Week 6: Data Anonymisation

Charlotte Hadley

Topics for today

this is an edit for Rin3 2022

1. Why is data anonymisation important?
2. What are specific risks of deanonymisation of health data?
3. Anonymity measures: k-anonymity and l-diversity
4. Case studies of deanonymisation

> … and why anonymity measures are often not enough

5. R packages for working with anonymouse data and sampling

Why is data anonymisation important?

Data isn’t always captured knowingly

Mostly during this course we’ve been talking about surveys or studies where data is explicitly being collected - and participants willingly submit their data.

But that’s often not the case.

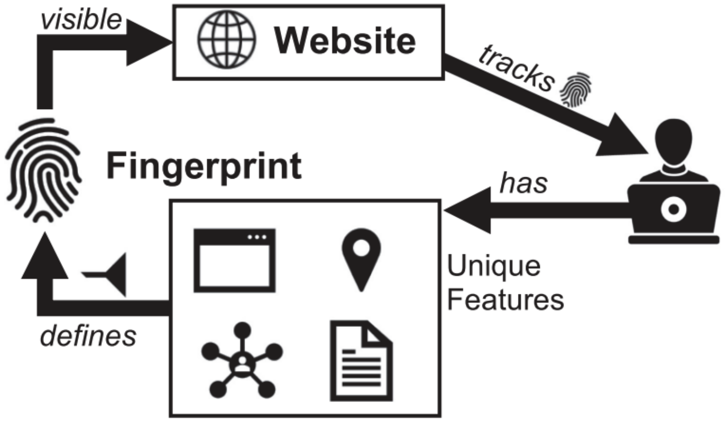
> Data is collected continuously about individuals without their explicit consent - and often without implicit consent.

Data tracking across websites

Cookies [and similar technologies] are ever present in the modern web.

They allow websites to track both the websites that we visit **and** how we engage with websites.

> … but what about all those cookie popups?!



(b) A representation of how device fingerprinting enables the tracking of users on the web. The specific fingerprint may vary across different websites, since it can include different features unique to a particular device, which are requested by the website.

Source: Kretschmer et al 2021[1]

GDPR and Cookies

The GDPR Policy[2] included mention of "cookie identifiers" which ultimately led to the cookie consent popups you see everywhere.

For a common sense description of what the policy requires see gdpr.eu/cookies/[3]

The policy came into effect in 2018 - and the cookie avalanche started.

The policy was well meaning, and necessary.

In 2016 it was demonstrated that 70% of the top million websites used some form of tracking[4].

However, there's clear evidence that the policy hasn't materially improved our privacy[1].

Cookies: Fingerprinting, nudging and dark patterns

These are the primary ways websites circumvent the GDPR cookie policy:

- The biggest issue with targeting cookies is they’re difficult to define and more modern tools like fingerprinting are harder to track.
- User’s are *nudged* to accept cookies by auto selecting “Accept”
- Dark patterns are employed to prevent users from refusing to accept cookies[5].

Overall, the policy has probably made things worse.

> GDPR impacted smaller advertisement companies considerably more than large brands, such as Google and Facebook, leading to a higher market concentration for these companies, which, in turn, may increase the privacy threat, rather than decrease it Source: Kretschmer et al 2021[1]

Why do websites want to track us?

There are simple uninteresting answers to this:

- Selling tracking data to advertising networks
- Using tracking data to target products at users

But there are interesting answers!

- Anticipatory shipping predicts future purchases
  - Amazon first patented this process in 2013[6], thoroughly explained by Eva-Maria Nyckel[7]
  - Used in the agro-food supply chains[8]

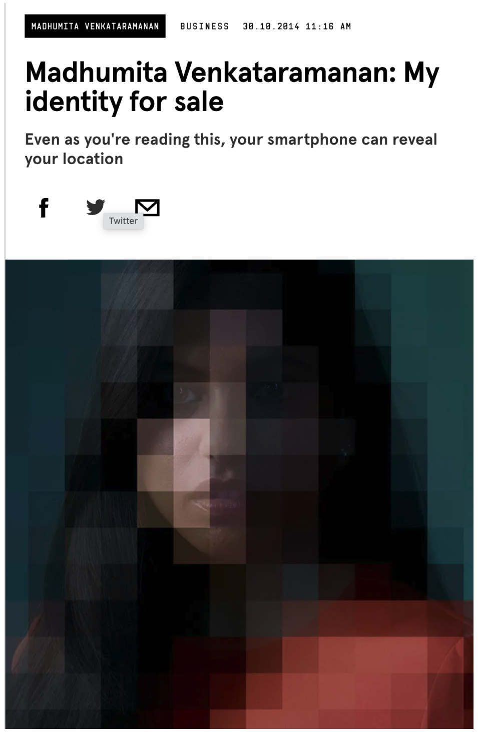
> Can we get some examples of what they track?

Wired: Madhumita Venkataramanan: My identity for sale

I strongly recommend reading all of this article - wired.co.uk/article/my-identity-for-sale[9].

- The article is from 2014, but as we've discussed the GDPR cookie policy has if anything made this situation worse.
- The article also provides a wealth of other examples of data tracking beyond cookies. **These would be good examples for your assignment**.
- The quote below is from the article and summarises descriptions of Madhumita's life tracked without consent.

> I’m a 26-year-old British Asian woman, working in media and living in an SW postcode in London. I’ve previously lived at two addresses in Sussex and two others in north-east London. While I was growing up, my family lived in a detached house, took holidays to India every year, donated to medical charities, did most of the weekly shopping online at Ocado and read the Financial Times. Now, I rent a recently converted flat owned by a private landlord and have a housemate. I’m interested in movies and startups, have taken five holidays (mostly to visit friends abroad) in the last 12 months and I’m going to buy flights within 14 days. My annual income is probably between £30,000 and £39,999. I don’t have a TV or like watching scheduled television…



MADHUMITA VENKATARAMANAN BUSINESS 30.10.2014 11:16 AM

Madhumita Venkataramanan: My identity for sale

Even as you're reading this, your smartphone can reveal your location

Data tracking during hospital visits

There's an A&E waiting time survey that's sent to folks that attend A&E.

This survey is sent without patients opting into it.

> The Section 251 of the NHS Act 2006[10] provides for the use of confidential patient data without consent for a specific purpose by the HRA or the Secretary of State for Health and Social Care.

That's an exception to the Data Protection Act[11]!

This means that there are people looking at this data and deciding who to target.

This exception was also used during the COVID-19 pandemic, it's not just used for surveying hospital wait times!

Note that the NHS makes sure there's an opt out [once you've been invited]

> However, as has always been the case, patients/service users must be given the opportunity to opt-out.

> # Data anonymisation is important because data is collected everywhere all of the time

… how does that match up with GDPR and the Data Protection Act?

Individual rights from the DPA

The DPA[11] provides 8 rights for individuals:

- The right to be informed
- The right of access
- The right to rectification
- The right to erasure
- The right to restrict processing
- The right to data portability
- The right to object
- Rights in relation to automated decision making and profiling.

If we don’t know that data is being collected - or by which organisations - our individual rights are not being protected.

This opens up lots of ethical questions. We’ll discuss these in the next lecture about data ethics.

In this lecture we’re going to focus on the specific risks to deanonymisation

What are the risks of de-anonymisation?

Who can be at risk of de-anonymisation?

Individuals

It’s the dangers to individuals that we should primarily be concerned with.

There are significant risks to individual liberty, livelihood and life from deanonymisation.

Organisations

However, organisations also suffer if data they store/process is deanonymised.

- Organisations might suffer reputational damage
- Organisations might suffer legal difficulties, including fines

> Let’s focus on the individual for now. The ICO provides a useful guide to managing data protection risk designed for organisations[12]

Specific risks of de-anonymisation (I)

First and foremost, there is a risk of:

- Information about someone’s private life ending up in the public domain.

This in and of itself should be of concern, but more specifically

- Individuals might suffer distress, embarrassment, or anxiety due to sensitive information being in the public domain

There is also a significant risk from sensitive information being in the public domain that:

- Individuals might suffer harassment, attack and/or injury
- Individuals might suffer persecution

Specific risks of deanonymised health data (II)

Sensitive information might be sold to third-party organisations resulting in a change in service options, costs or other loss.

