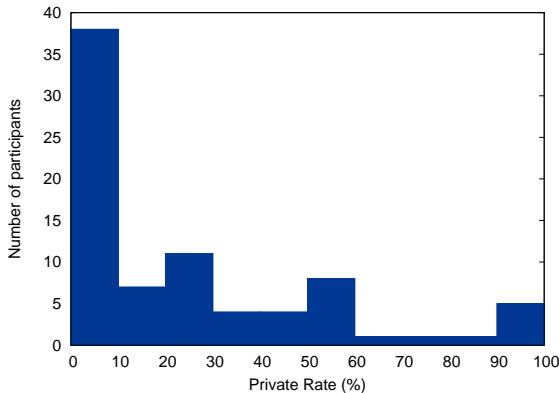


Facebook



# What is the effect of poor data?

*Private rate*: proportion of sharing activities that were private (i.e., not shared with anyone)

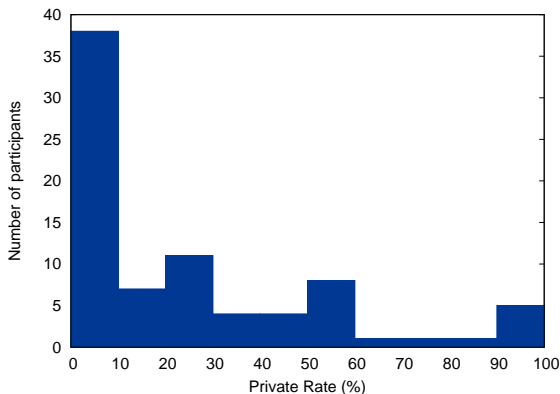


ESM



# What is the effect of poor data?

*Private rate*: proportion of sharing activities that were private (i.e., not shared with anyone)



ESM

ESM lets you distinguish between shared, public and private



# What other data do we miss?

Can ask users about attitudes<sup>[7]</sup>

<i>Group</i>	<i>Responses to location-sharing requests</i>	<i>Locations that were shared</i>
Fundamentalist	109	76.1%
Pragmatic	168	66.7%
Unconcerned	276	64.5%

---

[7] Louis Harris and A. F. Westin. E-commerce and privacy: What net users want. Sponsored by Price Waterhouse and Privacy & American Business, June 1998. Online at <http://www.privacyexchange.org/survey/surveys/ecommsum.html>



# Pros and cons of ESM

- ✓ Richer data
- ✓ Otherwise hard-to-get data
- ✓ Able to more easily obtain informed consent



# Pros and cons of ESM

- ✓ Richer data
- ✓ Otherwise hard-to-get data
- ✓ Able to more easily obtain informed consent
  
- ✗ Sparser data (in terms of number of users)
- ✗ Expense (time, money)



## Problems with using OSN data #2: Science!

---

[8] D. Patterson. How to have a bad career in research/academia, Nov. 2001. Online at <http://www.cs.berkeley.edu/~pattsrn/talks/BadCareer.pdf>



# Problems with using OSN data #2: Science!

## Obsolete Scientific Method

1. Hypothesis
2. Experiments
3. Change 1 parameter
4. Prove/disprove hypothesis
5. Document for others to reproduce

---

[8] D. Patterson. How to have a bad career in research/academia, Nov. 2001. Online at

<http://www.cs.berkeley.edu/~pattsrn/talks/BadCareer.pdf>



## Problems with using OSN data #2: Science!

### Obsolete Scientific Method

1. Hypothesis
2. Experiments
3. Change 1 parameter
4. Prove/disprove hypothesis
5. Document for others to reproduce

### Computer Scientific Method<sup>[8]</sup>

1. Hunch
2. 1 experiment and change all parameters
3. Discard if it doesn't support hunch
4. Why waste time? We know this

---

[8]

D. Patterson. How to have a bad career in research/academia, Nov. 2001. Online at <http://www.cs.berkeley.edu/~pattsrn/talks/BadCareer.pdf>





