Fingerprinting Relational Data Sets

Tanja Šarčević and Rudolf Mayer

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.

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

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.

Fingerprinting Categorical Data

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.

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].

Robustness Evaluation

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.