IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 4, 2020, accepted March 1, 2020, date of publication March 9, 2020, date of current version April 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2979659

A Double Fragmentation Approach for Improving Virtual Primary Key-Based Watermark Synchronization

MAIKEL LÁZARO PÉREZ GORT[1], CLAUDIA FEREGRINO-URIBE[1], AGOSTINO CORTESI[2], AND FÉLIX FERNÁNDEZ-PEÑA[3]

[1]Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla 72840, Mexico
[2]Department of Environmental Sciences, Informatics and Statistics, Università Ca' Foscari di Venezia, 30172 Venice, Italy
[3]ESAI, Universidad Espíritu Santo, Guayaquil 092301, Ecuador

Corresponding author: Claudia Feregrino-Uribe (cferegrino@inaoep.mx)

This work was supported in part by the CONACyT, Mexico, under Grant 714270 and Grant PN 2017-01-7092.

ABSTRACT Relational data watermarking techniques using virtual primary key schemes try to avoid compromising watermark detection due to the deletion or replacement of the relation's primary key. Nevertheless, these techniques face the limitations that bring high redundancy of the generated set of virtual primary keys, which often compromises the quality of the embedded watermark. As a solution to this problem, this paper proposes double fragmentation of the watermark by using the existing redundancy in the set of virtual primary keys. This way, we guarantee the right identification of the watermark despite the deletion of any of the attributes of the relation. The experiments carried out to validate our proposal show an increment between 81.04% and 99.05% of detected marks with respect to previous solutions found in the literature. Furthermore, we found out that our approach takes advantage of the redundancy present in the set of virtual primary keys. Concerning the computational complexity of the solution, we performed a set of scalability tests that show the linear behavior of our approach with respect to the processes runtime and the number of tuples involved, making it feasible to use no matter the amount of data to be protected.

INDEX TERMS Double fragmentation, duplicate problem, redundancy, relational data, virtual primary key, watermarking.

I. INTRODUCTION

Digital watermarking consists of a set of methods of information hiding techniques created to verify and validate the authenticity of a digital content, tracking its source, or proving the identity of its owner without restringing the data distribution [1]. Contrary to cryptography, the protected content must preserve its appearance, being unnoticed the hidden information [2]. On the other hand, watermarking techniques aim to guarantee the protection of the content for an unlimited amount of time and not only using it as a carrier of the hidden information, as steganography does [3].

According to the literature, watermarking techniques do not only protect digital content by embedding marks on it (e.g., Unnikrishnan and Pramod [4], Ahmad *et al.* [5], and Tufail *et al.* [6]) but also can extract information from the content for its identification (e.g., Halder and Cortesi [7], Bhattacharya and Cortesi [8], and Haider *et al.* [9]). In general, watermarking techniques that embed information are designed for ownership proof or traitor tracing. These techniques must be resilient to attacks that are focused on compromising the detection of the marks, that is why they are classified as robust techniques [10].

The changes caused by watermarking techniques that embed information have to be unnoticeable and cannot compromise the usability of the digital content, avoiding the degradation of its quality [11]. That means the usability constraints defined over the content should not be violated, but also the results obtained by using the watermarked content should be the same as those obtained with the original unwatermarked version of the data. That is why the tolerable amount of changes during watermark embedding varies depending on the purpose of the data and its nature.

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

The data representing the information to be hidden using watermarking techniques are defined as the watermark (WM), which is composed of indivisible items called marks, usually represented by bits. Each of the n marks composing a watermark WM (i.e., $n = length(WM) : n \in \mathbb{Z}^{+}$) is identified by $m_k : k \in \mathbb{Z} : 0 \leq k \leq (n-1)$ where $m \in \{0, 1\}$.

Watermarking techniques that embed information are composed of two processes: (i) embedding, and (ii) extraction of the WM [12]. The correct functioning of the technique requires the parameters having the same value in the embedding and the extraction processes; otherwise, the identification of the WM fails. Finally, the use of at least one parameter is required, being that the case of the secret key (SK), whose value should be known only by the data owner [13].

An important matter when working with watermarking techniques is the synchronization of the WM. According to Cox *et al.* [14], synchronization is the process of aligning two signals in time or space. In this case, those signals are the embedded and the extracted WMs, which is why the synchronization definition is directly related to the processes of embedding and extraction. Since low synchronization is caused by losing marks in the watermarking processes, it is critical a proper design of the technique in order to avoid that problem and to guarantee the WM detection; otherwise, the proposal lacks sense and cannot be applied in real scenarios.

Initially proposed for multimedia data, watermarking techniques were applied later on to relational data. It has been a challenging task considering the nature of relational data (e.g., high frequency of updates, lack of fixed order, among others). Even so, relational data watermarking techniques have been growing in diversity, yet some issues to be addressed remain, being one of them the use of the Primary Key (PK) of the relation for deciding where and how to embed the marks.

Most watermarking techniques use *pseudo-random selection* to choose the mark to be embedded and the embedding place in the relation of the database. That way, WM synchronization cannot be compromised by reordering the attributes or the tuples of the relation (i.e., *subset reverse order attack* [15]). Considering that the PK stores unique values, its use by watermarking techniques increases the probability of selecting different marks each time during the WM embedding despite *pseudo-random selection*.

The selection of the PK to perform this task gives reliability to the process because there is no way to delete or replace the PK without compromising the referential integrity of the database. Potential attackers trying to compromise the WM detection in the stolen data, also care about the quality of the data, that is why they might not attack the PK. However, if the relation is distributed separately from the rest of the database and the attacker has no interest in using it restored with the rest of the data, it is free to delete or update the PK making the WM undetectable despite the presence of the marks in the relation. In this case, the quality of the watermarked data remains whereas the attack entirely compromises WM synchronization.

To avoid the vulnerabilities that emerge by using the PK in watermarking processes, some schemes have been proposed to generate Virtual Primary Keys (VPK) to perform the embedding and extraction of the WM. Unfortunately, a new problem comes out because of using data sets of limited elements for generating the VPK values. This problem is due to the presence of duplicate values in the set of VPKs because of the intrinsic characteristics of the data in the attributes of tuples which are used for generating VPKs. Formally defined as the *duplicate problem* [16], it causes the embedding of some marks multiple times whereas others are completely ignored, compromising WM synchronization. As a result, a high number of marks are excluded in the embedded signal, causing a similar effect to the attacks based on deleting or modifying data (e.g. subset attacks [17]). For example, if the WM synchronized using the watermarking technique by Sun *et al.* [18] and the relation's PK is embedded using VPKs instead, its quality will be degraded in a range between 95.94% and 3.7% with respect to the original signal used to generate the WM, depending on the VPK scheme. It happens similarly with Sardroudi & Ibrahim [19]'s technique, where the use of VPK-based synchronization degrades the WM synchronized using PKs, approximately in a range between 98.94% and 6.7%. This behavior happens no matter the watermarking technique, due to the low quality that the VPK set presents compared to the set composed by the PKs.

Most of the VPK schemes try to avoid the *duplicate problem* by eliminating the redundancy of the set of VPKs. Nevertheless, the data stored in the relation are limited by the attributes' domains. So, this is a goal hard to accomplish. In this paper, we introduce a different approach to achieve the WM synchronization despite using low-quality sets of VPKs. The soundness of our contribution relies on the soundness of the combining VPK schemes and of the proposed technique. Our approach contributes to robustness against data elimination attacks, taking advantage of the presence of duplicate values in the VPK set instead of being affected because of it. The experiments carried out show that our goal is accomplished, increasing the number of embedded marks and improving their detection in comparison to previous approaches. We achieved an increment of the number of embedded marks in a range between 81.04% and 99.05% compared to previous solutions. We also performed a set of scalability tests showing the linear behavior of our approach with respect to the watermarking processes runtime and the number of tuples being protected. Obtained results prove that our proposal is feasible to be used no matter the amount of data to be watermarked.

The rest of this paper is organized as follows. Section II introduces the basics related to our work and a review of VPK schemes associated to our proposal. Section III presents the formalization of our approach, and Section IV shows the experimental results used to validate our work. The conclusions are given in Section V.

II. RELATED WORK

The main theoretical fields related to this work are the structure of relational data according to the relational model, the relational data watermarking techniques, and the schemes created for the VPK generation. In this section, we describe the main elements involved in each one of those fields.

