

vanessa@thinkingcybersecurity.com
chris.culnane@gmail.com
ben.rubinstein@unimelb.edu.au

De-identified data probably isn't

Technion Summer School on Cyber and Computer Security

Privacy in Challenging times

September 2020

Vanessa Teague

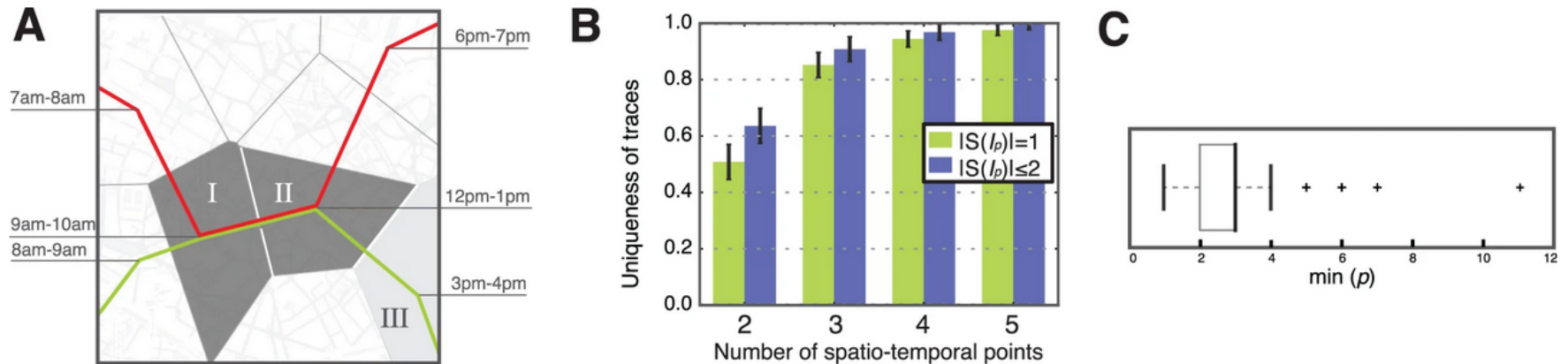
Thinking Cybersecurity Pty Ltd and the Australian National University

vanessa@thinkingcybersecurity.com

Based on joint work with Chris Culnane and Ben Rubinstein (Uni of Melbourne)

We're all individuals

From: [Unique in the Crowd: The privacy bounds of human mobility](#)



(A) $I_{\rho=2}$ means that the information available to the attacker consists of two 7am-8am spatio-temporal points (I and II). In this case, the target was in zone I between 9am to 10am and in zone II between 12pm to 1pm. In this example, the traces of two anonymized users (red and green) are compatible with the constraints defined by $I_{\rho=2}$. The subset $S(I_{\rho=2})$ contains more than one trace and is therefore not unique. However, the green trace would be uniquely characterized if a third point, zone III between 3pm and 4pm, is added ($I_{\rho=3}$). (B) The uniqueness of traces with respect to the number ρ of given spatio-temporal points (I_ρ). The green bars represent the fraction of unique traces, i.e. $|S(I_\rho)| = 1$. The blue bars represent the fraction of $|S(I_\rho)| \leq 2$. Therefore knowing as few as four spatio-temporal points taken at random ($I_{\rho=4}$) is enough to uniquely characterize 95% of the traces amongst 1.5 M users. (C) Box-plot of the minimum number of spatio-temporal points needed to uniquely characterize every trace on the non-aggregated database. At most eleven points are enough to uniquely characterize all considered traces.

by Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel.
<https://www.nature.com/articles/srep01376>



We're all individuals

- A few ordinary data points suffice to make almost anyone unique
 - *eg.* Where you live, where you went on the weekend, where you travel



Example 1: MBS-PBS open data

- 10% of the Australian population
- For each selected patient, all Medicare & Pharmaceutical Benefits bills 1984-2014
- Published as open data, August 2016
- Supplier (doctor) IDs were "encrypted"
 - But easily decrypted
- Patient data was de-identified
 - By randomly perturbing dates up to +/- 14 days
 - and removing rare *events*
 - But easily re-identified
 - By querying for known medical events e.g. childbirth



Lots of people share health information online

www.heraldsun.com.au/sport/af/more-news/injury-curse-who-is-the-afis-unluckiest-player/news-story/8b816172e64254cfce

Fedora Documentation Fedora Project Red Hat Free Content Sign In FC18: Program

Who has been hit by the injury jinx a your club?

