International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 5 (2019) pp. 1177-1185

Preserving Privacy of Sensitive Itemsets using Controlled Perturbation of Closed Itemsets

[1]Surendra H*
Research Scholar, Department of Information Science and Engineering,
SJB Institute of Technology
Bangalore, India.

[2]Dr. Mohan H S
Professor & Head, Department of Information Science and Engineering,
SJB Institute of Technology
Bangalore, India.

Abstract
Data perturbation is one of the famous techniques in privacy preserved data mining. It is considered relatively easy and effective approach for preserving sensitive information in the released data. In this paper, the authors propose an improved version of their previous work which uses value distortion-based data sanitization algorithm to safely perturb the original support of sensitive itemsets without generating any side effects. The data sanitization method used in the previous work suffers from spurts of uncontrolled loss of information and support accuracy of itemsets in sparse datasets. To overcome this limitation, in this paper, the authors propose an improved technique which randomly distorts the support of sensitive itemsets in the closed itemset lattice within specified error limit or threshold, also keeping the relationship between itemsets unchanged. Experimental results show that the proposed improved approach is more efficient in perturbing the data to preserve privacy when compared with the previous work and other well-known distortion-based approaches.

Keywords: Privacy Preserving Data Mining (PPDM), Privacy-Preserving Data Publishing (PPDP), Data Sanitization, Data Masking, Data Perturbation, Value Distortion, Frequent Itemset Masking, Closed Itemsets.

1. INTRODUCTION

Privacy Preserving Data Mining (PPDM) [1, 2] has received more attention in recent years due to the rising concern over the privacy of the data. Data Mining and Big Data analytics are being used by many organizations to improve their business. Interesting patterns are discovered from raw data to make better business decisions. These patterns may contain sensitive private information which needs to be preserved from disclosing during the mining process. Removal of sensitive attributes or key identifiers from the data is not adequate for preserving privacy as many countermeasures have been developed to breach privacy. The goal of the privacy-preserving data mining is to facilitate mining of quality information from the data without disclosing sensitive information.

Frequent Itemset (FI) mining is a prominent data mining technique used to retrieve interesting patterns present in the data. It is also the preliminary step in mining association rules and correlations in the data. Some frequent itemsets may contain sensitive information which the data owner does not want to disclose. So, the data need to be sanitized by either hiding or distorting the original information of sensitive frequent itemsets. Also, the problem of frequent itemset hiding has been proved to be NP-hard [3].

This paper is an enhancement to our previous work on preserving the privacy of sensitive association rules by distorting closed itemsets [4]. In this paper, an improved value distortion type of data perturbation technique is used as the sanitization method to perturb the original support values of sensitive itemsets without any side effects. The distortion is controlled within a defined range which protects the original support from disclosing and also keeping the relationship between the itemsets intact, preserving the utility of the data. The algorithm can be used to protect the whole database or selected sensitive frequent itemsets.

This technique is significantly useful in applications where the data needs to be shared with any third party for research or business purposes. In order to preserve privacy or, protect sensitive or confidential information from revealing to the third party, the data needs to be perturbed. Perturbing the data usually results in side effects which reduce the usability of the data. The alternate technique is to share privacy preserved patterns instead of data. The proposed method uses a pattern sharing technique instead of data sharing where closed frequent patterns are shared with their support values distorted within the specified error range. So, the shared perturbed closed frequent patterns do not reveal the original support value preserving privacy. The controlled perturbation, i.e., induction of error in the support values of the patterns increase the usability of the shared patterns. Also, sharing privacy preserved patterns to public protects the data owner from breach of privacy law and policies set by government or institutions. Some of the applications are market basket analysis and customer profile information sharing.

The paper is organized as follows: In Section 2, existing related work on different data sanitization methods used for preserving the privacy of frequent itemsets is presented. Section 3 formulates the problem. The proposed method is explained in detail in Section 4. Experimental results are presented in Section 5. The paper is concluded in Section 6.

