Full-Exact Approach for Frequent Itemset Hiding

Tolga Ayav Belgin Ergenc

Abstract

This paper proposes a novel, exact approach that relies on integer programming for association rule hiding. A large panorama of solutions exists for the complex problem of itemset hiding: from practical heuristic approaches to more accurate exact approaches. Exact approaches provide better solutions while suffering from the lack of performance and existing exact approaches still augment their methods with heuristics to make the problem solvable. In this case, the solution may not be optimum. This work present a full-exact method, without any need for heuristics. Extensive tests are conducted on 10 real datasets to analyze distance and information loss performances of the algorithm in comparison to a former similar algorithm. Since the approach provides the *optimum* solution to the problem, it should be considered as a reference method.

Keywords Association rule hiding, itemset hiding, exact approach, cost model, side effect

---

1 Introduction

Data mining field that aims to find interesting patterns from huge amounts of data attracted the increasing interest of organizations, researchers and

practitioners. However this growing use of data mining technology in different domains increased the concern for privacy leading to an active research area named privacy preserving data mining. From a general point of view, privacy issues related to the application of data mining can be classified into two main categories, namely data hiding and knowledge hiding. Data hiding aims to remove confidential or private information from data prior to its publication. Knowledge hiding, on the other hand is concerned with the sanitization of data leading to disclosure of confidential and private knowledge (Gkoulalas-Divanis and Verykios, 2010). The problem of knowledge hiding requires sanitizing the input database $\mathcal{D}$ in such a way that a set of sensitive knowledge K^S is hidden, while most of the information in $\mathcal{D}$ is maintained.

Association rule mining or frequent itemset mining is a major data mining methodology, mostly known in the area of market basket analysis and aims to capture relationships present among items in a transactional database (Agrawal et al., 1993) (Agrawal and Srikant, 1994). Despite its benefit in modern business, frequent itemset mining can also pose a threat to privacy and security in a database sharing environment when precautions are not taken in its implementation (Atallah et al., 1999) (Oliveira and Zaiane, 2002). Frequent itemset hiding is specialization of the generic knowledge hiding problem where the main requirement asks for lowering the support of sensitive itemsets in the input database $\mathcal{D}$ so that sanitized database $\mathcal{D}^{'}$ can be produced. Secondary requirements are minimization of deleted items and loss of nonsensitive frequent itemsets (Bonchi and Ferrari, 2011).

Large body of research emerged in the field of itemset hiding, since the database owners are in need of sharing data with their competitors, for their mutual benefit without revealing strategic patterns in the form of sensitive itemsets. Due to combinatorial nature of the problem of itemset hiding, proposed sanitization methodologies span from simple, time and memory efficient heuristics (Oliveira and Zaiane, 2002) (Oliveira and Zaiane, 2003a) (Verykios et al., 2004) (Amiri, 2007) (Wu et al., 2007) (Keer and Singh, 2012) (Yildiz and Ergenc, 2012), border-based approaches (Sun and Yu, 2005) (Sun and Yu, 2007) (Moustakides and Verykios, 2008) and reconstruction based approaches (Mielikainen, 2003) (Guo, 2007) (Lin and Liu, 2007) (Boora et al., 2009) (Mohaisen et al., 2010) to exact hiding (Menon et al., 2005) (Gkoulalas-Divanis and Verykios, 2006) (Gkoulalas-Divanis and Verykios, 2008) (Gkoulalas-Divanis and Verykios, 2009b) algorithms that offer guarantees on the quality of the computed hiding solution at an increased computational complexity cost. Whatever the technique used in sanitiza-

tion; different attributes are used in selecting the transaction, itemset in the transaction or the item in the itemset to modify. Sanitization techniques try to minimize distance and/or information loss while the number of sensitive itemsets, characteristics of the data sets or user defined support vary.

The motivation for this work is to find a full exact solution to the problem of itemset hiding. Although exact approaches claim to model the itemset hiding as optimization problem whose objective is to find the optimum solution, the ones proposed by Menon (Menon et al., 2005) and Divanis (Gkoulalas-Divanis and Verykios, 2008) cannot be considered as entirely exact since they rely not only on integer programming but also some heuristics. To the best of authors' knowledge, this work is the first one that models the entire problem as integer programming so that an optimum solution can be achieved without requiring any heuristics. Evaluation tests are carried out to measure distance and information loss performance of the proposed approach in comparison to an existing algorithm.

Organization of the paper is as follows; Section 2 gives preliminary information related to the problem definition of itemset hiding, metrics for side effects and a motivating example that will be used to explain the algorithms in the following sections. Section 3 starts with the explanation of the full exact approach of itemset hiding problem. Section 4 gives detailed performance analysis of the proposed method on UCI benchmark datasets (Coenen, 2003) while changing the number of sensitive itemsets in comparison to a similar exact algorithm. Section 5 is dedicated to the survey of existing approaches. Finally, Section 6 covers the conclusion remarks.

2 Preliminaries and Motivating Example

In this section, formulation of the problem of itemset hiding and metrics used in minimizing the side effect during sanitization process will be described. In the last part, a motivating example is presented, that example will be used in the following section to better explain exact itemset hiding algorithms.

