Sharing Network Data: Bright Gray Days Ahead

John Heidemann
University of Southern California / Information Sciences Institute
Passive and Active Measurements Conference keynote
10 March 2014

Copyright © 2014 by John Heidemann
Release notice: UIUC-ACDC/International

Data makes the Internet…

IPv4: 648-770M active addresses
---ASTF Project, http://www.isi.edu/amr/address/ current_address_counts_2014-03-14

Data makes the Internet…

world wide web: more than 1 trillion URLs
---Alessandro e Hanjalic, Google Blog, July 2008

world wide web: about 16 billion indexed web pages

global IP traffic: 55EB/month
---Cisco Visual Networking Index, May 2013

5700 tweets/s mean and 144Ks peak

IPv4: 648-770M active addresses
---ASTF Project, http://www.isi.edu/amr/address/ current_address_counts_2014-03-14

Data makes Internet Research

the research community values it:
ACM IMC “best paper award for the top paper that makes its data sets publicly available…”

the scientific method:
measure, hypothesize, predict, experiment, repeats with data

journals expect it:
Public Library of Science, Nature, Science: all require an explicit statement about data availability

the U.S. National Science Foundation demands it:
proposals must include “plans for data management and sharing”: types of data, standards, policies for privacy, re-distribution, archiving…

Collecting Internet Data…

• 20 years ago: ask your buddy down the hall
  – sure no problem, here it is
• 10 years ago: ask the network admin nicely
  – ok, but be careful with it
• 5 years ago: ask they lawyers upstairs
  – “no”
  – (more likely: delay, meeting, delay… “no”)
...Good Reasons It’s Harder

- internet data is real: your bank account, contacts, photos
- misuse of real data has real consequences
  - Dec. 2013 Target data breach: $613M expenses
  - 2005 Sears, Coca-Cola, Pepsi ad fraud: $400M
  - Aug. 2006 AOL Search Data Release: CTO resigns, 2 fired
    (emails regarding need for sharing of search data, “illegal”, and “illegal” claims, 25 Aug. 2006)
  - 2009 Netflix Contest: second contest canceled due to privacy
    (Netflix Contest After Contestants Are Rewarded Most Privacy”, LA Times, 15 Nov. 2009)
- and use of real data generates real value
  - $37B U.S. digital ad spending in 2012
    (Internet Advertising Bureau, U.S. Internet Advertising Revenue Report 2012 Full-Year Results)
  - consider internet analytics, telecom planning, …

Result: Black and White Evaluation

- data availability risks being just “yes” or “no”
- conference committees:
  - is the data public?
  - yes… but requires an agreement
  - NO… not truly public
- researcher:
  - can I download the data now?
  - yes… but, must document use and institution
  - NO… too much bother
- potential data provider asking her institution
  - can I release this data?
  - yes… but it has IP addresses in it
  - NO… IP’s are private, sharing creates risk

Sharing Beyond Black and White

too limiting for sharing to be only black and white:
  - private or public
  - protected or sharable
- the world is complex
- the world is gray
- the world must be gray

Bright Gray Sharing

- need for sharing
  - gray: why don’t people share?
- understanding the shade
- brighter alternatives
- towards a brighter gray future

Background and Biases

- studying the Internet since late-1990s
  - protocols, modeling and simulation
  - (and some time in sensornets)
- network data collection and analysis since 2004
  - DHS PREDICT program
  - some related support from NSF and DARPA
- goals
  - new measurement methods
  - provide data to others
  - with strong legal and ethical basis

A Case for Sharing?

- why share?
- direct costs of sharing
- indirect costs: risk
Sharing Helps Others

- sharing helps others… but what about you?
- altruism is nice, but it has opportunity cost
  - spend your time doing something else
- are there any direct benefits?
  - goodwill
  - academic citations
- others find bugs and improve results
  - well known for software $1.48$ “value” of Linux kernel
  - what value are we missing in undershared network data?

Sharing Has Direct Costs

- (beyond opportunity costs)
- sharing requires effort
  - documentation… hard to quantify, but apparent
  - distribution when packaging something (listen for the complaints)
- ongoing effort — answering questions

Anecdote #1: AOL

- in 2006 AOL released a search engine dataset
  - 20M queries, 650k users, 3 months
- why? support research
- anonymized: users identified by unique IDs
- but…

