SBA
Research

COMET

Competence Centers for
Excellent Technologies

Fingerprinting Relational Data

Bundesministerium
Klimaschutz, Umwelt,
Energie, Mobilität,
Innovation und Technologie

Bundesministerium
Digitalisierung und
Wirtschaftsstandort

FFG
Forschung wirkt.

wirtschafts
agentur
wien
Ein Fonds der
Stadt Wien

European
Commission

FWF
Der Wissenschaftsfonds.

netidee
OPEN INNOVATIONS

Agenda

- Background on fingerprinting relational data sets
  - Motivation
  - Requirements & workflow
- Current research directions
  - Fingerprinting categorical data
  - Evaluating fingerprinting schemes
- Challenges and future research

Fingerprinting Relational Data

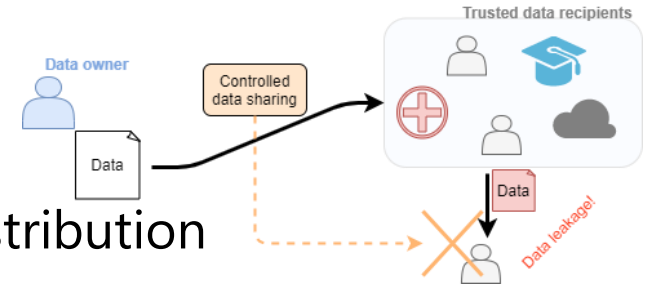
Background

Fingerprinting - motivation

- Why protecting the shared data?
  - Data owner used a lot of resources to collect/create the data (money, human experts, time…)
  - Sensitive data (e.g. medical data)
    - Privacy implications: only the trusted parties get the data and should not share it further

- Anonymising data? Reduces utility!

✓ Controlled data sharing → the goal
  - Share full data
  - Trace the unauthorised data re-distribution



Controlled data sharing via fingerprinting

- Embedding owner‘s signature and recipient‘s identification into the data

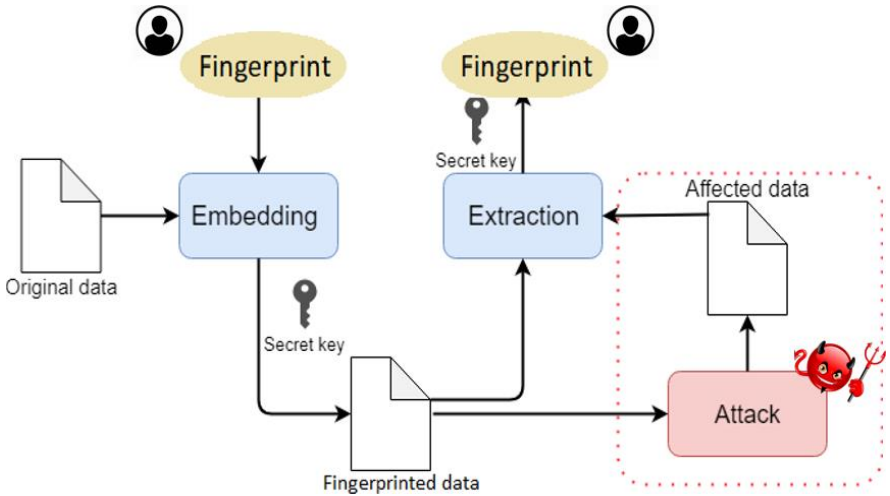
Age	Blood Pressure	Diabetes
32	64	1
31	66	0
50	72	1
48	70	0

Age	Blood Pressure	Diabetes
33	64	1
31	68	0
50	72	1
47	70	0

Fingerprinting: requirements & workflow

Requirements:

1. Recognisable by the owner
2. Not detectable and consequently removable by the recipients
3. Robust to the attacks
4. Does not change the utility of the data too much



Current research directions

Fingerprinting categorical data &

Evaluation of fingerprinting schemes

Challenge: categorical data

- *Minor* alterations in categorical data?

age	sex	employed	Alzheimers stage
59	Male	No	late
67	Male	No	late
…	…	…	…

age	sex	employed	Alzheimers stage
59	Male	Yes	late
68	Female	No	late
…	…	…	…

- kNN-based fingerprinting: „*modify the value to something that is likely to occur in the original dataseet*“ [1]

[1] Šarčević, Tanja, and Rudolf Mayer. "A Correlation-Preserving Fingerprinting Technique for Categorical Data in Relational Databases." *IFIP International Conference on ICT Systems Security and Privacy Protection*. Springer, Cham, 2020.

