

Elsevier Editorial System(tm) for Expert
Systems With Applications + OA Mirror
Manuscript Draft

Manuscript Number: ESWA-D-19-01580

Title: HQR-Scheme: A High Quality and Resilient Virtual Primary Key
Generation Approach for Watermarking Relational Data

Article Type: Full length article

Keywords: duplicate problem; deletion problem; quality measuring;
relational data; robust watermarking; virtual primary key

Corresponding Author: Dr. Claudia Feregrino-Uribe, Ph.D.

Corresponding Author's Institution: National Institute for Astrophysics,
Optics and Electronics (INAOE)

First Author: Maikel L Pérez Gort, M.Sc.

Order of Authors: Maikel L Pérez Gort, M.Sc.; Claudia Feregrino-Uribe,
Ph.D.; Agostino Cortesi, Ph.D.; Félix O Fernández Peña, Ph.D.

April 4th, 2019

Dr. Binshan Lin,
Editor-in-Chief
Elsevier Expert Systems with Applications Journal

Dear Dr. Binshan Lin,

We are submitting the following article for your consideration to be published as a regular paper at Elsevier Expert Systems with Applications Journal, entitled “*HQR-Scheme: A High Quality and Resilient Virtual Primary Key Generation Approach for Watermarking Relational Data*” by Maikel Lázaro Pérez Gort, Claudia Feregrino-Uribe, Agostino Cortesi and Félix Oscar Fernández-Peña.

This work propose metrics to allow a precise measuring of the quality of the Virtual Primary Keys (VPK) generated by any VPK scheme proposed so far, without requiring to perform the watermark embedding, so wasting time can be avoided in case of low-quality detection. We also analyze the main aspects to design the ideal VPK scheme, seeking the generation of high-quality VPK sets adding robustness to the process. Finally, a new scheme is presented along with the experiments carried out to validate and compare the results with the rest of the schemes proposed in the literature.

We believe that these findings will be of interest to the readers of Elsevier Expert Systems with Applications Journal.

This manuscript is original, has not been published before and is not currently being considered for publication elsewhere.

In case of any questions please do not hesitate to contact us.

The corresponding author is:
Claudia Feregrino-Uribe
Computer Science Department, National Institute for Astrophysics Optics and Electronics
Luis Enrique Erro No. 1, Tonantzintla, Puebla México, CP 72480
Tel.: +52 222 2663100 ext. 8229
E-mail: cferegrino@inaoep.mx, cferegrino@gmail.com

Kind Regards,

Maikel Lázaro Pérez Gort, Claudia Feregrino-Uribe, Agostino Cortesi and Félix Oscar Fernández-Peña

Available online at www.sciencedirect.com

Expert Systems with Applications 00 (2019) 1–27

~~Expert~~
Systems
with Appli-
~~cations~~

HQR-Scheme: A High Quality and Resilient Virtual Primary Key Generation Approach for Watermarking Relational Data

Maikel Lázaro Pérez Gort^a, Claudia Feregrino Uribe^{a,*}, Agostino Cortesi^b, Félix Oscar Fernández-Peña^c

^a*Instituto Nacional de Astrofísica, Óptica y Electrónica. Luis Enrique Erro 1, Sta María Tonanzintla, 72840 Puebla, México*

^b*Università Ca Foscari di Venezia. Campus Scientifico Via Torino, 155 30172 Mestre (VE), Italy*

^c*Universidad Técnica de Ambato. Av. Los Chasquis y Río Payamino. Campus Huachi, 180207 Ambato, Ecuador*

Abstract

Most of the watermarking techniques designed to protect relational data often use the Primary Key (PK) of relations to perform the watermark synchronization. Despite offering high confidence to the watermark detection, these approaches become useless if the PK can be erased or updated. A typical example is when an attacker wishes to use a stolen relation, unlinked to the rest of the database. In that case, the original values of the PK lose relevance, since they are not employed to check the referential integrity. Then it is possible to erase or replace the PK, compromising the watermark detection with no need to perform the slightest modification to the rest of the data. To avoid the problems caused by the PK-dependency some schemes have been proposed to generate Virtual Primary Keys (VPK) used instead. Nevertheless, the quality of the watermark synchronized using VPKs is compromised due to the presence of duplicate values in the set of VPKs and the fragility of the VPK schemes against the elimination of attributes. In this paper, we introduce the metrics to allow a precise measuring of the quality of the VPKs generated by any scheme without requiring to perform the watermark embedding, so wasting time can be avoided in case of low-quality detection. We also analyze the main aspects to design the ideal VPK scheme, seeking the generation of high-quality VPK sets adding robustness to the process. Finally, a new scheme is presented along with the experiments carried out to validate and compare the results with the rest of the schemes proposed in the literature.

Keywords: Duplicate problem, Deletion problem, Quality measuring, Relational data, Robust watermarking, Virtual primary key

1. Introduction

Databases have been traditionally protected by security methods oriented to control the access of users according to their role in the system. This approach has been mostly implemented through authentication policies and has been backed by other operations of the database management systems like log recording and backups management. Nevertheless, with the increasing of data demands and internet services, portable data and remote accessing of the information are growing in popularity. As a result, the access of data by unauthorized users is even easier, being possible to make illegal copies or to perform data tampering without leaving traces through the network. For that reason, it can be said that traditional database security techniques are not enough to guarantee the protection of the data.

To face those problems, alternative security methods such as watermarking techniques have been proposed. Watermarking techniques allow the protection of the data without restraining their deployment or copying, being that

*Corresponding author

Email addresses: mlazaro2002es@inaoep.mx (Maikel Lázaro Pérez Gort), cferegrino@inaoep.mx (Claudia Feregrino Uribe), cortesi@unive.it (Agostino Cortesi), fo.fernandez@uta.edu.ec (Félix Oscar Fernández-Peña)

an important reason for their rising popularity. By means of watermarking, data can be protected from tampering, stealing, unauthorized distribution, and many other malicious operations. Watermarking techniques were first proposed for protecting multimedia data (Choudhary et al., 2017; Nematollahi et al., 2017; Wang, 2017) and then were extended to relational data (Ghogare and Junnarkar, 2017; Khanduja, 2017; Rani et al., 2017; Unnikrishnan and Pramod, 2017). Performing watermarking over relational data have not been an easy task due to their features and differences respect to multimedia data types (Halder et al., 2010; Iftikhar et al., 2015b,a).

The essential element of a watermarking technique is the Watermark (WM), which consists of a set of items called marks mostly represented by bits. The WM must be imperceptibly embedded in the digital asset being protected, preserving the data usability. In that way, malicious users cannot get clues of the WM presence, since they may try to remove it to claim the data ownership. There are some requirements (Halder et al., 2010; Khanduja et al., 2016a; Mehta and Aswar, 2014; Pérez Gort et al., 2017b; Shih, 2017; Xie et al., 2016) that watermarking techniques must fulfill to avoid compromising data quality and prevent other undesirable situations. Also, for the extraction, the marks should remain in the watermarked content, at least in a quantity enough to guarantee the WM recognition.

Among the multiple criteria used for WM classification (Halder et al., 2010; Patil and Yawalkar, 2015; Prajapati and Tiwari, 2015), there is one linked to their intent (see Fig. 1). Some watermarking techniques have been developed for copyright protection (Jiang et al., 2009; Khanduja et al., 2015; Rao et al., 2012; Zhang et al., 2011), while others named fingerprinting, are oriented to detect traitor users as well as to check the authenticity of the data copies (Gursale and Mohanpurkar, 2014; Iftikhar et al., 2014; Mohanpurkar and Joshi, 2015). These techniques are classified as robust, considering the severity of the attacks aimed to compromise them, and the expected resiliency against them. On the other hand, WMs can also be used for monitoring the data integrity and protect them against malicious modifications like tampering and fraud (Camara et al., 2014; Chang and Wu, 2012; Guo, 2011; Khan and Husain, 2013; Şahin et al., 2016). These WMs are classified as fragile since the owner of the data not only knows of its presence but also is benefited from it. It is for that reason that it is assumed it will not suffer attacks focused on compromising their functioning.

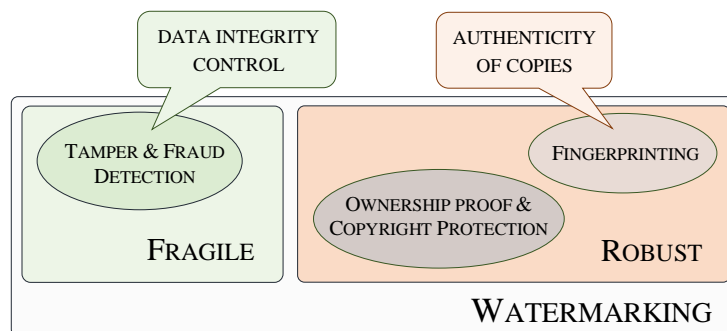


Fig. 1. Classification of watermarking techniques according to their intent.

Generally, watermarking techniques are composed of two processes: (A) WM embedding and (B) WM extraction (see Fig. 2). Embedding is composed of two sub-processes, WM generation and marks embedding. WM generation source could be a multimedia file, the database itself, among others (Halder et al., 2010; Agrawal et al., 2003). WM extraction is usually performed when it is requested as evidence in some litigation. For that, the WM should remain in the data no matter how long it has been embedded. For relational data, this is a major challenge compared to other data types since the information stored in databases are daily modified through operations called benign updates (addition, actualization, and elimination of data). Also, robust techniques must guarantee resilience against malicious attacks. The sub-processes composing the extraction are the detection of marks, their extraction, and WM reconstruction. Belonging to the extraction process there is also the WM enhancement optional sub-process that is very useful for cases when the WM signal is meaningful (e.g. when the WM is generated from a multimedia file). This operation makes the technique more resilient to attacks since can tolerate losing more marks and still achieve the WM identification once it is extracted and enhanced. Also, it can be used to cause a lower distortion during the embedding, helping to avoid compromising the data usability in the process. For both processes (the embedding and the extraction) the value of

the private parameters must be the same. Otherwise, WM synchronization will not be achieved. Finally, the simplest watermarking processes require at least the consideration of one parameter defined as the Secret Key (SK), which has to be only known by the data owner (Agrawal and Kiernan, 2002).

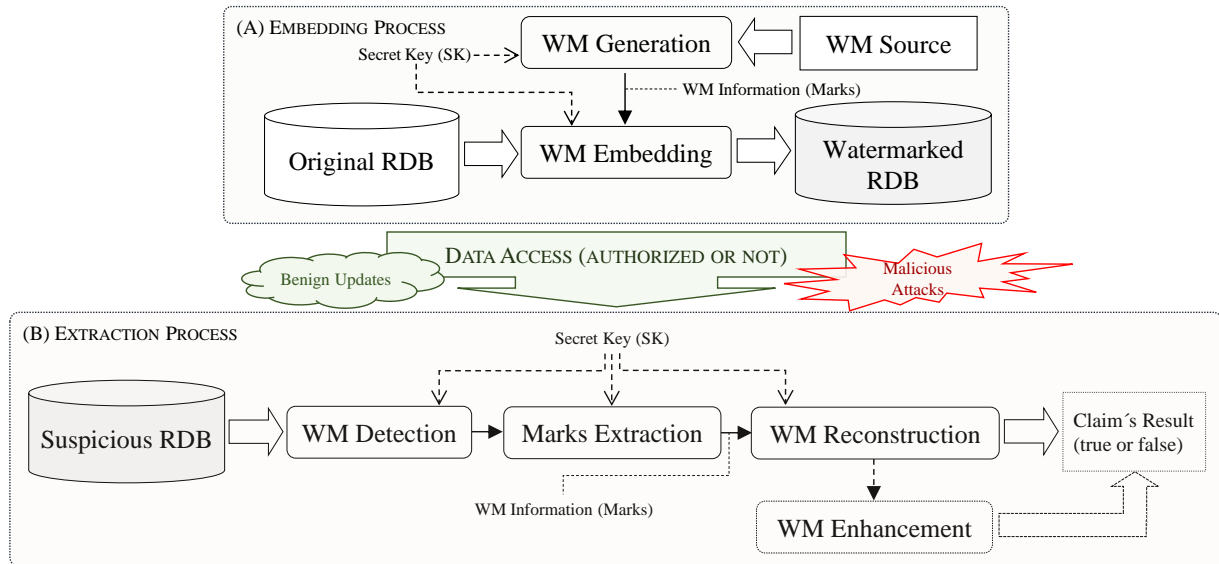


Fig. 2. Processes involved in watermarking techniques.