A. STRUCTURE OF A DATABASE RELATION

The relational databases are formed by different entities storing the data. Each entity is defined according to a conceptual weight from the modeled reality. In the relational model, the entities are named relations and are linked among them using their primary keys [20]. Relations are represented by tables where their columns are the entity's attributes and their records (also known as tuples) are the entity's instances where each attribute is combined. In this paper, we identify as R a generic relation, object of our study.

We use the notations introduced by Agrawal and Kiernan [21] in 2002 to identify each element composing R. According to that, the attributes of R are identified by $A_i : i \in \mathbb{Z} : 0 \leq i \leq (\nu - 1)$, being $\nu \in \mathbb{Z}^+$ the number of attributes available for marking. One particular attribute of R, excluded in the previous notation, is the primary key (PK),[1] which stores unique values used to identify each tuple. On the other hand, the tuples of R are identified by $r_j : j \in \mathbb{Z} : 0 \leq j \leq (\eta - 1)$, where $\eta \in \mathbb{Z}^+$ is the number of tuples. Finally, the relation is formally defined as $R(PK, A_0, \ldots, A_{\nu-1})$. In general, a particular element of R is identified by $r_j.A_i$ (meaning the attribute i of the tuple j). The primary key of a tuple is denoted by $r_j.PK$.

B. WATERMARKING TECHNIQUES FOR RELATIONAL DATA

The first watermarking technique for relational data was proposed in 2002 by Agrawal and Kiernan [21]. They considered the scenario of a single relation R with the presence of a PK out of the scope of the attacker's goals. Also, they focused on watermarking numeric attributes using their binary representation for embedding the marks in one of the less significant bits (*lsb*). For selecting the tuple, the attribute within the tuple, and the *lsb* of the attribute to embed the mark, a value is generated using the relation's PK and a secret key (SK) known only by the data owner. This value can be defined as a PK-based VPK.

Agrawal & Kiernan perform the selection of the tuple by using a parameter called *tuple fraction* defined as $\gamma \in \mathbb{Z} : 1 \leq \gamma \leq \eta$. If γ takes small values, more tuples are considered for marking. For the case when $\gamma = 1$, all tuples of the relation are marked if the usability constraints allow it. Since the publication of this approach, several techniques have been proposed, each one with its differences and particularities. One of those particularities has been generating VPKs excluding the relation's PK from the process.

C. EXISTING VPK SCHEMES

Also known as PK-dependent, watermarking techniques performing the WM embedding and extraction using the PK of the relation are useless if the attacker can attack the PK. Some researchers have proposed solutions defined as VPK schemes focused on avoiding the use of the PK, excluding them from the VPK generation. Until now, the VPKs values generated using the proposed schemes present high redundancy, compromising the WM synchronization. This problem was defined as the *duplicate problem* by Li *et al.* [16].

Another challenge faced by the VPK schemes is due to having no choice but to use the values stored in the attributes of the relation for generating the VPKs. When an attribute is erased, some of the VPK values generated to extract the WM do not match those used for its embedding. Due to that, some marks take false values, adding noise to the extracted WM, sometimes even compromising its identification. This was defined by Li *et al.* [16] as the *deletion problem*, and it is based on the version of the subset attack focused on attribute elimination [15].

1) THE S-SCHEME

The S-Scheme was presented in 2002 by Agrawal and Kiernan [21] as an alternative for watermarking relations with no PKs, but its name was given by Li *et al.* [16]. For cases when relations present a single attribute, this scheme creates the VPK by splitting the binary value of the attribute in two groups: the most significant bits (*msb*) and the less significant bits (*lsb*). The VPK is generated using the *msb* according to (1) where $\chi \in \mathbb{Z}^+$ represents the *msb* range. The *lsb* range (given by $\xi \in \mathbb{Z}^+$) is used to embed the marks.

$$vpk(r_j) = [VPK([r_j.A]_2, \chi)]_{10} \tag{1}$$

In (1), $[r_j.A]_2$ means the binary representation of the value of the attribute A for each tuple, and $[VPK([r_j.A)_2, \chi)]_{10}$ the integer value formed by the χ bits of the attribute in decimal notation. Taking into account that χ remains constant for all attributes, the number of duplicate values this scheme generates is extremely high.

In this scheme, watermarking techniques select the tuple to embed the mark if $H(SK \circ vpk(r_j))\ mod\ \gamma = 0$, where H is a one-way hash function taking as input the concatenation (represented by the symbol $\circ$) between the secret key SK known only by the data owner, and the virtual primary key of the tuple $vpk(r_j)$. Once the tuple is selected, mark embedding is performed in a bit pseudo-randomly selected out of the ξ bits based on the VPK value generated for the tuple according to the expression $H(SK \circ vpk(r_j))\ mod\ \xi$. Mark extraction is performed based on the same operations using the same parameter values.

When the relation is composed of more than one attribute, the VPK is generated using one attribute, and the rest of them is used for embedding the marks. Also, the attribute selected for the VPK generation can vary for each tuple. However, this scheme is highly vulnerable to *deletion problem* in all its variants.

[1] The PK can also be defined as the combination of more than one attribute taking unique values.

2) THE E-SCHEME

The E-Scheme was proposed in 2003 by Li *et al.* [16] as an alternative for embedding the marks in all attributes of every tuple. It works similar to S-Scheme, but uses each one of the ν attributes of R, generating a total of $\nu \times \eta$ VPKs, instead of just η. Also, for this scheme, the value of χ remains constant during the whole process.

One particularity of the E-Scheme is the use of two different hash functions (represented as H_1 and H_2): the first one to decide if the attribute will be marked and the last one for selecting the *lsb* to embed the mark [16]. The selection of the attribute is carried out if $H_1(SK \circ vpk(r_j.A_i))\ mod\ (\gamma \times \eta) = 0$, where $vpk(r_j.A_i)$ is a variation of (1) for its use on each attribute of every tuple. Once an attribute is selected, the *lsb* position to embed the mark is obtained according to the expression $H_2(SK \circ vpk(r_j.A_i))\ mod\ \xi$. Mark extraction is performed using the same expressions and the same parameter values.

For the E-Scheme, the VPKs are generated under the same conditions, which increase the number of duplicate values in comparison to the S-Scheme. On the other hand, this scheme is compromised by the *deletion problem*, since deleting any attribute directly compromises the value of η VPKs. However, the number of duplicate values is so high that, the VPKs never really contribute to embedding any mark without ambiguity.

3) THE M-SCHEME

The M-Scheme was proposed in 2003 by Li *et al.* [16] to generate the VPKs using different attributes on each tuple. The principle of the M-Scheme is to split each binary value into two fragments (*msb* and *lsb*), same as the S-Scheme. Then, for each tuple, it generates a VPK by concatenating the two values of $H_1(SK \circ vpk(r_j.A_i))$ closest to zero. If more attributes match the same lowest hash value, they can be selected for the VPK generation too.

The condition for considering the tuple for mark embedding is based on the same expression used by the S-Scheme. Once the tuple is selected, according to the AHK algorithm defined by Agrawal and Kiernan [21], the attribute where the mark is embedded is chosen using the expression $H(SK \circ vpk(r_j))\ mod\ \nu$. Then, the *lsb* to embed the mark is selected also as S-Scheme does. The steps for mark extraction are based on the same expressions used for the embedding and require the same parameter values.

This scheme can be used involving more than two attributes to generate the VPKs, with the condition that the number of attributes considered does not exceed the number of attributes of the relation. Same as when using the S-Scheme, one VPK per tuple is generated using the same value of χ every time, causing too many duplicate values.

Besides the M-Scheme proposed by Li *et al.* [16], there are also the VPK schemes proposed by Chang *et al.* [22] and by Khanduja *et al.* [23], very similar to the M-Scheme in principle. In particular, the proposal by Chang *et al.* [22] was created to work with textual attributes, simulating the S-Scheme by splitting each value into two fragments. The first fragment, denoted as A_i^1, contains the last word of the text and is used to embed the marks. On the other hand, the second fragment, denoted as A_i^{-1}, contains the rest of the text and is used to generate the hash value representing the attribute during the VPK generation process. Moreover, the main difference between the M-Scheme by Li *et al.* [16] and the proposal by Khanduja *et al.* [23], is that the later always uses two specific attributes to generate the VPK, chosen according to the data owner's criterion.