2.1 Problem Formulation

Let $\mathcal{I} = \{a_1, a_2, \ldots, a_m\}$ be a set of literals, called items. Any subset of $\mathcal{I} : I_k \subseteq \mathcal{I}$ is called an itemset. Let $\mathcal{D}$ denote a transactional database, where each transaction T_i is a tuple $< i, I_i >$ where I_i is an itemset and i

is the transaction ID. $\mathcal{D}$ can be represented as $\mathcal{D} = \{T_1, T_2, \ldots, T_n\}$. The database is represented in binary form as follows:

$$D = (d_{ij})_{n \times m} \tag{1}$$

such that

$$d_{ij} = \begin{cases} 1 & \text{if} \quad a_j \in I_i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

It can be simply said that frequent itemsets are the ones whose support values, *i.e.,* frequencies of co-occuring items in $\mathcal{D}$ are above the minimum support value. Frequent itemsets can be represented as

$$\mathcal{F}_{(\mathcal{D},\psi)} = \{I_k : \sigma(I_k) \geq \psi\} \tag{3}$$

where $\sigma(I_k)$ denotes the support value of the itemset I_k and ψ denotes the minimum support value. The set of sensitive frequent itemsets is denoted with $\mathcal{F}^S_{(\mathcal{D},\psi)}$, which is the subset of frequent itemsets. The set of non-sensitive frequent itemsets is also denoted with $\mathcal{F}^N_{(\mathcal{D},\psi)}$, *i.e.,*

$$\mathcal{F}_{(\mathcal{D},\psi)} = \mathcal{F}^S_{(\mathcal{D},\psi)} \cup \mathcal{F}^N_{(\mathcal{D},\psi)}$$

In order to represent sensitive frequent itemsets, another matrix is also defined such that:

$$S = (s_{kj})_{r \times m} \tag{4}$$

such that

$$s_{kj} = \begin{cases} 1 & \text{if} \quad a_j \in I_k \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where r is the number of sensitive frequent itemsets, *i.e.,* $r = |\mathcal{F}^S_{(\mathcal{D},\psi)}|$. The goal of hiding sensitive frequent itemsets is to transform database $\mathcal{D}$ into $\mathcal{D}'$ such that:

$$i) \quad \sigma(I_k) < \psi, \ \forall I_k \in \mathcal{F}^S_{(\mathcal{D}',\psi)},$$

$$ii) \quad \text{Side effect is minimized.}$$

The first requirement (privacy requirement) lowers the support of sensitive itemsets below the support threshold ψ, so that they are sanitized when the dataset is mined at support threshold ψ. The second requirement assures minimizing one or more distortion metrics explained next.

2.2 Metrics for Side Effect

There are two common performance metrics that can be used to minimize the side effect of sanitization. The first metric is *distance* or *data distortion*, and it relies on the number of items removed during the sanitization process:

$$\sum_{i,j} d_{ij} - d'_{ij} \tag{6}$$

The second metric, frequent pattern distortion gives an information about how many nonsensitive frequent itemsets in the original dataset become infrequent after the sanitization:

$$|\mathcal{F}^N_{(\mathcal{D},\psi)}| - |\mathcal{F}^N_{(\mathcal{D}',\psi)}| \tag{7}$$

Both metrics fulfill the property of "the smaller the better", although they measure different aspects of data/knowledge distortion.

2.3 Motivating Example

Throughout the text, a sample 10-transaction-database given in Table 1 will be used to present the FULL-EXACT method. This database has 62 non-singleton frequent itemsets under $\psi = 2$ (20%). These itemsets are given in Table 2. We mark 4 itemsets with bold as sensitive, *i.e.*, $\mathcal{F}^S_{(\mathcal{D},\psi)} = \{\{a,b,c\},\{c,h\},\{h,i\},\{f,g\}\}$ and strike the supersets of the sensitive itemsets. S matrix defines sensitive itemsets as shown in Table 3.

3 Full Exact Approach to Itemset Hiding Problem

In this section, full exact solution to the itemset hiding problem is presented. The solution satisfies the privacy requirement of the problem, *i.e.*, the support of sensitive itemsets are reduced below a given support threshold. This

Table 1: Sample dataset with 10 transactions and 10 items

d_{ij}	a	b	c	d	e	f	g	h	i	j
T_1	**1**	**1**	**1**	0	0	0	**1**	**1**	0	**1**
T_2	0	0	**1**	0	0	0	0	0	**1**	**1**
T_3	0	0	0	**1**	**1**	**1**	0	0	0	0
T_4	**1**	**1**	**1**	0	0	**1**	**1**	**1**	**1**	0
T_5	**1**	**1**	**1**	0	0	**1**	**1**	0	0	0
T_6	0	0	0	0	0	0	0	0	0	**1**
T_7	0	0	0	**1**	0	0	0	0	0	0
T_8	0	0	**1**	0	0	**1**	**1**	**1**	**1**	0
T_9	0	0	**1**	0	0	0	0	**1**	**1**	0
T_{10}	0	0	0	0	**1**	**1**	**1**	0	0	0