Since the first proposal, relational database watermarking techniques have been growing in diversity, backed by a broad theory. Nevertheless, most of the watermarking techniques use the Primary Key (PK) of the relation to decide where and how to embed each mark. The PK stores unique values that identify each tuple in the relation, which is why using it allows high synchronization of the WM. Watermarking techniques that use the PKs are backed by the assumption that there is no way to delete the PKs without compromising the referential integrity of the database, due to their role in the database design. That is why the attackers trying to compromise the WM detection are forced to modify the data, without tampering the PK, since they are also interested in maintaining the quality of the data. Based on that assumption, the majority of watermarking techniques proposed to protect relational data are PK-dependent (e.g. Franco-Contreras et al., 2014; İmamoğlu et al., 2015; Kamran et al., 2013; Melkundi and Chandankhede, 2015; Pérez Gort et al., 2017b).

Despite the PK relevance in database design, watermarked relations are often distributed separately from the database, which allows compromising the WM detection by erasing or updating the PK if the attacker has no intention of using the relation linked to the rest of the database. This has become the perfect attack since the attacker does not require the modification of any data beyond the PK. To avoid this vulnerability (or the absence of PK in the relation), some proposals have been published under the classification of Virtual Primary Key (VPK) schemes, oriented to generate VPKs to be used by watermarking techniques instead of the PKs. The work presented here is oriented to solve some weakness presented by most VPK schemes as well as to find a better way to measure the quality of the VPK set, giving so a more precise idea of how effective can be its use by watermarking techniques.

The rest of this paper is organized as follows. Section 2 introduces details concerning the relational data structure and the way watermarking techniques have approached WM embedding in relational data. Section 3 describes the challenges VPK schemes face to be functional and considered for watermarking relational data. Section 4 gives a critical review of the VPK schemes proposed so far. Section 5 presents the set of metrics that better describes the quality of the VPKs and the resilience of a VPK scheme against the elimination of attributes. Section 6 presents the elements describing the ideal design for a VPK scheme and our approach seeking the maximum matching to this design. Section 7 presents the experimental results allowing the comparison between previous schemes and our approach. The conclusion of this paper is given in Section 8.

2. Watermarking relational data

According to the relational model, relational databases are conceived as a collection of relations corresponding to the entities considered in its design (Date, 2006). The relations are linked using their primary keys, implementing in that way the referential integrity between their records. Each relation presents a table structure where the columns are the attributes and the rows are tuples representing the instances of the entities (see Fig. 3). Watermarking techniques for relational databases so far consider the simplest scenario of a single relation R with primary key PK . The attributes of R are identified by $A_i : i \in [0, \nu - 1]$, being ν the number of attributes available for marking, and the tuples by $r_j : j \in [0, \eta - 1]$, being η the number of tuples stored in the relation. The complete representation of the relation is given by $R(PK, A_0, \dots, A_{\nu-1})$. Also, the value stored in attribute i of tuple j is accessed by using the notation $r_j.A_i$. Finally, the primary key of each tuple is identified by $r_j.PK$.

		Attributes					
		A_0	A_1	A_2	...	$A_{\nu-1}$	
		ID	NAME	AGE	GENDER	...	HEIGHT
Tuples	r_0	1011	Alex	32	Male	...	175
	r_1	1012	Cindy	24	Female	...	160
	\vdots
	$r_{\eta-1}$	1015	Layla	30	Female	...	177

Fig. 3. Table structure of a database relation.

In 2002, Agrawal and Kiernan (2002) proposed the first watermarking technique for relational data. Their approach uses notations describing the relation structure presented before. The technique watermarks numeric attributes, which classifies it as a numerical cover-type WM. This technique uses the binary representation of the integer data type to embed the mark in one of the less significant bits (*lsb*). The *lsb* to store the mark is selected in the same way as the tuple and the attribute inside the tuple are selected, by using the value generated by the combination of the Secret Key (SK) chosen by the data owner and the PK of each tuple (See Algorithm 1). The value used to select the embedding place of the mark is a VPK generated involving the PK according to Eq. (1), being \circ the concatenation operator and H a common one-way hash function such as SHA-2, SHA-3, among others. This approach is also known in the literature as the AHK Algorithm.

$$F(r_j.PK) = H(SK \circ H(SK \circ r_j.PK)) \tag{1}$$

Algorithm 1: AHK Approach.

```

1 foreach tuple  $r \in R$  do
2   if  $F(r.PK) \bmod \gamma = 0$  then
3     attribute_index  $i \leftarrow F(r.PK) \bmod \nu$ 
4     bit_index  $b \leftarrow F(r.PK) \bmod \xi$ 
5      $r.A_i \leftarrow \text{mark}(r.PK, r.A_i, b)$ 

```

Agrawal & Kiernan define ω as the number of watermarked tuples after the embedding process is concluded. The AHK algorithm selects a tuple for embedding a mark if $F(r_j.PK) \bmod \gamma = 0$, being $\gamma \in [1, \eta]$ the tuple fraction parameter. With low values of γ more tuples are marked, and with high values, the number of marked tuples reduced. For the case when $\gamma = 1$, all tuples are marked if the usability constraints implemented over the relation allow it. Eq. (2) establishes the proportion between the number of tuples in R , the tuple fraction γ and the number of marked tuples ω . Also, once a tuple has been selected, $F(r_j.PK)$ is used for selecting the attribute to embed the mark according to the

expression $F(r_j.PK) \bmod v$. The selection of the bit for embedding the mark is done according to $F(r_j.PK) \bmod \xi$, where ξ is the range available to perform the *lsb* selection.

$$\omega \approx \eta/\gamma \quad (2)$$

The AHK algorithm is commonly used by other techniques to select the place for embedding the marks. On another hand, the way to generate the mark often varies for each technique, compared to the way that the approach of Agrawal & Kiernan does (see Algorithm 2).

Algorithm 2: AHK mark method.

```

1 Input:  $PK, A, b$ 
2  $hash\_val = H(SK \circ PK)$ 
3 if  $hash\_val$  is even then
4    $\lfloor$  set the  $b^{th}$  lsb of  $A$  to 0
5 else
6    $\lfloor$  set the  $b^{th}$  lsb of  $A$  to 1
7 return  $A$ 

```

In 2003, Li et al. (2003) named the tuple selection using the AHK algorithm as the T-Scheme. This approach has been proved to be robust against *set attacks* like the elimination or modification of tuples or attributes. This is mainly due to: (i) every tuple is marked independently; (ii) every mark is embedded multiple times, and (iii) a majority vote is used in the WM reconstruction (Li et al., 2003). Since it was published, the AHK algorithm has become the main model followed for PK-dependent techniques, which constitute the majority of watermarking techniques for relational data.

3. Synchronization problems

Watermarking techniques using the PK of the relation for performing WM embedding and its extraction achieve high synchronization considering that the PK stores unique values used to identify each tuple. Then, high entropy is added to the selection of marks as well as to the places to embed them, which also makes the technique more robust since the WM capacity increases while more marks are considered. The main fragility of this approach is that if the PK can be erased or updated, then WM detection is compromised with no need of attacking any other data of the relation. This allows the attackers to get good quality data, compromising the WM detection after performing a simple malicious operation. To avoid that fragility, VPK schemes have been proposed for generating VPK which are then used to perform the WM synchronization instead of the PK. Nevertheless, when VPKs are used by watermarking techniques other problems emerge. In this section the main challenges of using VPK for watermarking relational data are presented.

3.1. The duplicate problem

The set of VPKs generated using the tuples and attributes of the relation always present duplicate values which compromise WM synchronization compared to the results obtained when PKs are employed. The generation of duplicate values is due to VPK schemes only can use data stored in the relation, which are bounded by the domain of the attributes. Using duplicate keys cause embedding the same marks multiple times while others are completely ignored. This results in an incomplete WM embedding, compromising its recognition. Because of that, the degree of damage caused to the technique is so high that there is no need to perform an attack when the watermarked data is stolen. Mark exclusion is an issue that usually occurs, even in PK-based techniques, when the pseudo-random selection of the marks and of the place for mark embedding in the relation is implemented to make the technique resilient against the *subset-reverse order attack* (Pérez Gort et al., 2017b; Sardroudi and Ibrahim, 2010). The consequences of this problem get worst when there are duplicate values in the keys, causing the complete compromise of the WM detection.

Defined as the *duplicate problem* by Li et al. (2003) in 2003, the presence of duplicate values in the set of VPKs increases the damage to the WM synchronization caused by excluding marks in the WM embedding. Normally, embedding the same marks multiple times is something positive to avoid attacks based on updating or deleting data from the relation. The damage caused by those attacks is restored when a majority vote is performed after WM extraction, resolving the inconsistencies for the extraction of different values corresponding to the same mark. On the other hand, embedding the same marks multiple times at expenses of excluding others from the process compromises WM synchronization. This causes similar damage to the attacks performing extreme data elimination (e.g. tuple elimination, attribute elimination, among others) (Pérez Gort et al., 2017b; Sardroudi and Ibrahim, 2010).

3.2. The deletion problem

The main reason for using VPKs instead of PK of the relation for performing WM synchronization is that if the PK is erased, WM detection is compromised. This problem cannot be solved by using another attribute instead of the PK since the same risks remain as long as the attribute employed is irrelevant to the attacker. Even if more than one attribute is used to perform the VPKs generation, the attacker can erase one of them, compromising the VPKs and, affecting WM synchronization. This is due to the addition of noise to the extracted WM as a result of getting wrong values for some marks for using VPKs that do not match those used for performing WM embedding. This issue was defined as the *deletion problem* also by Li et al. (2003) in 2003.

4. Previous work

The first VPK scheme reported in the literature was proposed by Agrawal and Kiernan (2002) to apply the AHK algorithm in relations with no PK. Identified as the S-Scheme by Li et al. (2003) in 2003, this scheme is conceived to be used by watermarking techniques trying to protect relations composed by one or more than one numeric attributes. When only one attribute composes the relation, the VPK is created by splitting the binary value of the attribute into two fragments, the first one corresponding to the most significant bits *msb* used for the VPK generation and the second one corresponding to the less significant bits *lsb*, used for embedding the marks (see Section 2). When more than one attribute composed the relation, the VPK is generated by using only one attribute, and the rest of them are selected to perform the WM embedding. This scheme also considers varying the attribute selected for the VPK generation from one tuple to the other one.