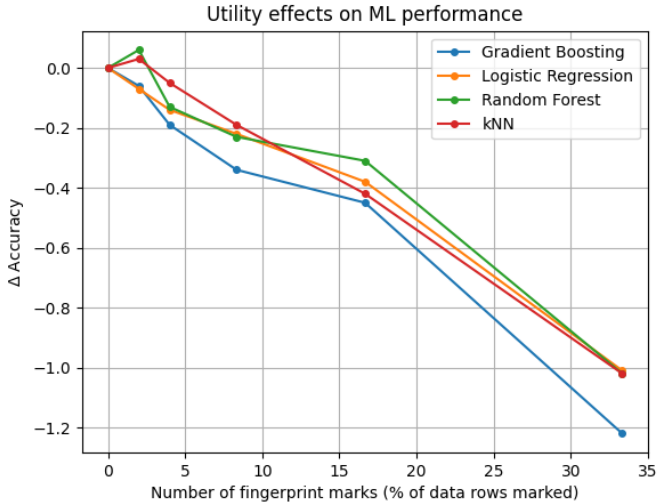
Quality estimation

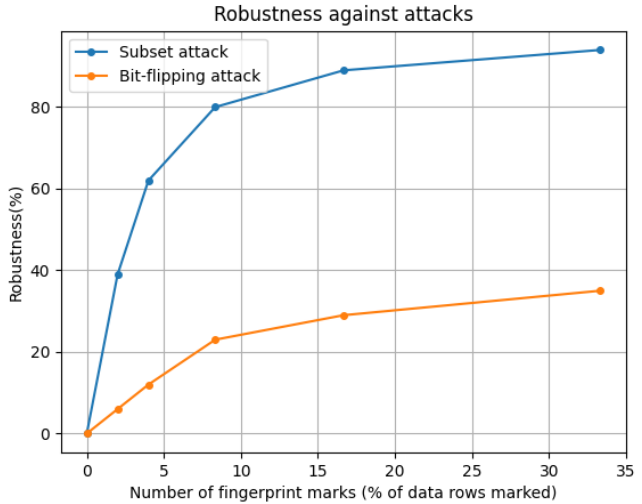
- Robustness against attacks
  - Attacker model: *white-box naive attacker*
  - Attacks: horizontal & vertical subset attack, flipping attack, …
- Utility of fingerprinted data
  - Statistical properties (mean, var, distribution, …)
  - **ML performance (accuracy, …)** [2]

[2] Šarčević, Tanja, and Rudolf Mayer. "An evaluation on robustness and utility of fingerprinting schemes." *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, Cham, 2019.

Quality estimation

- Robustness against attacks → **maximise** modifications
- Preserve data utility → **minimise** modifications





Quality estimation

- Data properties affect the robustness and utility of fingerprinted data, e.g. data size, type of attributes, entropy, etc.

Table Experimental results of **horizontal subset attack success** (fm) on the Forest Cover Type dataset (left; size: 581012 x 54) and Adult (right; size: 48842 x 14), where p' denotes the strength of the attack

$\%_marks$	$p' = 80\%$	$p' = 95\%$	$p' = 99\%$
17%	0	0	0.01
4%	0	0	1.0
2%	0	0.19	1.0
1%	0	0.99	1.0

$p' = 80\%$	$p' = 95\%$	$p' = 99\%$
0.20	0.95	1.0
0.99	1.0	1.0
1.0	1.0	1.0
1.0	1.0	1.0

[2] Šarčević, Tanja, and Rudolf Mayer. "An evaluation on robustness and utility of fingerprinting schemes." *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, Cham, 2019.

Summary & Future Challenges

Summary

- Fingerprinting wider range of attribute types – categorical data
- Effects of fingerprinting parameters on robustness and data utility
- Novel utility measurement – effects on a machine learning task

Future challenges

1. Designing tools for fingerprinting relational data in practice
   - Unifying fingerprinting scheme for relational data sets
   - Open-source implementation
   - OSSDIP – controlled data visiting infrastructure for sensitive data

2. Aiding the fingerprint parameter choice for the data holder
   - Bigger-scale analysis to capture pattern and trends in robustness and data utility depending on fingerprinting parameters
   - Analysis of effects of data properties on the quality of fingerprint
   - Parameter choice guidelines tailored for a data set and its intended usage scenario; (semi-)automated process

Competence Centers for Excellent Technologies

Tanja Šarčević

SBA Research

Floragasse 7, 1040 Vienna

tsarcevic@sba-research.org

[1] Šarčević, Tanja, and Rudolf Mayer. "A Correlation-Preserving Fingerprinting Technique for Categorical Data in Relational Databases." *IFIP International Conference on ICT Systems Security and Privacy Protection*. Springer, Cham, 2020.

[2] Šarčević, Tanja, and Rudolf Mayer. "An evaluation on robustness and utility of fingerprinting schemes "*International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 2019.

Bundesministerium Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie

Bundesministerium Digitalisierung und Wirtschaftsstandort

FFG Forschung wirkt.

wirtschaftsagentur wien Ein Fonds der Stadt Wien

European Commission

FWF Der Wissenschaftsfonds.

netidee OPEN INNOVATIONS