Table 2: Frequent itemsets of sample dataset

I_{1-15}	$\sigma(I)$	I_{16-30}	$\sigma(I)$	I_{31-45}	$\sigma(I)$	I_{46-60}	$\sigma(I)$	I_{61-62}	$\sigma(I)$
bc	3	ef	2	gh	2	~~ach~~	2	~~$abegh$~~	2
ac	3	cj	2	**ch**	2	~~afg~~	2	~~$efghi$~~	2
cf	3	ah	2	abh	2	~~fgh~~	2		
cg	4	ab	3	abf	2	~~cfh~~	2		
abc	3	bh	2	abg	2	~~cgh~~	3		
bcf	2	bf	2	bgh	2	~~$efgh$~~	2		
acf	2	bg	3	~~bch~~	2	$abgh$	2		
bcg	3	**fg**	4	~~bfg~~	2	~~$abch$~~	2		
acg	3	**hi**	3	~~fhi~~	2	~~$abfg$~~	2		
~~efg~~	3	fi	2	~~ghi~~	2	~~$bcgh$~~	2		
~~$abcf$~~	2	gi	2	~~chi~~	3	~~$fghi$~~	2		
~~$abeg$~~	3	ci	4	~~fgi~~	2	~~$cfhi$~~	2		
~~$bcfg$~~	2	af	2	cfi	2	~~$cghi$~~	2		
~~$acfg$~~	2	ag	3	cgi	2	~~$cfgi$~~	2		
~~$abcfg$~~	3	fh	2	agh	2	~~$acgh$~~	2		

optimization problem can be written as follows:

$$\min \quad \sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij}x_{ij} + \sum_{i=1}^{n}\sum_{k=1}^{o} y_{ik} \tag{8}$$

Table 3: Sensitive Itemsets

s_{kj}	a	b	c	d	e	f	g	h	i	j
I_5	**1**	**1**	**1**	0	0	0	0	0	0	0
I_{23}	0	0	**1**	0	0	0	0	**1**	0	0
I_{24}	0	0	0	0	0	0	0	**1**	**1**	0
I_{32}	0	0	0	0	0	**1**	**1**	0	0	0

Table 4: Nonsensitive Itemsets

s_{kj}	a	b	c	d	e	f	g	h	i	j
I_1	0	**1**	**1**	0	0	0	0	0	0	0
I_2	**1**	0	**1**	0	0	0	0	0	0	0
I_3	0	0	**1**	0	0	**1**	0	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_{45}	**1**	0	0	0	0	0	**1**	**1**	0	0

$$s.t. \quad \sum_{i=1}^{n} y_{ik} \geq \sigma(I_k) - \psi + 1, \quad \forall k : I_k \in \mathcal{F}^S_{(\mathcal{D},\psi)}, \tag{9}$$

$$s.t. \quad \sum_{j=1}^{m} s_{kj} x_{ij} \geq y_{ik}, \tag{10}$$
$$\forall\ (i,k) : i \in \{1,\ldots,n\}, I_k \in \mathcal{F}^S_{(\mathcal{D},\psi)},$$

$$s.t. \quad \sum_{i=1}^{n} z_{ik} \leq \sigma(I_k) - \psi, \quad \forall k : I_k \in \mathcal{F}^N_{(\mathcal{D},\psi)}, \tag{11}$$

$$s.t. \quad \sum_{j=1}^{m} n_{kj} x_{ij} \leq M_1\, z_{ik}, \tag{12}$$
$$\forall\ (i,k) : i \in \{1,\ldots,n\}, I_k \in \mathcal{F}^N_{(\mathcal{D},\psi)},$$

$$x_{ij},\ y_{ij},\ z_{ik} \in \{0,1\}, \tag{13}$$

where x_{ij} are decision variables indicating the items to be removed. p and

o represent the number of sensitive and nonsensitive itemsets respectively, *i.e.*, $p = |\mathcal{F}^{S}_{(\mathcal{D},\psi)}|$ and similarly, $o = |\mathcal{F}^{N}_{(\mathcal{D},\psi)}|$. $x_{ij} = 1$ means that item d_{ij} will be removed. y_{ik} are the decision variables that indicate the sensitive patterns to be removed. Similarly, z_{ik} are the decision variables that indicate the nonsensitive patterns to be removed. Since all the variables are binary integers, this problem falls into the category of *integer programming* and there are mature techniques to solve it (Jèunger et al., 2010).

The problem has several constraints. First, (9) defines a constraint for each sensitive itemset to ensure that the frequency of that itemset is below the support threshold. The number of constraints is p. Next, (10) defines one constraint for each occurrence of the sensitive itemsets over the whole database to ensure that at least one item in a sensitive itemset is removed if the itemset is set to be removed, *i.e.*, $y_{ik} = 1$. The total number of these constraints depend on the dataset, yet its maximum value would be $n \times p$.

Constraint (11) ensures that the frequency of nonsensitive itemsets are still above the threshold. The total number of these constraints is the number of nonsensitive itemsets, that is o.

Constraint (12) is known as big-M constraint and it ensures that $z_{ik} = 1$ if any item of the nonsensitive itemset is set to be removed. The constant M should be big enough for this purpose. In our case, M can be assigned the number of items, *i.e.*, $M = m$. The total number of these constraints again depend on the dataset, yet the maximum value would be $n \times o$. The size of the problem, that is the number of decision variables and the number of constraints highly depend on the size of the dataset.