The *msb* is selected from a given range identified by χ according to Eq. (3), where $[r_j.A]_2$ represents the value of the attribute A for the tuple r_j in binary notation. Then, the number generated from the selected range χ is represented in decimal notation, constituting the VPK for that tuple. Considering that the value assigned to χ is the same for the generation of the VPKs for all the tuples, the number of duplicated values obtained by applying this scheme is very high, being severely affected by the *duplicate problem*. Also, since so few attributes are involved in the VPK generation, this scheme is very fragile to the *deletion problem*.

$$vpk(r_j) = [VPK([r_j.A]_2, \chi)]_{10} \quad (3)$$

Another approach for the generation of VPKs is the E-Scheme. This approach was proposed by Li et al. (2003) in 2003 and works similarly to the S-Scheme using the same value of χ the whole generation process, which makes it also vulnerable to the *duplicate problem*. The main particularity of the E-Scheme is that is applied over each attribute of each tuple, to generate one VPK per each value stored in the relation. According to that, the number of VPK generated is $\eta \times \nu$ instead of η , which is the case for the S-Scheme. The E-Scheme also uses two different hash functions, one for deciding if the attribute will be watermarked and the other one for selecting the *lsb* to embed the mark once the attribute is selected for the embedding. Each hash function is identified as H_1 and H_2 respectively. The condition to select the attribute is given by $H_1(SK \circ vpk(r_j.A_i)) \bmod \gamma \times \nu = 0$, being $vpk(r_j.A_i)$ a variation of Eq. (3) to be applied over each attribute of each tuple. When an attribute is selected, the *lsb* being watermarked in the attribute is given by $H_2(SK \circ vpk(r_j.A_i)) \bmod \xi$. This scheme is more resilient to the *deletion problem* considering all attributes of the relation are involved. The issue with this scheme is that the *duplicate problem* compromises WM synchronization even more than for the S-Scheme since more VPK are generated using the same value of χ .

Li et al. (2003) also proposed the M-Scheme for generating the VPKs considering more than one attribute per tuple. This approach tries to select different attributes each time, being in principle more resilient to the *deletion problem*

in comparison to previous proposals. Nevertheless, if the number of attributes considered is too close to ν , then the *deletion problem* becomes a serious threat. Same as the S-Scheme, the M-Scheme splits each attribute involved in the VPK generation into two parts: the *msb* and the *lsb*. Then, for each tuple, VPK generation is performed by joining the results of $H_1(SK \circ vpk(r_j, A_i))$ closest to zero. The number of results being joined is two by default, but if more than two attributes match the same hash closest to zero, all of them can be considered. Finally, the default number of attributes involved can be settled by the data owner, under the condition that has to be equal or higher than ν . This scheme also uses the same value of χ each time, which makes it vulnerable to the *duplicate problem*.

Other VPK schemes are the proposals of Chang et al. (2014) and Khanduja et al. (2016b). Both of them are very similar to the M-Scheme, since they consider the two attributes that generate the closest hash value to zero. As a difference, the approach of Chang et al. uses textual attributes instead of numerical, splitting each attribute in the fragment A_i^1 (containing the last word of the text and used to embed the marks) and A_i^{-1} (containing the rest of the text that is input to the hash function for checking if the attribute will be involved in the VPK generation). On another hand, the approach of Khanduja et al. is oriented to generate VPKs using numeric attributes, similar to the M-Scheme. In this case, only two specific attributes are selected for the VPK generation, based on the data owner criterion.

Finally, there is the proposal by Pérez Gort et al. (2017a) focused on selecting each time different attributes and changing the value of χ for each one of them. We called this approach the Ext-Scheme considering the extension that was given to the variability of the selection of each element involved on it. The Ext-Scheme is designed for numerical attributes and analyzes each tuple, and each attribute per tuple, generating the value of χ for each attribute considering the values of the neighbors bits to the χ bit. The position for the χ bit is generated according to Eq. (4), being BL_{ji} the binary length of the value stored in the attribute r_j, A_i and $MSBF$ the Most Significant Bits Fraction, which takes the same value during the entire process.

$$\chi_{ji} = \lfloor BL_{ji}/MSBF \rfloor \quad (4)$$

For each tuple, the attributes considered for VPK generation are those that accomplish the condition $b_{ji1} \oplus b_{ji(\chi)} = 1$, being b_{ji1} the 1st *msb*, $b_{ji(\chi)}$ its (χ) th *msb* and \oplus the XOR operator. These considerations contribute to selecting different attributes for each tuple, and to generate different values of χ each time, making the approach more resilient to the *deletion* and *duplicate problems*. However, an important drawback for the Ext-Scheme is that sometimes the condition for selecting the attributes for VPK generation is not accomplished for any attribute of the tuple. Because of that, an important number of tuples may be excluded from the process, missing the opportunity to embed marks and affecting WM synchronization.

5. Measuring quality and resiliency

As far as we know, the only metric for quantifying the effects of the *duplicate problem* for WM synchronization was proposed by Li et al. (2003). They defined the equation $\delta = (w_{max} - w_{min})/w_{min}$, being δ the duplicate index, w_{max} the maximum number of times the same mark is embedded and w_{min} the minimum. According to this equation, the ideal WM embedding is accomplished when all marks are equally embedded, being $w_{max} = w_{min}$, resulting in $\delta = 0$. But when some mark is excluded from the embedding, $w_{min} = 0$, being $\delta = \infty$, the worst-case scenario no matter the number of marks excluded. This metric can only be applied once the WM is embedded but is useless for measuring the quality of the set of VPKs separately from the watermarking technique. Also, it will get the same result no matter the number of marks excluded by the embedding process, which affects the WM synchronization differently. Finally, as previously referred, the exclusion of marks usually happens even if no VPKs are being used, when pseudo-random selection is adopted to avoid the consequences of the *subset-reverse order attack*.

Considering the limitations of the metric by Li et al., we propose new metrics for measuring the quality of the VPK sets without performing the watermarking process, so if the quality of the VPK set is low, the time consumed by the watermarking processes can be employed in generating or selecting another VPK set. Besides measuring of the quality of the VPK set, we also measure the resiliency of the VPK scheme to the *deletion problem*, requiring for this case the embedding of the WM since the *deletion problem* consists in eliminating an attribute in the watermarked relation.

To quantify any feature of a VPK set it is essential to know its structure. In this work, we use S to identify a generic VPK set. Given the conditions for the generation of S , it is assumed the presence of duplicate and exclusive values on it. The subset E is formed by the exclusive values and G by the duplicate ones. Also, one VPK per duplicate group is selected and stored in the set D , allowing to know how many duplicate groups are in S . In this paper, we use the

lowercase letter corresponding to each set to identify their cardinality (e.g. $s = |S|$, $e = |E|$, and so on). Using that notation the rules $s = e + g$, and $d < g$ are established. Table 1 shows different samples of S and their subsets cardinality.

Although the elements in S do not present a fixed order, for making more comfortable its appreciation, we first present the exclusive values in bold style, then each group of duplicate values is highlighted with different color. The groups of duplicate values are ordered according to their cardinality.

Table 1. Different examples of S .

No.	$S, s = 10$	e	g	d
1	{ 12 , 34 , 21 , 45 , 13 , 37 , 15 , 48 , 22 , 82 }	10	0	0
2	{ 21 , 45 , 15 , 48 , 22 , 82 , 12, 12, 34, 34}	6	4	2
3	{ 22 , 82 , 12, 12, 12, 12, 12, 12, 12, 12}	2	8	1
4	{15, 15, 15, 34, 34, 34, 12, 12, 12, 12}	0	10	3
5	{15, 15, 34, 34, 12, 12, 12, 12, 12, 12}	0	10	3
6	{12, 12, 12, 12, 12, 12, 12, 12, 12, 12}	0	10	1

The cardinality of each subset is not enough to reflect the difference between the VPK sets. There are cases when the cardinality of some subsets is the same while S is different (see records 4 and 5 of Table 1). Then it is important to establish the difference of the groups of duplicate values stored in G . For that, we identify each group as the subsets $G_r \in [1, d]$ and the cardinality of each subset as g_r , where $g = \sum_{r=1}^d g_r$.

To know the quality of S for watermarking relational data is it important to know the number of different values (not just the exclusive ones) stored in the set. That is because the number of different values could be higher than the number of marks, allowing high WM synchronization despite the redundancy S presents. Then, being $n = length(WM)$, the desired scenario is that $(e + d) \gg n$. If S does not present duplicate values, the exclusion of marks is reduced at the same level that when the PKs of the relation are used by the watermarking technique. That is why exclusive values play a critical role in the quality of the set of VPK. Then, the best set of VPKs is defined according to the following order: (i) the set with more exclusive values, (ii) the set with more duplicate groups.

To highlight those differences properly we defined the index of exclusiveness of S , denoted by ρ , according to Eq. (5). This metric measures the different values presented in S as a percentage according to its cardinality. When only exclusive values are composing S , then $\rho = 100$, which is the highest possible value of the index. The summation is used to accumulate the inverse of the cardinality of each group of duplicate values, since the presence of bigger groups compromises the presence of different values in S , by reducing the space for other duplicate groups or exclusive values. Also, since each exclusive value is considered as a group of one item, the weight in the equation is 1. They are excluded for the summation and considered in the cardinality of the set of exclusive values e .

$$\rho = \frac{(\sum_{r=1}^d 1/g_r + e) \times 100}{s} \tag{5}$$

The higher ρ the better, giving more options to the watermarking scheme to select different marks each time. Knowing the value of the exclusiveness index it is possible to describe better the features of S . Table 2 shows the samples of S presented in Table 1 with their corresponding ρ .

Table 2. Value of ρ for the examples of S from Table 1.

No.	$S, s = 10$	e	ρ
1	{ 12 , 34 , 21 , 45 , 13 , 37 , 15 , 48 , 22 , 82 }	10	100
2	{ 21 , 45 , 15 , 48 , 22 , 82 , 12, 12, 34, 34}	6	70.0
3	{ 22 , 82 , 12, 12, 12, 12, 12, 12, 12, 12}	2	21.3
4	{15, 15, 15, 34, 34, 34, 12, 12, 12, 12}	0	9.17
5	{15, 15, 34, 34, 12, 12, 12, 12, 12, 12}	0	11.7
6	{12, 12, 12, 12, 12, 12, 12, 12, 12, 12}	0	1.00

Concerning to quantify the resiliency of a VPK scheme to the *deletion problem*, it is important to know that when one attribute of the relation is deleted, the set of VPKs generated will be different to the one obtained before the attack. To identify the VPK set generated after the attack we use S' . Then, there is $\vartheta = SS'$ where ϑ is a quality indicator of the differences between S and S' . The VPK scheme achieves the highest resiliency when $S = S'$.

There are different ways to implement the quality indicator ϑ . The first approach proposed in this work also uses the exclusiveness index ρ with additional considerations. Each group of values in S has associated a *useful load* that may be affected during the elimination of any attribute, varying the value of ρ . This *useful load* is the cause the numerator of the fraction $1/g_r$ in Eq. (5) is equal to 1, considering no perturbation has been caused by the *deletion problem*. The *useful load* is denoted as u , and it is measured by $u = (p - a)/p$, where p the number of elements in the group and a the number of elements affected by the attribute elimination. When none element is affected in a group, $a = 0$ and $u = 1$, which explains the use of 1 as the numerator. If all elements of the group are affected then $p = a$, resulting in $u = 0$. For the case when some exclusive values are affected, we subtract them from e . Considering the *useful load* concept, in Eq. (6) is given a more general approach of Eq. (5) which allows getting an idea of how critical can be the effects of the *deletion problem* over a VPK scheme

$$\rho = \frac{(\sum_{r=1}^d u_r/g_r + e) \times 100}{s} \tag{6}$$

The following examples allow a clear understanding of the way the value of ρ is obtained by using Eq. (6) breaking down the *useful load* calculation. The first example constitutes the original S with no values affected, used as the starting point of comparison. Then the same set is presented with different VPKs affected identified by shaded cells.

