Taxis, Tonics and Tokens

Turning the Alchemy of Deidentification into a Science.

Sarah Jamie Lewis

OPEN PRIVACY
RESEARCH SOCIETY

Executive Director, Open Privacy Research Society

Before:

- Independent Privacy & Anonymity Researcher & Book Publisher (Queer Privacy)
- Automated Systems Fraud / Security @ *Amazon*
- Computer Science Researcher @ *<Redacted> (British Government)*

What is Deidentification?

“**Deidentification** is the process used to prevent a person's identity from being connected with information.”

Deidentification is the process used to prevent a person's identity from being connected with information.

“Deidentification is the process used to prevent a person's identity from being connected with information”.

What is Anonymization?

Anonymization is the process of removing identity AND context from data such that it can NEVER be re-identified.

How much information do you need to uniquely identify a person?

Case Study: New York Taxis

https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1

Back in 2014...

A FOIA request resulted in a complete historical trip and fare logs from NYC taxis

173 million individual trips. Each trip record includes the pickup and dropoff location and time, *anonymized* hack licence number and medallion number and other metadata.

New York Taxi Dataset

```
medallion,hack_license,vendor_id,rate_
code,store_and_fwd_flag,pickup_datetim
e,dropoff_datetime,passenger_count,tri
p_time_in_secs
```

One Very Busy Taxi

```
CFCD208495D565EF66E
7DFF9F98764DA
```

Lesson:Data isn’t Random

In NYC, taxi licence numbers are 6-digit, or 7-digit numbers starting with a 5. **Only 3M possible taxi licence numbers.**

Similarly, medallion numbers conform to a very specific pattern:

```
[0-9][A-Z][0-9]{2}. For example: 5X55

[A-Z]{2}[0-9]{3} For example: XX555

[A-Z]{3}[0-9]{3}For example: XXX555
```

Bad Data

md5(0) =
CFCD208495D565EF66E7DFF9F98764DA

Only 22m possible hashes!

Complete Deanonymization!

“Modern computers are fast: so fast that computing the 22M hashes took less than 2 minutes”

Tokenization

When the underlying data is easy to guess, **hashing the data isn't sufficient to ensure deidentification.**

This applies to: ID numbers, email addresses, addresses, names, dates, locations...pretty much anything.

Takeaway: Hashes over constructed data are weak!

It Gets Worse…

…but we will come back later.

Is Tokenization a solution?

What is Tokenization?

Encrypt the attribute with a secret key.

OR

HMAC the attribute with a secret key.

OR

Associate the attribute with a randomly generated value.

WARNING

Tokenization preserves Structure!

A Tokenized Dataset is still vulnerable to pattern of life analysis if there is more than 1 entry per individual.

Simple Demographics Often Identify People Uniquely

Latanya Sweeney

1990 U.S. Census Summary

“**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}**.

53% (132 million of 248 million) are likely to be **uniquely identified by only {place, gender, date of birth}.**”

Simple Demographics Often Identify People Uniquely

https://dataprivacylab.org/projects/identifiability/paper1.pdf

Case Study: Netflix

Back in 2006...

Netflix launched the Netflix Prize, a competition to predict user ratings for films, based on previous ratings

100,480,507 ratings that **480,189 users** gave to **17,770 movies**

Netflix Prize Dataset

Movie ID, Customer ID, Rating, Movie Title, Year of Release, Date of Ratin

Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan & Vitaly Shmatikov

“How much does the adversary need to know about a Netflix subscriber in order to identify her record if it is present in the dataset, and thus learn her complete movie viewing history?”

8 movie ratings & dates they occurred! Even less when considering movies that are not blockbusters.

Lesson: Your Data isn’t the Only Data

Users who rated movies on Netflix also might rate the movies on IMDb, often under their real name.

They might talk about them on Facebook or Twitter.

Correlating datasets is very powerful

And for the most part, completely out of your control....

Case Study Revisited: NYX Taxis, Celebrities & Gentlemen's Clubs

Auxiliary Datasets

“celebrities in taxis in Manhattan in 2013”

Ability to reconstruct the journeys of celebrities by matching photographs of Taxis and timestamps with the dataset!

Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset

https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/

Lesson: Data isn’t Neutral

By isolating journeys from outside a particular “gentleman's” club in the early hours of the morning to less-dense areas of the city - it’s possible to infer the identities of those visiting the club!

This works in Reverse too!!

Once you have worked out the identity of someone one way, you can now map everywhere else they might have frequented!

Does Deidentification Actually Work?

"[Deidentificaion] is somewhat useful as an added safeguard, but it is not robust against near-term future re-identification methods. PCAST does not see it as useful basis for policy. Unfortunately [deidentificaton] is already rooted in law.".

PCAST. "Report to the President - Big Data and Privacy: A technological perspective"

Taxis, Tonics and Tokens

Turning the Alchemy of Deidentification into a Science.

Taxis, Tonics and Tokens

Turning the Alchemy of Deidentification into the Science of Anonymization.

Threat Models

Identity Disclosure Attack: When an adversary is able to assign an entity to an entry in your dataset.

Attribute Disclosure Attack: When an adversary is able to learn new information about an entity that they know is present in the dataset.

Many Laws ONLY require protection from Identity Disclosure Attacks

(Be Better than that, Please!)

Is k-anonymity a solution?

What is k-anonymity?

“A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least **k-1** individuals whose information also appear in the release.”

Samarati, Pierangela; Sweeney, Latanya (1998). "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression" . Harvard Data Privacy Lab.

What is k-anonymity?

Suppression: If a field in a dataset cannot be generalized (e.g name) then it is removed or redacted.

Generalization: Replace specific individual elements with broader categories until k-anonymity is reached (e.g. replace individual ages with an age range (25 becomes 25-30)

Samarati, Pierangela; Sweeney, Latanya (1998). "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression" . Harvard Data Privacy Lab.

Lesson: k-anonymized data can still be harmful

If I know John is 22 and in the dataset and all the 20-25 years olds in the dataset either have cancer or heart disease....then I know John has cancer or heart disease

Lesson: k-anonymity is really hard to achieve in relatively low-dimensional datasets

As the number of attributes increases the generalization needed to ensure uniqueness trends toward 100%

Making the dataset practically useless

Lesson: optimal k-anonymity is an **NP-hard** problem

Heuristical approaches exist but they are, by definition not perfect.

“Not perfect” is OK for packing a truck, not so great for individual privacy in a public dataset

Case Study: Australian Health Data

Australian de-identified open health dataset

Department of Health published the deidentified longitudinal medical billing records of **10% of Australians**, about **2.9 million people**.

You know where this is going….

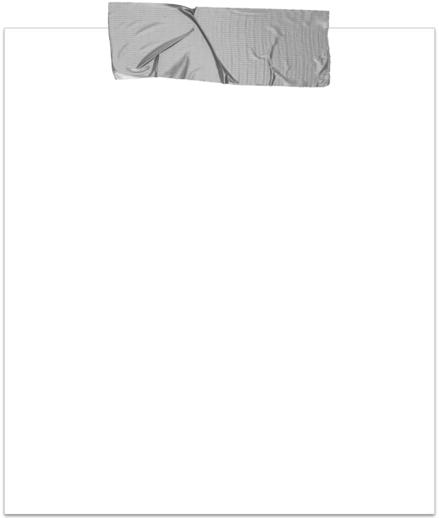
Health Data in an Open World (2017)

Chris Culnane, Benjamin I. P. Rubinstein, Vanessa Teague

Researcers used informaton of Child Birth patterns (which are often publicized in a number of ways)

Because young mothers (<18) and older mothers (>?) are rare, their data was trivial to reID

"In this paper we show that patients can be reidentified, without decryption, **by linking the unencrypted parts of the record with known information about the individual.**"



“In this paper we show that patients can be reidentified, without decryption, **by linking the unencrypted parts of the record with known information about the individual.**”

Location Data is very, very hard (i’d say impossible) to anonymize.

Does Dedentification Actually Work?

“The current evidence shows a high re-identification rate but is dominated by small-scale studies on data that was not de-identified according to existing standards. This evidence is insufficient to draw conclusions about the efficacy of de-identification methods.”