# What is reproducibility?

- Drummond<sup>[9]</sup> distinguishes between
  - replicability: exact repetition of an experiment as presented
  - reproducibility: building on an experiment and furthering science
  - both require suitable documentation

---

[9] C. Drummond. Replicability is not reproducibility: Nor is it good science. In *Proc. of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, Montreal, QC, Canada, 2009. Online at <http://cogprints.org/7691/>

[10] P. A. Thompson and A. Burnett. Reproducible research. *CORE Issues in Professional and Research Ethics*, 1(6), 2012. Online at <http://nationalethicscenter.org/content/article/175>



# What is reproducibility?

- Drummond<sup>[9]</sup> distinguishes between
  - replicability: exact repetition of an experiment as presented
  - reproducibility: building on an experiment and furthering science
  - both require suitable documentation
- Three components:<sup>[10]</sup>
  1. *code*: source code, tools, workflow
  2. *method*: scripts for analysis
  3. *data*: research artefacts such as papers and raw data

---

[9] C. Drummond. Replicability is not reproducibility: Nor is it good science. In *Proc. of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, Montreal, QC, Canada, 2009. Online at <http://cogprints.org/7691/>

[10] P. A. Thompson and A. Burnett. Reproducible research. *CORE Issues in Professional and Research Ethics*, 1(6), 2012. Online at <http://nationaalethicscenter.org/content/article/175>



# What is reproducibility?

- Drummond<sup>[9]</sup> distinguishes between
  - replicability: exact repetition of an experiment as presented
  - reproducibility: building on an experiment and furthering science
  - both require suitable documentation
- Three components:<sup>[10]</sup>
  1. *code*: source code, tools, workflow
  2. *method*: scripts for analysis
  3. *data*: research artefacts such as papers and raw data
- Let's look at these in the obvious order

---

[9] C. Drummond. Replicability is not reproducibility: Nor is it good science. In *Proc. of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, Montreal, QC, Canada, 2009. Online at <http://cogprints.org/7691/>

[10] P. A. Thompson and A. Burnett. Reproducible research. *CORE Issues in Professional and Research Ethics*, 1(6), 2012. Online at <http://nationaalethicscenter.org/content/article/175>



# What is reproducibility?

- Drummond<sup>[9]</sup> distinguishes between
  - replicability: exact repetition of an experiment as presented
  - reproducibility: building on an experiment and furthering science
  - both require suitable documentation
- Three components:<sup>[10]</sup>
  1. *code*: source code, tools, workflow
  2. *method*: scripts for analysis
  3. *data*: research artefacts such as papers and raw data
- Let's look at these in the obvious order
  - method, data, code

---

[9] C. Drummond. Replicability is not reproducibility: Nor is it good science. In *Proc. of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, Montreal, QC, Canada, 2009. Online at <http://cogprints.org/7691/>

[10] P. A. Thompson and A. Burnett. Reproducible research. *CORE Issues in Professional and Research Ethics*, 1(6), 2012. Online at <http://nationaalethicscenter.org/content/article/175>



# Method

- Searched venues for papers that collected data from OSNs
- Read each paper and determined how to reproduce

## Venues

ASONAM, CCS, Computers in Human Behavior, CHI, COSN, CSCW, EuroSys SNS, HotSocial, ICWSM, J. Computer-Mediated Communication, Nature, NDSS, Oakland, Science, Social Networks, SOUPS, Ubicomp, WebSci, WOSN, WPES

## Search term

```
abstract CONTAINS (facebook OR twitter OR sns OR  
osn OR foursquare OR linkedin OR friendster OR  
weibo OR flickr OR livejournal OR myspace OR  
“online social network” OR “social network site” OR  
“social networking site”) AND publication-date  
BETWEEN (2011-01-01, 2013-12-31)
```



# Reproducible OSN measurement

## Method:

- ✓ Source OSN
- ✓ Sampling strategy
- ✓ Length of study
- ✓ Number of participants/users
- ✓ Data processing
- ✓ Consent
- ✓ Participant briefing
- ✓ Ethics

## Data:

- ✓ Data shared

## Code:

- ✓ Code shared



## Some numbers

- 811 papers matched search string
- 487 papers used OSN data

How many papers matched all of our reproducibility criteria?