4) THE EXT-SCHEME

The Ext-Scheme was proposed in 2017 by Pérez Gort *et al.* [24] with the goal of varying the elements involved in the VPK generation. Same as the S-Scheme and the M-Scheme, the Ext-Scheme generates one VPK for each tuple, but the attributes involved can vary depending on their values. Also, the value of χ varies depending on the binary length of the value of the attribute analyzed.

The Ext-Scheme analyzes all attributes of each tuple to decide which of them is involved in the VPK generation. Only those attributes accomplishing the condition $b_{ji1} \oplus b_{ji(\chi-1)} = 1$, where b_{ji1} and $b_{ji(\chi-1)}$ are the first and $(\chi-1)^{th}$ bits of the attribute i of the tuple j, and $\oplus$ is the XOR operator, are considered.

On the other hand, the variation of χ is carried out according to a temporary value obtained by using (2), where BL_{ji} ($BL \in \mathbb{Z}^+$) is the binary length of the value of the attribute $r_j.A_i$, $MSBF \in \mathbb{Z}^+$ is the fraction of *msb* (parameter that remains the same during the whole process), and the symbols $\lfloor\ \rfloor$ represent the floor function. The value of χ is selected according of the neighboring bits to the $\_\chi^{th}$ bit ($\_\chi \in \mathbb{Z}^+$).

$$\_\chi_{ji} = \lfloor BL_{ji}/MSBF \rfloor \tag{2}$$

This scheme was created to perform embedding and extraction of marks using the VPKs by watermarking techniques based on the AHK algorithm. According to that, the selection of the tuples, attributes, and *lsb*s to embed the marks in R is performed following the same expressions used by the M-Scheme. Also, WM extraction is performed by the same rules using the same parameter values.

The Ext-Scheme generates a high number of exclusive values compared to previous schemes, and it is also more resilient to the *deletion problem*. Even though, this scheme does not always guarantee the embedding of enough marks, depending on the size of the WM. Also, due to establishing too many conditions for the attribute involvement in the VPK generation, often too many tuples are wasted by the process.

5) THE HQR-SCHEME

The HQR-Scheme was proposed in 2019 by Pérez Gort *et al.* [10] to generate one VPK per tuple. It is based on the cyclic model of the attribute order in R to change the way the attributes are analyzed for the VPK generation. Fig. 1 depicts how the first and last attribute of the relation can

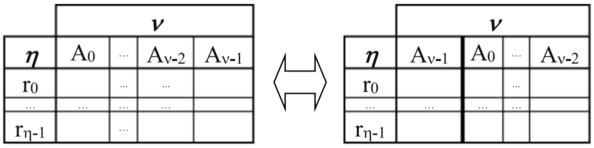


FIGURE 1. Cyclic model of the attribute order in R [10].

be seen as neighbors according to this model. This scheme also controls the maximum number of attributes that may be involved in the VPK generation (given by $\ell \in \mathbb{Z} : 0 < \ell \leq \nu$), and the minimum distance between them (given by $p \in \mathbb{Z} : 1 \leq p \leq \lfloor(\nu - \ell)/\ell\rceil$).

The HQR-Scheme was designed using the Ext-Scheme as a base, considering a higher value of χ by excluding from the binary string being analyzed for each attribute, only the bits used as candidates to embed the marks (i.e., the *lsb* range). Also, the attribute selected to begin the analysis, the direction of the analysis, and the attributes considered to generate the VPK vary for each tuple depending on the values stored in it.

This scheme does not waste tuples for the VPK generation. On the other hand, despite generating a higher number of exclusive values than the Ext-Scheme, it produces too many duplicate values compared to the number of tuples in R. Finally, as a variation of the Ext-Scheme, this approach generates the VPKs to perform embedding and extraction of the WM similarly to M-Scheme and to Ext-Scheme, and presents high resilience against the *deletion problem*.

III. THE DOUBLE FRAGMENTATION APPROACH

Despite the variation of the results obtained by applying the different VPK schemes, the *duplicate problem* always causes the exclusion of marks from the embedding process. On the other hand, the way the attributes are selected makes some schemes too vulnerable to the *deletion problem*. Facing these situations is a complicated task since the source to generate the VPKs is limited to the data stored in the relation. This is even more problematic for the E-Scheme that can only use the χ bits of one attribute as the source.

A. STRUCTURE OF THE VPK SET

For a better understanding of the solution proposed in this work, we first define the elements composing a generic set of VPK identified as S. Considering the effects of the *duplicate problem*, the presence of duplicate values in S is assumed. The set of duplicate values is identified as G and it is composed of groups of VPKs formed according to their values, analyzed from the group with the minimum length to the maximum one. Another set defined due to the presence of duplicate values is D, which is similar to G, but it is composed only by one item per duplicate group. According to that, the cardinality of G will always be higher than the cardinality of D (i.e., $|G| > |D|$). Some examples of the sets involved in the representation of the duplicate values are shown in Table 1. To highlight the different groups of duplicate values different colors are used. The groups of duplicate values are referenced as G_r where $G_r \subset G$ and $r \in \mathbb{Z} : 0 \leq r \leq (|D|-1)$. We do not define how to identify a particular item of a group considering that once in the same group, all items store the same value.

TABLE 1. Different examples of the sets *G* and *D* inside *S*.

S	$\|G\|$	$\|D\|$
$\{2, 2, 2, 2, 2, 2, 2, 2, 2, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5\}$	19	2
$\{6, 6, 6, 6, 2, 2, 2, 2, 2, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5\}$	19	3
$\{8, 8, 2, 2, 2, 6, 6, 6, 6, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5\}$	19	4

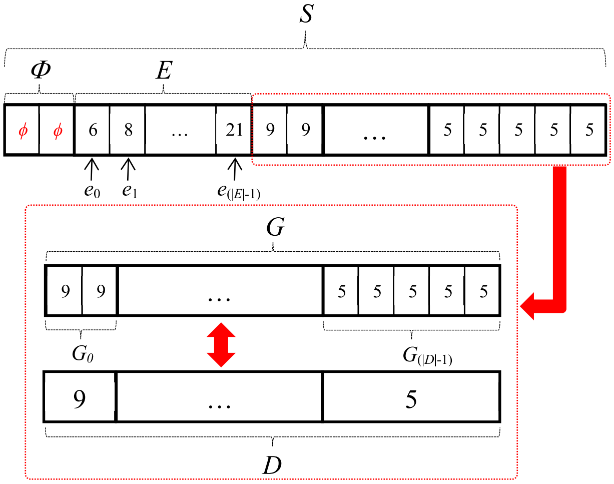


FIGURE 2. Structure of the set *S* of VPK values.

The presence of exclusive values in S is also considered. All the exclusive VPK values are represented by the set E and each one of them is identified by e_x where $x \in \mathbb{Z} : 0 \leq x \leq (|E| - 1)$. The elements of this set are organized ascending according to their values. Finally, the set of VPK *null* values is represented as Φ. All the sets G, D, E, and Φ are subsets of S. A graphical view of the structure of one example of S is shown in Fig. 2 where the differences between the sets G and D are also represented.

It is important to highlight that $\Phi \neq \emptyset$, since the set is formed by elements storing null values, and does not represent the empty set. This set is the result of the relation's attributes and tuples that do not accomplish the conditions established by the scheme for the VPK generation (e.g., when the range of the *msb* is higher than the length of the attribute binary value).

B. PROPOSED SCHEME

We exclude the E-Scheme from our approach considering it is the only scheme that generates more than one VPK per tuple, and this work is focused on generating only one. This decision is made for preventing the aggressive distortion that may be caused by embedding multiple marks into the same tuple. Based on the argument of generating one VPK per tuple, it is set that $\eta = |S|$.

