Fingerprinting Relational Data Sets

Tanja Šarčević and Rudolf Mayer

COMET
Competence Centers for Excellent Technologies
www.ffg.at/comet
SBA Research

Problem & Motivation

Fingerprinting techniques, which can be seen as a personalized version of generic watermarks applied to a digital object, can be utilized as a mechanism enabling ownership attribution. They generally embed a pattern in the data, i.e., they distort the original data set to a certain extent. A good fingerprint should (i) be recognizable by the original owner of the data, (ii) not be detectable (and consequently, removable) by recipients of the data, (iii) be robust to intentional or unintentional modifications of the data, and (iv) not lower the utility of the data too much.
The type of data in the dataset can be the 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.

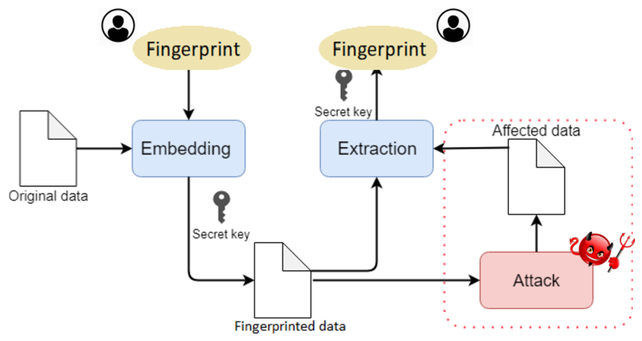

Figure 1: Fingerprinting workflow

Fingerprinting Numerical Data

- **AK Scheme**[2]: pseudo-random marking pattern
- **Block Scheme**[3]: binary image used as fingerprint information
- **Two-level Scheme**[4]: separate patterns for owner and the recipient

Utility Evaluation

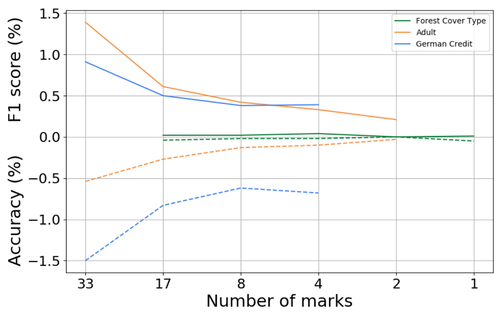

Figure 2: Classification performance of fingerprinted datasets

Data utility may be measured via its effect on machine learning model performance [5]. 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 as well as for small datasets.

Data Utility Under Malicious Attacks

The attacks are additionally decreasing dataset's utility. The analysis shows the decrease in utility of 5 different classifiers under attacks. The results show that modifying data such that the fingerprint is not likely to be extracted anymore, the data loses on its utility significantly [6].

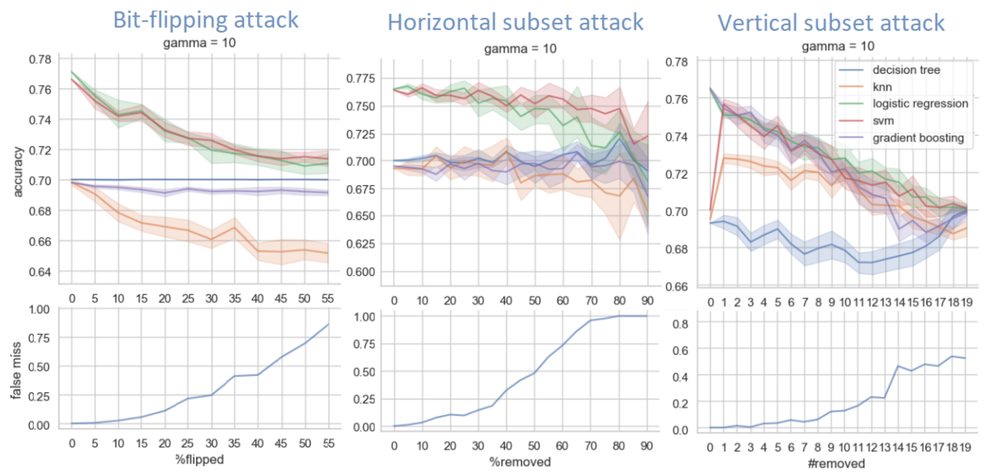

Figure 3: Data utility decrease by strengthening the attacks

Fingerprinting Categorical Data

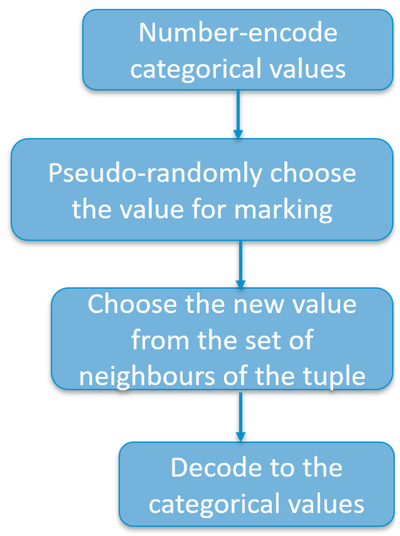

Figure 4: Fingerprinting scheme based on neighbourhood search

A novel scheme for fingerprinting categorical data in relational datasets is proposed in [1]. The scheme focuses on preserving the semantic relations between attributes, and thus limiting the perceptibility of marks, and the effects of the fingerprinting on the data quality and utility.

