Chapter 7.16
A Multiple-Bits Watermark for Relational Data

Yingjiu Li
Singapore Management University, Singapore

Huiping Guo
California State University at Los Angeles, USA

Shuhong Wang
University of Wollongong, Australia

ABSTRACT

At the heart of the information economy, commercially and publicly useful databases must be sufficiently protected from pirated copying. Complementary to the Database Protection Act, database watermarking techniques are designed to thwart pirated copying by embedding owner-specific information into databases so that the ownership of pirated copies of protected databases can be claimed if the embedded information is detected. This article presents a robust watermarking scheme for embedding a multiple-bits watermark to numerical attributes in database relations. The scheme is robust in the sense that it provides an upper bound for the probability that a valid watermark is detected from unmarked data, or a fictitious secret key is discovered from pirated data. This upper bound is independent of the size of the data. The scheme is extended to database relations without primary-key attributes to thwart attribute-related attacks. The scheme is also extended to multiple watermarks for defending additive attacks and for proving joint ownership.

INTRODUCTION

With the development of information technology, databases are becoming increasingly important in a wide variety of applications such as parametric specifications, surveys, and life sciences. While demand for the use of databases is growing, pirated copying has become a severe threat to such databases due to the low cost of copying and

the high values of the target databases. To fight against pirated copying, database watermarking techniques are designed to embed owner-specific information into database relations; when a pirated copy is found, the owner can extract the embedded information and use the detection process to assert the ownership of data. This complements the effort of the Database Protection Act (Vaas, 2003) as people realize that the law does not provide sufficient protection to valuable databases (Gray & Gorelick, 2004).

While watermarking multimedia data has long been rigorously studied (Cox, Miller, & Bloom, 2001; Johnson, Duric, & Jajodia, 2000; Katzenbeisser & Petitcolas, 2000), the approaches developed for multimedia watermarking cannot be directly applied to databases because of the difference in data properties. In general, database relations differ from multimedia data in significant ways and hence require a different class of information-hiding mechanisms. Unlike multimedia data whose components are highly correlated, database relations consist of independent objects or tuples. The tuples can be added, deleted, or modified frequently in either benign updates or malicious attacks. No existing watermarking techniques for multimedia data are designed to accommodate such tuple operations.

Perhaps the most well-known scheme for watermarking relational data is the one proposed by Agrawal and Kiernan (2002). For convenience, we call it the AK scheme. The main idea of the AK scheme is to change a small portion of numerical data according to a secret key such that the change can be detected for ownership proof. Without access to the secret key, a pirate cannot localize exactly where the change is made. It is difficult for a pirate to confuse the ownership detection unless he or she introduces an intolerable error to the underlying data. The AK scheme can be used in many real-world applications such as watermarking parametric specifications, surveys, and life-science data (Agrawal, Haas, & Kiernan, 2003; Agrawal & Kiernan).

Consider a database relation R that has a primary key P and ν numerical attributes $A_0,..., A_{\nu-1}$. Let there be η tuples. A portion of tuples is selected for embedding watermark information according to a control parameter γ ($\gamma < \eta$). The selection is also determined by a secret key K, known only to the owner of the data, as well as the primary key. Any tuple r is selected if $S_1(K, r.P) \bmod \gamma = 0$, where $S_1(K, r.P)$ is the first number generated by $S(K, r.P)$, and $S(K, r.P)$ is a cryptographic pseudorandom sequence generator seeded with a secret key K and the primary key $r.P$ of tuple r. Given a sequence of numbers $S_1, S_2,...$ generated by S, it is computationally infeasible to derive the secret key or to predict the next number in the sequence. Due to the uniqueness of the primary key, roughly one out of every γ tuples is selected for embedding watermark information.

For each selected tuple r, the AK scheme selects exactly one least significant bit j from attribute A_i and replaces it with a mark bit x, where $i = S_2(K, r.P) \bmod \nu$, $j = S_3(K, r.P) \bmod \xi$, and $x=0$ if $S_4(K, r.P)$ is even and $x=1$, otherwise. Here, ξ is another control parameter determining the range of least-significant bits of each value that may be modified.

For ownership detection, the mark bits are located using the same process provided that the secret key is known and the primary key remains unchanged. Let ω be the number of mark bits being localized ($\omega \approx \eta/\varphi$). To increase the robustness of the detection process, the ownership is claimed if more than $\tau\omega$ of the localized bits are as expected, where $\tau \in [0.5, 1)$ is a control parameter that is related to the assurance of the detection process.

The AK scheme has the following advantages: It is (a) key based, meaning all aspects of the scheme are determined by a secret key and a primary key, (b) blind, that is, the detection process does not require the knowledge of the original database or the embedded information, (c) incrementally updatable, where each tuple is marked independently of all other tuples, (d)