Anecdote #0: IPv4 Censuses

<table>
<thead>
<tr>
<th>methodology (at USC)</th>
<th>analysis (at USC)</th>
<th>external work</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980s</td>
<td></td>
<td>survey, partial topology studies</td>
</tr>
<tr>
<td>2000</td>
<td>statistical census work (Gonzalez, Prade, Botan)</td>
<td>long time to develop solid methodology; earlier sharing can perhaps help</td>
</tr>
<tr>
<td>2005</td>
<td>revised methodology and forum</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>2nd revision</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>finally published (Hilfinger et al., IMC 2008)</td>
<td>IPv4 Usage (Juniper Networks) and IPv4 Usage (ICANN)</td>
</tr>
<tr>
<td>2009-2010</td>
<td>continued debugging of census with</td>
<td>IPv6 (Juniper Networks, ICANN)</td>
</tr>
<tr>
<td>2013</td>
<td>2nd generation measurement: usage detection (Hilfinger et al., ICANN 2013)</td>
<td>Cerna Honorat (2013) and ZMap (Hilfinger et al., Computer Security. 2012)</td>
</tr>
</tbody>
</table>

Sharing Has Risks

- gives others an advantage
  - (I hope we agree research is not a zero-sum game)
- could be used against you
  - reveals your practices
    - business (or academic) “trade secrets”?
  - reveals about the subject
    - privacy expectations (something often in flux)

AOL Deanonymization

- 20M records (with specific search terms)
- + some hours reporter time
- => 1 person publicly (others outside print)
  - Thaddeus Arnold, No. 4417749
  - “A Face is Exposed. . .” Barbara and Zeller | NY Times 9 Aug. 2006
- => big trouble for AOL
  - horrible PR, CTO resigned, two others fired
- take away: with enough records, something always gives
  - like password cracking: >50% of md5-ed passwords cracked in a few hours
Annecdot #2: Netflix Prize

- in 2006, Netflix held a contest with customer data
  - 100M ratings+dates for 18k movies from 480k users
- why?
  - improve their ranking system… profit!
  - $1M to winning team with 10% improvement
- anonymized: only ratings and dates, no PII
- huge interest in the data and prize
  - 3 years of work by 20k teams from 150 countries!
- but…

Netflick Deanonymization

- 480k records
- anything else on the Internet
- => identified 2 people
  - using Internet Movie Database ratings for “a few dozen”
  - Namjoo and Shmatikov, “Robust De-anonymization of Large Datasets”, May 2008
- => non-zero risk for Netflix
  - enough concern that they canceled a second contest
  - investigation by the U.S. FCC
- take away: with enough other data, something gives
  - like password cracking: external information (mother’s maiden name, birthday, home address) is huge

Annecdot #3: Devil Advocating Details

- in 2005, ICIS researchers released anonymized enterprise data
  - “A First Look at Modern Enterprise Traffic”, Pang, Allman, Bennett, Lee, Paxson, Tieney, IMC 2005
- why?
  - support research—the first enterprise dataset (yes!)
  - develop strong anonymization methods
- anonymized:
  - very thoughtful job
  - when in doubt, truncate it!
- but…

Enterprise Deanonymization

- 871k flows over 60 hours for 6k hosts
- patterns in the data itself
  - => assertions about hosts
  - “Playing Devil’s Advocate: Inferring Sensitive Information from Anonymized Network Traces”, C Wald et al., NDSS 2007
- => non-zero risk for the enterprise
  - claim “nearly all IPs were incorrect de-anonymized”
  - Issues and Etiquette Concerning Use of Shared Measurement Data”,
  - Allman and Paxson, IMC 2007 [Allman07a]
  - but the threat was enough—a chilling effect on release
- take away: with enough data itself, something may give
  - like password cracking: with time and effort, risk is real

Risks Everywhere

- many records
- with other data
- or with enough internal information
- => risk
- our adversary:
  - may only need to break a few
  - brings other data
  - may expend effort
  - may not follow expectations

Is the Future Grayer?

“Attacks always get better, they never get worse”
  - Schneier quoting “an NSA adage” in RFC-4270

- once public, data is forever
- challenge of future, better de-anonymization
Why Share Data?

why do this again?