11	17	28	30	32	32	44	44	44	15	15	15	15	15
----	----	----	----	----	----	----	----	----	----	----	----	----	----

For this set, the value of ρ is calculated according to Eq. (6) as follows:

$$\begin{aligned} \rho &= \frac{(\sum_{r=1}^d u_r/g_r + e) \times 100}{s} = \frac{(\sum_{r=1}^d \frac{(p_r - a_r)/p_r}{g_r} + e) \times 100}{s} \\ &= \frac{\left(\left(\frac{1}{2} + \frac{1}{3} + \frac{1}{5}\right) + 4\right) \times 100}{14} \approx 35.95 \end{aligned}$$

When a duplicate value is affected, then the value of ρ is given by

11	17	28	30	32	32	44		44	15	15	15	15	15
----	----	----	----	----	----	----	--	----	----	----	----	----	----

$$\begin{aligned} \rho &= \frac{(\sum_{r=1}^d \frac{(p_r - a_r)/p_r}{g_r} + e) \times 100}{s} = \left(\left(\frac{(2 - 0)/2}{2} + \frac{(3 - 1)/3}{3} + \frac{(5 - 0)/5}{5}\right) + 4\right) \times 100/14 \\ &= \left(\left(\frac{1}{2} + \frac{2}{9} + \frac{1}{5}\right) + 4\right) \times 100/14 \approx 35.16 \end{aligned}$$

As it is appreciated, the change of the quality of S is clearly represented by the change of the value of ρ . Let us examine the example when one exclusive value is affected instead.

11		28	30	32	32	44	44	44	15	15	15	15	15
----	--	----	----	----	----	----	----	----	----	----	----	----	----

$$\rho = \frac{\left(\left(\frac{1}{2} + \frac{1}{3} + \frac{1}{5}\right) + (4 - 1)\right) \times 100}{14} \approx 28.81$$

For this case, it is clear that the consequences of the attack are higher, which gives the clear idea of the role the exclusives value represents to guarantee the WM synchronization.

The next example shows the case when all the elements of one group are affected. In that situation, the group affected is not considered for the evaluation of ρ .

11	17	28	30	32	32	44	44	44	15	15	15	15	15
----	----	----	----	----	----	----	----	----	----	----	----	----	----

$$\begin{aligned} \rho &= \left(\left(\frac{(2 - 0)/2}{2} + \frac{(3 - 3)/3}{3} + \frac{(5 - 0)/5}{5}\right) + 4\right) \times 100/14 \\ &= \left(\left(\frac{1}{2} + \frac{0}{3} + \frac{1}{5}\right) + 4\right) \times 100/14 \approx 33.57 \end{aligned}$$

Then, when one group is completely affected, this reduces the quality of the VPK set, but not as seriously as when one exclusive element is affected. Finally, there is the case when more than one group and exclusive values are affected.

11	17	28	30	32	32	44	44	44	15	15	15	15	15
----	----	----	----	----	----	----	----	----	----	----	----	----	----

$$\begin{aligned} \rho &= \left(\left(\frac{(2 - 0)/2}{2} + \frac{(3 - 2)/3}{3} + \frac{(5 - 1)/5}{5}\right) + (4 - 2)\right) \times 100/14 \\ &= \left(\left(\frac{1}{2} + \frac{1}{9} + \frac{4}{25}\right) + 2\right) \times 100/14 \approx 19.79 \end{aligned}$$

For this case, it is obvious that the value of ρ reduces more compared to the other examples since more elements are affected by the attack.

The way the *deletion problem* affects S is not only given by the elimination of the VPKs. When an attribute is deleted, the number of VPKs generated is equal to the original set before the elimination. That statement is settled by $s = s'$ being s' the cardinality of S' . Then, S 's quality change cannot be represented only by ρ since the same values of VPK can be generated after deleting one attribute, but identifying different elements of the relation (see Fig. 4). After an attack even a higher number of exclusive values can be generated, giving a higher value for ρ while the set is useless to perform the WM synchronization since the values do not match with the original ones used for embedding. According to that, ρ alone cannot be used for measuring the resiliency of a VPK scheme.

	S										
values	32	15	30	15	44	28	32	15	44	15	15
index	22	23	24	25	26	27	28	29	30	31	32
	S'										
values	30	32	15	44	15	15	15	28	15	32	44
index	22	23	24	25	26	27	28	29	30	31	32

Fig. 4. Example of S and S' formed by identical values but with different indexing.

Since the values added after the attack were not present during WM embedding, they will be responsible to add noise to the WM signal during its extraction. To establish the precision of ρ over S' we use Eq. (7).

$$\varphi = m/o \tag{7}$$

The symbol φ represents the precision of ρ , which is computed by comparing the number of VPKs m matching in index and value, with the number of original values o used for the embedding.

The next example illustrates how involving φ benefits the measure of ρ . Both sets S and S' are shown with their respective value of ρ calculated with no reference. The index on both subsets is given by the position of the keys in the set.

$$S = \{12, 34, 21, 45, 12, 34, 15, 48, 22, 82\}, \rho = 70.0$$

$$S' = \{12, 12, 23, 45, 12, 17, 15, 44, 67, 81\}, \rho = 73.0$$

At first view, the value of ρ for S' is higher due to it is composed of more *exclusive values*. But actually, only four keys in S' (values underlined) preserve the index they had in S since they are the only keys that remain in the same position for both sets. Considering that, $\rho = 73.0$ for a given $\varphi = 0.4$. But, to evaluate ρ with high precision, then the condition $\varphi = 1.0$ is established. To accomplish that we can only use the trusted values for calculating ρ , ignoring the rest of the keys from the operation. For that case, $\rho = 25.0$, giving a better appreciation of the quality of S' respect to S.

Identifying as ρ_T the value of ρ when $\varphi = 1.0$ (meaning the trusted ρ) and as ρ_O the value of ρ considering all elements of S' (not just those matching), the noise added to S' respect to S is determined by $\vartheta = |\rho_O - \rho_T|$. In the previous example, the noise added to S was $\vartheta = |73 - 25| = 48$. This metric can be appreciated as the quantitative measure of the difference of the qualities of S and S', previously defined as ϑ .

The second way to evaluate the resiliency to the *deletion problem* is not from the perspective of the structure of S but how the selection of the elements of R involved in the VPK generation is performed. According to that, the structure of the relation of Figure 3 plays a crucial role. The perturbations in S when one attribute of R is deleted can be reduced if: (i) the attributes used for the VPK generation uniformly vary from one tuple to another, selecting all of them the same number of times if possible, (ii) the number of attributes involved in the generation of VPKs for each tuple (defined as ℓ) is not too close to the number of attributes ν involved in the watermarking process. It is important to clarify that ℓ refers to the number of attributes involved in the VPK generation per tuple and not per VPK. So, when for a tuple 4 attributes are selected to generate 2 VPKs, $\ell = 4$, being the same value when a VPK is generated for the tuple considering 4 attributes.

Varying the attributes involved in the VPK generation from one tuple to another trying to equally involve them in the process will prevent the use of one attribute more than others. The idea is that no attribute should play the main role in the generation of S since with its deletion more VPKs are affected, becoming the attribute the main target of attackers. Then, if all attributes are used equally, no one constitutes the main vulnerability of the VPK scheme from the *deletion problem* perspective. Considering that $\eta \gg \nu$ is accomplished, the mean of the number of times each attribute

is selected is given by Eq. (8). Then, the number of times each attribute is involved in the generation of S is $\ell \times \eta / \nu \pm \varepsilon$. An equally selection of each attribute of R requires the reduction of ε , meaning getting a standard deviation as close as zero as it is possible.

$$\mu_A = \ell \times \eta / \nu \tag{8}$$

The standard deviation of the number of times each attribute is selected σ_A is obtained according to Eq. (9) where $is_sel()$ is a function with output 1 if the attribute is selected for VPK generation and 0 otherwise.

$$\sigma_A = \sqrt{\frac{\sum_{i=0}^{\nu-1} (\sum_{j=0}^{\eta-1} is_sel(r_j.A_i) - \mu_A)^2}{\nu}} \tag{9}$$

To avoid getting a high value of σ_A , from one tuple to another, the attributes selected to generate the VPK should not be the same. The challenge is that since tuples have not a fixed order, the only way to know the number of times each attribute is selected is by performing this calculation once all tuples are considered. Then, changing the selected attributes among the neighboring tuples is not a trivial task. By now we avoid this issue just to perform the graphical representation of the attribute selection and to show the expected values of the metrics when the highest resiliency to the *deletion problem* is achieved by the VPK scheme. All issues concerning the implementation of the proposed scheme are given in the next section. On another hand, the ordering of the attributes can be performed by using different criteria (e.g. attribute names, the domain of their values, etc.). According to this, we interpret the order of the attributes of R as fixed points of a circle, then the last attribute will precede the first one, using a cyclic indexing approach for moving through them (see Fig. 5).

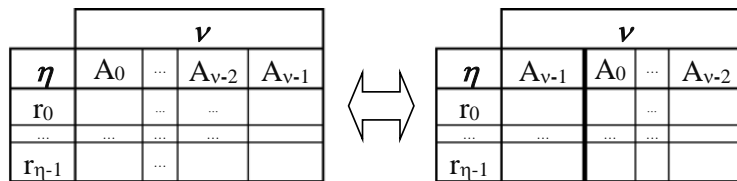


Fig. 5. Cyclic model of the attribute order in R .

Considering the cyclic order for the attributes and highlighting the attributes selected for the VPK generation with the gray color, we present some examples of how σ_A is affected using different attribute selections. For the first example, the number of attributes involved in the VPK generation is $\ell = 2$. According to Figure 5, when the attribute selected is the last one, if the index of the selected attribute is shifted once to the right, for the next tuple the first attribute will be chosen. Nevertheless, despite the selected attribute index is shifted from tuple to tuple, there is the risk to achieve an irregular attribute selection if the attributes selected between neighboring tuples overlap (see Fig. 6).

η	ν							$\ell = 2$
	1	2	3	4	5	6	7	
1								2
2								2
3								2
4								2
5								2
6								2
7								2
8								2
9								2
Total:	3	4	3	2	2	2	2	$\sigma_A = 0.73$

Fig. 6. Example of attribute uniformly selected with overlapping ($\ell = 2$).

In the previous example, the way the attribute selection is carried out results in a higher relevance of some attributes over others. Particularly, attribute 2 is selected 4 times, turning into a weak point to the scheme. The effects of

overlapping attribute selection are reduced when η increases, but still it is important its consideration to measure the VPK scheme resiliency against the *deletion problem*. In the next example, Figure 7 shows for the same conditions of Figure 6 a non-overlapping attribute selection for neighboring tuples.

η	ν							$\ell = 2$
	1	2	3	4	5	6	7	
1	■	■						2
2			■	■				2
3					■	■		2
4	■						■	2
5		■	■					2
6				■	■			2
7						■	■	2
8	■	■						2
9				■	■			2
Total:	3	3	3	3	2	2	2	$\sigma_A = 0.49$

Fig. 7. Example of attribute uniformly selected avoiding overlapping ($\ell = 2$).

