
PrivE: privacy evaluation of synthetic data

Florimond Houssiau, in collaboration with James Jordon, Callum Mole, Camila Rangel-Smith, James Geddes, Andrew Elliott, Lukasz Szpruch

A brief history of privacy-preserving data analysis

Data is collected at (very large) scale



British Airways fined £20m for data breach affecting 400,000 customers

The fine is far smaller than an initial estimate of £183m

First person identified from AOL Data: Thelma Arnold

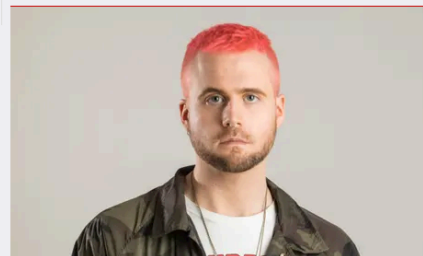
Michael Arrington @arrington?lang=en / 1:56 PM GMT+1 • August 9, 2006

On Sunday the news broke that AOL purposefully released 20 million partially anonymized search queries. On Monday AOL apologized, and later that evening the first web interface to the data went up.



The Cambridge Analytica Files

Key stories



Revealed / 50 million Facebook profiles harvested for Cambridge Analytica in major data breach

Traditional approach: de-identification

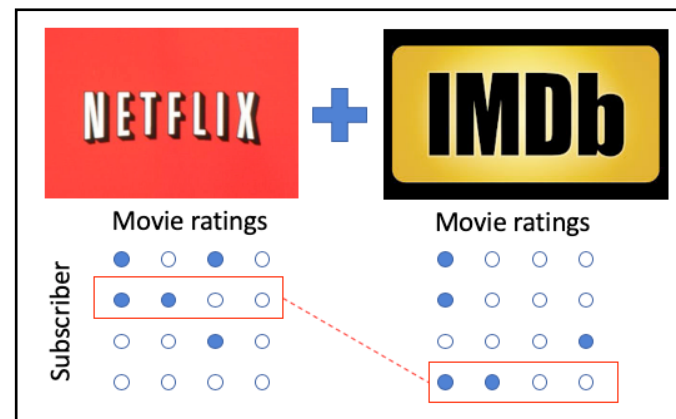
Make data “less identifiable”

- Remove identifiers,
- Coarsen data,
- Remove some entries/columns,
- Data swapping, etc.

But in many cases, this doesn't work:

- **Re-identification attacks.**

Name	Age	Postcode
f118abc2	20-30	W14***
13abf1h2	40-50	NW1***
e42eacb8	60-70	M11***



Ohm, P., 2009. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA L. Rev.*, 57, p.1701.
Narayanan, A. and Shmatikov, V., 2008, May. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)* (pp. 111-125). IEEE.
Image from <https://ai.plainenglish.io/ahh-the-computer-algorithm-still-can-find-you-even-there-is-no-personal-identifiable-information-6e077d17381f>

Synthetic Data: evading Re-identification attacks

Key promise: no 1-1 link, so no re-identification

Quite some enthusiasm:

Synthetic data are free of privacy issues when well designed and quality checked. Primary (e.g. population) microdata cannot be shared. Anonymized data and data aggregates lose too much information. Pseudonymized data are prone to de-anonymization attacks.

Synthetic data **change everything from privacy to governance** and need a serious research invested in order to understand, pilot and implement them into a normal policy cycle.

But: does synthetic data provides utility + privacy?

How do we measure privacy?

Option 1: Differential Privacy

“Gold standard” definition of privacy:

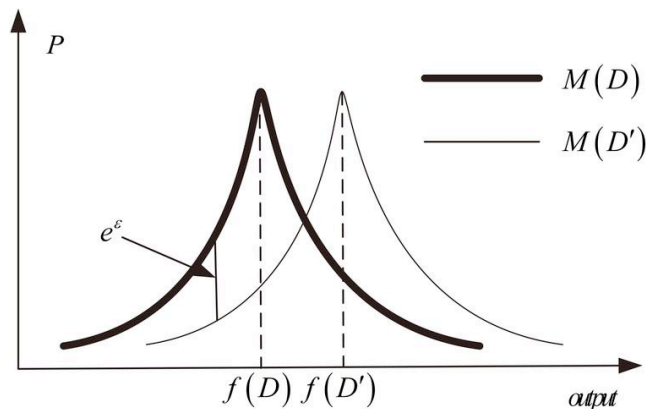
- Mathematical definition.
- Very robust against attacks.
- Many useful properties.

But not perfect:

- Tricky to implement with good utility.
- Magic parameters (ϵ, δ) .
- Bugs can invalidate guarantees!

A randomized algorithm $M : \mathcal{X}^n \rightarrow \mathcal{Y}$ is (ϵ, δ) -differentially private if, for all $x, x' \in \mathcal{X}^n$ differing on a single entry and all measurable $E \subseteq \mathcal{Y}$, we have

$$\mathbb{P}[M(x) \in E] \leq e^\epsilon \cdot \mathbb{P}[M(x') \in E] + \delta.$$



Dwork, C., 2008, April. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation* (pp. 1-19). Springer, Berlin, Heidelberg.

How do we measure privacy?

Option 2: Adversarial Approaches

Idea: use **attacks** to (try and) detect information leakage.

Pros:

- Easy to explain / analyse / understand.
- Can analyse diverse setups + assumed attackers.
- Can detect implementation bugs.