- Private healthcare data might be used by insurers to increase product fees or terminate existing products.
- Employers could potentially use this information in employment decisions.

> Remember that the DPA (and GDPR) provides specific rights (or protections) for individual's data. If data is sold without knowledge these rights cannot be guaranteed.

Specific risks of deanonymised health data (III)

These are the “protected characteristics” defined in the Equality Act 2010[13]

- Age
- Disability
- Gender reassignment
- Marriage and Civil Partnership
- Pregnancy and Maternity
- Race
- Religion
- Sex
- Sexual Orientation.

Frustratingly, and inhumanely there is prejudice against individuals in all of these groups.

This prejudice can be found in individual actions, from hate groups, as well as institutional policies and practices.

De-anonymisation of health data can realistically [and often easily] expose individual’s protected characteristics.

All individual data should be considered private, but there are significant risks to the de-anonymisation of sensitive healthcare data.

What is health data again?

Recall how the Data Protection Act[11] identifies three types of health data:

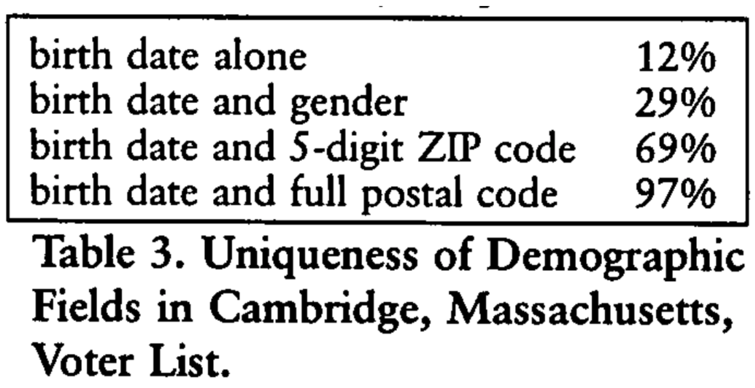
- **“biometric data”** means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of an individual, which allows or confirms the unique identification of that individual, such as facial images or dactyloscopic data;
- **“data concerning health”** means personal data relating to the physical or mental health of an individual, including the provision of health care services, which reveals information about his or her health status;
- **“genetic data”** means personal data relating to the inherited or acquired genetic characteristics of an individual which gives unique information about the physiology or the health of that individual and which results, in particular, from an analysis of a biological sample from the individual in question;

A potted history of de-anonymisation

Early evidence for de-anonymisation (I)

In the late 90s there was a rapid conceptualisation of how easy it is to de-anonymisation large, public datasets.

Latanya Sweeney showed in 1997[14] that using public data and the Massachusetts voting list (n=54,805) it was extremely easy to **uniquely** identify individuals from only 2 pieces of information.



birth date alone	12%
birth date and gender	29%
birth date and 5-digit ZIP code	69%
birth date and full postal code	97%

Table 3. Uniqueness of Demographic Fields in Cambridge, Massachusetts, Voter List.

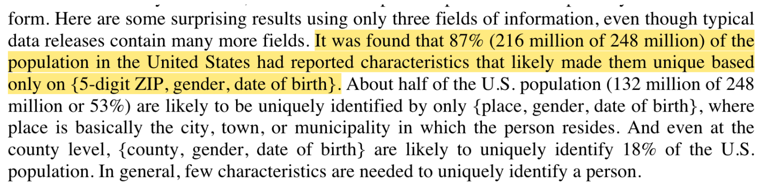
Source: Latanya Sweeney in 1997[14]

Early evidence for de-anonymisation (II)

In the late 90s there was a rapid conceptualisation of how easy de-anonymisation is of large, public datasets.

Three years later in 2000 Latanya Sweeney[15] demonstrated

> 87% of the US population can be uniquely identified from only ZIP, gender and date of birth.



form. Here are some surprising results using only three fields of information, even though typical data releases contain many more fields. It was found that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}. About half of the U.S. population (132 million of 248 million or 53%) are likely to be uniquely identified by only {place, gender, date of birth}, where place is basically the city, town, or municipality in which the person resides. And even at the county level, {county, gender, date of birth} are likely to uniquely identify 18% of the U.S. population. In general, few characteristics are needed to uniquely identify a person.

Source: Latanya Sweeney in 2000[15]

k-anonymity: A Model for Protecting Privacy

Two years later in 2002[1] a statistical technique called k-anonymity was introduced to measure the risk of re-identification.

… by Latanya Sweeney[16]!

Prof. Latanya Sweeney
Harvard Kennedy School

Originally her research into de-anonymisation was poorly received.

> "Even my Weld example[14] and related demographic analyses, despite making significant contributions to privacy regulations worldwide, were refused publication by more than 20 academic publications at the time."
>
> Only You, Your Doctor, and Many Others May Know[17]

k-anonymity: A definition (I)

k-anonymity is a property of a dataset that has been subject to anonymisation.

k-anonymity is an **integer value** that guarantees *internal* uniqueness of individuals amongst `k - 1` individuals.

Unfortunately, **a lot** of the material written about k-anonymity is confusing because people don’t declare their assumptions in calculating k values.

k-anonymity: A definition (II)

Let’s consider a simple pretend dataset.

We consider each column in the data to be an *attribute*.

These attributes can be categorised into two types of identifer:

- Unique identifiers
  - These are attributes that uniquely identify individuals. These **have** to be removed for anonymisation.
- Quasi-indentifiers
  - These attributes could be used used to identify individuals, even after anonymisation.

name	region	age_range	dise
Saindhavi	England	20-30	Hea
Enio	England	20-30	Hea
Daury	England	20-30	Hea
Alphus	England	20-30	Pan
Balian	England	20-30	Pan
Kenyea	England	20-30	Pan
Gracielynn	Wales	40-50	Live
Aliye	Wales	40-50	Live
Kadince	Wales	40-50	Live
Asaph	Wales	40-50	Live

k-anonymity: A definition (III)

Now we’ve thrown away the unique identifiers we need to decide which attributes are **sensitive**.

Sensitive attributes are medical/healthcare data that we need to protect from in the anonymisation process.

Non-sensitive		**Sensitive**
region	**age_range**	**disease**
England	20-30	Heart
England	20-30	Heart
England	20-30	Heart
England	20-30	Pancreatic
England	20-30	Pancreatic
England	20-30	Pancreatic
Wales	40-50	Liver
Wales	40-50	Liver
Wales	40-50	Liver
Wales	40-50	Liver

k-anonymity: A definition (IV)

There are now 3 different choices about how we calculate k-anonymity for our data.

- Combining together all non-sensitive attributes compared to **each** sensitive attributes.
- Combining together all attributes
- For each individual attributes

Let’s go through each of these in turn.

Non-sensitive		**Sensitive**
region	**age_range**	**disease**
England	20-30	Heart
England	20-30	Heart
England	20-30	Heart
England	20-30	Pancreatic
England	20-30	Pancreatic
England	20-30	Pancreatic
Wales	40-50	Liver
Wales	40-50	Liver
Wales	40-50	Liver
Wales	40-50	Liver

k-anonymity: A definition (V)

> Combining together all non-sensitive attributes compared to **each** sensitive attributes.

In toy example like this we can go through manually any count how many individuals belong to each group.

Using this measure the dataset is 4-anonymous as all individuals are guaranteed anonymity amongst 3 others (ie $k-1$).

disease	region	age_range	n_in
Heart	England	20-30	6
Heart	England	20-30	6
Heart	England	20-30	6
Pancreatic	England	20-30	6
Pancreatic	England	20-30	6
Pancreatic	England	20-30	6
Liver	Wales	40-50	4
Liver	Wales	40-50	4
Liver	Wales	40-50	4
Liver	Wales	40-50	4

k-anonymity: A definition (VI)

> Combining together all attributes

When we measure across **all** attributes the k-anonymity of the dataset is reduced.