ANDY OTTEN (ADELAIDE)

Drafted: 2007





Games: 79 (average 11 per year)

Drafted with pick 27 in the 2007 national draft, Otten has had an injury-cursed career after a breakout season in 2009. In a cruel twist of fate, it was at the end of that year that he suffered his first serious knee injury and required a knee reconstruction. Since then he has had a number of knee worries and before again rupturing his ACL in August last year.

He's not the only Crow who's a regular in the rehab group, with young gun Brad Crouch suffering a broken leg, broken foot and Achilles worries in his first three seasons.

Note: average games does not include 2015 season.

MOST VIEWED

-  'I was bawling my eyes out': Owners' pain over trashed house
-  Viewers slam MKR final as 'rigged'
-  'Ridiculous, joke': Ex-players slam ban
-  Shock as Tigers star backs mate in court

ADVERTISEMENT

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 Moonee Ponds in Melbourne.

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



Natalie Terese **Abberfield** was born in 1969.¹ She is the daughter of **Kenneth Beresford Abberfield** and **Melissa Judith Dutton**.¹ She married **Barnaby Joyce** in 1993.¹

Children of Natalie Terese Abberfield and **Barnaby Joyce**

1. **Bridgette Maree Joyce**¹ b. 1996
2. **Julia Frances Joyce**¹ b. 1998
3. **Caroline Wilga Joyce**¹ b. 2000
4. **Odette Honora Joyce**¹ b. 2002

The Daily Telegraph

Search

NEWS SPORT NRL ENTERTAINMENT OPINION BUSINESS LIFESTYLE RE/

Tanya's Plibersek juggle of love

By ROSIE SQUIRES, The Sunday Telegraph
October 10, 2010 12:12am

AT only 10 days old, Labor Minister Tanya Plibersek's newborn carries the weight of a nation.

The Sydney Morning Herald

YOUR SHORTLIST IS EMPTY
Add stories using the  button

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".

De-identification doesn't work on detailed records

- about 70 queries
 - Based on online, public info
 - Most had no matches in the sample
 - 10 returned a unique match
 - In some cases, there's a fair chance of a coincidental resemblance to someone else
 - In others, we're very confident it's the same person
 - Notified DoH December 2016



Can we be confident?

- Aus govt also releases aggregated group statistics of MBS bills
- The whole population, not a 10% sample

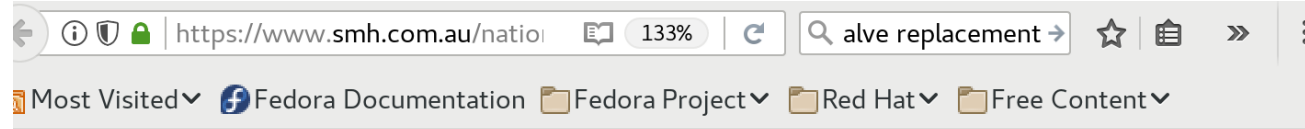
- **No**

Age range	45—54
State	Qld
Month	August
Year	2011
Gender	M
Item code	38556
Price reimbursed by Medicare	\$2240



Re-identification is possible in the aggregate data

- But you can't retrieve the patient's other records
- We didn't learn anything we didn't already know