The definitive solution to avoid the consequences of the *duplicate problem* is by obtaining S such that $|E| = |S|$, which means that S is composed only of exclusive values (e.g., the set of the relation's PKs). Getting a VPK set with that feature is hard to achieve considering that VPK schemes can only use data stored in the relation, which are bounded by the domain of the attributes. Because of that, VPK schemes try to reduce the impact of the *duplicate problem* by increasing the number of attributes involved in the VPK generation, which contributes to get a higher number of exclusive values in S, or reducing the size of the groups of duplicate values. The downside of that strategy is that the scheme's vulnerability to the *deletion problem* will be higher since the probability of compromising more VPKs when any attribute is deleted also increases. On the other hand, increasing the resilience to the *deletion problem* by reducing the attributes involved in the VPK generation will bring as consequence the increment of duplicate values.

Since the improvement of the quality of S cannot be achieved only by managing the number of attributes involved in the generation of the VPKs, we propose another approach. Our technique does not discard the idea either of the VPK schemes nor the traditional watermarking techniques but adds new elements to those solutions. The proposal is not focused on trying to improve the quality of S, but on using the duplicate values of the set to split the WM signal. Defined as the Double Fragmentation Technique (DF-Technique), we perform two times the fragmentation of the WM to increase the number of embedded marks regardless of the quality of S (see Fig. 3).

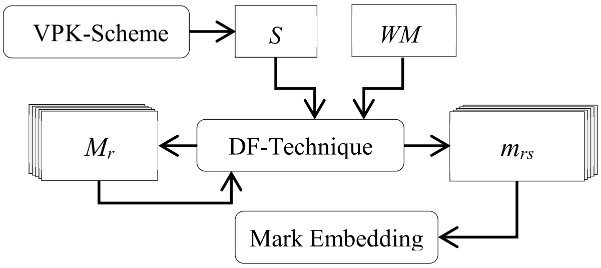


FIGURE 3. **Elements involved in the DF-technique.**

By applying our approach, the presence of duplicate values in S contributes to the improvement of the WM synchronization. The details about the way fragmentation processes are carried out are given in the next subsection.

C. FRAGMENTATION PROCEDURES

The first fragmentation process generates elements defined as *first fragments* corresponding to *WM* segments composed of more than one mark. *First fragments* are identified by M_r where $r \in \mathbb{Z} : 0 \leq r \leq (|D| - 1)$, being in number equal to the number of groups of duplicate values in S. The second fragmentation process generates the *second fragments*, being each one a mark extracted from one of the *first fragments*. *Second fragments* are identified by m_{rs}, where $s \in \mathbb{Z} : 0 \leq s \leq (l(M_r) - 1)$, being M_r the corresponding *first fragment* and l the function that returns the length of M_r (see Fig. 4).

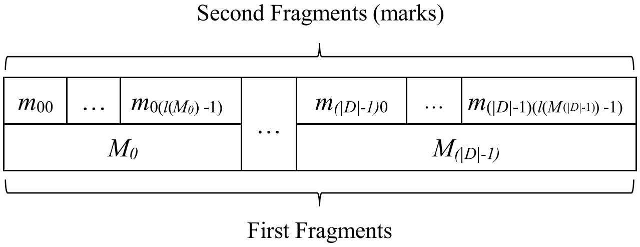


FIGURE 4. **Generic view of a fragmented *WM*.**

Contrary to the equality between the number of *first fragments* and the number of groups of duplicate values, the number of *second fragments* stored in a particular M_r is lower than the cardinality of its correspondent subset G_r. That is based on the principle that the number of tuples to be marked should not be equal, but much higher than the length of the WM (i.e., $\eta \gg n$), to allow the embedding of each mark multiple times and handle the distortion without compromising the inclusion of any mark during the embedding process. According to that, the condition $|G_r| > l(M_r)$ is a rule to be accomplished.

To generate the *first fragments*, the values of *WM*, $|D|$ and the cardinality of each subset G_r are required. The cardinality of each subset representing the groups of duplicate values G_r is used to assign the number of marks belonging to each first fragment M_r, in other words, to set $l(M_r)$. Since $|G_r| > l(M_r)$, instead of equality, the proportion between those values has to be settled. Fig. 5 gives a clear idea of the differences between the length of the *first fragments* and the cardinality of the groups of duplicate values in S.

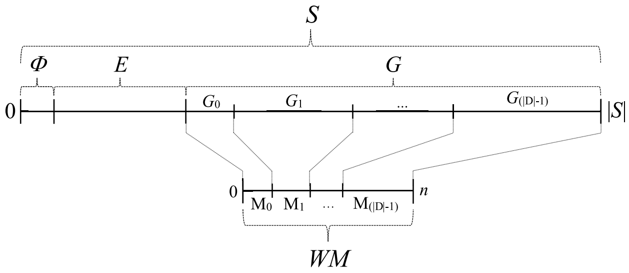


FIGURE 5. **Proportions between the groups of duplicate values in *S* and the first fragments of *WM*.**

The proportion between the length of M_r and the cardinality of its correspondent subset G_r of S is given according to (3), where the symbols $\lfloor \, \rceil$ represent the use of the rounding function.

$$l(M_r) = \lfloor \frac{l(M_r)\% \times n}{100} \rceil \tag{3}$$

Being $|G_r|\%$ the percentage of the group's size with respect to the size of the set of duplicate values of S obtained with (4), and considering that for establishing a fair proportion between all the fragments of WM and G, the rule $l(M_r)\% = |G_r|\%$ must be followed, the substitution of $l(M_r)\%$ by $|G_r|\%$ is carried out.

$$|G_r|\% = \frac{|G_r| \times 100}{|G|} \tag{4}$$

Once (4) is substituted in (3) we get (5), establishing the proportion between $l(M_r)$ and $|G_r|$. Then, once a duplicate VPK value generated from a tuple is known, it is possible to know its correspondent group G_r, the group cardinality $|G_r|$, and finally, to which fragment M_r of the WM that VPK value will be focused on to extract the mark to be embedded in the tuple.

$$l(M_r) = \lfloor \frac{|G_r| \times n}{|G|} \rceil \tag{5}$$

Algorithm 1 is used to select the mark to be embedded in the tuple. Once the mark is selected, the embedding is carried out depending on the conditions established by the watermarking technique and by considering the usability constraints defined in the database.

Algorithm 1 Selection of the Mark Given the VPK

```
1 for $j = 0;\ j < \eta;\ j++$ do
2     $G_r = get\_group\_of(VPK_j)$
3     $c_{Gr} = |G_r|$
4     $c_G = |S| - |E| - |\Phi|$
5     $seed = generate\_seed(j)$
6     $m_j = WM(H(VPK_j, \lfloor c_{Gr} \times n / c_G \rceil, seed) + L)$
```

In line 4 of the algorithm, the cardinality of G is obtained, which can be done either by using $|S| - |E| - |\Phi|$ or $\sum_{r=0}^{|D|-1} |G_r|$. In line 6, H represents a one-way hash function that takes as input the VPK duplicate value, the range allowed to perform the pseudo-random selection of the mark (the length of the *first fragment* according to (5)), and a *seed* that is generated for the current tuple. The generation of the seed is performed for each tuple (line 5 of the algorithm) to increase the entropy of the mark selection process, so the number of embedded marks increases. The term L represents a value proportional to the sum of the cardinalities of the groups of duplicate values located before G_r since the position of the selected mark obtained as the result of H is relative to its group. Finally, since exclusive values may also be presented in S, for those cases the marks are selected similarly as the PK-Dependent watermarking techniques do by using the relation's PKs.

Considering that choosing a process featuring a high entropy in the generation of the seeds is an important factor for our approach to succeed, we proposed a version of the Ext-Scheme [24] to perform this task despite having been initially conceived for VPK generation. Compared to the other solutions, this scheme is more robust against the *deletion problem* and generates less duplicate values. Even though by itself not always guarantee the embedding of enough marks, its use as a seed generator for our approach contributes to achieving a drastic increment of WM synchronization.

One of the drawbacks of the Ext-Scheme is the number of tuples that are not used due to generating *null* VPK values from them, given the complicated and high number of conditions to be satisfied in the scheme. Looking to duck the consequences of this limitation we introduced the constant $V \in \mathbb{Z}^+$, so all tuples contribute to the process. To avoid the collusion with the rest of the values generated, V will depend on the maximum seed value. Once all the seeds are generated, we add the maximum value to V and then assign the result of the sum to all the *null* seeds. Despite the presence of duplicate values in the set of seeds and in S, their combination in our approach allows the embedding of enough marks to guarantee the WM recognition, overcoming even the damage that the deletion of any attribute of the relation could provoke.