## Some numbers

- 811 papers matched search string
- 487 papers used OSN data

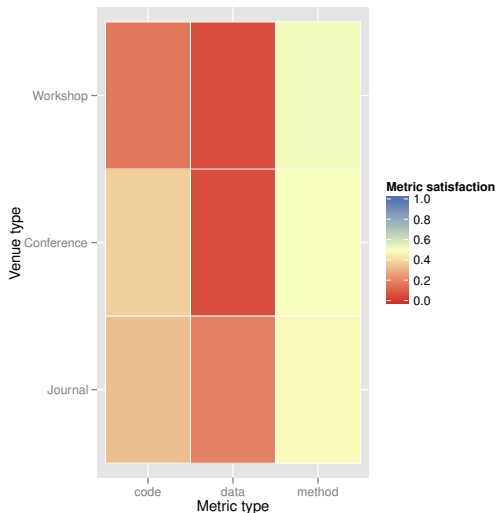
How many papers matched all of our reproducibility criteria?

1

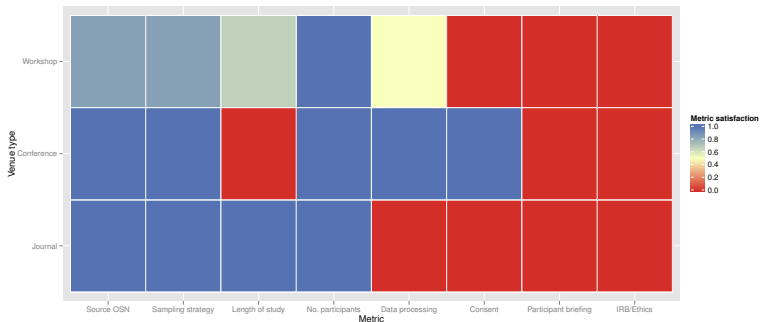


# Does type of venue make a difference?

Not much relationship between venue or length of paper



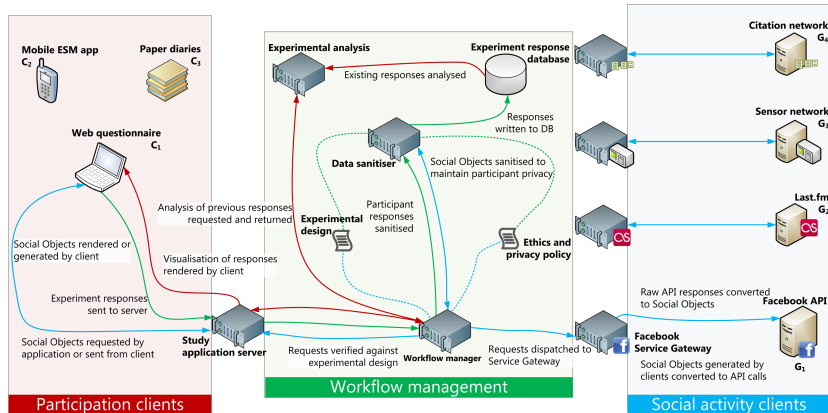
# Are particular bits of method better described?