Constraints (9) and (11) are contradictory and in almost all cases, there is no solution to this integer programming problem. Thus, solution can be found by relaxing or simply removing some of the constraints. This relaxation can be done through constraints (11) since constraints (9) are essential for the privacy preservation. Thus, solution can be found by relaxing or simply removing this constraint. Such a relaxation was also proposed by (Gkoulalas-Divanis and Verykios, 2006). However, removing both constraint (11) and its associated constraint (12) yields to ignoring the preservation of nonsensitive itemsets in this case. Instead, this constraint can be relaxed by introducing a new decision variable $\mathcal{R}_k$ in constraint (11) and adding another big-M constraint associated to this altered constraint:

$$s.t. \quad \sum_{i=1}^{n} z_{ik} \leq \sigma(I_k) - \psi + \mathcal{R}_k, \quad \forall k : I_k \in \mathcal{F}^{N}_{(\mathcal{D},\psi)}, \tag{14}$$

$$s.t. \quad \mathcal{R}_k \leq M_2\, u_k, \quad \forall I_k \in \mathcal{F}^N_{(\mathcal{D},\psi)}, \tag{15}$$

$$u_k \in \{0, 1\}, \tag{16}$$

$$0 \leq R_k \leq n, \qquad R_k \in \mathbb{N}. \tag{17}$$

Note that for any nonsensitive itemset I_k, it is desired that $\mathcal{R}_k$ remains always zero to ensure that the frequency is above the support threshold. If $\mathcal{R}_k > 0$, one can say that the nonsensitive itemset is also hidden as a side-effect. If all R_k is zero, in most cases, there is no solution to the above problem due to constraint (11).

Constraint 15 is another big-M constraint and it guarantees that $u_k = 1$ when $\mathcal{R}_k > 0$. Since $\mathcal{R}_k$ cannot exceed n, M_2 can be assigned $n + 1$. The objective function given in (8) becomes

$$\min \quad \sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij}x_{ij} + \sum_{i=1}^{n}\sum_{k=1}^{o} y_{ik} + \sum_{i=1}^{n}\sum_{k=1}^{o} u_k \tag{18}$$

For our sample database, the inequality (9) constructs the first group of constraints as follows:

$$y_{1,1} + y_{4,1} + y_{5,1} \geq 3 - 2 + 1 \tag{19}$$
$$y_{1,2} + y_{4,2} + y_{8,2} + y_{9,2} \geq 4 - 2 + 1 \tag{20}$$
$$y_{4,3} + y_{8,3} + y_{9,3} \geq 3 - 2 + 1 \tag{21}$$
$$y_{4,4} + y_{5,4} + y_{8,4} + y_{10,4} \geq 4 - 2 + 1 \tag{22}$$

where a_1, for example, represents the decision variable x_{11} in the context of this example. The inequality (10) constructs the second group of constraints as follows:

$$
\begin{aligned}
a_1 + b_1 + c_1 &\geq y_{1,1}, \\
c_1 + h_1 &\geq y_{1,2} \\
a_4 + b_4 + c_4 &\geq y_{4,1}, \\
c_4 + h_4 \geq y_{4,2}, \quad h_4 + i_4 &\geq y_{4,3}, \\
f_4 + g_4 &\geq y_{4,4} \\
a_5 + b_5 + c_5 &\geq y_{5,1}, \\
f_5 + g_5 &\geq y_{5,4} \\
c_8 + h_8 &\geq y_{8,2}, \\
h_8 + i_8 \geq y_{8,3}, \quad f_8 + g_8 &\geq y_{8,4} \\
c_9 + h_9 &\geq y_{9,2}, \\
h_9 + i_9 &\geq y_{9,3} \\
f_{10} + g_{10} &\geq y_{10,4}
\end{aligned}
$$

the inequality (11) constructs the first group of constraints as follows:

$$
\begin{aligned}
z_{1,1} + z_{4,1} + z_{5,1} &\leq 3 - 2 + \mathcal{R}_1 \\
z_{1,2} + z_{4,2} + z_{5,2} &\leq 4 - 2 + \mathcal{R}_2 \\
&\vdots \\
z_{1,45} + z_{4,45} &\leq 4 - 2 + \mathcal{R}_{45}
\end{aligned}
$$

The inequality (12) constructs the second group of constraints as follows:

$$
\begin{aligned}
b_1 + c_1 \leq 10z_{1,1}, \quad a_1 + c_1 &\leq 10z_{1,2}, \cdots, \\
a_1 + g_1 + h_1 &\leq 10z_{1,45} \\
b_4 + c_4 \leq 10z_{4,1}, \quad a_4 + c_4 &\leq 10z_{4,2}, \cdots, \\
a_4 + g_4 + h_4 &\leq 10z_{4,45} \\
&\vdots \\
e_{10} + f_{10} &\leq 10z_{10,16}
\end{aligned}
$$

The inequality (15) constructs the second group of constraints as follows:

$$\mathcal{R}_1 \leq 11u_1, \quad \mathcal{R}_2 \leq 11u_2, \quad \cdots, \quad \mathcal{R}_{45} \leq 11u_{45} \tag{23}$$

Thus, for our sample database, the objective function is:

$$\begin{aligned} \min \quad & a_1 + b_1 + c_1 + h_1 + \ldots + h_9 + i_9 + f_{10} \\ & + y_{1,1} + y_{1,2} + y_{4,1} + \ldots + y_{8,4} + y_{10,4} \\ & + u_1 + u_2 + u_3 + \ldots + u_{45} \end{aligned} \tag{24}$$

The solution of the problem is, $c_1 = b_5 = g_5 = f_8 = h_8 = h_9 = g_{10} = 1$ where the rest is zero. The distance is 7 and the number of frequent itemsets