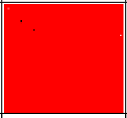
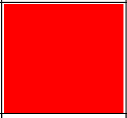
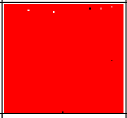
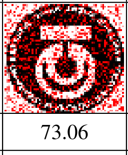
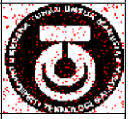
The success of the proposed technique highly depends on the WM fragmentation and the seed generator. Considering the source to generate the VPKs is restricted to data stored in the tuple (same as to the VPK schemes), the set of generated seeds will always present duplicate values. Nevertheless, the combination between a generator achieving low duplicates and the VPK set, increases the variability of the approach, avoiding to have the same pair of duplicated values (VPK + seed) each time. From this point of view, generating exclusive values for the seeds is the perfect solution, but considering it is unexpected, a generator with low duplicate seeds is a good option to our approach as well. For this reason, the version of the Ext-Scheme we proposed constitutes a good choice for our work.

IV. EXPERIMENTAL RESULTS

Our approach was validated through experiments performed using the numeric relational data set *Forest Cover Type* [25]. We mostly worked with the first 30,000 tuples out of 581,012 of this data set, and 10 of its 54 attributes in order to make a fair comparison against results previously reported in other papers. On the other hand, we worked with higher numbers of tuples to perform the experiments related to the scalability tests. The generation of the VPKs was done using the schemes that generate one VPK per tuple, excluding the E-Scheme from our experimental setup according to that criterion.

Our proposal is focused on improving the WM synchronization, often compromised due to the *duplicate problem*. Table 2 shows the quality of the embedded WM using the different VPK sets generated with the VPK schemes described in this work without applying the double fragmentation approach. The source of the WM was the image UTM (see Fig. 6 a)) and the watermarking technique employed was the technique proposed by Sardroudi and Ibrahim [19]. Despite our proposal contributes to increment the quality of the synchronized WM using the VPK sets generated by any scheme, we validated our approach using the S and M schemes, since those are the options that generate one VPK

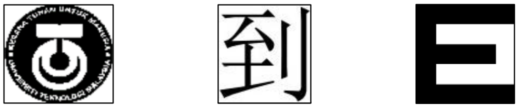
TABLE 2. WM embedded using each VPK scheme [10], [24].

VPK Scheme:	S-Scheme	E-Scheme	M-Scheme	Ext-Scheme	HQR-Scheme
WM:					
% Marks:	0.06	0	0.11	73.06	92.30

per tuple and present serious synchronization problems due to the *duplicate problem* (see Table 2).

The low WM synchronization shown in Table 2 is a direct result of the bad quality of the VPK sets, but it is also because of the set of requirements watermarking techniques demand from the data to avoid compromising its quality because of mark embedding. For example, giving the nature of the E-Scheme, just few VPKs will be generated several times. Because of that, if all VPKs do not accomplish the conditions stated by the watermarking scheme when mixed with the data being watermarked, the WM embedding will be entirely compromised. So, this scheme must be considered depending on the parameters, the WM size, and the data to be protected.

The watermarking technique proposed by Sardroudi & Ibrahim [19] uses an image as the WM source, which classifies it as Image-Based Watermarking (IBW) according to the watermark information classification criteria [15]. In this case, the image is binary, being each pixel value the simplest possible one (1 for black color and 0 for white), reducing the amount of data to embed, causing less distortion during the WM embedding. According to this watermarking technique, each pixel is pseudo-randomly selected by using the tuple's VPK. Then, the mark is generated xoring the pixel value and the value of one *msb* pseudo-randomly selected from the attribute to be watermarked. In this way, the mark will not only depend on the pixel value but also on the attribute value. Finally, the mark is embedded in one of the *lsb* of the same attribute, selected depending on the VPK scheme as it was stated in Section II-C. In this document, we also use the red color to identify the pixels missed due to incomplete embedding or the *deletion problem*.



a) Universiti Teknologi Malaysia logo

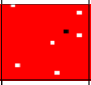
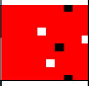
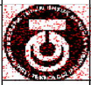
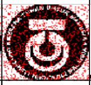
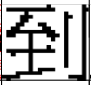
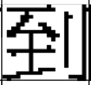
b) Chinese character "dào"

e) Character "E"

FIGURE 6. Images used as WM sources.

To analyze the role played by the WM length in the experiments, images of different sizes were used to generate the WM. The images were: the Universiti Teknologi Malaysia (UTM) logo (82 × 80 pixels), the Chinese character dào (20 × 21 pixels), and the E character (10 × 10 pixels) (see Fig. 6).

TABLE 3. WM UTM extracted using S and M VPK schemes, Sardroudi & Ibrahim's technique, with and without DF-Technique.

WM:	UTM Logo		Dào Character		E Character	
Scheme:	S-Scheme	M-Scheme	S-Scheme	M-Scheme	S-Scheme	M-Scheme
Not Frag.						
	0.06	0.11	0.95	1.66	4	6
Fragmented						
	86.05	81.15	100	100	100	100

The use of an image as the WM source allows the visual appreciation of the quality of the synchronized WM. On the other hand, correlation-based metrics can be applied when this data type is used to generate the WM. In this work we use the Correction Factor (CF) first applied on IBW by Sardroudi & Ibrahim [19] in 2010, given by (6), to compare the pixels of the image used to generate the WM (given by I_O) with the pixels of the image generated from the extracted WM (given by I_E). In the equation, $\varkappa \in \mathbb{R}^+$ represents the value of CF, which constitutes a percentage for which 100 means the exact similarity between both images, and 0 the absence of similarity. Also, the terms $h \in \mathbb{Z}^+$ and $w \in \mathbb{Z}^+$ refer to the height and width of both images, considering they have the same dimensions.

$$\varkappa = \frac{\sum_{i=1}^{h}\sum_{j=1}^{w}(I_O(i,j) \oplus \overline{I_E(i,j)})}{h \times w} \times 100 \tag{6}$$

The approach was implemented using Java 1.8 as the programming language for the client-side and Oracle Database 12c for the database server. The runtime environment was a 2.20 GHz AMD PC with 6.00 GB of RAM running on Windows 10 OS.

A. WM SYNCHRONIZATION IMPROVEMENT

The first experiment performed is oriented to detect how DF-Technique contributes to improving the quality of the WM synchronization. Since we are focused on analyzing the changes in the WM capacity, for this case we do not simulate any attacks or update operations on the database. The embedding of the same WM is done under similar conditions, first with no fragmentation and then applying our approach.

Table 3 shows the quality of WM synchronized by using the selected VPK schemes and the watermarking technique of Sardroudi & Ibrahim. For all cases, the image built from the extracted WM with their respective CF value is shown. The parameters used for the S-Scheme were: *attribute* = "ELEVATION", $\chi = 3$; for the M-Scheme: *no. of attributes* = 2, $\chi = 3$; and for the watermarking technique: $\gamma = 1$, $\chi = 3$, and $\xi = 1$.

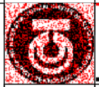
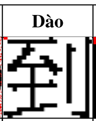
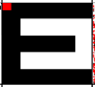
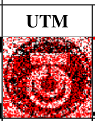
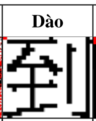
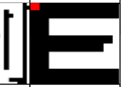
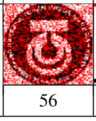
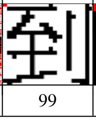
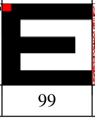
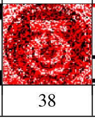
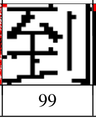
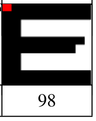
From Table 3, it can be appreciated that the reduction of the WM length has a positive impact on its synchronization since the probability of excluding marks due to the random selection is reduced. Also, the quality of the WM signals, once the DF-Technique is applied, improves considerably thanks to the increase in the number of embedded marks.

