Primary Key-Free Watermarking: A Robust Framework for Ownership Protection in Databases

Xin Che, Qiqi Zhang, Lingyang Chu

Computing and Software
McMaster University
chex5,zhangq16,chul9@mcmaster.ca

1 Introduction

In the emerging data economy, high-quality tabular databases, such as those used in marketing and healthcare, represent highly valuable digital assets. These datasets are often the result of significant investment in data collection and cleaning. However, digital data can be easily copied, shared, or resold without authorization, which threatens copyright protection.

Watermarking provides an effective solution by embedding imperceptible but verifiable information into the data, enabling owners to assert and verify their rights even when unauthorized copies are distributed.

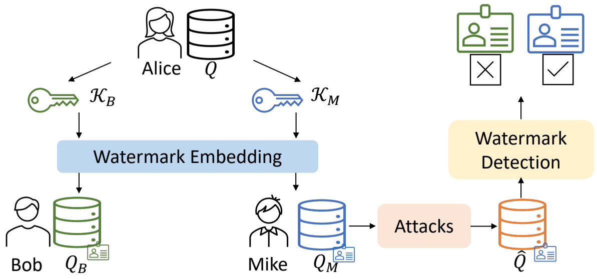

Figure 1: An application scenario of watermarking.

As shown in Figure 1, Alice owns a database Q, and she wants to sell it to Bob and Mike. Before selling Q, Alice uses two different secret keys $\mathcal{K}_B$ and $\mathcal{K}_M$ to embed two different watermarks in Q. This produces two watermarked datasets, denoted by Q_B and Q_M, which are sold to Bob and Mike, respectively. Mike uses attack to modify Q_M into an illegal copy $\hat{Q}$ and put $\hat{Q}$ on the market for sale. When Alice sees the suspicious dataset $\hat{Q}$ on the market, she uses each of $\mathcal{K}_B$ and $\mathcal{K}_M$ to detect watermark from $\hat{Q}$. The detection results show that the watermark linked to Mike's key but not Bob's, allowing Alice to identify its source and claim ownership.

Traditional database watermarking methods rely on primary key (PK) or virtual primary key (VPK) to organize and identify data records [1, 2]. However, these methods are vulnerable in machine learning contexts, since PKs can be modified or removed without affecting data utility, and VPKs are easily altered under attacks. Therefore, there is a pressing need for a primary key-free watermarking method that can ensure reliable ownership protection for databases without relying on key-based identifiers.

2 Method Overview

This presentation introduces a primary key-free (PK-free) watermarking framework for databases [1]. The key idea is to map tabular data instances into a discrete-time signal space and embed ownership information as a sinusoidal watermark. Because the embedding process treats the database as a whole rather than identifying individual records, it does not rely on primary keys or virtual keys, making it truly primary key-free. The proposed approach offers three main advantages: (1) primary key-free, ensuring robustness against key modification or removal; (2) blind detection [2], allowing watermark detection without access to the original dataset, thereby enhancing data confidentiality; and (3) robust to various attacks, such as noise addition, row or column deletion, PCA transformation, and re-watermarking.

By combining signal processing with database watermarking, this framework offers a practical and secure solution for database ownership protection in the era of AI-driven data trade.

References

[1] Xin Che, Mohammad Akbari, Shaoxin Li, David Yue, Yong Zhang, and Lingyang Chu. Primary key free watermarking for numerical tabular datasets in machine learning. In *International Conference on Pattern Recognition*, pages 254–270. Springer, 2024.

[2] Muhammad Kamran and Muddassar Farooq. A comprehensive survey of watermarking relational databases research. *arXiv preprint arXiv:1801.08271*, 2018.