FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Diplomarbeitspräsentation

Fingerprinting Relational Databases; Quality Evaluation and Impact on Learning Tasks

Tanja Šarčević

Technische Universität Wien
Institut für Informationssysteme
Arbeitsbereich: Information & Software Engineering Group
Betreuer: Ao.Univ.-Prof. Dr. Dipl.-Ing. Andreas Rauber

Masterstudium:

Computational Intelligence

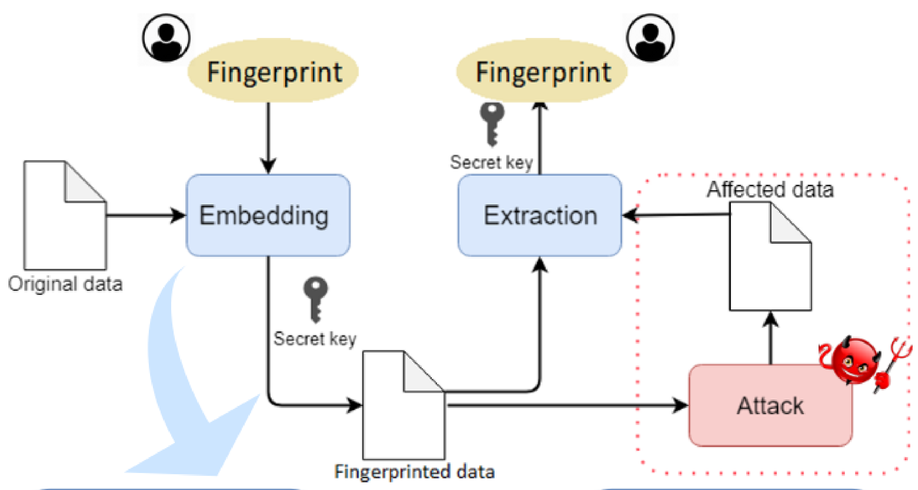
Motivation and Problem Statement

Fingerprinting digital data is an ownership attribution mechanism which can be seen as a personalised version of generic watermarking. The fingerprinting techniques generally embed a pattern in the data, i.e. they distort the original data set to a certain extent. The pattern identifies the owner and the receiver of the specific data copy. The type of data in the data set can be a crucial point for evaluating fingerprinting scheme effectiveness. Categorical data are shown to give rise to more problems with embedding the fingerprint compared to numerical data, yet the appropriate fingerprinting scheme for categorical data is necessary; otherwise, the domain of fingerprinting applications is very limited. We propose two techniques for categorical data and perform a detailed robustness and utility analysis for three known techniques.

Methodology

Literature Review → Design novel schemes → Implementation → Robustness Evaluation → Utility Evaluation → Comparison& Summary

General workflow and design of fingerprinting schemes



AK Scheme[1]

- Create the fingerprint
- For each tuple
  - Choose the tuple with probability 1/γ
  - Choose the attribute and LSB
  - Choose the fingerprint mark
  - Mark&update the data

Numerical data

Block Scheme[2]

- Create the fingerprint
- Create a binary image out of the data
- Divide the binary image into blocks of size βxβ
- For each block
  - Choose a bit to mark
  - Choose a fingerprint mark
  - Mark&update the data
- Decode the binary image

Fingerprint requirements

1. Recognisable by the owner
2. Non-detectable by the recipients
3. Robust to data modifications
4. Not lowering the utility of the data too much

Two-level Fingerprinting Scheme[3]

- Create a fingerprint
- For each tuple
  - Choose the tuple for marking with fingerprint bit with probability $1/\gamma_1$
  - Choose the tuple for marking based on the entire fingerprint with probability $1/\gamma_1$
  - Mark&update

Marking based on a pseudo-random choice

Number-encode categorical values → Apply fingerprinting scheme for numerical data → Correction → Decode to the categorical values

Categorical data

Marking based on the correlation between values

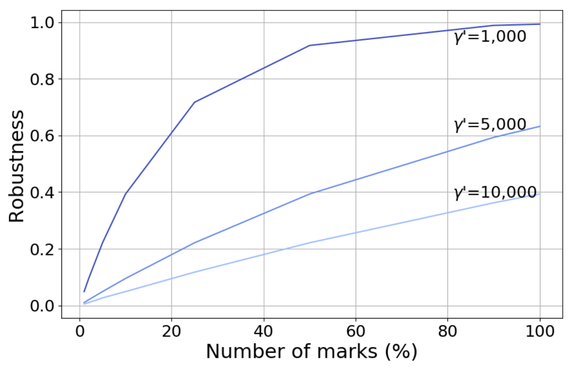
Number-encode categorical values → Pseudo-randomly choose the value for marking → Choose the new value from the set of neighbours of the tuple → Decode to the categorical values