Table 5: Sanitized database by FULL-EXACT

d_{ij}	a	b	c	d	e	f	g	h	i	j
T_1	**1**	**1**	~~1~~	0	0	0	**1**	**1**	0	**1**
T_2	0	0	**1**	0	0	0	0	0	**1**	**1**
T_3	0	0	0	**1**	**1**	**1**	0	0	0	0
T_4	**1**	**1**	**1**	0	0	**1**	**1**	**1**	**1**	0
T_5	**1**	~~1~~	**1**	0	0	**1**	~~1~~	0	0	0
T_6	0	0	0	0	0	0	0	0	0	**1**
T_7	0	0	0	**1**	0	0	0	0	0	0
T_8	0	0	**1**	0	0	~~1~~	**1**	~~1~~	**1**	0
T_9	0	0	**1**	0	0	0	0	~~1~~	**1**	0
T_{10}	0	0	0	0	**1**	**1**	~~1~~	0	0	0

of the distorted dataset is 20. One should recall that when we exclude 4 sensitive itemsets and its 28 supersets from the initial 62 frequent itemsets, 30 frequent itemsets are expected to survive after the hiding process. However, the 10 of 30 itemsets could not be preserved in the optimum solution. On the other hand, according to Menon's approach, the solution would be $c_1 = c_4 = g_4 = h_4 = g_8 = h_8 = g_7 = 1$ where the rest is zero. The distance is again 7 yet the number of frequent itemsets is 11 in this case resulting in the loss of 9 more non-sensitive itemsets compared to FULL-EXACT method. The resulting database is shown in Figure 6.

4 Performance Evaluation

In this section, a performance evaluation is performed using 10 real datasets arbitrarily chosen from UCI repository (Coenen, 2003). The characteristics of these datasets are given in Table 7. The FULL-EXACT method is performed together with a popular existing algorithm chosen from the literature for comparison purposes. The reference method is Menon's exact approach (Menon et al., 2005). The first part of the method proposed in Menon (Menon et al., 2005) decides about the minimum number of transactions that have to

Table 6: Sanitized database by Menon's method

d_{ij}	a	b	c	d	e	f	g	h	i	j
T_1	**1**	**1**	✱	0	0	0	**1**	**1**	0	**1**
T_2	0	0	**1**	0	0	0	0	0	**1**	**1**
T_3	0	0	0	**1**	**1**	**1**	0	0	0	0
T_4	**1**	**1**	✱	0	0	**1**	✱	✱	**1**	0
T_5	**1**	**1**	**1**	0	0	**1**	**1**	0	0	0
T_6	0	0	0	0	0	0	0	0	0	**1**
T_7	0	0	0	**1**	0	0	0	0	0	0
T_8	0	0	**1**	0	0	**1**	✱	✱	**1**	0
T_9	0	0	**1**	0	0	0	0	**1**	**1**	0
T_{10}	0	0	0	0	**1**	**1**	✱	0	0	0

Table 7: DB Parameters

DB	n	m	ψ	$\lvert\mathcal{F}_{(D,\psi)}\rvert$
Dermatology	366	49	164 (45%)	98
Auto	205	137	102 (50%)	63
Waveform	5000	101	2000 (40%)	50
Soybean	683	118	498 (73%)	98
Hepatitis	155	56	100 (65%)	85
Breast	699	20	384 (55%)	45
Glass	214	48	42 (20%)	132
Zoo	101	42	60 (60%)	40
Congres	435	34	174 (40%)	71
HorseColic	368	27	147 (40%)	43

be sanitized, with the objective of minimizing accuracy using linear programming. The results are then used by the heuristic part for actual sanitization trying to make minimum harm on actual database. Tests are done to see the performance of the algorithms in terms of distance, information loss and time while the number of sensitive itemsets vary.

All evaluations are performed on an Intel® Core™ $i7-2600$ @ 3.40GHz CPU, 16 GB RAM, Linux version 2.6.32-35-server (gcc version 4.4.3) standard computer. For the evaluation of the methods presented in this work, we developed various utilities that constitute the toolkit PPDM that is entirely available as a GNU open source project (Ayav, 2013). This toolkit has two dependencies, Borgelt's apriori utility (Borgelt, 2003) and the GNU Linear Programming Kit (GLPK) package (GLPK, 2000). The apriori utility is used to compute the frequent itemsets of a given dataset and the GLPK package allows us to solve the linear programming models. PPDM contains software that allows us to select the sensitive itemsets according to the specified criteria, to compute the cost matrix, to generate various matrices that GLPK and other programs need to work.

The results of the tests carried out on the benchmark datasets are given in Table 8. The table shows the distance and the number of nonsensitive itemsets in the distorted datasets after performing Full-Exact and Menon's approach. Note that time results are not included in the table since time performance issue is out of the scope of this work. It is known that exact approaches are quite expensive, which makes them usually impractical yet achieving the optimum solution may be quite useful especially when comparing the newly introduced algorithms. As seen, in all experiments, Full-Exact achieves better results in terms of both distance and information loss. For example, in Dematology dataset, the number of nonsensitive itemsets is 98. Ideally, this number minus the number of sensitive itemsets, *i.e.,* $98 - 2 = 96$ itemsets should be preserved after distortion. Full-Exact's distortion reduces this number to 73 whereas Menon's distortion reduces it to 45. On the other hand, distance gives the number of items removed during distortion process and should as small as posssible. Again, Full-Exact removed 60 items whereas Menon's method removed 98 items.