Using this metric, the data has 3-anonymity.

disease	region	age_range	n_in
Heart	England	20-30	3
Heart	England	20-30	3
Heart	England	20-30	3
Pancreatic	England	20-30	3
Pancreatic	England	20-30	3
Pancreatic	England	20-30	3
Liver	Wales	40-50	4
Liver	Wales	40-50	4
Liver	Wales	40-50	4
Liver	Wales	40-50	4

k-anonymity: A definition (VII)

> For each individual attribute

When we measure the anonymity of each individual variable the dataset has 2-anonymity

We always use the **smallest** value of k for our specific measure.

disease	region	age_range
Heart	England	20-30
Heart	England	20-30
Heart	England	20-30
Pancreatic	England	20-30
Pancreatic	England	20-30
Pancreatic	England	20-30
Liver	Wales	40-50
Liver	Wales	40-50
Liver	Wales	40-50
Liver	Wales	40-50

k-anonymity: A definition (VIII)

As we've seen, each of these methods gives a different measure of the anonymity of the data.

1. Combining together all non-sensitive attributes compared to **each** sensitive attributes.
2. Combining together all attributes
3. For each individual attributes

> **Frustratingly** it's quite rare for authors to explicitly state which combination of attributes they use.
>
> The methods are listed roughly in terms of the frequency that I've seen them in the literature.

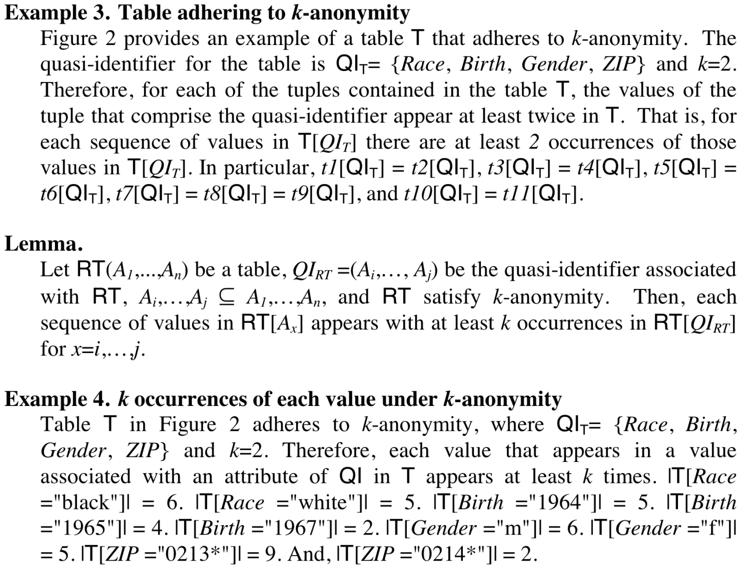
Non-sensitive		**Sensitive**
region	**age_range**	**disease**
England	20-30	Heart
England	20-30	Heart
England	20-30	Heart
England	20-30	Pancreatic
England	20-30	Pancreatic
England	20-30	Pancreatic
Wales	40-50	Liver
Wales	40-50	Liver
Wales	40-50	Liver
Wales	40-50	Liver

k-anonymity: A definition (IX)

The definition I’ve given you is sufficient and precise.

But be aware that you’ll often[16] see a more technical definition that uses set theory notation.

Ignore it. If a dataset is described as having “k-anonymity 10” that means for any row in the dataset there are at least 9 other rows identical to it.



Example 3. Table adhering to *k*-anonymity

Figure 2 provides an example of a table T that adheres to *k*-anonymity. The quasi-identifier for the table is $QI_T = \{Race, Birth, Gender, ZIP\}$ and k=2. Therefore, for each of the tuples contained in the table T, the values of the tuple that comprise the quasi-identifier appear at least twice in T. That is, for each sequence of values in $T[QI_T]$ there are at least 2 occurrences of those values in $T[QI_T]$. In particular, $t1[QI_T] = t2[QI_T]$, $t3[QI_T] = t4[QI_T]$, $t5[QI_T] = t6[QI_T]$, $t7[QI_T] = t8[QI_T] = t9[QI_T]$, and $t10[QI_T] = t11[QI_T]$.

Lemma.

Let $RT(A_1,\ldots,A_n)$ be a table, $QI_{RT} = (A_i,\ldots, A_j)$ be the quasi-identifier associated with RT, $A_i,\ldots,A_j \subseteq A_1,\ldots,A_n$, and RT satisfy *k*-anonymity. Then, each sequence of values in $RT[A_x]$ appears with at least *k* occurrences in $RT[QI_{RT}]$ for $x=i,\ldots,j$.

Example 4. *k* occurrences of each value under *k*-anonymity

Table T in Figure 2 adheres to *k*-anonymity, where $QI_T = \{Race, Birth, Gender, ZIP\}$ and k=2. Therefore, each value that appears in a value associated with an attribute of QI in T appears at least *k* times. $|T[Race = \text{"black"}]| = 6$. $|T[Race = \text{"white"}]| = 5$. $|T[Birth = \text{"1964"}]| = 5$. $|T[Birth = \text{"1965"}]| = 4$. $|T[Birth = \text{"1967"}]| = 2$. $|T[Gender = \text{"m"}]| = 6$. $|T[Gender = \text{"f"}]| = 5$. $|T[ZIP = \text{"0213*"}]| = 9$. And, $|T[ZIP = \text{"0214*"}]| = 2$.

Source: Formal definition for k-anonymity from Sweeney 2002[16]

k-anonymity: How is it achieved?

We are responsible for manipulating our dataset to achieve a desirable k-anonymity level.

> Even though a minimum k value of 3 is often suggested, a common recommendation in practice is to ensure that there are at least five similar observations (k = 5)[19]

We have two tools available to us:

Generalisation

We generalise a dataset through coarsening.

- Convert exact ages to age ranges
- Convert DOB to year of birth
- Trimming data, eg BS16 instead of BS16 6AB
- Creating new groups
  - Combine “Married”, “Divorced”, “Widowed” to “Been Married” and all

Suppression

Suppression removes data from a dataset.

We might suppress an attribute or set some specific values to “missing”.

Care must be taken to not suppress variables that are required for analysis.

📝 Task: Setup our project

SLIDE 1 OF 1

1. Create a new project for `week-6`
2. Create a new RMarkdown document called `data-anonymisation.Rmd`
3. Install the `{wakefield}` and `{faux}` package

{wakefield}

The `{wakefield}` package is very useful for creating random datasets of categorical variables.

The package has 49 different built-in variables with pre-defined distributions:

age	dice	hair	military	sex_inclusive
animal	dna	height	month	smokes
answer	dob	income	name	speed
area	dummy	internet_browser	normal	state
car	education	iq	political	string
children	employment	language	race	upper
coin	eye	level	religion	valid
color	grade	likert	sat	year
date_stamp	grade_level	lorem_ipsum	sentence	zip_code
death	group	marital	sex	

wakefield::r_data_frame()

We generate datasets with the `r_data_frame()` function:

```r
r_data_frame(10,
             id,
             name,
             dob,
             income,
             smokes,
             death)
```

```
# A tibble: 10 × 6
   ID    Name      DOB        Income Smokes Death
   <chr> <chr>     <date>      <dbl> <lgl>  <lgl>
 1 01    Dreniyah  2008-05-24 42462. TRUE   FALSE
 2 02    Alexiana  2008-11-06 60950. FALSE  TRUE
 3 03    Ashauria  2008-10-16 17069. FALSE  TRUE
 4 04    Azelea    2007-12-11 19028. FALSE  FALSE
 5 05    Krystof   2009-09-11 49067. TRUE   FALSE
 6 06    Loudes    2008-10-24 65993. FALSE  FALSE
 7 07    Rebeckah  2008-07-21 50860. FALSE  FALSE
 8 08    Jarek     2009-01-26 21266. FALSE  FALSE
 9 09    Gavrielle 2008-08-03 50405. FALSE  TRUE
10 10    Tarajah   2008-05-07 28798. FALSE  FALSE
```

> Can you explain why you see different data on your machine?

Pseudorandomness

When programming we use pseudorandom number generators to generate random numbers.