El Emam K, Jonker E, Arbuckle L, Malin B (2011) A Systematic Review of Re-Identification Attacks on Health Data. PLoS ONE 6(12): e28071. https://doi.org/10.1371/journal.pone.0028071

Is Aggregation a solution?

What is Aggregation?

Summarizing data at a high level e.g. averages and summations.

Answer: Not Really

Lesson: Aggregation is Vulnerable to Differential Attacks

We can gain information through changing the sets we are aggregating and noting the changes to the output.

Unless your aggregation sets are very large, and the degree of freedom is small - aggregation is very vulnerable.

Example: Average Income

If we know the average income of a group of people,and then know that someone leaves the group and take the average again - we gain information.

Is Differential Privacy a solution?

What is Differential Privacy?

Adding “noise” to an **aggregate query result** to protect individual entries **without significantly changing the result.**

Algorithms aim to ensure that an **attacker can learn nothing more about an individual than they would learn if that person’s record were absent from the dataset.**

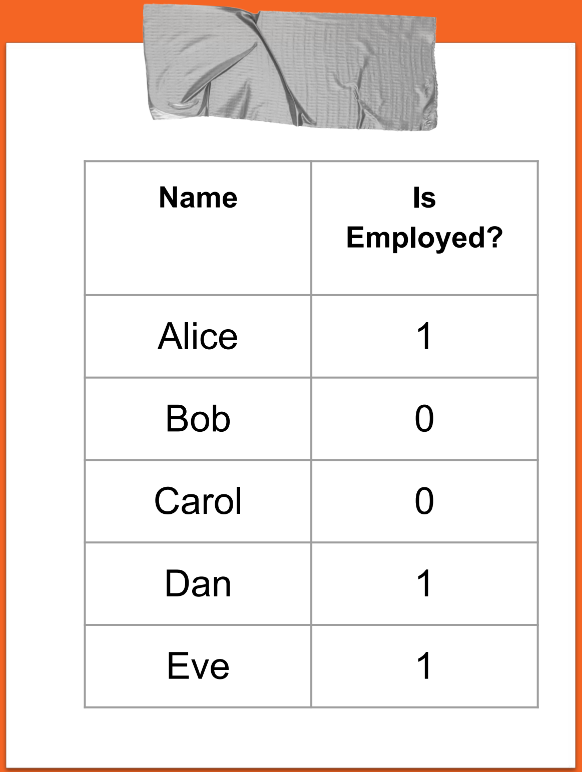
We are not going to go too deep into the maths!

Maths Ahead!

Differential Privacy aims to **maximize the accuracy of queries** from statistical databases **while minimizing the chances of identifying its records.**

$$\Pr[\mathcal{A}(D_1) \in S] \leq e^{\epsilon} \times \Pr[\mathcal{A}(D_2) \in S],$$

Where $\mathcal{A}$ is a randomized algorithm, and S is a subset of the image of A.



Name	**Is Employed?**
Alice	1
Bob	0
Carol	0
Dan	1
Eve	1

Differential Privacy isn’t an Algorithm

It is what we use to measure the privacy properties of a randomized algorithm; **it is not an algorithm itself.**

Critically: Methods used to protect one type of data may not be appropriate for another kind of data. There are no true off-the-shelf solutions.

Differential Privacy for Dummies

https://github.com/frankmcsherry/blog/blob/master/posts/2016-02-03.md

Is Differential Privacy a solution?

Maybe

Aggregation & Deidentification is NOT Anonymization

—

OK, Sarah...This is all great, but wtf do I actually do?

1. Do I really need to release this data? Can I publish aggregate statistics instead?

2. Release the minimum subset of the data required.

3. Remove Location Information. If you can.

4. Apply k-anonymity to suppress and generalize fields that could uniquely identify a participant.

5. Tokenize data that isn’t needed for reporting (e.g id numbers).

6. Actively attempt reidentificaton attacks and evaluate the results.

The End!

Open Privacy Research Society is a non-profit dedicated to researching and building privacy-enhancing technologies that benefit marginalized communities.

Please support our work:
https://openprivacy.ca/donate

OPEN PRIVACY
RESEARCH SOCIETY