International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Effective Watermarking Techniques on Structured Datasets

Mini Joswin[1], Deeksha Bhardwaj[2]

[1]M.E, Department of Computer Science and Engineering, G.H.Raisoni College of Engineering & Technology, Wagholi, Pune-11, Maharashtra

[2]HoD & Professor, Department of Computer Science and Engineering, G.H.Raisoni College of Engineering & Technology, Wagholi, Pune-11, Maharashtra

Abstract: ***The data that floats on the web has a wide exposure to authorized and unauthorized users who may copy and reuse or manipulate the data. This can lead to serious breach in the security of the individual or organization who owns the data. Apart from unauthorized regeneration, the data may be subjected to various types of attacks. Watermarking serves as an effective measure to ensure ownership protection and attack resilience of digital data. In this paper we survey the current state-of-the-art methodologies advocated for watermarking relational databases and their effectiveness against different categories of data attacks.***

Keywords: Watermarking, Robust , Reversible , Usability constraints, Data attacks

1. Introduction

In today's era, data of different types and volume are generated profusely over a wide range of computing and non-computing devices. This data maybe broadly classified into structured, semi-structured and unstructured. Structured data account for 10% of the data floating on the web whereas semi-structured and unstructured data account for the remaining 90%. Structured data typically refer to formatted, constrained data represented in the form of tables and managed by Database Management Softwares. Oracle, mySQL, MS SQL Server and IBM DB2 are few of the most popularly used RDBMS. Semi-structured data, on the other hand, does not adhere to strict forms of representation and are constrained through the usage of predefined and user-defined tags. Document and content management systems fall into the category of technical solutions provided for semi-structured HTML and XML are the popularly used standards for developing semi-structured datasets. Unstructured datasets are wide spectrum spanning over scanned documents, emails, social media data, instant and short messages, online intellectual assets and so on. These demarcations in datasets though based on structural conformity are sometimes compatible with each other. There can be a significant paradigm shift in the outcomes of decision support systems if the data generated is appropriately shared and analyzed.

Sharing of digital contents faces serious issues of unauthorized redistribution, piracy, ownership claims, and data attacks which may lead to its modification or damage. To overcome the issues the owner can hide distinct information in the data being shared. The popular techniques of information hiding used are cryptography, steganography and watermarking. In cryptography, the original message is encrypted and thus locked using secret keys known only to the sender and receiver(s).In steganography, a stego-gramme is produced by embedding secret information in non-secret data. A watermark is also secret information embedded into the original message by the data owner. Considering the level data modification, we can say that cryptography changes the data into a non-readable format and thus the properties of the original data are altered considerably. Also the fact that the sender and receiver are in a secret communication may be revealed to an attacker who may then dedicatedly attempt to hack the message. Steganography, on the other hand, hides secret information such that it's indistinguishable to a third party and the secret communication thus remains hidden. Watermarking is subtly different from steganography as it is not only difficult to both detect and but also to remove. Also the existence of secret information is hidden in steganography and if this message is detected then steganography fails [10].

Various researches have been conducted on watermarking techniques. These techniques primarily depend on the data being watermarked as the attack is correlated to the type of dataset. Still images, video, audio, VLSI design were the types of data being watermarked till over a decade ago.[8] But with the infiltration of data mining and knowledge engineering, relational databases too have moved into the spotlight. Relational databases differ from multimedia data as it is independent and discrete and the latter is highly correlated and continuous [1]. Earlier techniques of watermarking were irreversible which aimed at embedding watermark into the cover data for ownership protection. A significant amount of change was brought about in the original data. Reversible watermarking on the other hand allows the recovery of the original data and the watermark as well as prove the ownership rights over the data. Reversible watermarking provide robustness, imperceptibility, high embedding capacity and readily retrieving capacity [9]. Watermarking techniques are classified as distortion-based and distortion-free according to the amount of change they bring about in the original data. Further they are classified as robust and fragile depending on the ease of detection. [2] .Fig1 illustrates the basic technique of watermarking.