- minimal benefit
- some direct costs
- uncertain risks
- that can only grow

Bright Gray Sharing

- need for sharing
- gray: why don’t people share?
  - **understanding the shades**
  - brighter alternatives
  - towards a brighter gray future

What Is the Problem Again?

want to share data (perhaps)

**need to sanitize it**

so it’s useful

and we’re all happy

Data Sanitization Challenges

- IP anonymization: Cryptopan
  - structure means breaking one IP leaks some about others
- **k-anonymization**
  - vulnerable to dataset combination; sometimes aggregation varies
- differential privacy
  - challenge of assigning and managing privacy budget
- payload removal
  - missing can be bad, see Google Street Views wifi
    - “Google Huang to Share Its Content for Privacy”, Arisoff and Miller, NY Times 14 Mar 2013
- releasing anything releases something

Data Sanitization: Anonymization and Friends

- IP anonymization: Cryptopan
- **k-anonymization**
- differential privacy
- coupled with payload removal

**sanitization is building a “box” around the data**

Data Must Be Useful

<table>
<thead>
<tr>
<th>attribute</th>
<th>data revealed</th>
<th>example</th>
<th>consequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>user + payloads</td>
<td>website password</td>
<td>John’s password is “ahh”</td>
<td>others get in</td>
</tr>
<tr>
<td>user + website</td>
<td>conversing parties</td>
<td>John’s at monster.com</td>
<td>general topics</td>
</tr>
<tr>
<td>user + protocol</td>
<td>action taken</td>
<td>John’s browsing</td>
<td>not working on talk 12</td>
</tr>
<tr>
<td>anon IP + protocol</td>
<td>apps at site</td>
<td>browsing and running Tor, Riptorrent, nap...</td>
<td>embarrassing applications?</td>
</tr>
<tr>
<td>anon IP + payload data</td>
<td>site has network connectivity</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3/10/2014
**Data Utility Challenges**

- every step of anonymization destroys some research value, and yet there’s always some risk
- harder still
  - research is by definition new
  - so anonymization must keep changing
  - often need clear subset when developing new approaches

**So We’re All Happy**

- many expectations
  - data providers
  - researchers
  - end-users
- when unstated

**Towards Brighter Alternatives**

- general goal: balance benefits vs. risk
  - inspiration: The Belmont Report (1979):
    - ethics in medical research
  - see also: The Menlo Report (2011)
    - applying Belmon guidelines to networking research
- insight: anonymization is just a box around the data

**Some Brighter Alternatives**

- insight: anonymization is just a box around the data
- broadening the box:
  - more than technical: legal
  - end-to-end: output
  - third parties: data

**More Than Technical: Legal**

- technical methods are fundamentally limited
  - data includes something (or it’s useless)
  - something always has some risk
- recommendations:
  - explicit expectations [Allen07a]
  - and formal legal agreements
Why Technical and Legal

- technical is essential, but not sufficient (80%?)
- legal can bridge the gap
- the trade-offs
  - expectations and legal are not perfect
  - a bad actor can ignore them
  - accidental release can circumvent them
  - but expectations and legal have consequences
  - both social and legal
  - but there are consequences
  - social and possibly legal
- but necessary to balance utility and protection
- when technical methods alone do not usefully get to 100%

Implications of Legal

- things will move slower
  - reviewing legal agreements and getting signatures
- don’t apply everywhere
  - across jurisdictions
  - not everyone is in “an institution”
- everyone is “special”
  - a lawyer’s job is not done until some clause has changed
- need to educate our institutions and our expectations

End-to-End Privacy: Output

- technical methods (on data) ignore two-thirds of the space
- recommendation:
  - consider end-to-end privacy and the research output
  - mechanism: data enclaves

Data Enclaves

- concept: sensitive data in a room; control in and out
- result: limited access for researchers, privacy stays in room, safe results come out

- status
  - common in social sciences and in the networking industry
  - work-in-progress by several networking groups (UMich, USC, CSU, PCH)

- term identified by Michael Bailey (UMich)

Implications of Data Enclaves

- enabler for research on more sensitive data
- more work to run, and to use
- but a balanced point to explore

- future work:
  - experience with open enclaves for networking
  - virtual enclaves: replacing the locked room with monitoring

Third Parties and Data

- solution to any problem in computer science?
  - …another level of indirection
Other Requests for a Third Party

- president’s review group on intelligence and communications
  — Clark et al., Dec. 2013
- “We recommend...access to [telephony meta-data] should be held ...by a private third party. Access...permitted only under FISA court order.”

Roles of a Third Party

- neutral: they don’t do anything (themselves)
- auditing: they can observe what is done
- transparency: they can report it
- shift emphasis from perfection to risk management

Bright Gray Sharing

- need for sharing
- gray: why don’t people share?
- understanding the shade
- brighter alternatives
- towards a brighter gray future

Bright Gray Sharing

- sharing network data is important
- gray: it’s hard today, and getting harder
- brightness: need to look beyond technical means alone
  — legal, process (data enclaves), and auditing (third parties)
- we should expect more effort to share (because data has value)
- perhaps you can help?
  — share your data
  — use available data (ex: http://predict.org and www.stei.edu/art/imagex/)
  — tolerate more process when getting others’ data
  — but consider if bright gray ways can help