The other important issue to consider the VPK resiliency to the *deletion problem* is concerning the selection of the value of ℓ . It is important that ℓ does not take a value too close to ν since this will cause the overlapping of the attributes among the tuples, affecting all those VPKs that have in common that attribute in their generation. Figure 8 shows the consequences of selecting a value of ℓ too close to ν . Even getting a reduction of σ from 0.49 to 0.35, deleting one attribute compromises more VPK values according to the rate of compromised keys for one attribute elimination given by $\sum_{j=0}^{\eta-1} is_sel(r_j, A_i) / \eta$. For example, comparing the attribute selection of Figure 7 vs. Figure 8, when attribute 2 is deleted, the rate of compromised keys is 3/9 vs. 5/9 (0.33 vs. 0.56).

η	ν							$\ell = 4$
	1	2	3	4	5	6	7	
1	■	■	■	■				4
2	■	■	■	■	■	■	■	4
3		■	■	■	■	■		4
4	■	■	■	■	■	■	■	4
5			■	■	■	■	■	4
6	■	■	■	■	■	■	■	4
7				■	■	■	■	4
8	■	■	■	■	■	■	■	4
9	■	■	■	■	■	■	■	4
Total:	6	5	5	5	5	5	5	$\sigma_A = 0.35$

Fig. 8. Selection of a value of ℓ too close to ν .

On another hand, it is not convenient to select a low value for ℓ , this increases the risk of being affected by the *duplicate problem*, considering that reducing the source of VPK generation increases the duplicate values in S . To avoid the fragility due to the high or low value of ℓ it is important to choose its value proportionally to ν . Then, the increment of ℓ is only allowed if ν also increases. Following that principle, the selection of ℓ is performed according to Eq. (10), being K the constant that establishes the proportion.

$$\ell = \lfloor \nu \times K \rfloor \tag{10}$$

Since the use of σ_A to compare the resiliency of the VPK scheme depends on ℓ , the number of attributes involved in each case should remain the same. Also, the number of tuples must be constant since adding more tuples increase the number of times each attribute is selected. To consider both η and ν in the evaluation, we propose Eq. (11) being q the quality of the spreading experimented by the scheme.

$$q = \frac{\eta \times (\nu - \ell) \times \nu^{(\nu - \ell)}}{\sigma_A} \tag{11}$$

According to Eq. (11), if $\nu = \ell$ the quality will be 0, the lowest value possible. If ℓ increases under the same conditions, q gets closer to 0. Finally, the lowest σ_A the better, since $\sigma_A = 0$ represents the perfect expected distribution of ℓ . By

definition, when $\sigma_A = 0$ then $q = \infty$, being the best possible result unless the numerator of the fraction in the equation became 0, giving as result $q = 0$, the worst result.

Table 3 shows some examples of the value of q , due to different attribute selection over the same relation. Each experiment corresponds to the examples presented in Figures 6, 7, and 8. On each case, the number of tuples involved was equally incremented.

Table 3. Value of q for different scenarios.

η	Figure 6	Figure 7	Figure 8
9	1.04×10^6	1.53×10^6	2.65×10^4
18	1.82×10^6	4.32×10^6	4.10×10^4
36	6.70×10^6	6.70×10^6	7.49×10^4
2000	2.31×10^8	3.40×10^8	5.88×10^6

Table 3 shows the experiment of Figure 7 as the favorite candidate according to the attributes selection quality. The outcome should be proportional to σ_A if both the elements of the relation and ℓ remain constant, and only the attributes selected are different.

6. HQR-Scheme: The high quality and resilient VPK generation approach

According to the literature published, the VPK schemes proposed so far generate too many duplicate values, are vulnerable to the *deletion problem*, or waste elements of R , denigrating the quality of S and due to that, the WM synchronization. For those reasons, it is important to design a VPK scheme that generates a higher number of exclusive values, varying the attributes involved in the process for each tuple. In this section, we present the main aspects to consider in the design of a VPK scheme with high quality of S and the highest possible resiliency to the *deletion problem*. Also, an implementation that considers the challenges of working with relational data for this task, is presented.

The first aspect we consider to design a VPK scheme is the selection of ℓ so a fair trade-off between the response given to the *duplicate problem* and the *deletion problem* can be carried out. The selection of ℓ brings the formation of a structure involving the number of attributes in between each one of those selected for the VPK generation. To identify the role played by each attribute in the structure, we denote as $y_i \in \{Y\} \mid i \in [0, \ell - 1]$ the set that contains the index of the attributes selected for the VPK generation, and as $b_j \in \{B\} \mid j \in [0, \ell - 1]$ the set that contains the number of attributes separating those stored in Y .

Figure 9 shows an example of the elements forming Y and B in a relation with $\nu = 12$. The order of the attributes of the relation is settled according to the cyclic attribute order presented in Figure 5. Considering the selected attributes are shifted from tuple to tuple, the combinations for two consecutive tuples is performed using the same value of the parameters and a right-shift of one unit. The idea is that the same attributes can play a different role according to the shift. Independently of the attributes combinations, the statement $\nu = \sum_{j=0}^{\ell-1} b_j + \ell$ is always accomplished.

	v											
η	A ₀	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇	A ₈	A ₉	A ₁₀	A ₁₁
r _z	y ₀	b ₀			y ₁	b ₁			y ₂	b ₂		
r _(z+1)	b ₂	y ₀	b ₀			y ₁	b ₁			y ₂	b ₂	

Fig. 9. Role per tuple for each attribute once ℓ is defined ($\ell = 3, \nu = 12$).

According to Eq. (10), we use $K = 0.2$ to obtain ℓ since this represents not involving too many attributes in the VPK generation (maximum 20% of ν) getting an extensive cover to variate the values of the VPKs forming S . On the other hand, a low value of ℓ is not enough to guarantee high resiliency to the *deletion problem*. Taking as an example the S-Scheme, when the same attribute is used to generate the VPKs, if that attribute is erased, the scheme is compromised. For that reason, it is also important the way the ℓ attributes are distributed over R .

Figure 10 shows the possible attribute combinations when $\nu = 7$ and $\ell = 2$. Each combination impacts in a different way according to Eq. (11). Under the same conditions, working with the same relation and the same value of ℓ , the differences among the results in Eq. (11) are given by σ_A .

η	ν							ν							ν							ν							ν							ν						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Total	3	4	3	2	2	2	2	3	3	3	3	2	2	2	3	3	2	3	3	2	2	3	3	2	2	3	3	2	3	3	2	2	2	3	3	4	3	2	2	2	2	3
	$\sigma_A = 0.728$							$\sigma_A = 0.495$							$\sigma_A = 0.495$							$\sigma_A = 0.495$							$\sigma_A = 0.495$							$\sigma_A = 0.728$						
	Case A							Case B							Case C							Case D							Case E							Case F						

Fig. 10. Different combinations to generate the VPK ($\ell = 2, \nu = 7$).

The value of σ_A varies according to the values in B . The standard deviation of the values in B denoted as σ_B , is obtained according to Eq. (12) where μ_B is the mean of the values stored in B . The lowest σ_A is achieved when the value of σ_B is the nearest to zero, which is the combination of attributes promising higher resilience to the *deletion problem* given the role of σ_A in Eq. (11).

$$\sigma_B = \sqrt{\frac{\sum_{j=0}^{\ell-1} (b_j - \mu_B)^2}{\ell}} \tag{12}$$

Figure 11 shows the different compositions of B used in the experiments performed to analyze the relationship between σ_A and σ_B . In those experiments, there is always an equivalent combination corresponding to the change of the values b_1 and b_2 . For example, C_1 represented by $b_1 = 0$ and $b_2 = 19$ is equivalent of C_{20} where $b_1 = 19$ and $b_2 = 0$. According to that, similar pairs of combinations are also $(C_2, C_{19}), (C_3, C_{18}), (C_4, C_{17}),$ etc...

ℓ	B	Combination of Attributes																			
		C ₁	C ₂	C ₃	C ₄	C ₅	C ₆	C ₇	C ₈	C ₉	C ₁₀	C ₁₁	C ₁₂	C ₁₃	C ₁₄	C ₁₅	C ₁₆	C ₁₇	C ₁₈	C ₁₉	C ₂₀
2	b ₁	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	b ₂	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0

Fig. 11. Different combinations of attribute forming B ($\ell = 2, \nu = 21$).

Figure 12 shows how the value of σ_A changes for each combination presented in Figure 11 using a different number of tuples. The number of combinations with lower σ_A varies depending on the number of tuples, but there is a set of combinations that represents the best options for all cases, which is when the values in B presents lower standard deviation. In Figure 12, those combinations are from C_8 to C_{13} .

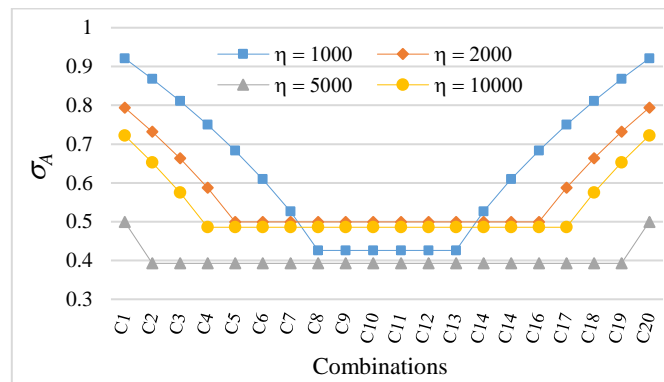


Fig. 12. Value of σ_A for the combinations of Figure 11.

Figure 13 shows the relationship between σ_A and σ_B for the combinations shown in Figure 11, also involving a different number of tuples. Since σ_B and σ_A describe a proportional behavior, to get higher quality q the goal is to use a combination that allows getting the lowest σ_B .

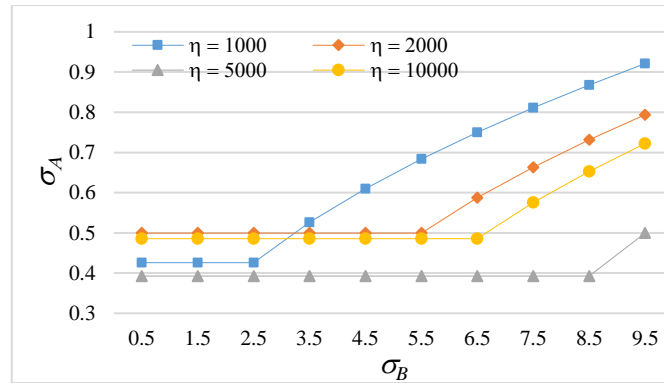


Fig. 13. Relationship between σ_A and σ_B for the combinations of Figure 11.

From Figure 13 on, the results shown consider the equivalence between combinations. The same relationship between σ_A and σ_B is appreciated for different values of ℓ (see Fig. 14 and Fig. 15).

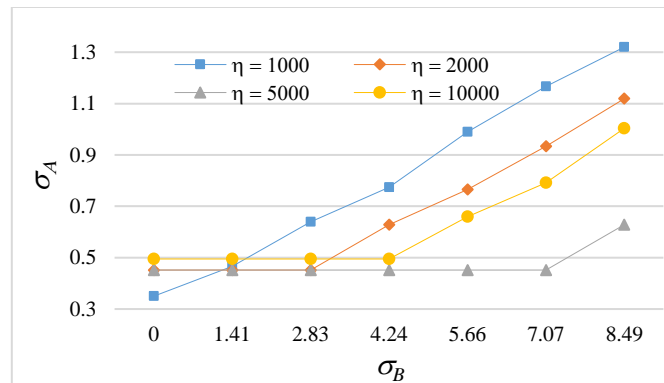


Fig. 14. Relationship between σ_A and σ_B for the possible combinations when $\ell = 3$ and $\nu = 21$.