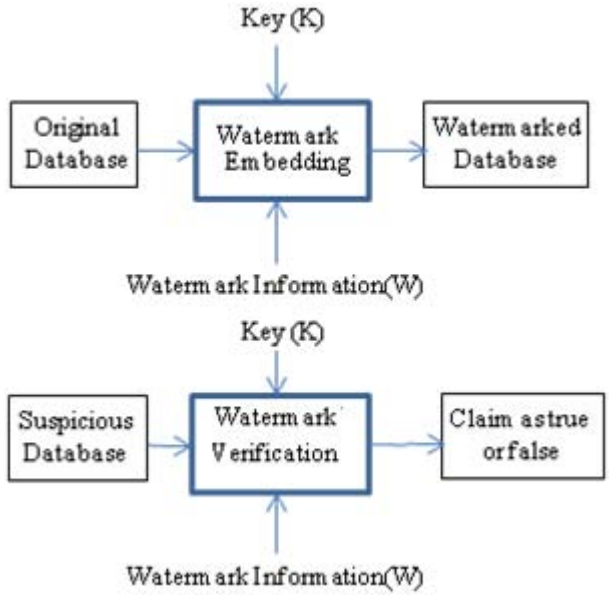

Figure 1: Basic Watermarking Technique [9]

2. Related Work

Techniques that have been implemented for watermarking databases include:

1. Circular Histogram Modulation: This is a robust and reversible watermarking modulation for images in order to protect relational databases. [3] The resulting scheme modulates the relative angular position of the circular histogram center of mass of one numerical attribute for message embedding. It can be used for verifying database authentication as well as for traceability when identifying database origin after it has been modified.

2. Distance-Based Mining: This mechanism considers the database as a collection of objects. The watermark embedding is based on the nearest – neighbor (NN) and Minimum Spanning Tree (MST) of the dataset. These important distance relationships in the original topology are unaltered during the embedding. The watermarked database can be effectively used during mining operations that depends on the ordering of distances between objects, such as NN-search and classification. A spread-spectrum approach that embeds the watermark across multiple frequencies of each object and across multiple objects of the dataset is employed. As such, it renders the removal of the watermark difficult without substantially compromising the data utility. [5]

3. Usability Constraints: These constraints define the maximum amount of change or distortion that can be encoded into the watermark without significant changes to the original data. They are defined by the data owner to ensure the preservation of knowledge within the data. The system takes the dataset as input, models the "usability constraints" to be enforced during the watermark embedding in the dataset [4, 6]. Later it uses three different optimizers to find an optimum watermark that meets the relative constraints. A novel watermark decoding algorithm which: a) ensures that its decoding accuracy is independent of the usability constraints (or available bandwidth); and b) enables "once-for-all" usability constraints definition by providing the maximum robustness with the least possible distortions is used in the approach. The technique is proposed to be highly resilient against insertion, deletion, alteration, and multi-faceted attack yet it results in minimum distortions in the original data set. Regardless of the severity of malicious attack on the watermarked data, the watermark bits are successfully decoded with 100% accuracy because the decoding accuracy of the proposed approach is independent of the usability constraints.

4. Group-based watermarking: The paper [7] proposes an algorithm for watermarking numeric relational data. The algorithm sorts the bits of each tuple in a secret order and selects some of its data bits to route the tuple to a specific watermark bit and one data bit to be marked by the value of the assigned watermark bit**.** **T**he watermark is thus embedded on the group basis. The tuples are uniformly divided into |W| groups, using a mixed sequence of $\log_2|W|$ msbs and lsbs of the attribute that will be watermarked and, afterwards, one bit of watermark information is stored in each group. Therefore, the only information which needs to be saved in a safe storage regarding this process is $\log_2|W| + 1$ short integer values. There is no need to store large quantities of information related to the constructed groups of tuples, like the number of groups, the number of tuples in each group, the tuples that define the borders of each group, the parameters of the function which distributed the tuples in the groups, and any other related information regarding the groups' content. Therefore the method offers an almost blind decoding process.

5. SHA 512 Signature Generation : A database DB is transformed into a watermarked version W by applying a watermark embedding function that also takes an input of public key PK only known and used by the admin of the database. Watermarking changes the original data but these changes are tolerable. In the watermark encryption stage the database Db is partitioned into x number of partitions. The SHA512 [11] algorithm is used to generate the signature that is used for further encryption. The encryption is done using RSA algorithm with the public key of the admin or the owner. This encrypted signature is sent to the other end along with the original data. Further, a watermark bit is embedded in the tuple. The watermarked version W is then delivered to the desired recipient. Watermark decryption is the process of extracting the watermark using the database DB and the private KP and the signature generated. The decryption algorithm is not blind as the original database DB and the signature generated is required for the successful decryption of the embedded watermark.

The watermark decryption is divided into three main steps:
Step 1. Tuple wise partitioning: The database Db is partitioned: by using the data partitioning algorithm used in Encryption stage, the data partitions are generated.
Step 2. Checksum evaluation: The signature is extracted using private key of the receiver and tuples of each partition are checked and checksum is evaluated.
Step 3. Verification: The watermark bits are verified and tampering is indentified.