Rudd heart surgery a success: doctor

1 August 2011 – 9:57am



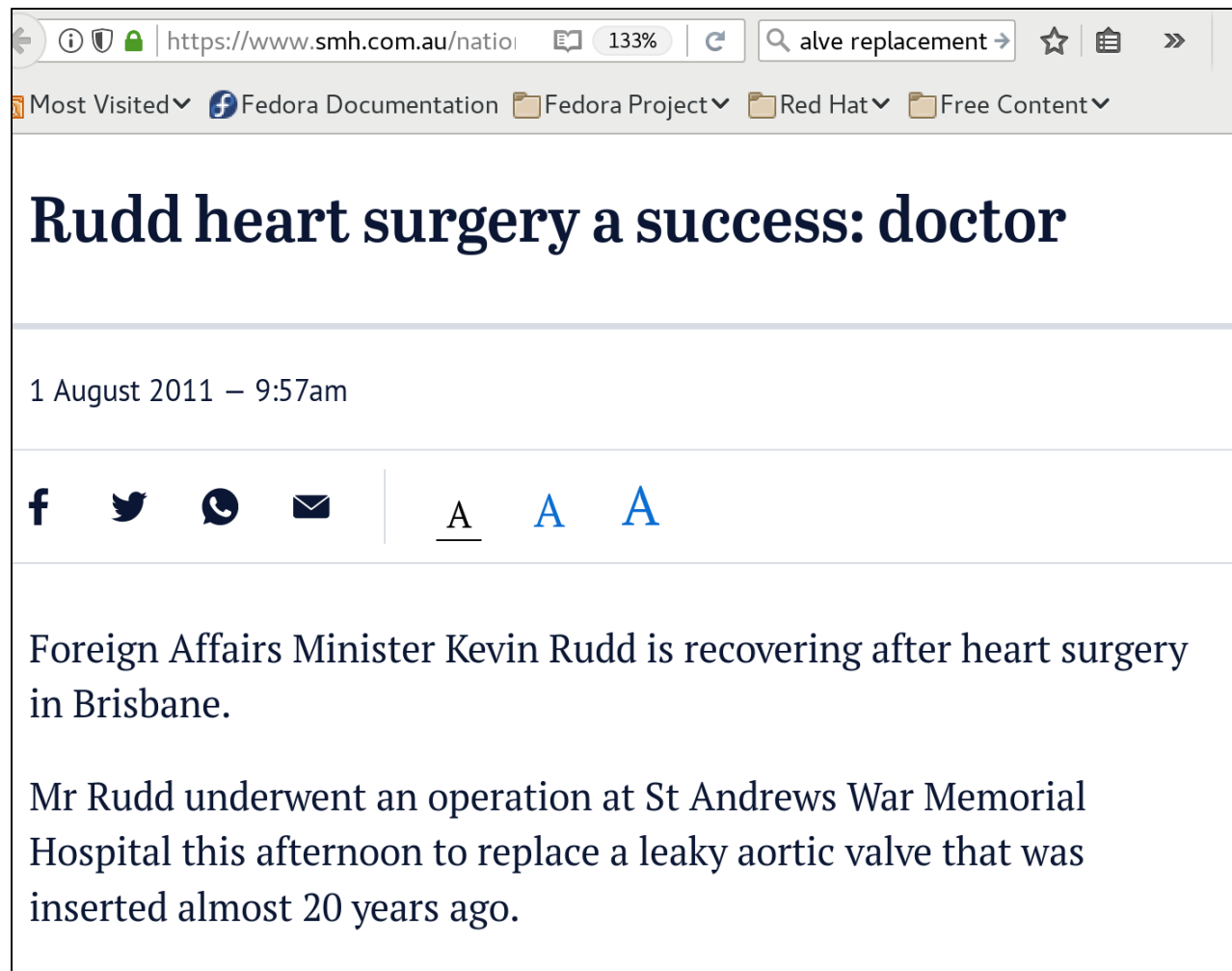
Foreign Affairs Minister Kevin Rudd is recovering after heart surgery in Brisbane.

Mr Rudd underwent an operation at St Andrews War Memorial Hospital this afternoon to replace a leaky aortic valve that was inserted almost 20 years ago.



Re-identification in DHS data -> Confidence in MBS-PBS 10% sample re-identification

- Mr Rudd's record (fortunately) is **not** in the 10% sample
- But if it was, we could be certain of correct re-identification based on the DHS data



The screenshot shows a web browser window displaying a news article. The address bar shows the URL <https://www.smh.com.au/natio>. The page title is "Rudd heart surgery a success: doctor". The article is dated "1 August 2011 - 9:57am". The main text reads: "Foreign Affairs Minister Kevin Rudd is recovering after heart surgery in Brisbane. Mr Rudd underwent an operation at St Andrews War Memorial Hospital this afternoon to replace a leaky aortic valve that was inserted almost 20 years ago." The browser interface includes a search bar with the text "alve replacement", a star icon, and a list of bookmarks such as "Fedora Documentation", "Fedora Project", "Red Hat", and "Free Content".

← <https://www.smh.com.au/natio> 133% ☆

Most Visited Fedora Documentation Fedora Project Red Hat Free Content

Rudd heart surgery a success: doctor

1 August 2011 – 9:57am

| A A A

Foreign Affairs Minister Kevin Rudd is recovering after heart surgery in Brisbane.

