pyCANON: A Python library to check the level of anonymity of a dataset

JUDITH SÁINZ-PARDO DÍAZ (sainzpardo@ifca.unican.es)

Instituto de Física de Cantabria (IFCA), CSIC-UC

Motivation

The unstoppable improvements in data analysis techniques for knowledge extraction and decision-making make necessary the evolution of techniques for the secure publication of data. Moreover, the need for collaboration between different institutions, research centers or companies makes it necessary to be able to share data with certain security guarantees. There are numerous attacks that can be carried out on databases: re-identification, linkage, skewness and semantic attacks among others. For this the implementation of *pyCANON*, a Python library and CLI that can be used to know the level of anonymity of a dataset (and thus publish or share it while being aware of the risks involved), is presented. Nine different techniques will be used for this purpose.

Key concepts

- **Identifiers:** variables that allow a person to be unequivocally identified (e.g. name, ID number, email).
- **Quasi-identifiers (QI):** set of variables accessible to an attacker that allow to identify an individual (e.g. city, sex, age).
- **Sensitive attributes (SA):** variables that contain information that must not be disclosed (e.g. diseases, political opinion).
- **Equivalence class (EC):** partition of a database in which all the quasi-identifiers have the same value. Users in the same EC are all indistinguishable with respect to the QI.

Name	Age	Sex	ZIP code	Stroke
Alice	32	F	28105	Yes
Bob	37	M	28305	No
Charles	44	M	28520	Yes
Dianne	25	F	28025	No

Table 1. Simple example of a database with one identifier, three QI and one SA.

Name	Age	Sex	ZIP code	Stroke
*	[25, 35)	F	28**5	Yes
*	[25, 35)	F	28**5	No
*	[35, 45)	M	28***	No
*	[35, 45)	M	28***	Yes

Table 2. Anonymized version of Table 1.

- **Linkage attack:** consists of combining at least two anonymized databases in order to reveal the identity of some individuals.
- **Re-identification attack:** occurs when the anonymization process is reversed.
- **Homogeneity attack:** occurs when all the values for a SA in an EC are identical.
- **Background knowledge attack:** the adversary has some foreknowledge about the target of the attack.
- **Skewness attack:** can occur when a SA is not really frequent in the whole database, but it is extremely frequent in an EC.
- **Similarity attack:** may occur when the values of a SA in an EC are different but semantically similar.
- **Inference attack:** consists of applying data mining techniques to extract information from a database.

Anonymity tools implemented

pyCANON allow to check the following nine anonymization techniques:

- **k-anonymity.**
- **(α,k)-anonymity.**
- **ℓ-diversity.**
- **Entropy ℓ-diversity** (more restrictive than ℓ-diversity).
- **Recursive (c,ℓ)-diversity.**
- **t-closeness.**
- **Basic β-likeness.**
- **Enhanced β-likeness** (more robust privacy than basic β-likeness).
- **δ-disclosure privacy.**

	Principal attack which prevents						
Technique	Linkage	Re-identif cation	Homogeneity	Background	Skewness	Similarity	Inference
k-anonymity	✓	✓					
(α,k)-anonymity	✓	✓	✓				
ℓ-diversity			✓	✓			
Entropy ℓ-diversity			✓	✓			
Recursive (c,ℓ)-diversity			✓	✓			
t-closeness					✓	✓	
Basic β-likeness					✓		
Enhanced β-likeness					✓		
δ-disclosure privacy					✓		✓

Table 3. Anonymization techniques and principal attacks that prevent.

Impact

GOAL: given a dataset, a list of QI and a list of SA, check for which parameters the aforementioned techniques are satisfied.

pyCANON helps to detect failures in the anonymization process.

Allows to know how different properties scale as a function of others. Example: evolution of *t* and log(β) when varying *k* (for *t-closeness, basic β-likenes* and *k-anonymity* respectively).