5 Related Work

Privacy preserving association rule hiding problem was first introduced in (Atallah et al., 1999). The authors proposed heuristic algorithms and gave the proof of NP-Hardness of optimal sanitization. Since then, many approaches have been proposed to preserve privacy for sensitive patterns or sensitive association rules in database. Different categorizations of rule hiding or itemset hiding approaches can be done according to various attributes of the solution process. Their difference is their ability to hide i) single or multiple itemsets, ii) exclusive or overlapped itemsets, iii) itemsets of single or multiple thresholds, iv) according to support or confidence, v) depending on no pre-hiding process, vi) by removal of sensitive transaction or not. Besides these categorizations, widely known classification is done according to the nature of the base algorithm and following classes appear; heuristic based approaches, border based approaches, exact approaches, reconstruction based approaches and cryptography based approaches.

Table 8: Experimental Results

DB Name	Sensitive itemsets	Sensitive transactions	Non-sensitive itemsets	Full Exact Approach		Menon's Exact Approach	
				# of Nonsensitive itemsets	Distance	# of Nonsensitive itemsets	Distance
Dermatology	2	285	98	73(74.5%)	60(1.3%)	45(45.9%)	98(2.0%)
Auto	2	110	63	63(100%)	5(0.1%)	61(96.8%)	5(0.1%)
	3	165	62	62(100%)	23(0.5%)	56(90.3%)	24(0.5%)
Waveform	2	2555	50	50(100%)	394(0.36%)	47(94%)	399(0.36%)
Soybean	2	554	64	63(98.4%)	35(0.18%)	54(84.4%)	35(0.18%)
Hepatitis	2	107	85	67(68.4%)	8(0.29%)	58(68.2%)	8(0.29%)
	3	114	77	59(76.6%)	15(0.54%)	54(70.1%)	15(0.54%)
Breast	2	396	45	36(80%)	217(3.1%)	24(53.3%)	341(5.0%)
Glass	2	44	132	113(85.6%)	3(0.15%)	110(83.3%)	3(0.15%)
	4	64	130	108(83%)	23(1.1%)	95(73.1%)	23(1.1%)
Zoo	2	79	43	40(93%)	13(0.76%)	28(65.1%)	16(0.93%)
	3	95	40	34(85%)	21(1.2%)	28(70%)	21(1.2%)
Congres	2	176	71	52(73.2%)	3(0.06%)	52(73.2%)	3(0.06%)
	4	247	82	48(58.5%)	39(0.75%)	45(54.9%)	44(0.85%)
HorseColic	2	321	43	27(62.8%)	174(3.16%)	19(44.2%)	176(3.2%)

Heuristic Based Approaches: Hiding problem is generalized as to consider the hiding of both sensitive frequent itemsets and sensitive association rules in (Dasseni et al., 2001). The authors propose three single rule heuristic hiding algorithms that are based on the reduction of either the support or the confidence of the sensitive rules, but not both. The sanitization framework and four different sanitization algorithms that follow similar steps but differ in selecting the item in victim itemset are proposed by (Oliveira and Zaiane, 2002). Two more sanitization algorithms are proposed in (Oliveira and Zaiane, 2003a); Round Robin Algorithm and Random Algorithm which try to distort items equally while decreasing of the support of the items in sensitive rules. The algorithm in (Oliveira and Zaiane, 2003b), called SWA, is an efficient, scalable, one-scan heuristic which aims at providing a balance between the needs for privacy and knowledge discovery in association rule hiding. Three effective, multiple association rule hiding heuristic algorithms are proposed by (Amiri, 2007) and shown that they outperform SWA by offering higher data utility and lower distortion. Five heuristic algorithms based on two strategies are proposed in (Verykios et al., 2004); first approach prevents rules from being generated, by hiding the frequent sets from which they are derived whereas the second approach reduces the importance of the rules by setting their confidence below a user-specified threshold. An interesting heuristic proposition comes from (Wu et al., 2007) where they classify all the valid modifications such that every class of modifications is related with the sensitive rules, nonsensitive rules, and spurious rules that can be affected after the modifications. Template based hiding strategy of (Kuo et al., 2008) is interesting since sensitive frequent patterns with multiple sensitive thresholds are sanitized aiming fulfill user requirements in real applications. Co-occuring frequent itemset hiding framework with different heuristic algorithms is the contribution of (Abul, 2009). Pattern Inversion Tree is used to store related information in (Wang et al., 2008). DSC (Decrease Support and Confidence) decreases confidence of a rule and/or decrease the support of the large itemset of a rule. Similarly a matrix structure is used in (Yildiz and Ergenc, 2012).They propose an integrated itemset hiding algorithm that eliminates the need of pre-mining and post-mining and uses a simple heuristic in selecting the itemset and the item in itemset for distortion.

All above techniques are data distortion based; they try to hide association rules by decreasing or increasing support (or confidence). To increase or decrease support (or confidence), they replace 0's by 1's or vice versa in selected transactions. Majority of research belongs to this group since they

are efficient and scalable but they produce side effect in new database (i.e. lost rules, ghost rules).There is another group of heuristic techniques named data blocking based; they replace the 0's and 1's by unknowns ("?") in selected transaction instead of inserting or deleting items (Saygin et al., 2001; Yeh and Hsu, 2010). So it is difficult for an adversary to know the value behind "?". Although blocking based techniques minimize the side effects, it is difficult to reproduce the original dataset.