- Most (but not all!) said which network was being studied
- Very few discussed ethics/consent/people
  - despite the aforementioned debate on this

# Our increment: PRISONER

## Privacy-Respecting Infrastructure for Social Online Network Experimental Research<sup>[11]</sup>



[11] L. Hutton and T. Henderson. An architecture for ethical and privacy-sensitive social network experiments. *ACM SIGMETRICS Performance Evaluation Review*, 40(4):90–95, Apr. 2013. doi:10.1145/2479942.2479954



# Architectural details

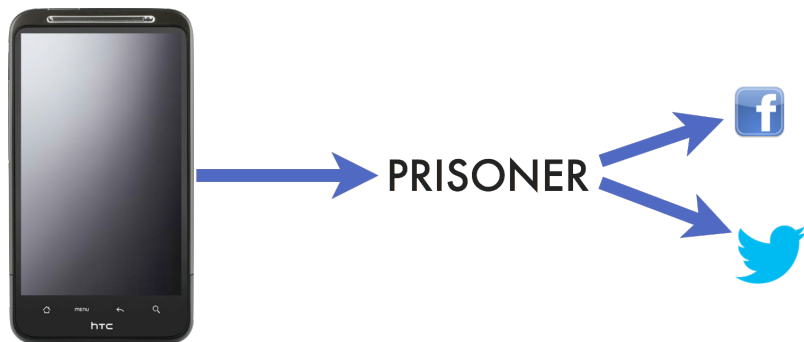
- Workflow management:
  - collect data according to policy
  - store data according to policy
  - sanitise data according to policy
  - share data according to policy
- Social activity clients:
  - abstraction for various OSNs
    - and other sources of network data: citation networks, sensor networks
  - use standard *Social Objects*<sup>[12]</sup>
- Participation clients:
  - abstraction for various research methods (mobile, web, paper, ESM,...)

---

[12] <http://activitystrea.ms/>

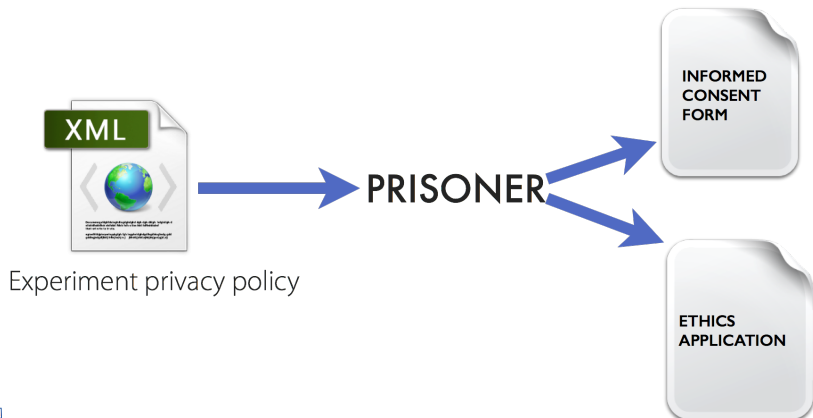
# PRISONER features [1]

- Abstract experiments from specific social networks
  - encourage reproducibility
  - easy to add support for other sites through plugins
  - Facebook, twitter, last.fm already implemented



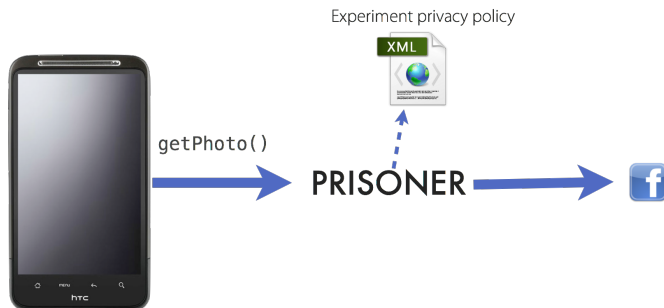
## PRISONER features [2]

- Encapsulate workflow
  - experimental designs and privacy policies can be shared for replication/further research
  - workflows can generate readable documentation, e.g., for (initial prototypes of) consent forms, ethics applications



## PRISONER features [3]

- Real-time validation of experiments
  - ensures experiments can only handle data permitted by privacy policy
  - dynamically sanitises data



# Does it work?

- Hmm...
- We can reproduce the *one* reproducible paper<sup>[13]</sup> from our literature study...
- We have used it for lots of *our* user studies...
  - always looking for volunteers
- What have we learned?
  - main difficulty with reproducibility was changes to Facebook API
  - how to capture all interactions with all relevant systems?
  - Docracy<sup>[14]</sup> tracks changes in Terms of Service; who tracks APIs (and how?)
  - code, method, data, other?