These are algorithms that **deterministically** give random numbers when given an input. We can therefore always get the **same** random numbers by setting the *seed* of the algorithm.

```r
set.seed(1)
r_data_frame(10,
             id,
             name,
             income,
             dna,
             smokes,
             death)
```

```
# A tibble: 10 × 6
   ID    Name      Income DNA      Smokes Death
   <chr> <chr>      <dbl> <fct>    <lgl>  <lgl>
 1 01    Donaldeen 48108. Cytosine FALSE  FALSE
 2 02    Martiqua  36496. Guanine  FALSE  FALSE
 3 03    Juliaann  12130. Thymine  FALSE  TRUE
 4 04    Poyraz    53488. Cytosine FALSE  TRUE
 5 05    Boleslaus 46311. Adenine  FALSE  FALSE
 6 06    Duc       56733. Adenine  FALSE  FALSE
 7 07    Hadeer    52217. Thymine  TRUE   TRUE
 8 08    Camilya   31854. Thymine  FALSE  TRUE
 9 09    Ashlay    77153. Adenine  TRUE   TRUE
10 10    Dutch     62490. Adenine  FALSE  FALSE
```

k-anonymity for our data (I)

Let's pretend our dataset is from a study on the effect of income on smoking morbidity[1].

When thinking about making this anonymous…

- Are there any variables we should **suppress**?
- How could we generalise the remaining variables?

ID	Name	DOB	Income	Sm
01	Donaldeen	2009-01-22	31854.26	FAL
02	Martiqua	2009-08-02	77153.29	FAL
03	Juliaann	2008-10-11	62490.28	TRU
04	Poyraz	2009-05-18	15462.24	TRU
05	Boleslaus	2008-08-16	19430.90	FAL
06	Duc	2008-08-20	24504.18	FAL
07	Hadeer	2009-04-01	27535.05	TRU
		2008-		

📝 Task: k-anonymity calculation

SLIDE 1 OF 2

1. Use this code to create a dataset:

```r
library(wakefield)
library(tidyverse)
library(lubridate)

set.seed(1)
smoke_data <- r_data_frame(
  50000,
  id,
  name,
  dob(start = ymd("1950-01-01"),
      k = abs(as.integer(days(ymd("1950-01-01") - Sys.Date()) - 365*18))),
  income,
  smokes,
  death)
```

2. Suppress inappropriate columns from the dataset.

📝 Task: k-anonymity calculation

SLIDE 1 OF 2

1. Generalise the remaining variables as follows:

- Extract year of birth
- Split income into 4 categories
    - "< £30,000"
    - "£30,000 - £70,000"
    - "£70,000 - £100,000"
    - "£100,000+"

2. Calculate the k-anonymity of the dataset

How to attack k-anonymised datasets (I)

There are 3 known attacks for attempt to de-anonymise datasets with k-anonymity:

- Unsorted matching attack
  - If an anonymised dataset is ordered in the same order that observations were recorded this is a potential attack vector. Dependent on the attack this might provide either an additional quant-identifier or (worst case) a unique identifer.
  - It’s a good practice to randomise the order of observations in anonymised data releases.

How to attack k-anonymised datasets (II)

There are 3 known attacks for attempt to de-anonymise datasets with k-anonymity:

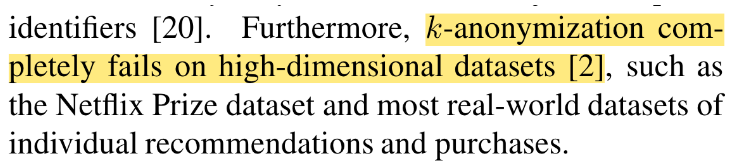
- Unsorted matching attack
- Subsequent release attacks
  - Large healthcare datasets might be used in multuple studies and be subject to multiple k-anonymised releases.
  - Temporal attacks are possible by analysing the modification of rows (eg health intervention) or the removal of data (eg morbidity).
  - Protecting against these attacks requires care and consideration. It is wise to consider the k-anonymity of combined releases.

How to attack k-anonymised datasets (III)

There are 3 known attacks for attempt to de-anonymise datasets with k-anonymity:

- Unsorted matching attack
- Subsequent release attacks

- Background knowledge
  - This is the most common attack vector and the Achilles' heel of k-anonymity.
  - In these attacks background knowledge of relationships between quanti-identifiers is used to reduce anonymity.
  - High dimensionality means lots of quasi-identifiers.



identifiers [20]. Furthermore, k-anonymization completely fails on high-dimensional datasets [2], such as the Netflix Prize dataset and most real-world datasets of individual recommendations and purchases.

Source: Narayanan and Shmatikov 2008[20]

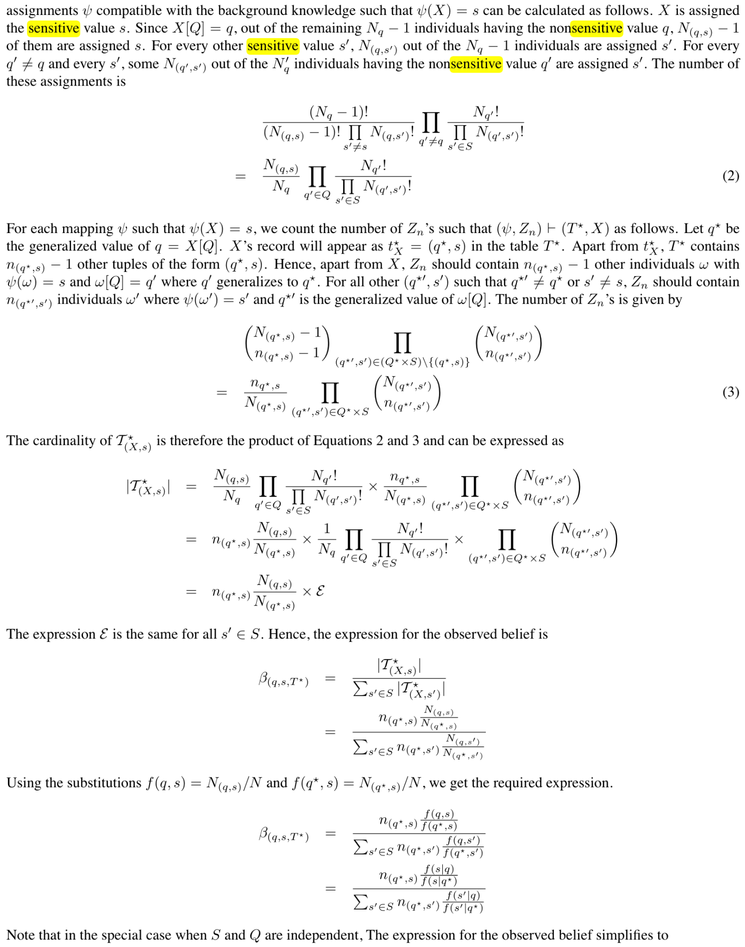
l-diversity

l-diversity (I)

l-diversity is a more sophisticated measure of the anonymity of sensitive variables in an anonymised dataset - introduced by Machanavajjhala et al in 2006[21]

This method depends on a Bayesian model of background knowledge.

We are **not** going to cover Bayesian statistics. Read statswithr.github.io/book/ if you're interested.



assignments ψ compatible with the background knowledge such that $\psi(X) = s$ can be calculated as follows. X is assigned the sensitive value s. Since $X[Q] = q$, out of the remaining $N_q - 1$ individuals having the nonsensitive value q, $N_{(q,s)} - 1$ of them are assigned s. For every other sensitive value s', $N_{(q,s')}$ out of the $N_q - 1$ individuals are assigned s'. For every $q' \neq q$ and every s', some $N_{(q',s')}$ out of the N'_q individuals having the nonsensitive value q' are assigned s'. The number of these assignments is

$$\begin{aligned} & \frac{(N_q - 1)!}{(N_{(q,s)} - 1)! \prod_{s' \neq s} N_{(q,s')}!} \prod_{q' \neq q} \frac{N_{q'}!}{\prod_{s' \in S} N_{(q',s')}!} \\ = & \frac{N_{(q,s)}}{N_q} \prod_{q' \in Q} \frac{N_{q'}!}{\prod_{s' \in S} N_{(q',s')}!} \end{aligned} \tag{2}$$

