De-Anonymising the Australian Medicare Data Release

Ben Rubinstein with Chris Culnane & Vanessa Teague
The University of Melbourne

You call *that* a data release?
Cybersecurity problems are engineering problems

Think carefully about what problem we’re trying to solve.
If the first attempt doesn’t work, try to understand why not.
Redesign with new understanding.
Security researchers spend a lot of time breaking each other’s designs.

not the best city in the world but it does have a nice bridge
1 August 2016 – MBS/PBS dataset is released

Medicare/Pharmaceutical Benefits Schedules

- “Medicare is Australia’s universal health insurance scheme... It guarantees all Australians (and some overseas visitors) access to a wide range of health and hospital services at low or no cost.”
- “PBS provides timely, reliable and affordable access to necessary medicines for Australians... Under the PBS, the government subsidises the cost of medicine for most medical conditions.”

3 billion lines of data

- 10% (2.9 million) of the population
- MBS 30 years: 1984 – 2014
- PBS 11 years: 2003 – 2014
- Almost*: Demographics + billing records

Infographic from the Dept of Health
*What protections were in place?

- Encrypted Provider IDs and Patient PINs
- Collapsed locations into 4 geographic regions
  - patient MBS enrolment
  - provider location
- Year of birth
- Removal of centenarians
- Service & supply date perturbed by up to 14 days
  - Constant perturbation within a patient’s records
- Extremely low service volume items removed

<table>
<thead>
<tr>
<th>(Encrypted) patient ID</th>
<th>0345952108</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>F</td>
</tr>
<tr>
<td>Year of birth</td>
<td>1963</td>
</tr>
<tr>
<td>(Encrypted) patient ID</td>
<td>0345952108</td>
</tr>
<tr>
<td>State</td>
<td>Vic-Tas</td>
</tr>
<tr>
<td>Date</td>
<td>7 Aug 1992</td>
</tr>
<tr>
<td>(Encrypted) supplier ID</td>
<td>2340981234</td>
</tr>
<tr>
<td>Item code</td>
<td>00023 (GP visit)</td>
</tr>
<tr>
<td>Price paid by patient</td>
<td>$85</td>
</tr>
<tr>
<td>Price reimbursed by Medicare</td>
<td>$60</td>
</tr>
<tr>
<td>Various other details</td>
<td>...</td>
</tr>
</tbody>
</table>
Decrypted Provider IDs in full
- described pseudo RNG insecure in setting
- reversible by guesswork

Aust Privacy Commissioner: Encryption dates back...
- 2005: DHS for 10% PBS releases to select groups
- mid-1990s: former Health Insurance Commission

Instead: should have used RSA, AES, random IDs

We responsibly disclosed to Dept
- providers can be recovered
- high risk of patient reID also
Privacy Amendment (Re-identification Offence) Bill 2016

One day before agreed announcement day...

Attorney-General memo:
Intention to amend Privacy Act 1988
- Legislative instrument permitting retroactive prosecution

The Bill
- Criminalises reID of Commonwealth data
- Up to 2 year jail term
- Reverses burden of proof
- Retrospective legislation (see memo above)
- Exemptions for AG or Dept-sanctioned investigations
- Apparently* might not apply to all* academics

Brandis swings his golden hammer, misses mark
Criminalising the re-identification of de-identified government data will hinder legitimate researchers and do nothing to improve citizens’ privacy.

By Stilgherrian for The Fell TBF | September 29, 2016 – 06:52 GMT (08:52 AEST) | Topic: Security

Life of a ReID Bill
12 Oct 2016 – Intro to the Senate
10 Nov 2016 – Ref to Senate Legal Committee
16 Dec 2016 – Consultation Period Ended
7 Feb 2017 – Senate Committee Report
... 6 Jun 2019 – Zombie Bill!
Overwhelmingly critical response

Law Council of Australia
- “The reverse onus provisions should be removed …”
- “definition of ‘de-identified’ … not always clear cut”

Australian Bankers Associations
- “re-identification ... might accidentally occur, without the bank intending”

14/15 submissions critical

Final Senate Committee report:

“The committee notes the concerns ... However...the bill provides a necessary and proportionate response”
“Health Minister Sussan Ley insists the data, which was loaded onto the internet, does not identify patients.”
But, we’re abound with health data

This is the tender moment that shows the modern family of Bill and Chloe Shorten after weeks on the federal election campaign trail.