---

[13] J. King, A. Lampinen, and A. Smolen. Privacy: Is there an app for that? In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, Pittsburgh, Pennsylvania, 2011. doi:10.1145/2078827.2078843

[14] <https://www.docracy.com/tos/changes>





# Data sharing

- Data sharing poor in our OSN survey
- Some small scale data sharing efforts, e.g., ICWSM
- Data sharing is good for science<sup>[15]</sup>
  - Indeed it is now required by RCUK<sup>[16]</sup>
- Can we learn from other fields?

---

[15] 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

[16] <http://www.rcuk.ac.uk/research/DataPolicy/>





## 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!)
  - 7,424 users from 108 countries (as of April 2015)
  - 119 datasets and tools used in over 1,700 papers (that we know of)
- Some popular datasets:
  - Cambridge Bluetooth encounters: 381 papers
  - Dartmouth WLAN data: 285 papers
  - MIT Reality Mining: 161 papers
  - EPFL taxi cabs: 157 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, ...

# 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)
  - DOIs coming soon (surprisingly messy)
- 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<sup>[17]</sup>

---

[17] <http://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/IEEEExplore/...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 <sup>[18]</sup>

---

[18] T. Henderson and D. Kotz. Data citation practices in the CRAWDAD wireless network data archive. *D-Lib Magazine*, 21(1/2), Jan. 2015. doi:10.1045/january2015-henderson





# Tracking usage in practice

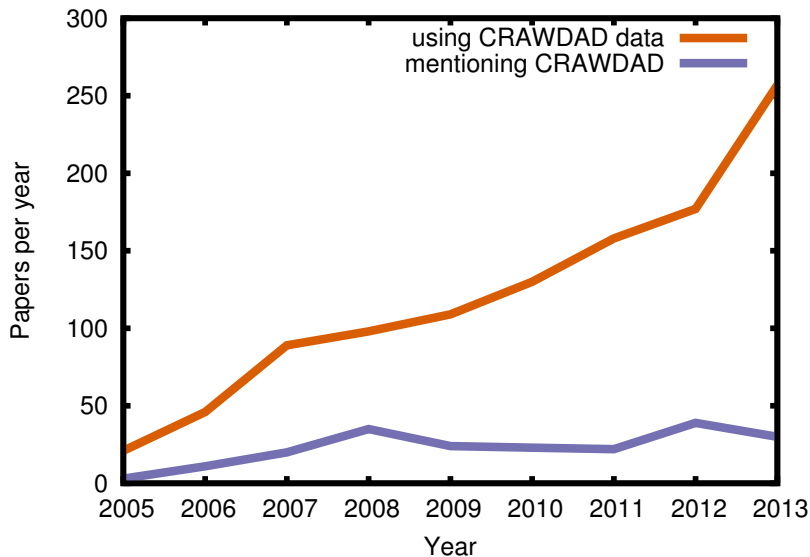
1. Google Scholar/ScienceDirect/IEEEExplore/...searches for “CRAWDAD”
  2. filter out all the references to shellfish, CRAWDAD text analysis tool, CRAWDAD neurophysiology tool
  3. check paper manually to determine which (if any) datasets were used <sup>[18]</sup>
- There *must* be a better way!