For each mapping ψ such that $\psi(X) = s$, we count the number of Z_n's such that $(\psi, Z_n) \vdash (T^\star, X)$ as follows. Let $q^\star$ be the generalized value of $q = X[Q]$. X's record will appear as $t^\star_X = (q^\star, s)$ in the table $T^\star$. Apart from $t^\star_X$, $T^\star$ contains $n_{(q^\star,s)} - 1$ other tuples of the form $(q^\star, s)$. Hence, apart from X, Z_n should contain $n_{(q^\star,s)} - 1$ other individuals ω with $\psi(\omega) = s$ and $\omega[Q] = q'$ where q' generalizes to $q^\star$. For all other $(q^{\star\prime}, s')$ such that $q^{\star\prime} \neq q^\star$ or $s' \neq s$, Z_n should contain $n_{(q^{\star\prime},s')}$ individuals ω' where $\psi(\omega') = s'$ and $q^{\star\prime}$ is the generalized value of $\omega[Q]$. The number of Z_n's is given by

$$\begin{aligned} & \binom{N_{(q^\star,s)} - 1}{n_{(q^\star,s)} - 1} \prod_{(q^{\star\prime},s') \in (Q^\star \times S) \setminus \{(q^\star,s)\}} \binom{N_{(q^{\star\prime},s')}}{n_{(q^{\star\prime},s')}} \\ = & \frac{n_{q^\star,s}}{N_{(q^\star,s)}} \prod_{(q^{\star\prime},s') \in Q^\star \times S} \binom{N_{(q^{\star\prime},s')}}{n_{(q^{\star\prime},s')}} \end{aligned} \tag{3}$$

The cardinality of $\mathcal{T}^\star_{(X,s)}$ is therefore the product of Equations 2 and 3 and can be expressed as

$$\begin{aligned} |\mathcal{T}^\star_{(X,s)}| &= \frac{N_{(q,s)}}{N_q} \prod_{q' \in Q} \frac{N_{q'}!}{\prod_{s' \in S} N_{(q',s')}!} \times \frac{n_{q^\star,s}}{N_{(q^\star,s)}} \prod_{(q^{\star\prime},s') \in Q^\star \times S} \binom{N_{(q^{\star\prime},s')}}{n_{(q^{\star\prime},s')}} \\ &= n_{(q^\star,s)} \frac{N_{(q,s)}}{N_{(q^\star,s)}} \times \frac{1}{N_q} \prod_{q' \in Q} \frac{N_{q'}!}{\prod_{s' \in S} N_{(q',s')}!} \times \prod_{(q^{\star\prime},s') \in Q^\star \times S} \binom{N_{(q^{\star\prime},s')}}{n_{(q^{\star\prime},s')}} \\ &= n_{(q^\star,s)} \frac{N_{(q,s)}}{N_{(q^\star,s)}} \times \mathcal{E} \end{aligned}$$

The expression $\mathcal{E}$ is the same for all $s' \in S$. Hence, the expression for the observed belief is

$$\begin{aligned} \beta_{(q,s,T^\star)} &= \frac{|\mathcal{T}^\star_{(X,s)}|}{\sum_{s' \in S} |\mathcal{T}^\star_{(X,s')}|} \\ &= \frac{n_{(q^\star,s)} \frac{N_{(q,s)}}{N_{(q^\star,s)}}}{\sum_{s' \in S} n_{(q^\star,s')} \frac{N_{(q,s')}}{N_{(q^\star,s')}}} \end{aligned}$$

Using the substitutions $f(q,s) = N_{(q,s)}/N$ and $f(q^\star,s) = N_{(q^\star,s)}/N$, we get the required expression.

$$\begin{aligned} \beta_{(q,s,T^\star)} &= \frac{n_{(q^\star,s)} \frac{f(q,s)}{f(q^\star,s)}}{\sum_{s' \in S} n_{(q^\star,s')} \frac{f(q,s')}{f(q^\star,s')}} \\ &= \frac{n_{(q^\star,s)} \frac{f(s|q)}{f(s|q^\star)}}{\sum_{s' \in S} n_{(q^\star,s')} \frac{f(s'|q)}{f(s'|q^\star)}} \end{aligned}$$

Note that in the special case when S and Q are independent, The expression for the observed belief simplifies to

l-diversity (II)

In order to estimate l-diversity we need to once again consider the structure of our dataset.

We split our data into:

- Sensitive variables - the medical variables.
- Keys - the non-sensitive variables.

A dataset is l-diverse if for each unique combination of key attributes there are at least *l* “well-represented” values for each sensitive variable.

Non-sensitive attributes **Keys**		**Sensitive**
region	**age_range**	**disease**
England	20-30	Heart
England	20-30	Heart
England	20-30	Heart
England	20-30	Pancreatic
England	20-30	Pancreatic
England	20-30	Pancreatic
Wales	40-50	Liver
Wales	40-50	Liver
Wales	40-50	Liver
Wales	40-50	Liver

l-diversity (III)

> A dataset is l-diverse if for each unique combination of key attributes there are at least *l* “well-represented” values for each sensitive variable.

For very simple datasets we can calculate l-diversity by hand.

However, for real-world applications there are 3 different methods for estimating “representative” values:

- Distinct l-diversity. This is the most commn and simplest method, it’s what we’ve just used. It requires there are l distinct values.
- Entropy l-diversity. This is a more sophisticated measure and goes beyond the scope of this lecture.
- Recursive l-diversity. This is a compromise between the two methods.

region	age_range
England	20-30
England	20-30
England	20-30
England	20-30
England	20-30
England	20-30
Wales	40-50
Wales	40-50
Wales	40-50
Wales	40-50

📝 Task: Calculate l-diversity

SLIDE 1 OF 2

1. Install the `{sdcMicro}` package
2. Add this dataset to your `.Rmd`

```r
data_diseases <- tibble(
  name = c("Saindhavi", "Enio", "Daury", "Alphus", "Balian",
           "Kenyea", "Gracielynn", "Aliye", "Kadince", "Asaph"),
  region = c(rep("England", 6), rep("Wales", 4)),
  age_range = c(rep("20-30", 6), rep("40-50", 4)),
  disease = c(rep("Heart", 3), rep("Pancreatic", 3), rep("Liver", 4))
)
data_diseases
```

```
# A tibble: 10 × 4
   name       region  age_range disease
   <chr>      <chr>   <chr>     <chr>
 1 Saindhavi  England 20-30     Heart
 2 Enio       England 20-30     Heart
 3 Daury      England 20-30     Heart
 4 Alphus     England 20-30     Pancreatic
 5 Balian     England 20-30     Pancreatic
 6 Kenyea     England 20-30     Pancreatic
 7 Gracielynn Wales   40-50     Liver
 8 Aliye      Wales   40-50     Liver
 9 Kadince    Wales   40-50     Liver
10 Asaph      Wales   40-50     Liver
```

📝 Task: Calculate l-diversity

SLIDE 2 OF 2

1. Compute the `ldiversity()` of the `disease` attribute

```r
ld_diseases <- data_diseases %>%
  mutate(disease = as_factor(disease)) %>%
  createSdcObj(keyVars = c("region", "age_range")) %>%
  ldiversity(ldiv_index = "disease")
```

1. Extract the l-diversity value

```r
ld_diseases@risk$ldiversity
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    1.0     1.0     2.0     1.6     2.0     2.0 
```

Summarising k-anonymity and l-diversity

Summarising k-anonymity and l-diversity

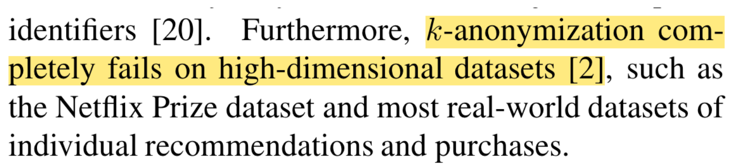
k-anonymity has well known vulnerabilities for high-dimensional datasets.

However. It is still worthwhile establishing at least 5-anonymity[20] in released datasets containing sensitive attributes.

---

l-diversity is a much more sophisticated tool that provides stronger privacy protections. It's verifiable NP-hard to re-identify individuals in l-diverse datasets.

However, l-diversity does not guarantee against re-identification. The coarsening of data might also degrade the usability of released data.



identifiers [20]. Furthermore, k-anonymization completely fails on high-dimensional datasets [2], such as the Netflix Prize dataset and most real-world datasets of individual recommendations and purchases.