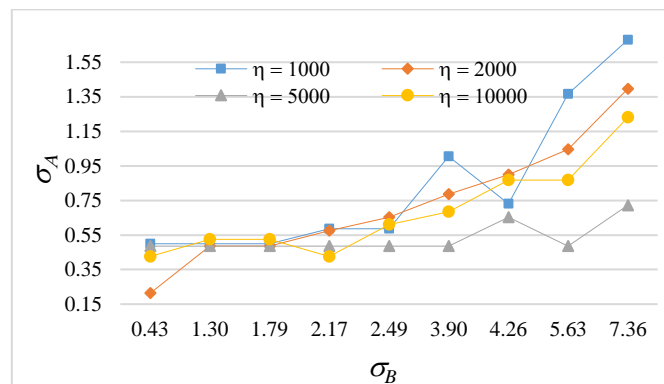


Fig. 15. Relationship between σ_A and σ_B for the possible combinations when $\ell = 4$ and $\nu = 21$.

Eventually, there are some irregularities in selecting the elements forming B . For example, when $\nu \bmod \ell = 0$ there is the possibility of $\sigma_B = 0$ since all the elements in B can take equal value, but when $\nu \bmod \ell \neq 0$, it is not possible. To select the number of attributes in between $Y_{(\ell-2)}$ and $Y_{(\ell-1)}$ allowing getting the closer value of σ_B to zero when $\nu \bmod \ell \neq 0$ the best option is to increment the attribute index in $\lfloor (\nu - \ell) / \ell \rfloor$ until $\ell = |Y|$. By doing so, when $\nu \bmod \ell \neq 0$, only the last element of B will be different from the others giving the most convenient value of σ_B .

On the other hand, another approach to select the value of ℓ considers the expected robustness of the VPK scheme. This directly involves the resilience of the scheme to the *deletion problem*, but there is one condition that must be fulfilled, which is $\sigma_B \leq 0.5$. According to that, and if $\eta \gg \nu$, we can select the value of ℓ using the mean given by Eq. (8). Since μ_A represents the average number of times each attribute is selected to generate the VPK, and considering the relationship between σ_A and σ_B , we can assure that the difference between μ_A and the number of times each attribute is selected is reduced if σ_B gets closer to zero. Since the value of ℓ is unknown, we cannot use σ_A , that is why it is used σ_B instead. Then, μ_A represents the expected number of times S will be affected once one attribute is deleted, then Eq. (8) can be used to obtain ℓ by $\ell = \mu_A \times \nu / \eta$.

None of the previous VPK schemes have considered the aspects discussed above. Nevertheless, the nature of relational data is featured by the lack of physical order among tuples and attributes. For that reason, implementing previous consideration results in a challenging task. Ordering attributes can be achieved by using their names, statistical measures of their values or considering their domain. The order of the tuples, on the contrary, is harder to settle.

The order of the tuples is critical to shift the attributes among neighboring tuples, so the desired values for the metrics previously presented can be obtained with the highest precision. One criterion for ordering tuples can be by using the hash of their values, an approach easily compromised if some attributes are deleted. That is why the only way to guarantee its success is by involving only those attributes essential to both, the data owner and the attacker. Also, since $\eta \gg \nu$ is generally accomplished, the hash values will present high redundancy, being more useful for partitioning R than for ordering tuples.

Considering the features of relational data, and seeking to accomplish previous considerations, we propose an implementation that simulates attribute shifting without linking the scheme to the order of the tuples, in other words, without requiring any knowledge of the neighboring tuples. To achieve that, each tuple is analyzed independently, adding a pseudo-random shifting to the attribute selection, seeking the highest variation of the attributes involved from one tuple to the other. Our approach generates one VPK per tuple, involving a maximum of ℓ attributes in the process, separating each one at least in $\lfloor (\nu - \ell) / \ell \rfloor$ attributes.

In this work, the list of attributes is ordered alphabetically by their names. Then, each tuple is independently analyzed, and for each tuple, each attribute. To avoid the impact of benign updates and update attacks, it is used the decimal value generated from the *msb* range of each value stored in the attributes. The value of χ is selected differently to the other schemes, excluding just the *lsb* bits and using the rest of the positions of the binary length, increasing the length of the source to avoid the *duplicate problem* (see line 5 of Algorithm 3). For this case, the value generated is identified as A_{s-vpk} , being a VPK temporal substitute.

Algorithm 3: HQR-Scheme Approach.

```

1 Input:  $AL, \ell, SK, \xi, p$ 
2 sort  $AL$  by name
3 foreach tuple  $r \in R$  do
4   foreach attribute  $A \in AL$  do
5      $A_{s-vpk} = [VPK([r.A]_2, BL(r.A) - \xi)]_{10}$ 
6      $A_{vv} = H(A_{s-vpk} \circ SK)$ 
7    $A_{max} = MAX(A_{vv})$  in  $r$ 
8    $A_{avg} = AVG(A_{vv})$  in  $r$ 
9    $z =$  count zeros in  $[A_{max}]_2$ 
10  set  $A_{max}$  as the starting point of analysis
11  if  $A_{max}$  is even then
12    analyze  $AL$  increasing the index in  $p$  units
13     $vpk\_builder(z, \ell, A_{avg})$ 
14  else
15    analyze  $AL$  decreasing the index in  $p$  units
16     $vpk\_builder(z, \ell, A_{avg})$ 

```

To increase the secrecy and variation of the selection, we generate a new value using the results of line 5 and the SK chosen by the data owner as the inputs of a hash function (see line 6). Next, the analysis is carried out using all those new values, each one identified as A_{vv} , being the notation for attribute virtual value. The starting point of analysis for each tuple is the attribute with maximum value among those generated in line 6, identified as A_{max} (see line 7).

Another variation in this approach is the difference in the direction of the analysis for each tuple. The attributes of the tuple are analyzed starting from A_{max} . When A_{max} is even the analysis is performed increasing the attribute index by p units, otherwise (when A_{max} is odd) the analysis is done decreasing the attribute index by p units (see lines 11-16). The role played by each attribute for each tuple varies based on the cyclic conception of the attributes order given in Figure 5, so if A_{max} is odd, once the attribute index is 0, if $p = 1$, the next attribute to be analyzed will be A_{v-1} . For the case when A_{max} is even, the next attribute to be analyzed after A_{v-1} , is A_0 if $p = 1$. Parameter p increases the flexibility of the method so the data owner must choose a value of p between 1 and $\lfloor (v - \ell) / \ell \rfloor$.

The parameterization of p helps to vary the attributes involved without compromising obtaining the VPK of the tuple. For example, when p is too high, fewer attributes are checked, and there is a higher risk of none attribute accomplishing the conditions to be considered for the VPK generation. In this approach, once one attribute is selected, it is compared to the average of the set of attribute virtual values for the tuple. The virtual values average is identified as A_{avg} . According to that, despite moving the index in p units, is not guaranteed that the attribute is going to be included in the VPK generation. For establishing the condition to include an attribute in the VPK generation we get the number of zeros from the binary representation of A_{max} and then if the number of zeros is even, the attributes considered for the VPK generation are those that are above A_{avg} . Otherwise, if the number of zeros is odd, the attributes considered are those below A_{avg} . The attributes that accomplish those conditions will be added to the process while the number of attributes involved is less or equal to ℓ . All these operations are performed by the $vpk_builder$ method (see Algorithm 4).

Algorithm 4: Method vpk_builder of the HQR-Scheme.

```

1 Input:  $z, \ell, A_{avg}$ 
2 if  $z$  is even then
3   | selection of the attributes above  $A_{avg}$  until the number of attributes  $\leq \ell$ 
4 else
5   | selection of the attributes below  $A_{avg}$  until the number of attributes  $\leq \ell$ 
6 return A

```

The generation of A_{avg} involves all the considered attributes of the relation for the VPK generation. Since the values are normalized by the use of the hash function, the *deletion problem* does not compromise the value of A_{avg} at a level that may affect the attributes involved in the process for each tuple.

7. Experimental results




The experiments performed measure the quality of the sets of VPK generated by each VPK scheme and their resiliency against the deletion problem. The relation used to embed the WM consist in a subset of the dataset *Forest Cover Type* (Colorado State, 1999). This subset is formed by the first 30,000 tuples out of the 581,012 that the relation stores. Also, from the 54 attributes, only the first 10 of them were used. This subset is selected to fairly compare the results with others previously reported in the literature. The names of the numerical attributes selected for the experimentation are shown in Table 4 along features such as mean, standard deviation, and average of their binary length (BL), among others.

Table 4. Information of the subset used to perform the experimentation.

Name	BL(Avg)	Max	Min	Mean	StdDev
ELEVATION	11.99	3849	1863	2780.08	322.30
ASPECT	7.11	360	0	144.60	108.17
SLOPE	4.13	61	0	14.19	8.13
HOR_DIST_TO_HYDROLOGY	7.13	1343	0	207.04	183.73
VERT_DIST_TO_HYDROLOGY	4.33	554	-146	38.63	50.41
HOR_DIST_TO_ROADWAYS	11.40	7117	0	2643.32	1895.24
HILLSHADE_9AM	7.99	254	0	215.53	27.04
HILLSHADE_NOON	8.00	254	99	221.77	19.61
HILLSHADE_3PM	7.57	248	0	136.57	38.05
HOR_DIST_TO_FIRE_POINTS	11.67	7173	0	3210.02	2157.09

The previous dataset was used to generate the VPKs and to embed and extract different WMs, to allow a visual appreciation of the quality of the WMs. For WM embedding two image-based watermarking techniques were selected, which generate the WM from a binary image, obtaining from each pixel the simplest values to generate the marks (0 or 1). This type of technique was selected considering that this kind of images allows embedding fewer marks while the WM is visually comprehensible. The watermarking techniques used were the proposals of Sardroudi and Ibrahim (2010), and Pérez Gort et al. (2017b). Since in a binary image the pixels represent black or white colors, the pixels missed (as results of attacks or low synchronization) are highlighted by using the red color. To generate the WMs, different images were used to analyze the role of the WM size in the processes. The images used in the experiments are shown in Table 5.

Table 5. Images used as WM source.

Name	Sample	Size (pixels)
Logo of the Universiti Teknologi Malaysia (UTM)		82 × 80
Logo of the World Wildlife Fund (WWF)		40 × 45
Picture of the chinese character Dào		20 × 21

The Correction Factor (CF), given by Eq. (13), is used to compare each pixel of the image employed to generate the embedded WM with the ones forming the image generated from the extracted WM. The maximum value for CF is 100, which means that the extracted WM is identical to the embedded one. On the contrary, when $CF = 0$ the embedded and extracted WMs are completely different.

$$CF = \frac{\sum_{i=1}^x \sum_{j=1}^y (Img_{org}(i, j) \oplus \overline{Img_{ext}(i, j)})}{x \times y} \times 100 \quad (13)$$

All the experiments performed in this work were implemented using Java 1.8 as client-side technology and Oracle Database 12c as the database server.