---

[18] T. Henderson and D. Kotz. Data citation practices in the CRAWDAD wireless network data archive. *D-Lib Magazine*, 21(1/2), Jan. 2015. doi:10.1045/january2015-henderson



## CRAWDAD usage: healthy



## CRAWDAD usage: healthy?

- $\approx$ 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

---

[19] [force11.org/datacitation](https://force11.org/datacitation)



## CRAWDAD usage: healthy?

- $\approx 3,800$  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

---

[19] [force11.org/datacitation](http://force11.org/datacitation)



# 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
  - after the Force 11 Data Citation Principles<sup>[19]</sup>:
  - **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?

---

[19] [force11.org/datacitation](https://force11.org/datacitation)

# 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
  - after the Force 11 Data Citation Principles<sup>[19]</sup>:
  - **credit and attribution**: do the data citations appropriately credit the data providers?
    - **i.e., used our BibTeX**
  - **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?

---

[19] [force11.org/datacitation](http://force11.org/datacitation)



# 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 Dartmouth, <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

use the number of the observation districts with (1, 2) and the Fi network. We analyzed the SIGCOMM traces and other 802.11 datasets obtained from [crawdad](http://crawdad.org/) 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 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 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 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 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 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



# 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?
- $\approx 6,500$  users, but only  $\approx 1,200$  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) <sup>[20]</sup>
  - “evaluation research” is highlighted as a future topic of research <sup>[21]</sup>

---

[20] 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

[21] 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



# Reproduciblecomputable code

- My colleagues (Ian Gent *et al*) at [recomputation.org](http://recomputation.org)<sup>[22]</sup>
- “If we can compute your experiment now, anyone can recompute it 20 years from now”
- Using virtual machines to capture and enable the exact “recomputation” of an experiment that has been deposited in the repository

---

[22] I. P. Gent. The recomputation manifesto, 12 Apr. 2013. Online at <http://arxiv.org/abs/1304.3674>





# Reproducibility summer school

- “Summer School on Experimental Methodology in Computational Science Research”, Aug 2014
  - [blogs.cs.st-andrews.ac.uk/emcsr2014/](http://blogs.cs.st-andrews.ac.uk/emcsr2014/)
- Basic idea: get some students together, throw a bunch of reproducibility problems at them, and write a paper by the end of the week
- Speakers from MSR (Azure), Software Sustainability Institute, and a môtley crüe of academics



# Case studies

Students worked on four case studies:

1. ethics approval processes and reproducibility
  - can we create a specification for ethics approval that encodes sufficient details for others to reproduce an experiment involving human subjects?
2. parallel and distributed experiments
  - what are the problems in using multiple VMs to reproduce parallel computing experiments?
3. reproducibility in computational science outside of CS
  - how easy is it to recompute astrophysics and urban planning experiments?
4. can an author help others reproduce their own paper?
  - author might think the paper is reproducible, but do other people agree?



# Let's write a paper in a week



By Ohiopetwatch (Own work)  
[CC-BY-SA-3.0], via Wikimedia  
Commons

- Paper in (re)submission and on arXiv [23]
- Paper is reproducible!
  - Code on github<sup>[24]</sup>
  - Uses Sweave and R so that all plots in the paper can be regenerated from the source data
  - VMs containing all the code and data used to generate the paper on re computation.org<sup>[25]</sup> and Microsoft VM Depot<sup>[26]</sup>

---

[23] S. Arabas, M. R. Bareford, I. P. Gent, B. M. Gorman, M. Hajjarabderkani, T. Henderson, L. Hutton, A. Konovalov, L. Kotthoff, C. McCreesh, R. R. Paul, K. E. J. Petrie, A. Razaq, and D. Reijsbergen. Case studies and challenges in reproducibility in the computational sciences, 11 Sept. 2014. Online at <http://arxiv.org/abs/1408.2123>

[24] [github.com/larskotthoff/recomputation-ss-paper/](https://github.com/larskotthoff/recomputation-ss-paper/)

[25] [recomputation.org/emcsr2014/](http://recomputation.org/emcsr2014/)