Source: Narayanan and Shmatikov 2008[]

> Even though a minimum k value of 3 is often suggested, a common recommendation in practice is to ensure that there are at least five similar observations (k = 5)[19]

Case Study: Netflix Prize Dataset

Netflix Prize Dataset: What was it? (I)

In October 2006 Netflix created a competition with the intention of improving their recommendation engine[22].

Netflix sought an algorithm/methodology that would improve their recommendation algorithm.

> Let’s get into this a little bit more.

> Netflix provided over 100 million ratings (and their dates) from over 480 thousand randomly-chosen, anonymous subscribers on nearly 18 thousand movie titles. The data were collected between October, 1998 and December, 2005 and reflect the distribution of all ratings received by Netflix during this period. The ratings are on a scale from 1 to 5 (integral) stars.

user	movie	date_of_grade	grade
132	Alien	2006-01-01	5
132	Aliens	2006-01-02	4

> It withheld over 3 million most-recent ratings from those same subscribers over the same set of movies as a competition qualifying set.

Netflix Prize Dataset: What was it? (II)

Netflix had an algorithm called **Cinematch** which attempted to predict a user's rating of film **X** based on the user's existing movie ratings.

The intention was for **Cinematch** to provide more personalised recommendations than simply using the average rating for movie **X** across all users.

Netflix chose to use the root mean squared error (RMSE) of the **Cinematch** and **all other user ratings** compared to a user's actual rating as a measure of accuracy.

The **Cinematch** algorithm was roughly 10% more accurate than the **all other user rating**

The competition challenged participants to further improve this accuracy by at least an addition 10%.

Netflix Prize Dataset: What was it? (III)

The competition was extremely popular and it wasn't won until 2009.

In fact, it's quite a dramatic story with two winning entries submitted within a day of one another in 2009.

I'd recommend reading this thrillist.com article by Dan Jackson[23].

However.

16 days later Arvind Narayanan and Vitaly Shmatikov demonstrated their ability to re-identify users from the dataset[24].

This draft paper was finally published in 2008[20] and we'll look into how the re-identification was possible.

Netflix Prize Dataset: Privacy Breach (I)

With the announcement of a 2nd competition in 2009 a class-action suit was filed.

The suit described the Netlix Prize dataset as the biggest “voluntary privacy breach to date”. The Federal Trade Commission (FTC) also got involved.

Unfortunately, the case was settled privately so we don’t know the damages. The best I can get is this quote from a deleted blogpost[25]:

> In the past few months, the Federal Trade Commission (FTC) asked us how a Netflix Prize sequel might affect Netflix members’ privacy, and a lawsuit was filed by KamberLaw LLC pertaining to the sequel. With both the FTC and the plaintiffs’ lawyers, we’ve had very productive discussions centered on our commitment to protecting our members’ privacy. We have reached an understanding with the FTC and have settled the lawsuit with plaintiffs. The resolution to both matters involves certain parameters for how we use Netflix data in any future research programs.
>
> In light of all this, we have decided to not pursue the Netflix Prize sequel that we announced on August 6, 2009.” - Neil Hunt, Chief Product Officer @ Netflix

> “To some, renting a movie such as Brokeback Mountain or even The Passion of the Christ can be a personal issue that they would not want published to the world.”
>
> Jane Doe, a lesbian, who does not want her sexuality nor interests in gay and lesbian themed films broadcast to the world, seeks anonymity in this action

Netflix Prize Dataset: Re-identification (I)

The exact mechanics of the algorithm behind Arvind Narayanan and Vitaly Shmatikov[20] re-identification attack are beyond the scope of this course.

We're going to walk through the mechanics of the attack.

It's important to identify this paper introduces **a robust statistical de-anonymisation of large sparse datasets** that is not unique to the Netflix dataset.

> With 8 movie ratings (of which we allow 2 to be completely wrong) and dates that may have a 3-day error, 96% of Netflix subscribers whose records have been released can be uniquely identified in the dataset.
>
> Source: arXiv pre-print 2006[24]

Netflix Prize Dataset: Re-identification (II)

The de-anonymisation attack is powered by two data sources:

- IMDB Ratings
- Date of movie review

The authors have written an extremely useful FAQ[26].

IMDB provides a large, public database of movie ratings.

An assumed similarity between Netflix and IMDB ratings provides a re-identification attack vector.

The paper provides a proof of concept attack using **only 50 IMDB** users.

Despite this, the authors are confident of positively cross-matching at least 2 users between the datasets.

Netflix Prize Dataset: Re-identification (II)

Netflix Prize Dataset: Re-identification (II)

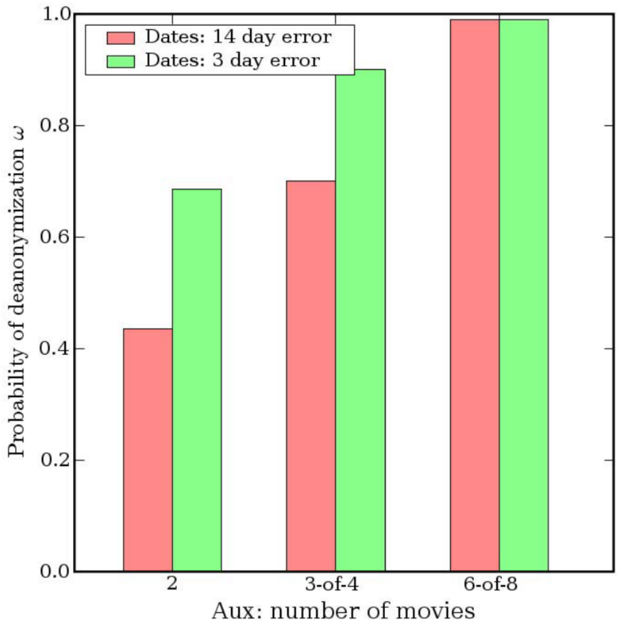
The de-anonymisation attack is powered by two data sources:

- IMDB Ratings
- Date of movie review

The authors have written an extremely useful FAQ[26].

Date of movie rating provides an additional attack vector.

Background knowledge might include account creation date, providing an attack vector for that user.



Selected other case studies

Selected other case studies

We’re going to look at a few additional case studies.

Please note that for most of these examples we are not explicitly talking about re-identification of users from de-anonymised datasets. Instead these case studies often breach privacy - sometimes publicly.

All of these case studies can be used in the data anonymisation section of your assessment. And please do share additional case studies with the group.

Remember I mentioned https://www.wired.co.uk/article/my-identity-for-sale as being a good source of additional case studies.

Case Study: Facebook beacon

Facebook has a long and awful history of privacy breaches and questionable activity.

Facebook Beacon is one of the oldest examples, all the way back from 2007.

It’s a rare example where Mark Zuckerberg open talks about it as a mistake[27].

> Discuss why Facebook Beacon breaches user privacy

Beacon was designed to automatically post purchases to your friend’s Facebook feeds.

The company originally claimed the service was “opt-in”, but there was clear evidence this was not true.

However, after criticism Facebook provided an opt-out. Users were required to turn off the service.

A class action suit in 2009 was settled for $9.5million.

Case Study: Google Buzz

Google Buzz was a very short lived social networking tool:

- Launched: February 9, 2010[28]
- Discontinued: December 15, 2011[29]

It shut down explicitly due to privacy violations and Google settled for $8.5million[30] **within one month** of the service launching.

Google automatically created **public** Google Profile pages using the Buzz service.