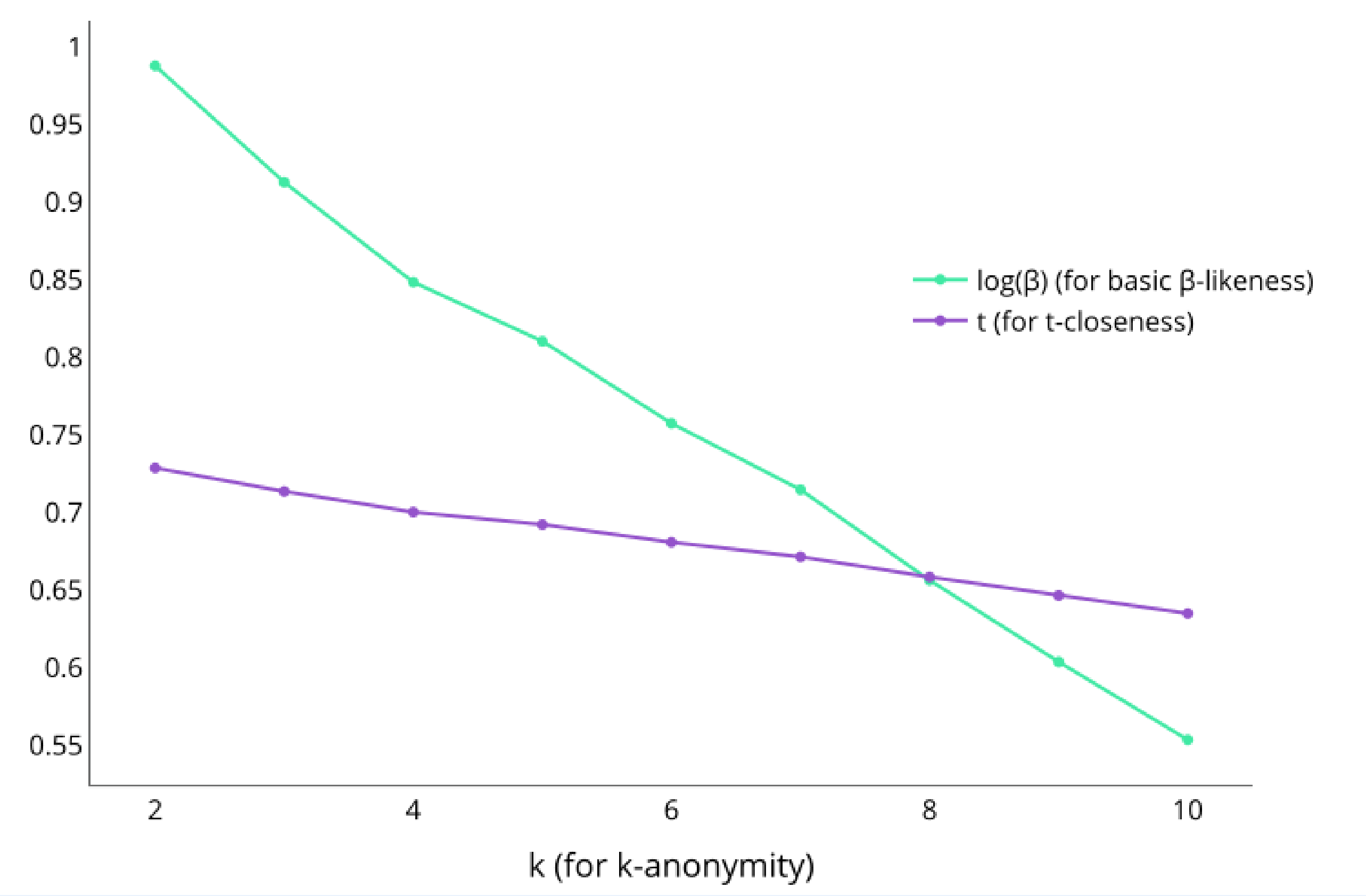

Figure 1. Use example of the *report* package of *pyCANON*.

Evaluate the level of anonymity depending on the selected QI.

Enable data to be published or shared with security guarantees.

Two approaches are allowed in case of more than one SA.

No previous knowledge of Python or anonymity techniques is required.

Customized reports in JSON and PDF formats can be easy obtained as in the following example:

```
import pandas as pd
from pycanon.report import pdf

FILE_NAME = "adult.csv"
DATA = pd.read_csv(FILE_NAME)
QI = ["age", "education", "occupation",
    "relationship", "sex", "native-country"]
SA = ["salary-class"]
FILE_PDF = "report_adult.pdf"
pdf.get_pdf_report(DATA, QI, SA, file_pdf = FILE_PDF)
```

Example code 1. Use example of the *report* package of *pyCANON*.

Documentation: https://pycanon.readthedocs.io/

PyPi (installation): https://pypi.org/project/pycanon/

Preprint: J. Sáinz-Pardo Díaz, Á. López García, *"pyCANON: A Python library to check the level of anonymity of a dataset"*, 2022, https://arxiv.org/abs/2208.07556

GitHub repository: https://github.com/IFCA/pycanon

Instituto de Física de Cantabria   CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS   UNIVERSIDAD DE CANTABRIA

Acknowledgements:

I would like to thank the funding through the European Commission - NextGenerationEU (Regulation EU 2020/2094) through CSIC's Global Health Platform (PTI Salud Global) and the support from the project AI4EOSC "Artificial Intelligence for the European Open Science Cloud" that has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement number 101058593.