[26] [vmdepot.msopentech.com/Vhd/Show?vhdId=44582](http://vmdepot.msopentech.com/Vhd/Show?vhdId=44582)



## Problems with using OSN data #3: consent

- PRISONER lets us document how we collect data
- But how can we collect data responsibly?
- Is informed consent *meaningful* consent?
- Relevant key actors: OSN users; friends of users; other users; researchers



# Informed consent

- The gold standard post-Nuremberg
- How to know if a participant is informed?
- What if information is too complex <sup>[27]</sup>
- “Secured” consent: checkbox/EULA at start of experiment <sup>[28]</sup>
- “Sustained” consent: ask over and over in a sustained process
- Goal: Can we achieve accuracy of sustained consent while approaching the burden of secured consent?

---

[27] E. Luger, S. Moran, and T. Rodden. Consent for all: revealing the hidden complexity of terms and conditions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2687–2696, Paris, France, 2013. doi:10.1145/2470654.2481371

[28] E. Luger. Consent reconsidered; reframing consent for ubiquitous computing systems. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 564–567, Pittsburgh, Pennsylvania, 2012. doi:10.1145/2370216.2370310



# Contextual integrity

- A commonly-used framework for detecting privacy violations <sup>[29]</sup>
- Look for violations in informational norms
- If norms are violated, then perhaps privacy is
- So can we detect norms in OSN usage?

---

[29] H. F. Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–157, Feb. 2004. Online at <http://ssrn.com/abstract=534622>



# Experiment

- Asked 81 participants about 100 pieces of information from their Facebook accounts and whether they would share them with researchers <sup>[30]</sup>
  - these were used to develop *norms*
- Asked 154 different participants about their Facebook information, and tried to see how “norm-compliant” each participant was
- Participants were divided into three conditions: secured, sustained, and “contextual integrity” consent (using norms to predict what information they would be willing to share with researchers)
- Then asked them to check our predictions <sup>[31]</sup>

---

[30] S. McNeilly, L. Hutton, and T. Henderson. Understanding ethical concerns in social media privacy studies. In *Proceedings of the ACM CSCW Workshop on Measuring Networked Social Privacy: Qualitative & Quantitative Approaches*, San Antonio, TX, USA, Feb. 2013. Online at <http://www.cs.st-andrews.ac.uk/~tristan/pubs/mnsp2013.pdf>

[31] L. Hutton and T. Henderson. “I didn’t sign up for this!”: Informed consent in social network research. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM)*, Oxford, UK, May 2015. Online at <http://tristan.host.cs.st-andrews.ac.uk/research/pubs/icwsm2015.pdf>



# Data acquisition


Facebook usage study [Leave the study](#)

Question 61

You like PIXIES.

Will you share that you like this with us?

YES NO

  
PIXIES  
Musician/band



# Data confirmation

Facebook usage study

Leave the study

This screen summarises all the data we will collect from your profile, based on what you have told us so far. Click any items you do not want us to use. When you're ready, confirm your selection to finish the experiment.

I'M DONE!

VISITED ON 16/05/14  
National Museums Scotland

POSTED ON 29/03/14

LIKES  
PIXIES

POSTED ON 17/06/13  
Well, I did note she's been getting increasingly nutty, so I shouldn't be surprised...

POSTED ON 14/01/14  
Dean gets it.

VISITED ON 12/02/12  
DCA

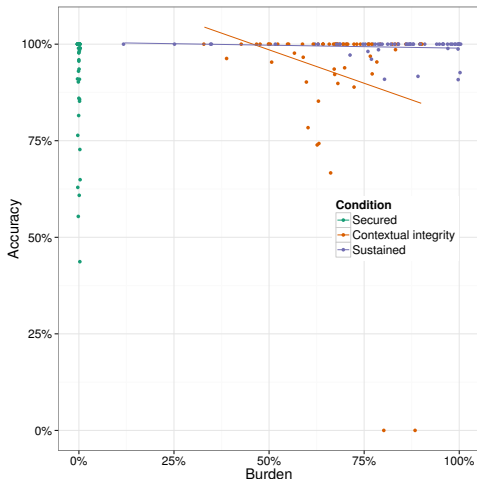
VISITED ON 10/09/13  
Office 13th

POSTED ON 12/09/14  
The most important television show ever is being repeated. Well caught, Tivo.