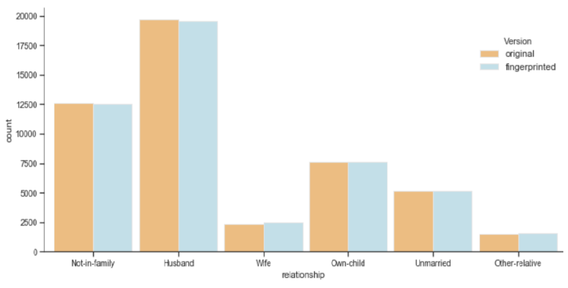

Figure 5: Distribution of a categorical attribute before and after fingerprinting

Robustness Evaluation

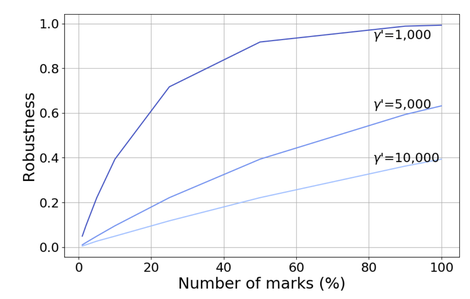

Figure 6: Additive attack

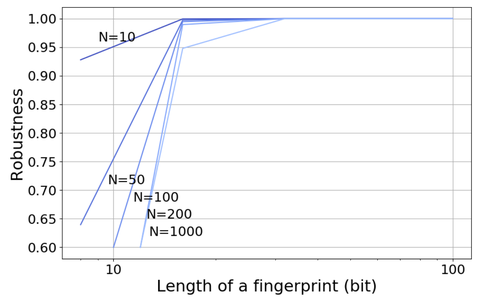

Figure 7: Misdiagnosis false hit

The schemes with less marks embedded in data are generally more susceptible to malicious attacks (actions on the dataset with the goal of removing the fingerprint). The main step for gaining robustness is choosing smaller fingerprint and embedding more marks.

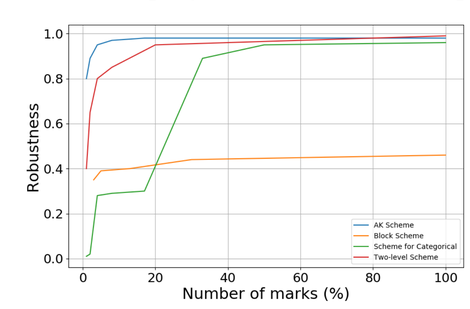

Figure 8: Subset attack

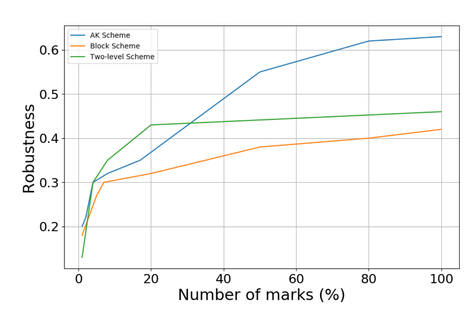

Figure 9: Bit-flipping attack

[1] Šarčević, T., Mayer, R.: A Correlation-Preserving Fingerprinting Technique for Categorical Data in Relational Databases. In 35th International Conference on ICT Systems Security and Privacy Protection (IFIP SEC 2020). Springer (2020).
[2] Li, Y., Swarup, V., Jajodia, S.: Fingerprinting relational databases: Schemes and specialities. IEEE Transactions on Dependable and Secure Computing 2(1), 34–45(2005).
[3] Liu, S., Wang, S., Deng, R.H., Shao, W.: A block oriented fingerprinting scheme in relational database. In: International Conference on Information Security and Cryptology. Springer (2004).
[4] Guo, F., Wang, J., and Li, D.: Fingerprinting Relational Databases. In ACM Symposium on Applied Computing (SAC). (2006).
[5] Šarčević, T., Mayer, R.: An Evaluation on Robustness and Utility of Fingerprinting Schemes. In Machine Learning and Knowledge Extraction: International Cross-Domain Conference (CD-MAKE). Springer (2019).
[6] Šarčević, T., Mayer, R.: Data Utility Assessment in Fingerprinted Datasets under Malicious Attacks. *Submitted for publication* (2020).

Bundesministerium Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie
Bundesministerium Digitalisierung und Wirtschaftsstandort
FFG Forschung wirkt.
wirtschaftsagentur wien Ein Fonds der Stadt Wien
FWF Der Wissenschaftsfonds.
netidee OPEN INNOVATIONS

SBA Research (SBA-K1) is a COMET Centre within the framework of COMET – Competence Centers for Excellent Technologies Programme and funded by BMK, BMDW, and the federal state of Vienna. The COMET Programme is managed by FFG.