B. RESILIENCE AGAINST THE DELETION PROBLEM

Applying the DF-Technique increases the number of embedded marks, improving the WM synchronization, but also it is important to know how the synchronization will be affected once the *deletion problem* is considered. As it was previously mentioned, this will depend on the features of the VPK Scheme and how the seeds are generated. Since the seeds are generated using the version of the Ext-Scheme we described in the previous section, the attributes considered for each tuple and the *msb* range vary, being the scheme resilient enough against attributes elimination.

Table 4 shows the best and the worst results by deleting one of the ten attributes composing the watermarked relation. The experiments were performed by deleting a different attribute each time. Each attribute presents different relevance to the VPK generation, watermarking processes, and data usability. We do not know the attribute relevance for the attacker, but we know how relevant each one of them is to the data owner. Then, that attribute can be selected to generate the VPK, and if it is attacked, not only the WM detection is compromised but also the data quality, causing the entire data set to lose relevance.

TABLE 4. Ranges of the quality of the detected WM once a deletion of one attribute is carried out.

Scheme:	S-Scheme			M-Scheme		
WM:	UTM	Dào	E	UTM	Dào	E
Best						
	69	99	99	49	99	98
Worst						
	56	99	99	38	99	98

According to Table 4, the resilience against the deletion problem seems higher using S-Scheme, but there is something to be remarked. For the case of the M-Scheme, the attribute deletion also involves those attributes selected for generating the VPK, not only the seed. For experiments using S-Scheme, only the attribute involved in seed generation was deleted. This was carried out considering that if the attribute used to generate the VPK is deleted, WM synchronization is compromised. Thus, the worst result with S-Scheme will be $\varkappa = 0$. On the other hand, if the attribute containing the most valuable data in the relation is known, we could use it to generate the VPKs with the S-Scheme, reducing the risk of being affected by the *deletion problem* and the consequences of the duplicate values thanks to the use of the DF-Technique.

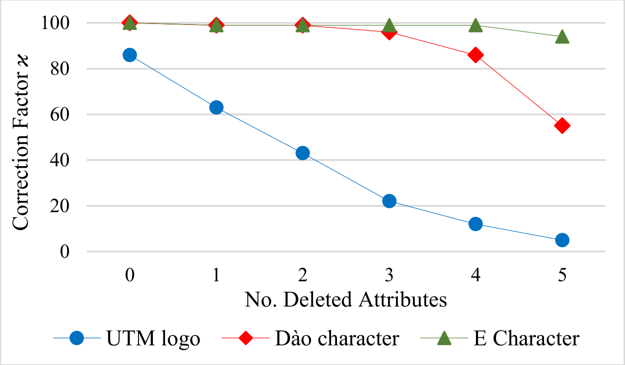


FIGURE 7. Resilience to the deletion problem for each WM, using S-Scheme and erasing more than one attribute.

We also performed an extra experiment for testing the resilience to the *deletion problem*. Despite the *deletion problem* being defined as deleting a single attribute, we stressed the conditions to analyze how the results behave according to the different elements involved. To do that, we deleted more than one attribute and registered an average of the CFs obtained. The firsts results shown are those obtained by using S-Scheme (see Fig. 7).

Once again, the resilience is higher when the WM is composed of fewer marks. It is important to understand that we are erasing more attributes than tolerated since the quality of the data set is compromised for both, the attacker and the data owner, when more than one attribute is deleted. For example, since the relation is composed of ten attributes when two of them are deleted the data lost is 20% of the total amount of data stored.

The second results shown are obtained by performing the experiment under the same conditions but by using M-Scheme. Fig. 8 also shows a direct proportion of the WM length and the resilience to the deletion problem when more than one attribute is deleted.

Finally, considering the differences between S-Scheme and M-Scheme, Fig. 9, 10, and 11 present a direct comparison for each WM. According to the results and the features of each scheme, the data owner will be able to select one of those combinations depending on the requirements to meet.

Fig. 12 shows the differences between the synchronization by using S-Scheme and M-Scheme for all WMs. This parameter is identified as $d_\varkappa$ and is given by the equation $d_\varkappa = |\varkappa_S - \varkappa_M|$, where $\varkappa_S$ is the CF obtained by using S-Scheme and $\varkappa_M$ is the CF obtained by using M-Scheme.

For the case of the WMs generated using Dào and E characters, there are no big differences when 1 or 2 attributes are

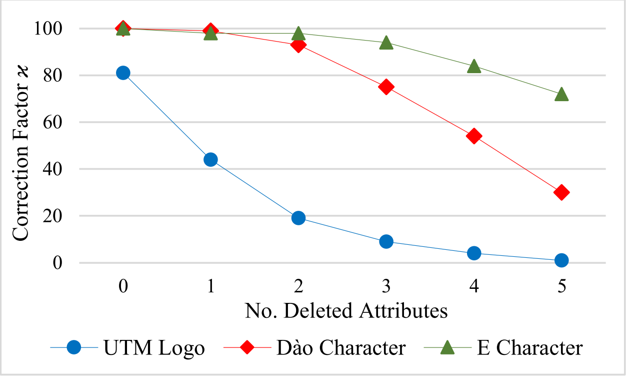


FIGURE 8. Resilience to the deletion problem for each WM, using M-Scheme and erasing more than one attribute.

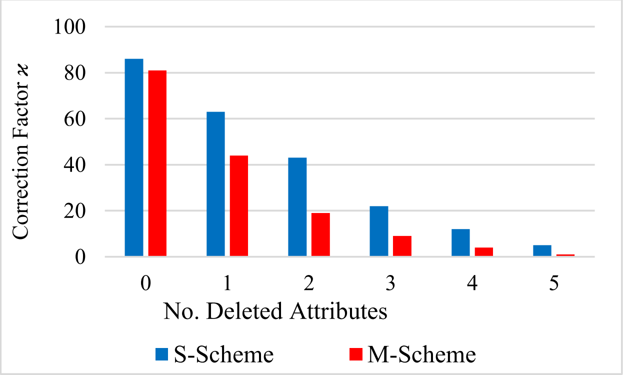


FIGURE 9. Resilience to the deletion problem by erasing more than one attribute for each VPK scheme using UTM as WM source.

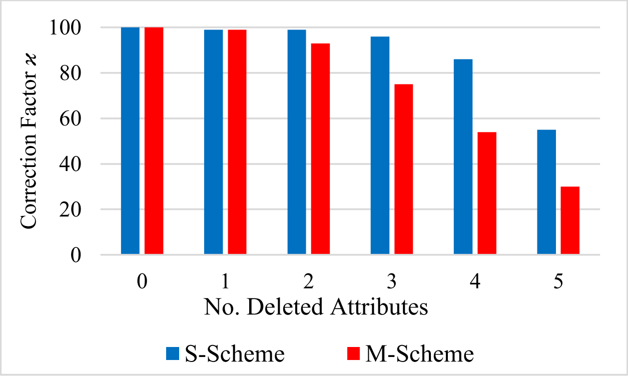


FIGURE 10. Resilience to the deletion problem by erasing more than one attribute for each VPK scheme using Dào chinese character as WM source.

deleted until the damage is too serious (deletion of 3 or more attributes). For the case of UTM watermark, the difference will be always noticeable, especially when 1 to 3 attributes are deleted. For all those cases, always the result using the S-Scheme will be superior, independently the degree of the attack.

C. RESILIENCE AGAINST OTHER THREATS

The improvement of WM synchronization by applying the DF-Technique contributes to increasing the WM robustness

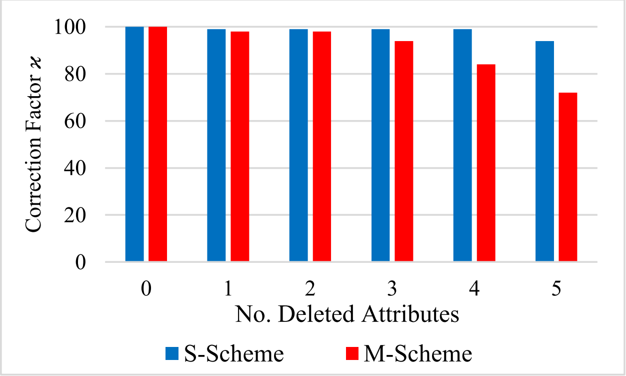


FIGURE 11. Resilience to the deletion problem by erasing more than one attribute for each VPK scheme using E character as WM source.