- Blind scheme
- Resulting in possible semantic inconsistencies in the

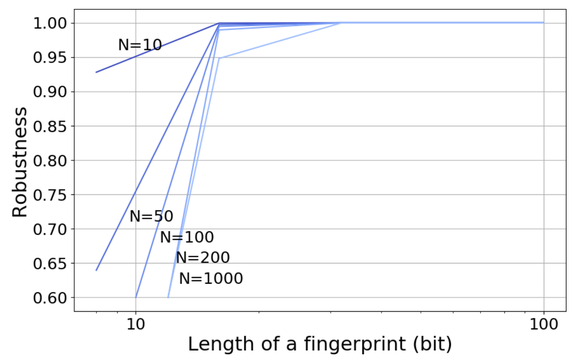
- Non-blind scheme
- Resulting data does not consist of additional outliers and impossible combinations of attribute values

Robustness evaluation

Additive attack

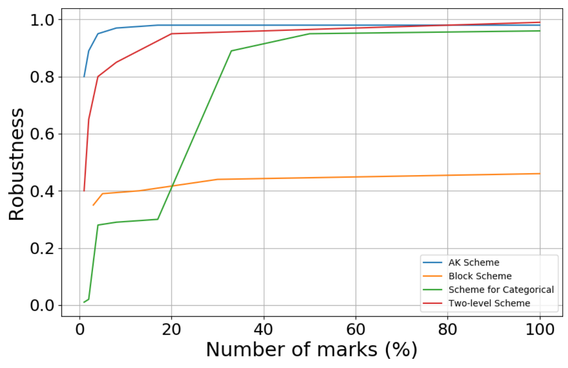


Misdiagnosis False Hit

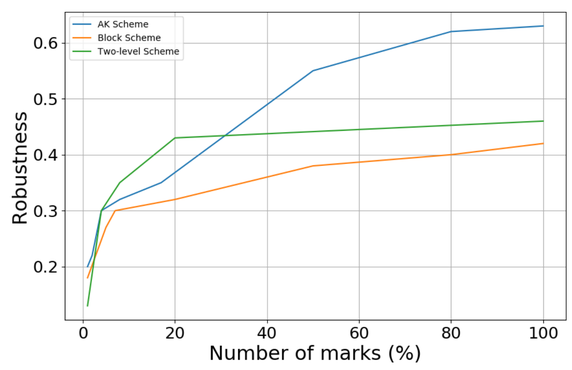


The schemes with less marks embedded in data are generally more susceptible to attacks. The main step for gaining robustness is choosing smaller fingerprint and more marks in a marking pattern.

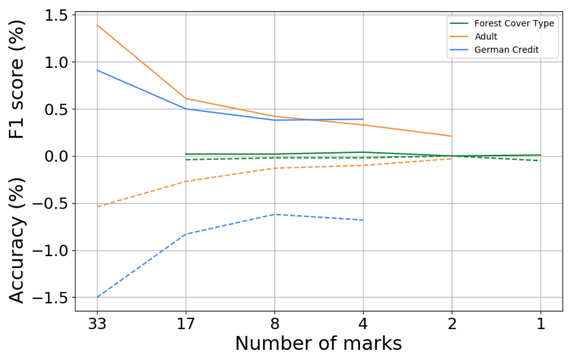
Subset attack



Bit-flipping Attack



Utility evaluation



EXPERIMENTAL SETUP

Datasets: Forest Cover Type, Adult (Census), German Credit Data
Classifiers: Decision Tree, Random Forest, Logistic Regression, Gradient Boosting, k-NN
Performance metrics: F_1 score, accuracy

The representative results with Random Forest show rather small performance decreases, up to 1.5%. The performance drop is bigger for datasets with more introduced marks, and for small datasets in general.

References

- Y. Li, V. Swarup, and S. Jajodia, "Fingerprinting relational databases: Schemes and specialties," 2005.
- S. Liu, S. Wang, R. H. Deng, and W. Shao, "A block oriented fingerprinting scheme in relational database," 2004.
- F. Guo, J. Wang, and D. Li, "Fingerprinting relational databases," 2006.
- T. Sarcevic and R. Mayer, "An evaluation on robustness and utility of fingerprinting schemes," 2019.

Kontakt: ta.sarcevic@gmail.com