Cons:

- (Almost) always a lower bound on the risk.

Groundhog attack (Stadler et al.)

First black-box attack against SDG

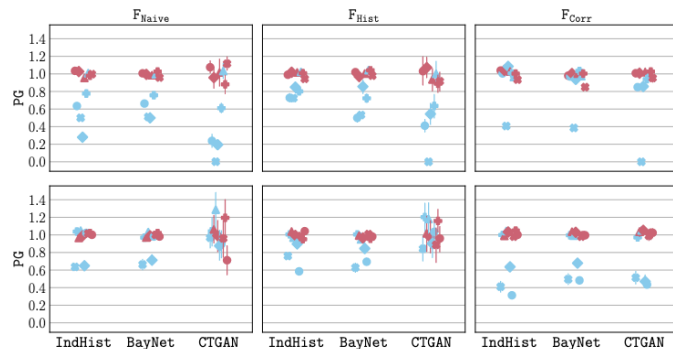
Uses shadow modelling + simple features.

Results:

- Many common SDGS have bugs that can be exploited.
- Outliers are more vulnerable.

Limitations:

- Only tabular data.
- Computationally expensive.
- Very simple attack.



P(rivacy)G(ain) for different SDG and versions of the attack F_* . PG=1 means that the synthetic data protects privacy, PG=0 means that it is equivalent to revealing the original data. Blue points are outliers, red points are inliers.

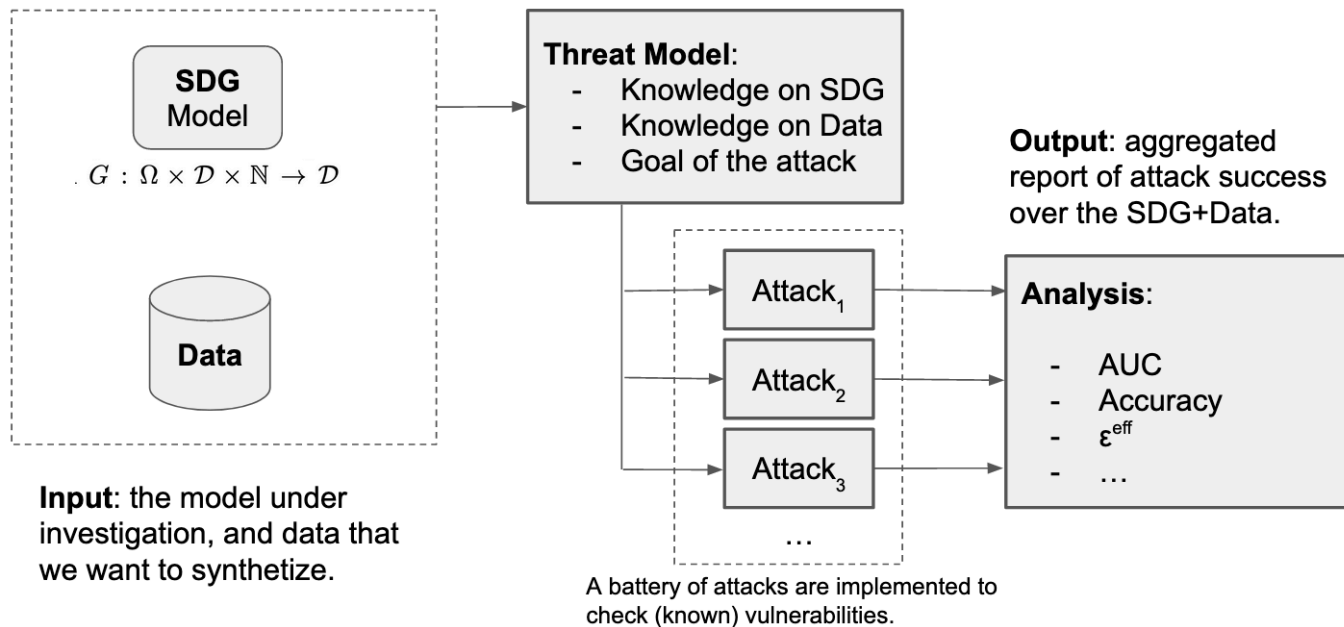
PrivE: A toolbox for adversarial evaluation

PrivE: A toolbox for adversarial evaluation

Goal: build a general toolbox of attacks to evaluate SDG

- General purpose:
 - Diverse data formats: time-series, networks, ...
 - Range of threat models for real-world situations.
 - Allowing *any* SDG model.
- Open source, (soon) open contributions:
 - SDG model developers (esp. at Turing).
 - Attack developers.

Components of PrivE



Research possibilities


- Developing new attacks, especially for non-tabular data.
- Studying different threat models:
 - Reconstruction attacks.
 - Different types of auxiliary knowledge.
 - ...
- Theoretical work:
 - Statistical significance.
 - Optimal attacks?

$$= \frac{1}{C_0} h_0(x)$$

$$\int \frac{h_0(x)}{h_\psi(x)} p_\psi(x) dx$$

$$-\frac{1}{n} \sum \frac{h_0(x_i)}{h_\psi(x_i)}$$

Thank you for your attention!
Don't hesitate to reach out at fhoussiau@turing.ac.uk



$$\sum_{i=1}^n \left(\frac{f(x_i)}{h_\psi(x_i)} - \frac{f(x_i)}{h_0(x_i)} \right)$$

$$A T = 0$$

$$T \geq 0$$