Border Based Approaches: Borders allow for a condense representation of the frequent itemsets in a database, effectively identifying those key itemsets in the lattice which separate the frequent patterns from their infrequent counterparts (Mannila and Toivonen, 1997). The process of border revision facilitates the minimum harm in the hiding of the sensitive itemsets.These approaches preprocess the sensitive rules so that minimum numbers of rules are given as input to hiding process. So, they maintain database quality while minimizing side effects. Border revision process is proposed by ((Sun and Yu, 2005) and (Sun and Yu, 2007)). Hiding process greedily selects those modifications that lead to minimal side effects by assigning dynamic weights to each itemset on pozitive border. On the other hand another methodology which relies on max-min criterion for hiding sensitive itemsets is given in (Moustakides and Verykios, 2008). Revised positive border of frequent itemsets is used to keep track of the impact of each tentative item modification. Border based approaches maintain data quality by greedily selecting the modification by minimum side effect. They bring improvement over pure heuristic approaches but still they are unable to identify optimal hiding solution.

Exact Approaches: Exact approaches formulate the hiding problem to constraint satisfaction problem (CSP) and solve it by using binary integer programming (BIP). They provide an exact (optimal) solution that satisfies all the constraints. However if no exact solution exists in database, some of the constraint are relaxed. The approach presented in (Menon et al., 2005) finds minimum number of transactions that have to be sanitized by formulating CSP. Set of transactions found by CSP are used by two different item selecting heuristic strategies. Higher sanitization granularity is proposed by (Gkoulalas-Divanis and Verykios, 2006); inline algorithm finds optimal solution for rule hiding problem without using any heuristics. Two-phase iterative algorithm of (Gkoulalas-Divanis and Verykios, 2009a) proposes a par-

titioning approach for the scalability of the inline algorithm. A framework that is suitable for decomposition and parallelization of the exact hiding algorithms given in ((Gkoulalas-Divanis and Verykios, 2006), (Gkoulalas-Divanis and Verykios, 2008) and (Gkoulalas-Divanis and Verykios, 2009b)) is presented in (Gkoulalas-Divanis and Verykios, 2008). A new exact approach is proposed by (Leloglu and Ergenc, 2014) where the idea is to use coefficients in the inequalities of integer programming to prevent the deletion of non-sensitive itemsets in sanitization process. Exact approaches guarantee quality for hiding sensitive information than other approaches but they require very high time comlexity due to integer programming.

Reconstruction Based Approaches: Reconstruction based approaches generate privacy aware database by extracting sensitive characteristics from the original database. These approaches generate lesser side effects in database than heuristic approaches. A FP tree based algorithm which reconstructs the original database by using non characteristic database is demonstrated in (Guo, 2007). A fake transaction randomization method is presented in (Abul, 2009). The method of (Lin and Liu, 2007) ensures the privacy of data by mixing real transactions with fake transactions. Another similar method is given by (Boora et al., 2009) where transaction randomization method is a combination of the fake transaction randomization method and a new per-transaction randomization method. These approaches provide less side effect than heuristic approaches but open problem is to restrict the number of fake transactions in the new database.

Cryptography Based Approaches: These approaches are used for multiparty computation in case database is distributed among sites. Multiple parties may wish to share their private data, without leaking any sensitive information at their end.In these approaches, instead of distortion, database is encrypted before sharing. A secure approach where database is vertically partitioned is proposed by (Vaidya, 2001). Communication for large datasets is high in this approach. Secure mining of association rules over horizontal partitioned data is addressed in (Kantarcioglu and Clifton, 2004). Communication and computation cost is claimed to be reasonable. Cryptography based approaches provide secure mining of association rules over partitioned databases but they fall short of providing complete answer to the problem of privacy preserving data mining.

After having in-depth analysis of existing itemset approach we can make following remarks: i) majority of the solutions are based on heuristics, ii) in most of them the objective is to minimize distance during hiding process, iii) distance and information loss is aimed to be provided by completely different objectives, iv) exact methods provide better results in terms of distance or information loss, v) there is no full exact method which aims to find the optimal solution without using heuristics.

6 Conclusion

This work presents a novel full exact approach for itemset hiding problem in the context of association rule hiding. The approach utilizes integer programming that optimize two common side effects, distance and information loss in the sanitization process. 10 real but relatively small datasets are used for performance evaluations. The results show that Full-Exact achieves similar results in terms of distance. However information loss performance of Full-Exact approach is distinctively better than the similar approach. This work disregards the time performance of the method since the size and nature of the problem makes it impractical yet this full exact approach should be considered as a reference that provides the optimum solution to the itemset hiding problem without any heuristics.

References

Abul, O. (2009). Hiding co-occuring frequent itemsets. In *2nd International Workshop on Privacy and Anonymity in the Information Society (PAIS)*, St. Petersburg, Russia.

Agrawal, R., Imilinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. Morgan Kaufmann Publishers.

Amiri, A. (2007). Dare to share: Protecting sensitive knowledge with data sanitization. *Decision Support Systems*, 43(1):181–191.

Atallah, M., Bertino, E., Elmagarmid, A., Ibrahim, M., and Verykios, V. S. (1999). Disclosure limitation of sensitive rules. In *Workshop on Knowledge and Data Engineering Exchange*, pages 45–52, Chicago, USA.

Ayav, T. (2013). PPDM: Privacy Preserving Data Mining Toolkit. `https://code.google.com/p/privacy-preserving-data-mining/`. Version 1.0.

Bonchi, F. and Ferrari, E. (2011). *Privacy-aware knowledge discovery novel applications and new techniques*. CRC Press.

Boora, R. K., Shukla, R., and Misra, A. (2009). An improved approach to high level privacy preserving itemset mining. *International Journal of Computer Science and Information Security*, 6(3):216–223.

Borgelt, C. (2003). Apriori - Association Rule Induction / Frequent Item Set Mining Toolkit. `http://www.borgelt.net/apriori.html`.