Mr Rudd underwent an operation at St Andrews War Memorial Hospital this afternoon to replace a leaky aortic valve that was inserted almost 20 years ago.

High confidence for ordinary people

- Childbirth is very common
- But if you have 2 or 3 children, the number of other mums who match *all* of them is very small
- Consider re-identification in MBS-PBS 10% sample based on childbirth dates...



k-anonymity is high

Query

count

Selector

1 ▾ {"GENDER": "F", "YOB": 197}

By default, valid **ObjectID** and **ISODate** (YYYY-MM-DD HH:mm:ss) strings are being
change this behaviour

Execute

count - patients

count - patients

count - patients

Tree ▾

: 19000

But re-identification is easy

Query count

Selector

```
1 ▾ {"GENDER": "F", "YOB": 197,
2   { $and: [
3
4   {"MBS": { $elemMatch: {"ITEM": { $in: ["00200", "00201", "00204", "00205", "00207", "00208", "00209",
5     {"DOS": { "$gte": "200- - 00:00:00", "$lte": "200- - 00:00:00" } }
6   }
7   }
8   }
9 ]
10 } }
11
```

By default, valid **ObjectID** and **ISODate (YYYY-MM-DD HH:mm:ss)** strings are being converted into MongoDB objects, **unchecked** right? change this behaviour

Execute

count - patients

count - patients

Tree

: 51



How many match both births?

Query

count

Selector

```
1  {"GENDER":"F", "YOB":197,
2    { $and: [
3
4    {"MBS": {$elemMatch: {"ITEM":{$in: ["00200", "00201", "00204", "00205", "00207", "00208", "00209", "0
5    {"DOS": {"$gte": "20- - 00:00:00", "$lte": "20- - 00:00:00"}}}
6    }
7      }
8    },
9    {"MBS": {$elemMatch: {"ITEM":{$in: ["00200", "00201", "00204", "00205", "00207", "00208", "00209
10   {"DOS": {"$gte": "20- - 00:00:00", "$lte": "20- - 00:00:00"}}}
11   }
12     }
13   }
14
15 ]
16 } }
17
```

By default, valid **ObjectID** and **ISODate (YYYY-MM-DD HH:mm:ss)** strings are being converted into MongoDB objects, **uncheck** right upper
change this behaviour

Execute

count - patients

count - patients

count - patients

Tree

: 1

3rd data point for added confidence

Selector

```
1 ▾ {"GENDER":"F", "YOB":197,
2   { $and: [
3
4   {"MBS": {$elemMatch: {"ITEM":{"$in": ["00200", "00201", "00204", "00205", "00207", "00208", "0020
5   {"DOS":{"$gte":"20- - 00:00:00", "$lte":"20- - 00:00:00"}}
6   }
7   }
8   },
9   {"MBS": {$elemMatch: {"ITEM":{"$in": ["00200", "00201", "00204", "00205", "00207", "00208", "
10  {"DOS":{"$gte":"20- - 00:00:00", "$lte":"20- - 00:00:00"}}
11  }
12  }
13  },
14  {"MBS": {$elemMatch: {"ITEM":{"$in": ["00200", "00201", "00204", "00205", "00207", "00208", "
15  {"DOS":{"$gte":"20- - 00:00:00", "$lte":"20- - 00:00:00"}}
16  }
17  }
18  } ] ]
19 }
```

By default, valid **ObjectID** and **ISODate (YYYY-MM-DD HH:mm:ss)** strings are being converted into MongoDB objects, **uncheck** right change this behaviour

Execute

count - patients

count - patients

count - patients


count - patients



Tree

: 1

Confidence from 3 equal shifts

- Remember that a patient's dates are perturbed randomly *by the same amount for all that person's events*
 - Mary's 3 children are all shifted by the same number of days.
 - What's the likelihood of coincidental resemblance to someone else, even someone else who matches all 3 28-day windows?
 - Depending on your assumptions, it's
- 

Plan: three Australian case studies

- The (Australian) Medicare-Pharmaceutical Benefits Scheme (MBS-PBS) 10% sample dataset
 - Can patients be identified?
 - Can re-identifications be confident?
- The (Victorian) Myki transport dataset
 - More easy re-identifications
- The (Queensland) open data portal

"Succinctly put, 'De-identified' data isn't, and the culprit is auxiliary information."