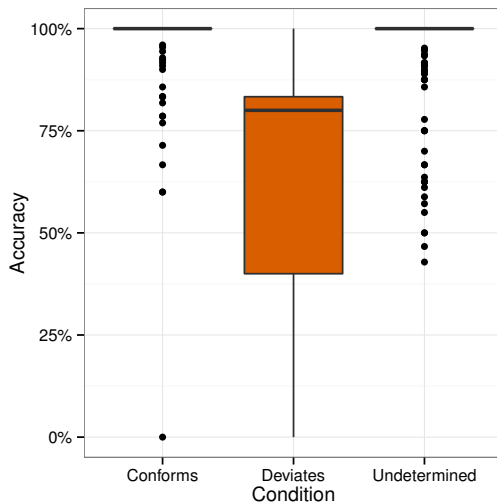
# Accuracy versus burden

Accuracy is most variable under secured consent; contextual integrity is slightly less accurate than sustained but has much lower burden



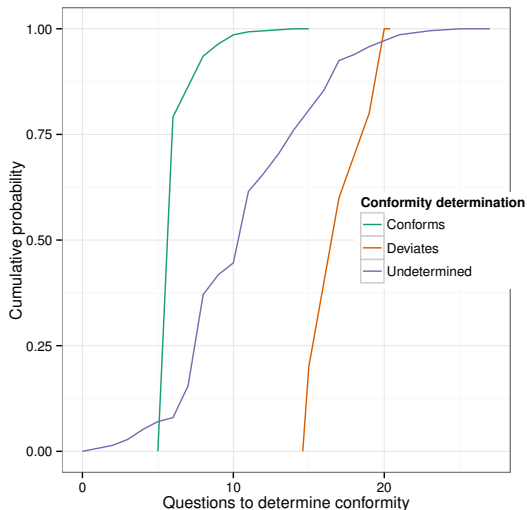
# Norm conformity

If a participant conforms with norms, then contextual integrity is useful



# How to determine norm conformity?

Around seven questions were sufficient



# Summary

- We need to be careful when using OSN data
- 1. Would be nice if your data were appropriate and reliable
- 2. Would be nice if your research was reproducible (method, data and code)
  - Sharing data introduces a whole new kettle of problems
- 3. Would be nice if your research was responsible (engage all actors, only collect what is needed)
  - Think about consent and how people might want you to use their “public” data
- We are beginning to address some of these problems, but have a long way to go!



# Thanks & contact

- my (current and ex) students: Luke Hutton, Sam McNeilly, Iain Parris
- CRAWDAD: Dave Kotz, Chris McDonald, Anna Shubina, Jihwang Yeo
- PVNets: Fehmi Ben Abdesslem, Angela Sasse, Sacha Brostoff
- summer school co-organisers: Ian Gent, Lars Kotthoff, Lakshita de Silva, and all the participants!
- funders and other helpful partners: EPSRC, NSF, Microsoft Azure, Software Sustainability Institute

🏠	<a href="http://tnhh.org">tnhh.org</a>	<a href="http://crawdad.org">crawdad.org</a>
✉	<a href="mailto:tnhh@st-andrews.ac.uk">tnhh@st-andrews.ac.uk</a>	<a href="mailto:crawdad@crawdad.org">crawdad@crawdad.org</a>
🐦	<a href="https://twitter.com/tnhh">@tnhh</a>	<a href="https://twitter.com/CRAWDADdata">@CRAWDADdata</a>