Coenen, F. (2003). `http://www.csc.liv.ac.uk/ frans/KDDD/Software/`.

Dasseni, E., Verykios, V. S., Elmagarmid, A. K., and Bertino, E. (2001). Hiding association rules by using confidence and support. In *4th International Workshop on Information Hiding*, pages 369–383.

Gkoulalas-Divanis, A. and Verykios, V. (2008). A parallelization framework for exact knowledge hiding in transactional databases. In *IFIP International Federation for Information Processing*, volume 278, pages 349–363.

Gkoulalas-Divanis, A. and Verykios, V. S. (2006). An integer programming approach for frequent itemset hiding. In *15th ACM International Conference on Information and Knowledge Management*.

Gkoulalas-Divanis, A. and Verykios, V. S. (2009a). Exact knowledge hiding through database extension. *IEEE Transactions on Knowledge and Data Engineering*, 21(5):699–713.

Gkoulalas-Divanis, A. and Verykios, V. S. (2009b). Hiding sensitive knowledge without side effects. *Knowledge and Information Systems*, 20(3):263–299.

Gkoulalas-Divanis, A. and Verykios, V. S. (2010). *Association rule hiding for data mining*. Springer.

GLPK (2000). GNU Linear Programming Kit. `http://www.gnu.org/software/glpk`.

Guo, Y. (2007). Reconstruction-based association rule hiding. In *SIGMOD Ph.D. Workshop on Innovative Database Research*.

Jèunger, M., Liebling, T., and Naddef, D. (2010). *50 Years of Integer Programming 1958-2008: The Early Years and State-of-the-art Surveys*. Springer Berlin Heidelberg.

Kantarcioglu, M. and Clifton, C. (2004). Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Trans. on Knowl. and Data Eng.*, 16(9):1026–1037.

Keer, S. and Singh, A. (2012). Hiding sensitive association rule using clusters of sensitive association rule. *International Journal of Computer Science and Network (IJCSN)*, 1(3).

Kuo, Y., Lin, P. Y., and Dai, B. R. (2008). Hiding frequent patterns under multiple sensitive thresholds.

Leloglu, E. Ayav, T. and Ergenc, B. (2014). Coefficient-based exact approach for frequent itemset hiding. In *eKNOW 2014 : The Sixth International Conference on Information, Process, and Knowledge Management*, pages 124–139.

Lin, J. and Liu, J. Y. C. (2007). Privacy preserving itemset mining through fake transactions. In *22nd ACM Symposium on Applied Computing, SAC*, pages 375–379, Seoul, Korea.

Mannila, H. and Toivonen, H. (1997). Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3):241–258.

Menon, S., Sarkar, S., and Mukherjee, S. (2005). Maximizing accuracy of shared databases when concealing sensitive patterns. *Information Systems Research*, 16(3):256–270.

Mielikainen, T. (2003). On inverse frequent set mining. In *IEEE ICDM Workshop on Privacy Preserving Data Mining*, pages 18–23.

Mohaisen, A., Jho, N., Hong, D., and Nyang, D. (2010). Privacy preserving association rule mining revisited: Privacy enchancement and resource efficieny. *IEICE Transactions on Information and Systems*, E93(2):315–325.

Moustakides, G. V. and Verykios, V. S. (2008). A maxmin approach for hiding frequent itemsets. *Data and Knowledge Engineering*, 65(1):75–89.

Oliveira, S. R. M. and Zaiane, O. R. (2002). Privacy preserving frequent itemset mining. In *International Conference on Data Mining, ICDM*, pages 43–54, Maebashi City, Japan.

Oliveira, S. R. M. and Zaiane, O. R. (2003a). Algorithms for balancing privacy and knowledge discovery in association rule mining. In *Seventh International Database Engineering & Applications Symposium*, pages 54–63.

Oliveira, S. R. M. and Zaiane, O. R. (2003b). Protecting sensitive knowledge by data sanitization. In *3rd IEEE International Conference on Data Mining (ICDM)*, pages 211–218.

Saygin, Y., Verykios, V. S., and Clifton, C. (2001). Using unknowns to prevent discovery of association rules. *ACM SIGMOD*, 30(4):45–54.

Sun, X. and Yu, P. (2007). Hiding sensitive frequent itemsets by a border-based approach. *Computing Science and Engineering*, 1(1):74–94.

Sun, X. and Yu, P. S. (2005). A border-based approach for hiding sensitive frequent itemsets. In *5th IEEE International Conference on Data Mining (ICDM)*, pages 426–433.

Vaidya, J. (2001). Privacy preserving association rule mining in vertically partitioned data. In *In The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 639–644.

Verykios, V. S., Emagarmid, A. K., Bertino, E., Saygin, Y., and Dasseni, E. (2004). Association rule hiding. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):434–447.

Wang, S. L., Maskey, R., Jafari, A., and Hong, T. (2008). Efficient sanitization of informative association rules. *Science Direct, Expert Systems with Applications*, 35:442–450.

Wu, Y. H., Chiang, C. M., and Chen, A. L. P. (2007). Hiding sensitive association rules with limited side effects. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):29–42.

Yeh, J. S. and Hsu, P. C. (2010). Novel algorithms for privacy preserving utility mining. *Science Direct, Expert Systems with Applications*, pages 4779–4786.

Yildiz, B. and Ergenc, B. (2012). Integrated approach for privacy preserving itemset mining.