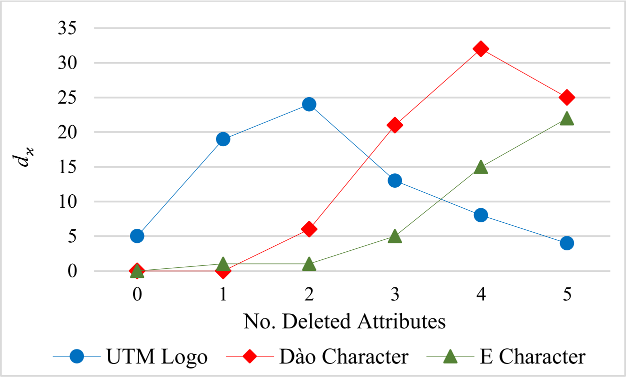


FIGURE 12. Difference between the synchronization achieved by using S and M schemes for each WM.

as well. Besides testing the resilience of our technique against the *deletion problem*, we performed a set of experiments against other threats or malicious operations focused on compromising WM detection. The results obtained showed how the WM signal remains detectable even attacking a considerable amount of data in the watermarked relation. Nevertheless, it is not expected that the watermarked relation will suffer attacks so severe as some of those performed in the experiments.

Since the experiments performed to analyze the resilience against the *deletion problem* are focused on deleting attributes, now we perform a set of experiments erasing a different number of tuples each time. In this case, we are testing the resilience against the *subset attack* based on tuple elimination. We performed the experiment by erasing a different percentage of tuples each time until the entire watermarked data is erased. The attacks were performed over two copies of the data, both marked with the WM generated with the UTM logo, first using the fragmentation technique with the VPK set generated by the S-Scheme, and second by embedding the WM using the same VPK set with no fragmentation. The results of these experiments are shown in Fig. 13.

The quality of the WM embedded with no fragmentation is so low due to the bad quality of the VPK set, that the attacks never get to damage the signal already unrecognizable. On the

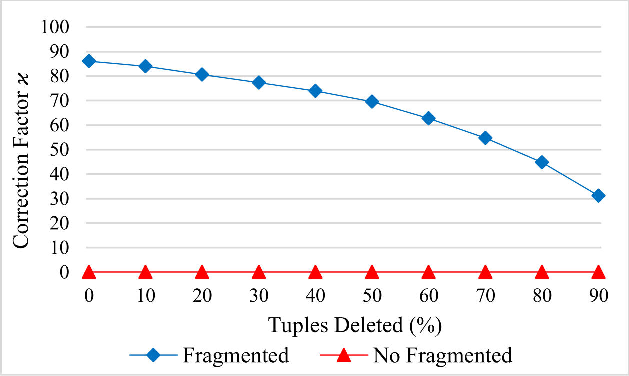


FIGURE 13. WM detected in a relation that suffered different degrees of tuple deletion attack.

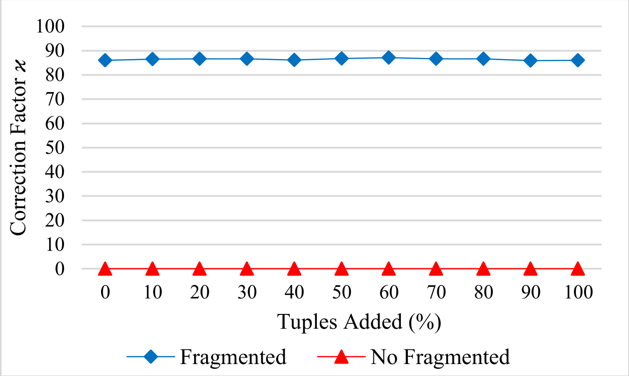


FIGURE 14. WM detected in a relation that suffered different degrees of tuple addition attack.

other hand, the quality of the WM synchronized using the fragmentation technique remains high, allowing its recognition despite the severity of the attacks. Notice that even deleting up to 80% of the tuples (which is not expected considering that the attacker pretends to use the stolen data) the WM presents a quality of $\varkappa = 44.83\%$, being clearly detectable.

We performed another set of experiments focused on detecting the WM in a relation suffering *superset attacks* based on tuple addition. In this case, also the number of added tuples was incremented each time, seeking to analyze the quality of the detected WM once attacks with different severity degrees are carried out. Since the values stored in the embedded tuples are generated randomly using data corresponding to the domains of the attributes of R, the WM signal suffers noise addition, but some of the marks also are enhanced given the random marks that match the right values, detected in the new tuples. This happens due to the probability of obtaining true marks in the tuples added during the attack is 1/2 (given the possible values that the bit selected to extract the mark may contain). In this case, no matter the number of tuples, the quality of the WM remains almost the same. On the contrary, the quality of the WM embedded using the VPK set with no fragmentation is so low that no attacks are needed to compromise it (see Fig. 14).

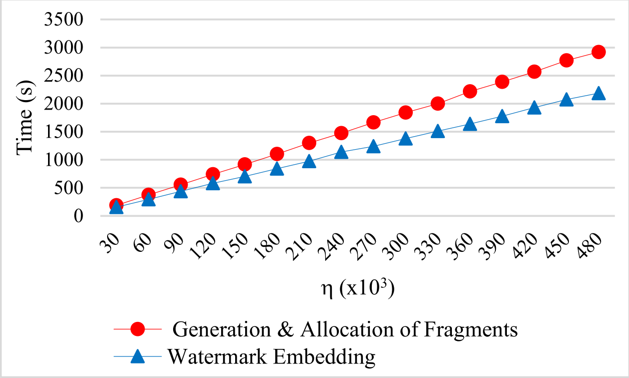


FIGURE 15. Time consumed by the approach's processes using different data amounts.

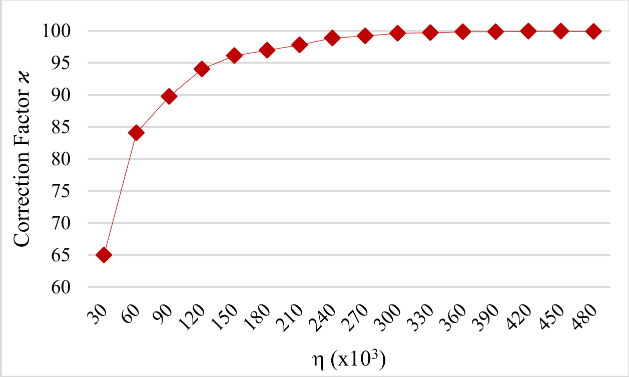


FIGURE 16. Variation of the CF when the amount of data to be watermarked increases.

Given the resilience of our technique against tuple deletion and tuple addition attacks, no experiments were performed for tuple update attacks, since they perform the selection of the data pseudo-randomly, and the probability of affecting attributes being watermarked is lower compared to tuple elimination attacks. On the other hand, to make VPK-based techniques resilient against false ownership claims based on *additive attacks*, we recommend considering the DF-Technique with the Secondary Embedding approach [26] based on creating organized backups for recovering the value of each mark.

D. SCALABILITY OF THE PROPOSAL

For testing the scalability of our approach, all previous experiments were also carried out with different amounts of data, from 30,000 to 480,000 tuples, by adding 30,000 each time. This gave as result the finding of a linear correlation between the time consumed by the execution of the processes and the number of tuples being watermarked (see Fig. 15). Then, the performance of the processes can be expressed by $O(\eta)$ given their linear complexity. According to that, our proposal is feasible independently the volume of the relational data to be protected.

CF values obtained on each one of the experiments performed for the scalability test are shown in Fig. 16.

Considering the increment of CF is not significant from the 120,000 tuples on, we recommend to perform less aggressive embedding by increasing the value of the tuple fraction γ (see Section II-B). The idea is to consider for marking a number of tuples equivalent to the cases previous to the amount of 120,000 shown in Fig. 16. This will allow embedding enough marks for WM recognition whereas the distortion caused under current circumstances is reduced.

V. CONCLUSION

In this paper, we presented a new approach to improve WM synchronization on VPK-based watermarking techniques. We accomplished our goal not by trying to improve the quality of the VPK set but by using the duplicate values stored on it to fragment the WM signal and to embed the fragments according to the particularities of the VPK set. The experiments carried out show how by applying our technique the extracted WM drastically improves. Also, despite the elimination of attributes, the quality of the detected WM remains, being its identification possible.