7.1. Quality of the VPK sets

In this work, we measure the quality of the VPK sets generated by each scheme by using the exclusiveness index ρ . The first VPK scheme we analyze is the S-Scheme generating the VPKs with a single attribute. The results of applying this scheme on every attribute of the dataset are shown in Table 6, where ρ and e values are presented. Considering that only one attribute of R is used to generate S and the same value of χ is used each time, too many duplicate values are generated, compromising the quality of the set.

Table 6. Value of ρ and e for the S generated by using S-Scheme over each attribute of R ($\chi = 3$).

Attribute	e	ρ
ELEVATION	0	1.04×10^{-5}
ASPECT	0	1.23×10^{-4}
SLOPE	0	1.41×10^{-4}
HOR_DIST_TO_HYDROLOGY	0	3.38×10^{-6}
VERT_DIST_TO_HYDROLOGY	0	1.90×10^{-5}
HOR_DIST_TO_ROADWAYS	0	1.11×10^{-3}
HILLSHADE_9AM	1	3.34×10^{-3}
HILLSHADE_NOON	0	1.37×10^{-5}
HILLSHADE_3PM	1	4.48×10^{-3}
HOR_DIST_TO_FIRE_POINTS	0	1.67×10^{-3}

The results shown in Table 7 correspond to a similar experiment than the previous one, but increasing the value of χ in the binary length of the value stored in the attribute with the exception of two lsb that are reserved to embed the mark. By doing this, the value of χ increases and even varies in some cases. Despite that, given the results is not recommended the use of the VPK set by any watermarking technique since the quality of S does not improve enough to allow the WM recognition.

Table 7. Value of ρ and e for the S generated using S-Scheme over each attribute of R ($\chi = BL - 2lsb$).

Attribute	$\chi(Avg)$	e	ρ
ELEVATION	9	7	8.00×10^{-2}
ASPECT	5	0	5.06×10^{-4}
SLOPE	2	0	1.27×10^{-4}
HOR_DIST_TO_HYDROLOGY	5	0	4.60×10^{-5}
VERT_DIST_TO_HYDROLOGY	2	0	8.45×10^{-6}
HOR_DIST_TO_ROADWAYS	9	5	5.74×10^{-2}
HILLSHADE_9AM	5	1	3.41×10^{-3}
HILLSHADE_NOON	6	0	1.73×10^{-3}
HILLSHADE_3PM	5	2	1.86×10^{-2}
HOR_DIST_TO_FIRE_POINTS	9	10	7.83×10^{-2}

Table 8 shows the metrics for the sets generated by the rest of the VPK schemes. Also the set generated by our proposal is included.

Table 8. Value of ρ and e for the VPKs sets generated by using the rests of the schemes.

VPK Schemes	e	ρ
E-Scheme	0	9.88×10^{-7}
M-Scheme (2 attributes)	0	4.97×10^{-4}
Ext-Scheme (MSBF = 3)	5513	21.91
Proposed Scheme	14946	52.96

According to the results shown in Table 8, there is a considerable quality increment in S for the two last schemes. By applying our approach, the number of exclusive values is increased more than two times compared to the number of exclusive values obtained by applying the second best choice, the Ext-Scheme. Figure 16 shows the number of exclusive values obtained by each scheme as a percentage of the total number of tuples involved in the experiment.

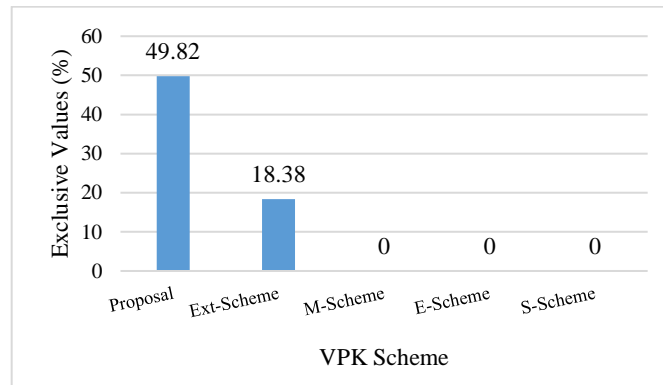


Fig. 16. Percentage of exclusive values generated by each scheme respect to the number of tuples in R .












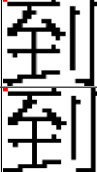



The quality of S is mostly determined by e , but it is important to know the different values generated even when they are not unique, since this is also considered to obtain ρ . Table 9 shows the number of groups on each set of VPKs along with the maximum and minimum size detected.

Table 9. Values of the metrics describing the groups of duplicate values in S for each scheme ($\chi = 3$).

VPK Schemes	No. Groups	Size range of groups	
		Maximum	Minimum
S-Scheme	4	17803	369
E-Scheme	8	80777	776
M-Scheme	7	15825	9
Ext-Scheme (MSBF = 3)	3125	4956	2
HQR-Scheme	2653	375	2

The number of marks considered in the WM embedding process is directly determined by the quality of S . The following results show the WM embedded by using the sets generated by each VPK scheme. Table 10 shows the image of the embedded watermark (column *Img*) and the percentage of the pixels embedded for each case (column % *E.M*). For this experiment, WM embedding was performed using the technique of Sardroudi & Ibrahim with parameters: $\gamma = 1, \chi = 3, \xi = 1$.

Table 10. WM embedded using the set of VPKs generated by each scheme.

VPK Schemes	WM UTM		WM WWF		WM Do	
	Img.	% E.M	Img.	% E.M	Img.	% E.M
S-Scheme (Attr.01)		0.06		0.22		0.95
E-Scheme		0		0		0
M-Scheme		0.11		0.38		1.66
Ext-Scheme (MSBF = 3)		73.06		98.61		99.76
HQR-Scheme		92.30		99.88		99.76

When S presents low quality, many marks are excluded by the embedding process, sometimes compromising the WM recognition. Figure 17 shows the percentage of embedded pixels for the cases when the WM recognition is achieved. Also, the results of using the relation PK for WM embedding are shown to establish a point of comparison with the results obtained by other VPK schemes. It can be appreciated from the figure how the difference between the results reduces when the size of the WM decreases.

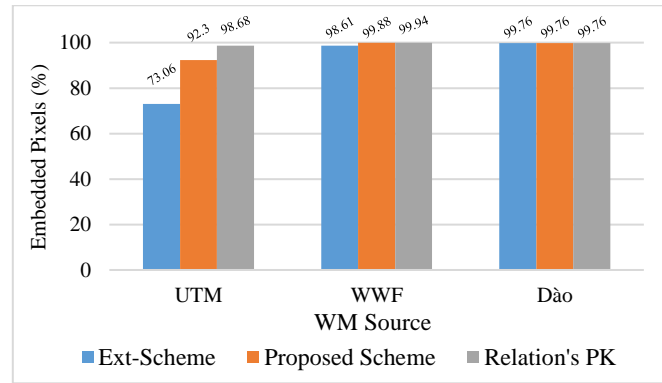


Fig. 17. Percentage of embedded marks for the different WMs by using Ext-Scheme, HQR-Scheme and the PK of R .

According to the results shown in Figure 17, it is possible to generate VPK sets with enough quality to achieve WM recognition despite the presence of redundancy. Also, the size of the WM plays an important role, since in some cases it can allow higher WM synchronization. Despite all of that, it is important to guarantee WM detection when some attribute of R is deleted. The results of the experiments performed concerning this issue are presented in the next section.

7.2. Resilience against the deletion problem

The resiliency of each VPK scheme against the *deletion problem* is tested using the same dataset, with values of ν and η constant for all the experiments. The first results showed are the values of ρ with precision $\varphi = 1.0$, allowing to know how the quality of S changes once one attribute is deleted. Table 11 shows the highest and lowest value of ρ for each scheme. This constitutes the range of quality obtained by deleting each one of the attributes of R .

Table 11. Range of value of ρ when each one of the attributes of R is deleted ($\chi = 3$).

VPK Schemes	Lowest ρ	Highest ρ
S-Scheme	0	4.48×10^{-3}
E-Scheme	4.24×10^{-7}	9.87×10^{-7}
M-Scheme (2 attributes)	6.29×10^{-5}	1.29×10^{-3}
Ext-Scheme (MSBF = 3)	8.61	12.99
Proposed Scheme	26.44	44.29

The results of Table 11 describing the resiliency of our approach against the *deletion problem* allow a clear understanding of the confidence incorporated to a VPK scheme when the issues previously analyzed in this paper are considered on the scheme design. Also, the quality of the schemes in terms of the distribution of the attributes selected to generate the VPKs is obtained. Table 12 shows the value of σ_A and q for the Ext-Scheme and our approach. Other schemes are excluded considering the low WM synchronization they allow.

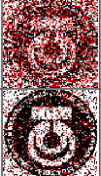
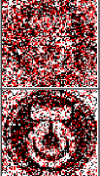










Table 12. Quality of the distribution of the attribute selection for the Ext-Scheme and our approach.

ℓ	σ_A		q	
	Ext	Proposal	Ext	Proposal
1	34.62×10^2	8.49×10^2	7.80×10^{10}	3.18×10^{11}
2	38.93×10^2	13.60×10^2	6.16×10^9	1.76×10^{10}
3	35.41×10^2	15.70×10^2	5.93×10^8	1.34×10^9

The resilience against the *deletion problem* of a VPK scheme increases if more than one mark is embedded each time one tuple is selected. Table 13 shows the pictures corresponding to the range of quality of the WMs extracted

after deleting each one of the attributes of R . For this case, the employed watermarking technique was the approach by Sardroudi and Ibrahim (2010) that embeds one mark per tuple.

Table 13. Resilience shown to the *deletion problem* using the technique of Sardroudi & Ibrahim.

VPK Scheme	WM UTM				WM WWF				WM Do			
	Highest Quality		Lowest quality		Highest Quality		Lowest quality		Highest Quality		Lowest quality	
	Img	CF	Img	CF	Img	CF	Img	CF	Img	CF	Img	CF
Ext-Scheme		54		47		84		69		91		86
Proposal		78		65		97		91		99		99

Also in Table 14 are presented the results of the experiments under the same conditions but embedding more than one mark per tuple using the watermarking technique by Pérez Gort et al. (2017b).

Table 14. Resiliency shown to the *deletion problem* using the technique of Pérez Gort et al.






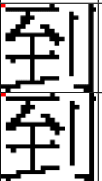






VPK Scheme	WM UTM				WM WWF				WM Do			
	Highest Quality		Lowest quality		Highest Quality		Lowest quality		Highest Quality		Lowest quality	
	Img	CF	Img	CF	Img	CF	Img	CF	Img	CF	Img	CF
Ext-Scheme		99		99		99		99		99		99
Proposal		99		99		99		99		99		99

Figure 18 shows the results of the stress tests performed over the Ext-Scheme and our proposal to detect how the WM degrades when more than one attribute is deleted. For this experiment, the WM synchronization was performed through the technique of Sardroudi & Ibrahim. In the figure, the higher resiliency of our approach when more than one attribute is deleted is appreciated.