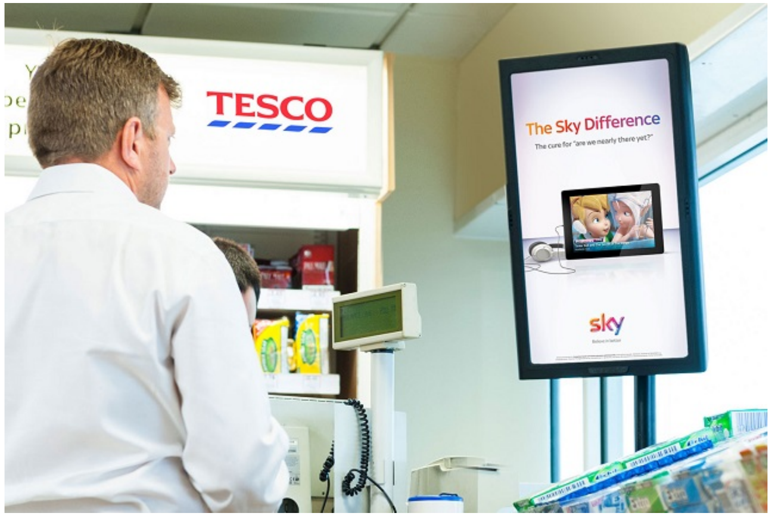
These **public** pages disclosed who the user most frequently communicated with via email or chat within GMail.

This was as designed. It wasn't a mistake. Google designed this product like this!

> When you first enter Google Buzz, to make the startup experience easier, we may automatically select people for you to follow based on the people you email and chat with most. Similarly, we may also suggest to others that they automatically follow you. You can review and edit the list of people you follow and block people from following you.
>
> **Your name, photo, and the list of people you follow and people following you will be displayed on your Google profile, which is publicly searchable on the Web**. You may opt out of displaying the list of people following you and who you're following on your profile.

Case Study: In-store screens

> Do you know what these machines are for in Tesco?



Case Study: In-store screens

These devices have embedded cameras that can track customer gaze - a proxy for the their attention.

The video feed from these devices is processed by software by Quividi that can estimate[31]:

- Gender
- Age
- Mood

Their privacy page makes for interesting reading.

> “Quividi’s software employs advanced facial detection software, not facial recognition technologies.” Quividi marketing

Anonymisation software?

Anonymisation software? (I)

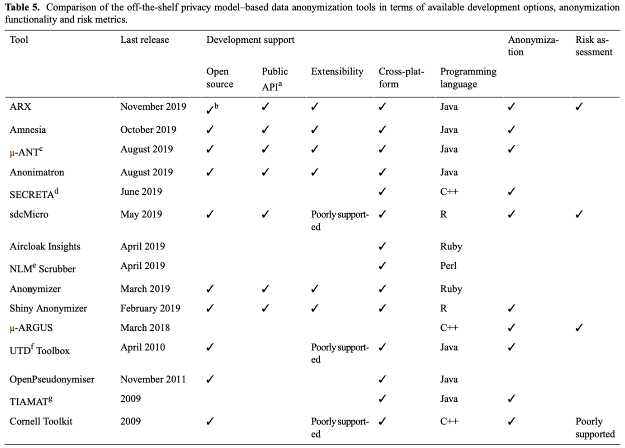


Table 5. Comparison of the off-the-shelf privacy model–based data anonymization tools in terms of available development options, anonymization functionality and risk metrics.

Tool	Last release	Development support					Anonymization	Risk assessment
		Open source	Public API[a]	Extensibility	Cross-platform	Programming language		
ARX	November 2019	✓[b]	✓	✓	✓	Java	✓	✓
Amnesia	October 2019	✓	✓	✓	✓	Java	✓	
µ-ANT[c]	August 2019	✓	✓	✓	✓	Java	✓	
Anonimatron	August 2019	✓	✓	✓	✓	Java		
SECRETA[d]	June 2019				✓	C++	✓	
sdcMicro	May 2019	✓	✓	Poorly supported	✓	R	✓	✓
Aircloak Insights	April 2019				✓	Ruby		
NLM[e] Scrubber	April 2019				✓	Perl		
Anonymizer	March 2019	✓	✓	✓	✓	Ruby		
Shiny Anonymizer	February 2019	✓	✓	✓	✓	R	✓	
µ-ARGUS	March 2018					C++	✓	✓
UTD[f] Toolbox	April 2010	✓		Poorly supported	✓	Java	✓	
OpenPseudonymiser	November 2011	✓			✓	Java		
TIAMAT[g]	2009				✓	Java	✓	
Cornell Toolkit	2009	✓		Poorly supported	✓	C++	✓	Poorly supported

In 2021 Zuo et al[32] performed a systematic review of de-anonymisation tools in digital health care.

There were two off-the-shelf tools identified that were built with R

- `{sdcMicro}`
  - This package contains many tools for measuring/exploring the anonymity of datasets via a {shiny} app.
- `{ShinyAnonymizer}
  - This package provides a {shiny} app for anonymising healthcare data

Anonymisation software? (II)

However. There isn’t a one-size fits all methodology or guarantee of privacy through anonymisation.

k-anonymity and l-diversity are useful metrics and provide some assurance of privacy. But background knowledge attacks might undermine these.

We need to be vigilant and careful when:

- Preparing data for release
- Designing services

Simulating fake datasets

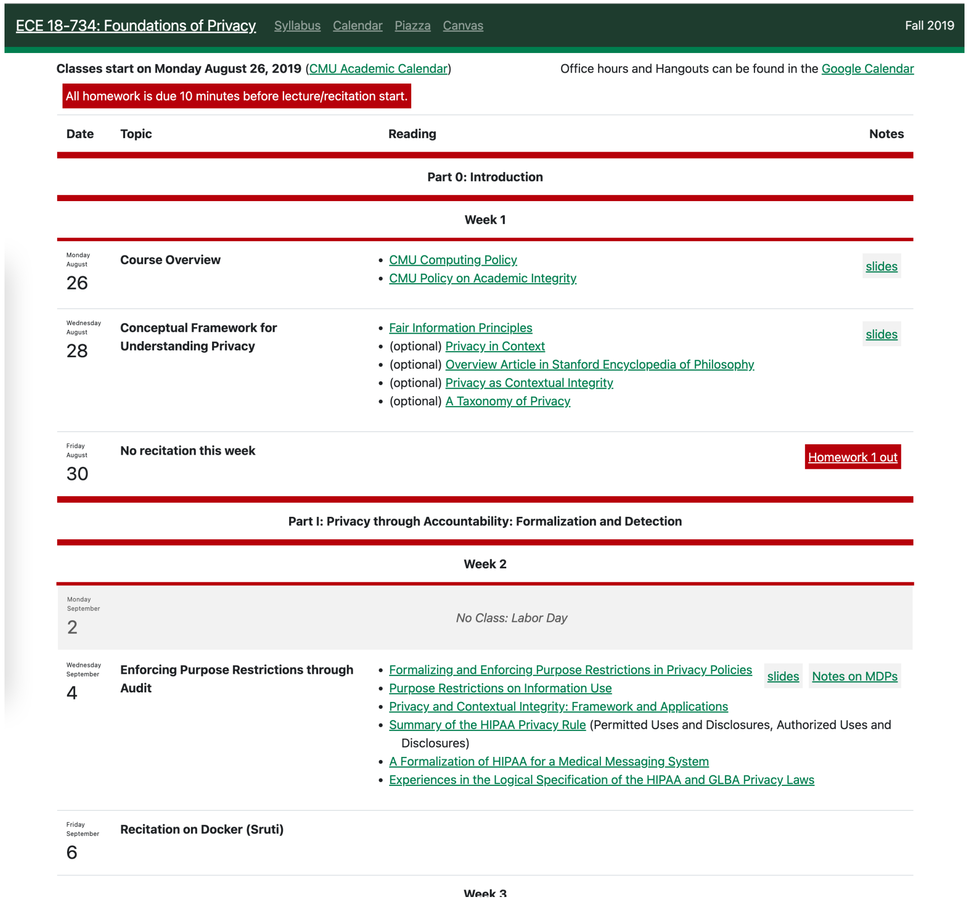
Simulating fake datasets

We've used the `{wakefield}` package to create fake datasets.

The `{faux}` package is a more sophisticated package for simulating datasets designed by Lise DeBruine.

Useful resources

Useful resources (I)



ECE 18-734: Foundations of Privacy Syllabus Calendar Piazza Canvas Fall 2019

Classes start on Monday August 26, 2019 (CMU Academic Calendar) Office hours and Hangouts can be found in the Google Calendar

All homework is due 10 minutes before lecture/recitation start.