The opposition leader was supported by his wife, their six-year-old daughter and his two step-children on Sunday as they enjoyed some family time in a park near their home at Moorabbin, Melbourne.

Mrs Shorten was pictured holding hands with her teenage children Rupert, 15, and George, 13, while her daughter Clementine, who she shares with Mr Shorten, was held tightly by her niece Alexandra.

Soon afterwards, Bernardi was diagnosed with tuberculosis and put in isolation, where he was asked to list all the people he had been in recent contact with.

‘I didn’t have the nerve to tell them about John Howard,’ Bernardi admits in the latest episode of ABC’s Kitchen Cabinet to air on Wednesday night.

‘I thought, let him go to the election and then we’ll deal with it.’

But when Howard was later hospitalised with lung problems, Bernardi was convinced he’d ‘done John Howard in’.
Searching for Vanessa

17,310 women share her birthyear

59 also had children born 2006, 2011 in Australia

23 also based in Victoria

0 with child DOB with perturbations

→ Not in dataset

Anyone could do this!!
Mothers unique in MBS-PBS-10%

Child birth years $[b, b + 4]$  
Mum birth years $[m, m + 5]$  
... for suppressed $m \gg b$

Similarly for small $m$

<table>
<thead>
<tr>
<th>Year of birth (mother\baby)</th>
<th>b</th>
<th>b+1</th>
<th>b+2</th>
<th>b+3</th>
<th>b+4</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>m+1</td>
<td>0</td>
<td>0</td>
<td>$1^3$ Dec b+2 or Jan b+3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>m+2</td>
<td>0</td>
<td>1</td>
<td>July/Aug</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>m+3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>m+4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
DeID as property of datasets... (compare to DP)

- **Unique in the Crowd: The privacy bounds of human mobility** - Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel

- **Not So Unique in the Crowd: a Simple and Effective Algorithm for Anonymizing Location Data** - Yi Song, Daniel Dahlmeier & Stephane Bressan

- **Unique in the shopping mall: On the reidentifiability of credit card metadata** - Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, Alex “Sandy” Pentland

- **Big Data and Innovation, Setting the Record Straight: De-identification Does Work** - Ann Cavoukian & Daniel Castro

- **No silver bullet: De-identification still doesn't work** - Arvind Narayanan & Edward W. Felten
  - ReID models overly narrow by assuming only access to public data
  - ReID does not require specialist skills
  - Quantification of reID probabilities at best lower bounds/fundamentally meaningless
It’s only a sample; bah humbug “confidence”!!

DHS whole-of-population statistics on MBS billing rates

- aggregated: 10yr age ranges, billing state, billing month
- 27% of codes are uniquely reported
- E.g. aortic valve replacement, Former PM in Brisbane, Aug 2011 (in pop, not sample)
Reidentifications

Wiki/news articles on 18 mums with 2+ births

- 13 had no matches, including
  - Gillian Triggs – Former Human Rights Commis.
  - Natasha Stott-Despoja – Former Senator
  - Cathy Freeman – Olympic Athlete
  - Tanya Plibersek – Current MP
- 2 rejected due to inconsistent information
- 3 returned a unique match

25 more queries

- Professional footballers with injuries
- Politicians with reported unusual surgeries
- 4 more unique matches
- One is an AFL team captain matching medical history, birth year, and interstate movements
- One has likely been reported as the oldest person in a state to have received a surgery, as confirmed by whole-of-population data

Other risks – only partially assessed
Fingerprinting by billing amounts
Melbourne Pharmaceutical Datathon with zipcodes
The Commissioner’s view is that the processes or steps for achieving re-identification are so extensive, and the risk of identification for any given patient so low, that the patients in the dataset are not reasonably identifiable for the purposes of the Privacy Act.

On the information before him, the Commissioner’s view is that the published dataset does not contain the personal information of patients, as defined in the Privacy Act, as patients cannot be reasonably identified.
Where to from here?

Data can hold huge potential societal benefit

Provide access to people who need it, whilst protecting privacy

Directions

- List what data is held – transparency in who holds what data
- Secure research environments – e.g. ONS Virtual Microdata Laboratory
- Crypto, SMC, private linkage
- *More* reID for assessing privacy risk
- **Differential privacy with many advantages** (much over the next 2 days)
Partners in it’s-not-a-crime

We also variously like
Diff. privacy (TransportNSW)
Applied crypto (Swiss sVote)
Bayesian record linkage (ABS)
...

Visitors welcome!!