6. Voice Based Watermarking: Voice of database holder is used to generate watermark by watermark generation

algorithm [12]. Voice is taken from microphone and it is being converted into bit format. Bits of voice along with a random string are used to generate watermark. The watermark w is then encrypted. A one-way hash function is used to decide which tuple and which bit to be marked. The relation is divided into groups of varied but similar sizes. The i^{th} bit selected for the watermark is being replaced by the 1^{st} bit of the voice file and so on. The watermark detection algorithm is then used to recover the watermark from the suspicious relation. The majority voting scheme is used to find the final watermark. Voice denoising is used to eliminate the impact of attacks after the final watermark is changed to voice by the inverse process of watermark generation.

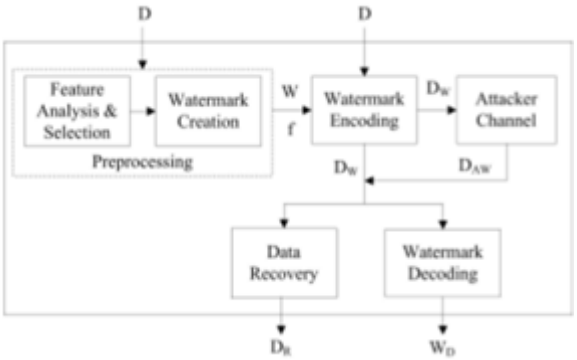

Figure 2: RRW architecture [1]

7. Using Genetic Algorithm: This paper[1] proposes a Robust and Reversible Watermarking techniques, which mainly comprises a (1) data preprocessing phase, (2) watermark encoding phase, (3) attacker channel, (4) watermark decoding phase and (5) data recovery phase. In data preprocessing phase, secret parameters are defined and strategies are used to analyze and rank features to watermark. An optimum watermark string is created in this phase by employing GA - an optimization scheme that ensures reversibility without data quality loss. In the watermark encoding phase, the watermark information is embedded in the selected feature(s). Two parameters, β the optimized value from the GA and η_r a change matrix are used in the watermark encoding and decoding phases. Finally, the watermarked data for intended recipients is generated. In the watermark decoding phase the embedded watermark is decoded from the suspicious data. In order to achieve this, the preprocessing step is performed again, and decoding strategies (feature selection on the basis of MI, β the optimized value from the GA and η_r the change matrix) are used to recover the watermark. Semi-blind nature of RRW is used mainly for data reversibility in case of heavy attacks (attacks that may target large number of tuples). Original data is recovered in data recovery phase, through post processing steps for error correction and recovery. Fig 2 shows the architecture used in RRW.

3. Types of Dataset Attacks

The watermarked database may suffer from various types of intentional and unintentional attacks which may damage or erase the watermark, as described below [8]:

1. Benign Update: In this case, the tuples or data of any watermarked relation are processed as usual. As a result, the marked tuples may be added, deleted or updated which may remove the embedded watermark or may cause the embedded watermark undetectable (for instance, during update operation some marked bits of marked data can be erroneously flipped). This type of processing are performed unintentionally.

2. Value Modification Attack:

- Bit Attack: This attack attempts to destroy the watermark by altering one or more bits in the watermarked data. More information about the marked bit position makes attack more successful. However, in this case usefulness of data is crucial: more alternation may result the data completely useless.
- Bit attack may be a) Randomization Attack: randomly assign random values to certain bit positions; b) by Zero Out Attack where the values in the bit positions are set to zero; c) Bit Flipping Attack: performed by inverting the values of the bit positions.
- Rounding Attack: Attacker changes high-precision data contained in a numeric attribute by rounding all the values of the attribute. Success of this attack depends on the estimation of how many bit positions are involved in the watermarking. Underestimation of it may cause the attack unsuccessful, whereas overestimation may cause the data useless.
- Transformation: An attack related to the rounding attack is one in which the numeric values are linearly transformed.

3. Subset Attack: The attacker may consider a subset of the tuples or attributes of a watermarked relation and by attacking (deleting or updating) them he may hope that the watermark has been lost.

4. Superset Attack: Some new tuples or attributes are added to watermarked databases which can affect the correct detection of the watermark.

5. Collusion Attack: This attack requires the attacker to have access to multiple finger-printed copies of the same relation.

- Mix-and-Match Attack: The attacker may create his relation by taking disjoint tuples from multiple relations containing similar information.
- Majority Attack: This attack creates a new relation with the same schema as the copies but with each bit value computed as the majority function of the corresponding bit values in all copies so that the owner cannot detect the watermark.

6. False Claim of Ownership: This type of attack seeks to provide a traitor or pirate with evidence that raises doubts about merchant's claim.

- *Additive Attack:* The attacker can add his watermark to the watermarked relation and try to claim ownership.
- *Invertibility Attack*: The attacker may launch an invertibility attack to claim his ownership if he can successfully discover a fictitious watermark which is in fact a random occurrence from a watermarked database.