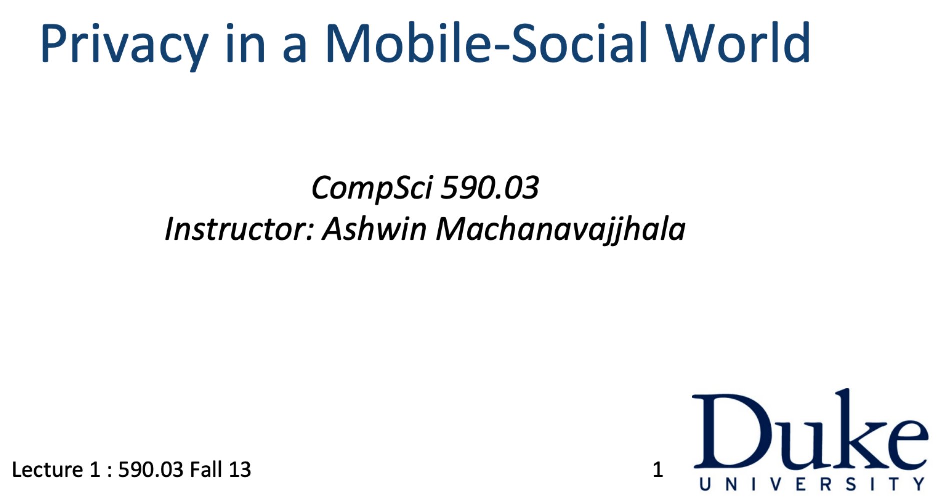
Date	Topic	Reading	Notes
Part 0: Introduction			
Week 1			
Monday August 26	Course Overview	• CMU Computing Policy • CMU Policy on Academic Integrity	slides
Wednesday August 28	Conceptual Framework for Understanding Privacy	• Fair Information Principles • (optional) Privacy in Context • (optional) Overview Article in Stanford Encyclopedia of Philosophy • (optional) Privacy as Contextual Integrity • (optional) A Taxonomy of Privacy	slides
Friday August 30	No recitation this week		Homework 1 out
Part I: Privacy through Accountability: Formalization and Detection			
Week 2			
Monday September 2	*No Class: Labor Day*		
Wednesday September 4	Enforcing Purpose Restrictions through Audit	• Formalizing and Enforcing Purpose Restrictions in Privacy Policies • Purpose Restrictions on Information Use • Privacy and Contextual Integrity: Framework and Applications • Summary of the HIPAA Privacy Rule (Permitted Uses and Disclosures, Authorized Uses and Disclosures) • A Formalization of HIPAA for a Medical Messaging System • Experiences in the Logical Specification of the HIPAA and GLBA Privacy Laws	slides Notes on MDPs
Friday September 6	Recitation on Docker (Sruti)		
Week 3			

This is an **excellent** module on privacy from Carnegie Mellon University.

You can find all lecture materials and even exercises here: https://course.ece.cmu.edu/~ece734/fall

This quickly becomes very technical.

Useful resources (I)



Privacy in a Mobile-Social World

CompSci 590.03
Instructor: Ashwin Machanavajjhala

Lecture 1 : 590.03 Fall 13 1

Ashwin Machanavajjhala has an excellent course about privacy in a mobile world - https://courses.cs.duke.edu/fall13/comps

This quickly becomes very technical.

Assessment

Assessment

In this section you must explain:

- What is an open dataset?
- What is “health data”?
- Why is it important to government, industry and academia that health datasets are made available?
- What are some of the benefits to individuals and groups in making health datasets open?
- What is data anonymisation and why is it important?
- What are the dangers to individuals and groups in health data being data deanonymised?
- What are some steps that can be taken to reduce the danger of deanonymisation?

> In answering these questions, you must include details of at least two case studies about data deanonymisation. Include as much technical information as possible about how the data was deanonymized.

References

1. Kretschmer, M., Pennekamp, J. & Wehrle, K. Cookie Banners and Privacy Policies: Measuring the Impact of the GDPR on the Web. *ACM Transactions on the Web* **15**, 1–42 (2021).
2. European Union. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. (2016).
3. Cookies, the GDPR, and the ePrivacy Directive. *GDPR.eu* (2019).
4. Englehardt, S. & Narayanan, A. Online Tracking: A 1-million-site Measurement and Analysis. in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* 1388–1401 (Association for Computing Machinery, 2016). doi:10.1145/2976749.2978313.
5. Habib, H., Li, M., Young, E. & Cranor, L. "Okay, whatever": An Evaluation of Cookie Consent Interfaces. in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* 1–27 (Association for Computing Machinery, 2022). doi:10.1145/3491102.3501985.
6. Spiegel, J. R., McKenna, M. T., Lakshman, G. S. & Nordstrom, P. G. Method and system for anticipatory package shipping. (2013).
7. Nyckel, E.-M. Ahead of Time: The Infrastructure of Amazon's Anticipatory Shipping Method. in *Media Infrastructures and the Politics of Digital Time* (eds. Volmar, A. & Stine, K.) 263–278 (Amsterdam University Press, 2021). doi:10.2307/j.ctv1xcxr3n.18.
8. Viet, N. Q., Behdani, B. & Bloemhof, J. Data-driven process redesign: Anticipatory shipping in agro-food supply chains. *International Journal of Production Research* **58**, 1302–1318 (2020).
9. Venkataramanan, M. Madhumita Venkataramanan: My identity for sale. *Wired UK* (2014).
10. UK Government. National Health Service Act 2006. (2006).
11. UK Government. Data Protection Act 2018. (2018).
12. Information Commissioner's Office. *Anonymisation: Managing data protection risk code of practice*. (2012).
13. UK Government. Equality Act 2010. *legislation.gov.uk* (2010).
14. Sweeney, L. Weaving Technology and Policy Together to Maintain Confidentiality. *The Journal of Law, Medicine & Ethics* **25**, 98–110 (1997).
15. Sweeney, L. Simple Demographics Often Identify People Uniquely. 0 Bytes (2000) doi:10.1184/R1/6625769.V1.
16. Sweeney, L. K-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10**, 557–570 (2002).
17. Sweeney, L. Only You, Your Doctor, and Many Others May Know. *Technology Science* (2015).
18. Samarati, P. & Sweeney, L. *Protecting Privacy when Disclosing Information: K-Anonymity and Its Enforcement through Generalization and Suppression*. (1998).
19. El Emam, K. & Dankar, F. K. Protecting Privacy Using k-Anonymity. *Journal of the American Medical Informatics Association : JAMIA* **15**, 627–637 (2008).
20. Narayanan, A. & Shmatikov, V. Robust De-anonymization of Large Sparse Datasets. in *2008 IEEE Symposium on Security and Privacy (sp 2008)* 111–125 (2008). doi:10.1109/SP.2008.33.

21. Machanavajjhala, A., Gehrke, J., Kifer, D. & Venkitasubramaniam, M. L-diversity: Privacy beyond k-anonymity. in *22nd International Conference on Data Engineering (ICDE'06)* 24–24 (2006). doi:10.1109/ICDE.2006.1.
22. Bennet, J. & Lanning, S. The Netflix Prize. (2007).
23. Jackson, D. The Netflix Prize: How a $1 Million Contest Changed Binge-Watching Forever. *Thrillist* (2017).
24. Narayanan, A. & Shmatikov, V. How To Break Anonymity of the Netflix Prize Dataset. (2006) doi:https://arxiv.org/abs/cs/0610105v1.
25. Waxman, S. Whoops! Netflix Gets Caught by FTC, Cancels Contest. (2010).
26. Narayanan, A. & Schmatikov, V. "How to Break Anonymity of the Netflix Prize Dataset" - FAQ. (2007).
27. Zuckerberg, M. Our Commitment to the Facebook Community. *Meta* (2011).
28. Ho, E. Google Buzz in Gmail. *Official Gmail Blog* (2010).
29. Google Blog. A fall sweep. *Official Google Blog* (2011).
30. BuzzClassAction.com. Google Buzz User Privacy Litigation Class Action Settlement Website. (2010).
31. Quividi. CONSUMER Privacy - Quividi - Insightful data with fully anonymous measurements. *Quividi* (2022).
32. Zuo, Z. *et al.* Data Anonymization for Pervasive Health Care: Systematic Literature Mapping Study. *JMIR Medical Informatics* **9**, e29871 (2021).