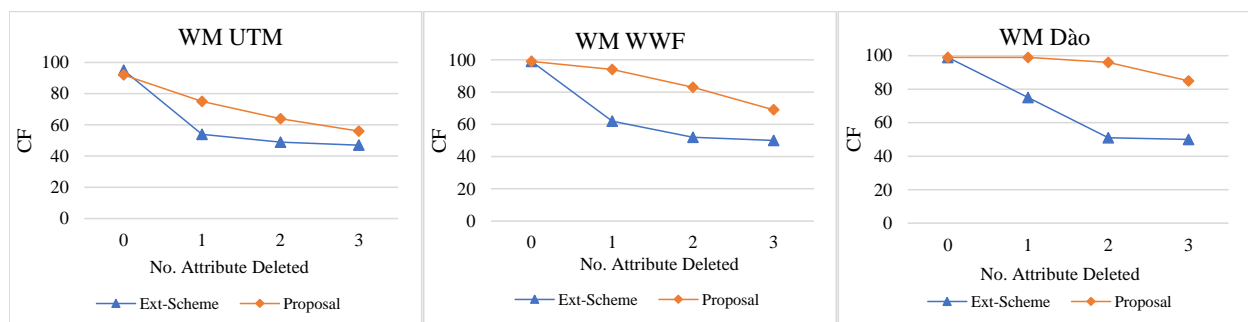


























Fig. 18. WM extracted for each WM after performing a multiple attribute deletion.

Finally, the images corresponding to the WM extracted in the previous experiment are presented in Table 15, when images in the middle of the range of the quality previously shown for each case, with its correspondent CF, are shown. It is appreciated how our approach allows higher resiliency than the Ext-Scheme when more than one attribute is deleted. Also, it is shown the relevance of the size of the WM in the resiliency to this type of operation.

Table 15. Images of the WMs extracted when more than one attribute is deleted when the technique of Sardroudi & Ibrahim is used to perform the WM synchronization.

No.	Ext-Scheme			Proposal		
	UTM	WWF	Do	UTM	WWF	Do
0						
	95	99	99	92	99	99
1						
	54	62	75	75	94	99
2						
	49	52	51	64	83	96
3						
	47	50	50	56	69	85

The quality of the detected WM shown in Table 15 increases when more than one mark per tuple is embedded, describing a similar behavior of the one appreciated between the quality of the WM shown in Table 13 and Table 14. In general, to deal with the duplicate problem it is recommended to generate the VPKs by using the attributes storing the more relevant data of R . That way, with each attempt of deletion attack, also the attacker will be paying the price of compromising the data.

8. Conclusions

In this paper, we present the metrics to measure the quality of a set of VPK values to be used for watermarking relational data as well as the main aspects to be considered to design a VPK scheme. Based on these considerations and the challenges that practical scenarios bring with them, it was proposed a novel VPK scheme to achieve higher WM synchronization despite the elimination of attributes in the watermarked relation. The experiments carried out show how previous VPK schemes do not guarantee the WM recognition or its persistence when some attribute of R is deleted. Also, the proposed VPK scheme allows watermark detection, showing resilience against the deletion problem, recognizing the signal even when more than one attribute of the relation is attacked.

Acknowledgements

This work was partially supported by a Ph.D. grant No. 714270 from CONACyT, Mexico.

References

- Agrawal, R., Haas, P. J., and Kiernan, J. (2003). Watermarking relational data: framework, algorithms and analysis. *The VLDB journal*, 12(2):157–169.
- Agrawal, R. and Kiernan, J. (2002). Watermarking relational databases. In *Proceedings of the 28th international conference on Very Large Data Bases*, pages 155–166. VLDB Endowment.
- Camara, L., Li, J., Li, R., and Xie, W. (2014). Distortion-free watermarking approach for relational database integrity checking. *Mathematical problems in engineering*, 2014.
- Chang, C.-C., Nguyen, T.-S., and Lin, C.-C. (2014). A blind robust reversible watermark scheme for textual relational databases with virtual primary key. In *International Workshop on Digital Watermarking*, pages 75–89. Springer.
- Chang, J.-N. and Wu, H.-C. (2012). Reversible fragile database watermarking technology using difference expansion based on svr prediction. In *Computer, Consumer and Control (IS3C), 2012 International Symposium on*, pages 690–693. IEEE.

- Choudhary, S., Nath, K., and Panda, J. (2017). Double layered audio zero-watermarking using dwt & dsss. In *Communication and Signal Processing (ICCSPP), 2017 International Conference on*, pages 0419–0423. IEEE.
- Colorado State, U. (1999). Forest covertype, the uci kdd archive.
- Date, C. J. (2006). *An introduction to database systems*. Pearson Education India.
- Franco-Contreras, J., Coatrieux, G., Cuppens, F., Cuppens-Boulahia, N., and Roux, C. (2014). Robust lossless watermarking of relational databases based on circular histogram modulation. *IEEE transactions on information forensics and security*, 9(3):397–410.
- Ghogare, G. R. and Junnarkar, A. A. (2017). Genetic algorithm based reversible watermarking approach for relational data. *IJETT*, 1(2).
- Guo, J. (2011). Fragile watermarking scheme for tamper detection of relational database. In *Computer and Management (CAMAN), 2011 International Conference on*, pages 1–4. IEEE.
- Gursale, N. and Mohanpurkar, A. (2014). Distortion minimization fingerprinting technique for relational database. *International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC)*, 2(6):1737–1741.
- Halder, R., Pal, S., and Cortesi, A. (2010). Watermarking techniques for relational databases: Survey, classification and comparison. *J. UCS*, 16(21):3164–3190.
- Iftikhar, S., Anwar, Z., and Kamran, M. (2014). A novel and robust fingerprinting technique for digital data based on genetic algorithm. In *High-capacity Optical Networks and Emerging/Enabling Technologies (HONET), 2014 11th Annual*, pages 173–177. IEEE.
- Iftikhar, S., Kamran, M., and Anwar, Z. (2015a). Rrw-a robust and reversible watermarking technique for relational data. *IEEE Transactions on Knowledge & Data Engineering*, 27(4):1132–1145.
- Iftikhar, S., Kamran, M., and Anwar, Z. (2015b). A survey on reversible watermarking techniques for relational databases. *Security and Communication Networks*, 8(15):2580–2603.
- İmamoğlu, M. B., Ulutaş, M., and Ulutaş, G. (2015). A watermarking technique for relational databases based on partitioning. In *Signal Processing and Communications Applications Conference (SIU), 2015 23th*, pages 2094–2097. IEEE.
- Jiang, C., Chen, X., and Li, Z. (2009). Watermarking relational databases for ownership protection based on dwt. In *Information Assurance and Security, 2009. IAS'09. Fifth International Conference on*, volume 1, pages 305–308. IEEE.
- Kamran, M., Suhail, S., and Farooq, M. (2013). A robust, distortion minimizing technique for watermarking relational databases using once-for-all usability constraints. *IEEE Transactions on Knowledge and Data Engineering*, 25(12):2694–2707.
- Khan, A. and Husain, S. A. (2013). A fragile zero watermarking scheme to detect and characterize malicious modifications in database relations. *The Scientific World Journal*, 2013.
- Khanduja, V. (2017). Database watermarking, a technological protective measure: Perspective, security analysis and future directions. *Journal of Information Security and Applications*, 37:38–49.
- Khanduja, V., Chakraverty, S., and Verma, O. (2016a). Ownership and tamper detection of relational data: framework, techniques and security analysis. *Embodying Intelligence in Multimedia Data Hiding*, 5:21–36.
- Khanduja, V., Chakraverty, S., and Verma, O. P. (2016b). Enabling information recovery with ownership using robust multiple watermarks. *Journal of Information Security and Applications*, 29:80–92.
- Khanduja, V., Verma, O. P., and Chakraverty, S. (2015). Watermarking relational databases using bacterial foraging algorithm. *Multimedia Tools and Applications*, 74(3):813–839.
- Li, Y., Swarup, V., and Jajodia, S. (2003). Constructing a virtual primary key for fingerprinting relational data. In *Proceedings of the 3rd ACM workshop on Digital rights management*, pages 133–141. ACM.
- Mehta, B. B. and Aswar, H. D. (2014). Watermarking for security in database: A review. In *IT in Business, Industry and Government (CSIBIG), 2014 Conference on*, pages 1–6. IEEE.
- Melkundi, S. and Chandankhede, C. (2015). A robust technique for relational database watermarking and verification. In *Communication, Information & Computing Technology (ICCICT), 2015 International Conference on*, pages 1–7. IEEE.
- Mohanpurkar, A. and Joshi, M. (2015). A fingerprinting technique for numeric relational databases with distortion minimization. In *Computing Communication Control and Automation (ICCUBEA), 2015 International Conference on*, pages 655–660. IEEE.
- Nematollahi, M. A., Vorakulpipat, C., and Rosales, H. G. (2017). Video watermarking. In *Digital Watermarking*, pages 67–80. Springer.
- Patil, V. B. and Yawalkar, P. M. (2015). A review on multi-attribute watermarking technique for relational data. *International Journal of Advanced Research in Computer and Communication Engineering (IJARCC)*, 4(12):474–479.
- Pérez Gort, M. L., Aparicio Díaz, E., and Feregrino Uribe, C. (2017a). A highly-reliable virtual primary key scheme for relational database watermarking techniques. In *Proceedings of the International Conference on Computational Science and Computational Intelligence, CSCI'17*. IEEE.
- Pérez Gort, M. L., Feregrino Uribe, C., and Nummenmaa, J. (2017b). A minimum distortion: High capacity watermarking technique for relational data. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, IH&MMSec '17*, pages 111–121, New York, NY, USA. ACM.
- Prajapati, S. and Tiwari, N. (2015). Image based relational database watermarking: a survey. *IOSR Journal of Computer Engineering (IOSR-JCE)*, pages 54–65.
- Rani, S., Koshley, D. K., and Halder, R. (2017). Partitioning-insensitive watermarking approach for distributed relational databases. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXXVI*, pages 172–192. Springer.
- Rao, U. P., Patel, D. R., and Vikani, P. M. (2012). Relational database watermarking for ownership protection. *Procedia Technology*, 6:988–995.
- Şahin, Y., Ulutaş, G., and İmamoğlu, M. B. (2016). An effective database watermarking method based on histogram of pairs. In *Signal Processing and Communication Application Conference (SIU), 2016 24th*, pages 353–356. IEEE.
- Sardroudi, H. M. and Ibrahim, S. (2010). A new approach for relational database watermarking using image. In *Computer Sciences and Convergence Information Technology (ICCIT), 2010 5th International Conference on*, pages 606–610. IEEE.
- Shih, F. Y. (2017). *Digital watermarking and steganography: fundamentals and techniques*. CRC press, second edition edition.
- Unnikrishnan, K. and Pramod, K. (2017). Robust optimal position detection scheme for relational database watermarking through holpsofa algorithm. *Journal of Information Security and Applications*, 35:1–12.
- Wang, T. (2017). Digital image watermarking using dual-scrambling and singular value decomposition. In *Computational Science and Engineering*

- (CSE) and Embedded and Ubiquitous Computing (EUC), 2017 IEEE International Conference on, volume 1, pages 724–727. IEEE.
- Xie, M.-R., Wu, C.-C., Shen, J.-J., and Hwang, M.-S. (2016). A survey of data distortion watermarking relational databases. *IJ Network Security*, 18(6):1022–1033.
- Zhang, L., Gao, W., Jiang, N., Zhang, L., and Zhang, Y. (2011). Relational databases watermarking for textual and numerical data. In *Mechatronic Science, Electric Engineering and Computer (MEC), 2011 International Conference on*, pages 1633–1636. IEEE.

The proposed metrics help to know if a VPK set will be useful to watermark the data.

There is no need to perform the WM embedding to know if the VPK set is good.

The cyclic idea of the attribute order helps to design an effective VPK scheme.

The WM is detected even after being deleted more than one attribute of R.