7. Subset Reverse Order Attack: Attacker enjoys this attack by exchanging the order or positions of the tuples or attributes in relation which may erase or disturb the

watermark.

8. Brute Force Attack: In this case, the attacker tries to guess about the private parameters (e.g. secret key) by traversing the possible search spaces of the parameters. This attack can be thwarted by assuming that the private parameters are long enough in size.

4. Drawbacks of the State-of-Art Methods

- The current state-of-the-art watermarking methods mainly deal with the ownership and copyright protection of multimedia and relational databases.
- Watermarking techniques for relational datasets mainly concentrate on numeric datasets.
- Many of the approaches discussed have been tested and proved successful against insertion, deletion and alteration attacks but not against brute-force or collusion attacks.
- While digital watermarking is a widely used measure to protect digital data from copyright offences, the complex and flexible construction of XML data poses a number of challenges to digital watermarking, such as insertion, deletion and alteration attacks.

5. Conclusion

In this paper we have reviewed the need for securing shared data and the different techniques of watermarking relational databases. We have also looked into the different attacks that can take place on the datasets.

References

[1] S.Iftikhar, M. Kamran and Z.Anwar-"RRW - A Robust and Reversible Watermarking Technique for Relational Data" - Knowledge and Data Engineering, IEEE Transactions (Vol.:27 , Iss. 4), 2015 pp. 1132 – 1145
[2] S. Iftikhar, M. Kamran and Z. Anwar - "A Survey on Reversible Watermarking Techniques for Relational Databases" - Security and Communication Networks Volume 8, Issue 15, Wiley Online Library1
[3] J.F-Contreras, F. Cuppens, C. Roux- "Robust Lossless Watermarking of Relational Databases Based on Circular Histogram Modulation"- IEEE Transactions On Information Forensics And Security, Vol. 9, No. 3, Mar 2014 pp.397-410
[4] M. Kamran, M. Farooq- " A Formal Usability Constraints Model for Watermarking of Outsourced Datasets" - IEEE Transactions On Information Forensics And Security, Vol. 8, No. 6, June 2013 , pp 1061-1072
[5] S.I.Zoumpoulis, M.Vlachos, N.M. Freris, C.Lucchese - "Right-Protected Data Publishing with Provable Distance-Based Mining" - IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 8, August 2014,pp.2014-2028
[6] M.Kamran, S.Suhail, M.Farooq - " A Robust, Distortion Minimizing Technique for Watermarking Relational Databases Using Once-for-All Usability Constraints " - IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 12, December 2013 pp 2694-2707
[7] T.Tzouramanis - "A Robust Watermarking Scheme for Relational Databases" - In Proceedings of6th International Conference on Internet Technology and Secured Transactions, 11-14 December 2011, Abu Dhabi, United Arab Emirates pp 783-790
[8] R.Halder ,S.Pal, A.Cortesi-"Watermarking Techniques for Relational Databases: Survey, Classification and Comparison" - Journal of Universal Computer Science, vol. 16, no. 21 (2010),pp 3164-3190
[9] P.A.Raj, Mrs. A. Umamageswari -" Enhancing Security In Medical Image Communication Using Digital Signature"- International Journal of Computer Network and Security (IJCNS) Vol 6. No.1 – Jan-March 2014 pp. 16-21
[10] H.V.Desai - "Steganography, Cryptography, Watermarking: A Comparative Study "- Journal of Global Research in Computer Science Volume 3, No. 12, December 2012
[11] R.W.Gore, R.Tare -" Database Watermarking Using SHA 512 Signature Generation Technique" - International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, pp 4065-4068
[12] S.B.Takmare, R.K.Gupta, G.S.Chandel - "Voice Based Watermarking Technique for Relational Databases" - International Journal Of Scientific & Technology Research Vol.1(10), Nov 2012 pp 65-67
[13] M.Thapa, Dr. S.K.Sood, A.P.M Sharma- "Digital Image Watermarking Technique Based on Different Attacks"- International Journal of Advanced Computer Science and Applications, Vol. 2, No. 4, 2011,pp14-19
[14] G.Kaur, Sukhwinderbir- "Text Watermarking Approaches for Copyright Protection" - International Journal Of Engineering And Computer Science Volume 4 Issue 7 July 2015, Page No. 13553-13558
M. G. Bhandare, K.P Thakur, M. R. More, J.D. Rokade - "Multimedia Piracy Detection Using Invisible Watermark"- International Journal of Advance Research in Computer Science and Management Studies Volume 3, Issue 9, Sept. 2015