The experiments were performed using schemes that generate low-quality VPK sets. Nevertheless, the WM synchronization and its resilience against the *deletion problem* increase by combining our approach with any other VPK set, independently the scheme used for its generation. Also, it is important to consider the proportion between the length of the WM to embed and the number of tuples to be watermarked. This may allow the approach to be more resilient, even when more aggressive versions of the *deletion problem* are carried out over the data.

REFERENCES

[1] F. Y. Shih, *Digital Watermarking and Steganography: Fundamentals and Techniques*. Boca Raton, FL, USA: CRC Press, 2017.

[2] Y. Xue, K. Mu, Y. Li, J. Wen, P. Zhong, and S. Niu, "Improved high capacity spread spectrum-based audio watermarking by Hadamard matrices," in *Proc. Int. Workshop Digit. Watermarking*. Cham, Switzerland: Springer, 2018, pp. 124–136.

[3] A. Yahya, *Steganography Techniques for Digital Images*. Cham, Switzerland: Springer, 2019.

[4] K. Unnikrishnan and K. V. Pramod, "Robust optimal position detection scheme for relational database watermarking through HOLPSOFA algorithm," *J. Inf. Secur. Appl.*, vol. 35, pp. 1–12, Aug. 2017.

[5] M. Ahmad, A. Shahid, M. Y. Qadri, K. Hussain, and N. N. Qadri, "Fingerprinting non-numeric datasets using row association and pattern generation," in *Proc. Int. Conf. Commun. Technol. (ComTech)*, Apr. 2017, pp. 149–155.

[6] H. Tufail, K. Zafar, and R. Baig, "Digital watermarking for relational database security using mRMR based binary bat algorithm," in *Proc. 17th IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun./12th IEEE Int. Conf. Big Data Sci. Eng. (TrustCom/BigDataSE)*, Aug. 2018, pp. 1948–1954.

[7] R. Halder and A. Cortesi, "Persistent watermarking of relational databases," in *Proc. IEEE Int. Conf. Adv. Commun., Netw., Comput. (CNC)*, Oct. 2010, pp. 4–5.

[8] S. Bhattacharya and A. Cortesi, "Database authentication by distortion free watermarking," in *Proc. ICSOFT*, 2010, pp. 219–226.

[9] W. Haider, M. Sharif, H. Bashir, M. Raza, and M. Yasmin, "Prefix oriented n4wa coding scheme for improved tampering detection in relational data," *Kuwait J. Sci.*, vol. 43, no. 2, pp. 121–138, 2016.

[10] M. L. Pérez Gort, C. Feregrino-Uribe, A. Cortesi, and F. Fernández-Peña, "HQR-scheme: A high quality and resilient virtual primary key generation approach for watermarking relational data," *Expert Syst. Appl.*, vol. 138, Dec. 2019, Art. no. 112770.

[11] N. Agarwal, A. K. Singh, and P. K. Singh, "Survey of robust and imperceptible watermarking," *Multimedia Tools Appl.*, vol. 78, no. 7, pp. 8603–8633, Apr. 2019.

[12] P. Bas, T. Furon, F. Cayre, G. Doërr, and B. Mathon, *Watermarking Security: Fundamentals, Secure Designs Attacks*. Singapore: Springer, 2016.

[13] S. Rani, D. K. Koshley, and R. Halder, "A watermarking framework for outsourced and distributed relational databases," in *Proc. Int. Conf. Future Data Secur. Eng.* Cham, Switzerland: Springer, 2016, pp. 175–188.

[14] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, *Digital Watermarking Steganography*. Burlington, MA, USA: Morgan Kaufmann, 2007.

[15] R. Halder, S. Pal, and A. Cortesi, "Watermarking techniques for relational databases: Survey, classification and comparison," *J. Universal Comput. Sci.*, vol. 16, no. 21, pp. 3164–3190, 2010.

[16] Y. Li, V. Swarup, and S. Jajodia, "Constructing a virtual primary key for fingerprinting relational data," in *Proc. ACM Workshop Digit. Rights Manage. (DRM)*, 2003, pp. 133–141.

[17] A. S. Alfagi, A. A. Manaf, B. Hamida, and M. G. Hamza, "A systematic literature review on necessity, challenges, applications and attacks of watermarking relational database," *J. Telecommun., Electron. Comput. Eng. (JTEC)*, vol. 9, nos. 1–3, pp. 101–108, 2017.

[18] J. Sun, Z. Cao, and Z. Hu, "Multiple watermarking relational databases using image," in *Proc. Int. Conf. Multi Media Inf. Technol.*, Dec. 2008, pp. 373–376.

[19] H. M. Sardroudi and S. Ibrahim, "A new approach for relational database watermarking using image," in *Proc. 5th Int. Conf. Comput. Sci. Converg. Inf. Technol.*, Nov. 2010, pp. 606–610.

[20] C. J. Date, *An Introduction to Database Systems*. London, U.K.: Pearson, 2006.

[21] R. Agrawal and J. Kiernan, "Watermarking relational databases," in *Proc. 28th Int. Conf. Very Large Data Bases*, 2002, pp. 155–166.

[22] C.-C. Chang, T.-S. Nguyen, and C.-C. Lin, "A blind robust reversible watermark scheme for textual relational databases with virtual primary key," in *Proc. Int. Workshop Digital Watermarking*. Cham, Switzerland: Springer, 2014, pp. 75–89.

[23] V. Khanduja, S. Chakraverty, and O. P. Verma, "Enabling information recovery with ownership using robust multiple watermarks," *J. Inf. Secur. Appl.*, vol. 29, pp. 80–92, Aug. 2016.

[24] M. L. P. Gort, E. A. Diaz, and C. F. Uribe, "A highly-reliable virtual primary key scheme for relational database watermarking techniques," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2017, pp. 55–60.

[25] U. Colorado State. (Jun. 1999). *Forest Covertype, the UCI KDD Archive*. [Online]. Available: http://kdd.ics.uci.edu/databases/covertype/covertype.html

[26] M. L. P. Gort, M. Olliaro, C. Feregrino-Uribe, and A. Cortesi, "Preventing additive attacks to relational database watermarking," in *Proc. Int. Conf. Res. Practical Issues Enterprise Inf. Syst.* Cham, Switzerland: Springer, 2019, pp. 131–140.

MAIKEL LÁZARO PÉREZ GORT received the M.Sc. degree in applied informatics from the Universidad Tecnológica de La Habana. He is currently pursuing the Ph.D. degree in computer science with the National Institute of Astrophysics, Optics, and Electronics (INAOE), Puebla, Mexico. His research interests include relational databases theory, information security and privacy, and data usability and authenticity.

CLAUDIA FEREGRINO-URIBE received the B.Sc. degree in computer systems engineering from the Querétaro Institute of Technology, the M.Sc. degree in electrical engineering with telecommunications option from CINVESTAV, Guadalajara, and the Ph.D. degree in electronic engineering in digital systems from Loughborough University, U.K. She is currently pursuing the M.Sc. (Tech.) degree in security program coord. She is also a Researcher with the Computer Science Department, National Institute of Astrophysics, Optics and Electronics (INAOE), Puebla, Mexico. Her research interests include cryptography, watermarking, and digital systems design.

AGOSTINO CORTESI is currently a Full Professor of computer science with Ca' Foscari University, Venice, Italy. He has extensive experience in the area of static analysis and software verification techniques. In particular, he contributed to the design and practical evaluation of abstract domains within the Abstract Interpretation framework. He coordinates the MAE Italy–India project 2017–2020 Formal Specification for Secured Software System.

FÉLIX FERNÁNDEZ-PEÑA received the M.Sc. degree in telematics and the Ph.D. degree in electronics and computer science from the Universidad Tecnológica de La Habana. He is currently the Director of the Applied Informatics Research Group, Universidad Técnica de Ambato, Ecuador, and a member of the Ecuadorian Research Network of Computer Science. He is currently a Professor of cyber-security with the Universidad Técnica de Ambato. His research interests include the use of machine learning in the detection of cyber-attacks, cryptography, and watermarking. His participation in this research was as a researcher of Universidad Espíritu Santo, Guayaquil, Ecuador.