Myki 'de-identified' data

- Tap-on and tap-off events for all Melbourne public transport users
 - Trains, trams and some buses
 - July 2015-June 2018
 - Exact route/stop/station numbers
 - Times to the second
 - All events for the same card are linked
 - No information about people
 - though there are different kinds of cards, e.g. children, MPs



Myki 'de-identified' data

<https://www.abc.net.au/news/2019-08-15/myki-data-spill-breaches-privacy-for-millions-of-users/11416616>

'Shocking' myki privacy breach for millions of users in data release

By [Mary Gearin](#)

Posted Thu 15 Aug 2019 at 3:25pm, updated Thu 15 Aug 2019 at 6:33pm



Researchers found they could identify commuters by their travel histories. (ABC News: Danielle Bonica)

Share   



Just a few taps on and off, and a couple of tweets — that's all it would take for a hacker or stalker to identify you and track down your movements with a myki.



Myki 'de-identified' data



- Chris identified himself
 - based on exact times
- Then he identified Ben
 - because they travelled together
- Then we looked for tweets re public transport
- 2-3 points suffices for uniqueness
 - even if you don't have an MPs card



Myki 'de-identified' data



Anthony Carbines MP  @ACarbinesMP · May 3, 2018 

 See you about 05.24AM tomorrow at Rosanna to catch the first train to town. Well done all. Thanks for hanging in there. Massive construction effort. Single track gone. Two level crossings gone. The trains! The trains! The trains are coming! 



 8

 15

 58



What did the Victorian govt do about it?

- Refused to acknowledge commuters were identifiable
- *but...*
- The Victorian Privacy Commissioner wrote a very detailed and damning report about it

https://ovic.vic.gov.au/wp-content/uploads/2019/08/Report-of-investigation_disclosure-of-myki-travel-information.pdf



Plan: three Australian case studies

- The (Australian) Medicare-Pharmaceutical Benefits Scheme (MBS-PBS) 10% sample dataset
 - Can patients be identified?
 - Can re-identifications be confident?
- The (Victorian) Myki transport dataset
 - More easy re-identifications
- The (Queensland) open data portal

"Succinctly put, 'De-identified' data isn't, and the culprit is auxiliary information."



The Queensland Open Data Portal

https://www.data.qld.gov.au/dataset/alcohol-and-other-drug-treatment-services-aodts-national-minimum-data-set-nmds

Organisations / Queensland Health / Alcohol and Other Drug Treatment Services (AODTS) National Minimum Data Set (NMDS)

Alcohol and Other Drug Treatment Services (AODTS) National Minimum Data Set (NMDS)

Organisation

Queensland Health

Social

Twitter

Facebook

License

[Creative Commons Attribution 4.0](#)

Dataset

Groups

Activity Stream

Alcohol and Other Drug Treatment Services (AODTS) National Minimum Data Set (NMDS)

PLEASE NOTE that the data sets available for the Queensland Alcohol and Other Drug Treatment Services (AODTS) National Minimum Data Set (NMDS) have temporarily been removed to review content and formatting. Once this review has been completed, the data sets will be republished on this site by 1 October 2020.

Data and Resources

This dataset has no data

health

- Postcode, age bracket, indigenous status, place of seeking treatment..
- Primary problematic drug, other drugs used...

The Queensland Open Data Portal: indigenous data sovereignty?

5

Data politics and Indigenous representation in Australian statistics

Maggie Walter

Introduction

Accepting the philosophical premise that numbers exist, as per Quine (1948), is ontologically different to accepting that numbers have a fixed reality. This differential is the essence of the reality of numbers as they are applied to indigenous populations. In First World colonised nations such as Australia, Aotearoa/New Zealand, Canada and the United States, the question is not just 'are these numbers real', but also 'how are these numbers deployed and whom do they serve'. The reality query is not of the numbers themselves but of what they purport to portray.




What did the Queensland govt do about it?

- Quietly took the data offline.
- Stay tuned...



What to do about future data sharing?

- Differential Privacy quantifies privacy loss
 - Against a powerful attacker with lots of auxiliary information
 - Good for basic aggregates; research continues for more complex data types
 - k-anonymity protects obvious identifiers
 - But fails if the adversary has other info
 - Sensitive unit-record level data belongs in a secure research environment
 - Further reading:
- 

What to do about future data sharing?

- Assume that detailed unit-record level data is identifiable
 - Even if someone tried to de-identify it
- Don't share data "on the basis that it is de-identified" if individuals are identifiable