2. RELATED WORK

S. Rizvi and J. Harisa [5] proposed a probabilistic distortion of data using random numbers generated from a pre-defined distribution function. The transaction converted to the Boolean database such as retail data is considered. The distortion was to performed by flipping 0 to 1 and vice versa with some probability p and retain original with probability 1-p. Mining the distorted database had become much complex and time-consuming. S. Agrawal, V. Krishnan and J. Harista [6] addressed these issues by selecting the distortion parameters appropriately and applying optimization methods based on set theory.

M. Atallah et al. [7] was the first to propose the algorithm for hiding sensitive association rules by reducing the support count of the itemsets. Greedy iterative traversal of subsets of the large sensitive itemset is performed, and the subset with maximum support is identified as candidate sensitive itemset for hiding. The itemsets are hidden by reducing their support which is achieved by removing itemsets from the transaction one by one. E. Dasseni et al. [8] proposed a generalized technique for hiding both sensitive itemsets and association rules. Frequent itemset hiding is achieved by pruning them from the transaction till their support drops below minimum support threshold. The rule hiding is performed by either increasing the support of rule antecedent or reducing the support of rule consequent so that the rule confidence drops below minimum confidence threshold. The technique worked for rules not having same itemsets and suffers from side effects. Y.-H. Wu et al. [9] remove this limitation. V.S. Verykios et al. [10] improved the work of [8]. The rules are hidden one by one in decreasing order of their size and support. The itemsets are removed from the transaction in round robin fashion until its support drops below a threshold. Oliveira S R M and Zaiane O R [11] first proposed an approach to hide multiple sensitive rules. The algorithm first scans the database to index the sensitive transactions and scans the database second time to remove the items from the transaction. They improved their earlier work in [12] which use a single scan of the database to identify the need for the hiding rules and hide them. Lin CW., Hong TP., Hsu HC [13] proposed hiding missing utility itemsets by removing them from the transactions. The missing itemsets and hiding failures are used to remove the itemset from the transactions optimally. The total number of transactions is maintained the same as the original data in the sanitized database.

Gregory Caiola and Jerome P. Reiter [14] proposed to model the data as Random Forests. The identified sensitive attributes are sanitized in the model. Then the privacy preserved synthetic data is generated from the sanitized model. The limitation of this approach is that the model has to be re-generated if a different class variable or the sensitive attribute is selected. Also, this approach is designed to work on only categorical data. Jun Zhang et al. [15] proposed to model the data as Bayesian Network. By learning the conditional probability distribution of attributes values, a differentially private Bayesian Network is built. The privacy of sensitive information is preserved by adding noise to selected distributions.

Tzung-Pei Hong et al. [16] proposed a Sensitive Items Frequency-inverse database frequency (SIF-IDF), a greedy-based approach. The SIF-IDF is based on Term Frequency-Inverse Document Frequency (TF-IDF) used in Text Mining. The SIF-IDF value for each transaction is found. The minimum count to which the sensitive itemsets are to be reduced is calculated. The transactions having best SIF-IDF value are identified, and the sensitive items are removed from those transactions.

Chirag Modi et al. [17] used a hybrid technique for preserving sensitive association rules from disclosure. The association rules are generated using the Apriori algorithm, and then the sensitive rules are sanitized by decreasing their support or confidence. The number of transactions to be modified is found, and respective itemsets are removed from the transactions. Hai QuocLe et al. [18] adopted the lattice theory and used intersection lattice of frequent itemsets to reduce the side effects due to the reducing the support of sensitive items. The support of sensitive items is reduced below minimum support by removing the items from selected transactions. Amirhosain Shahsavari and Shahram Hosseinzadeh [19] measured different factors to list the sensitive rules in the decreasing order based on their sensitivity levels. The rules are then hidden one by one from the top of the list by removing the sensitive items in the rule from the transactions. Peng Cheng et al. [20] used an evolutionary algorithm named multi-objective optimization (EMO)for hiding sensitive rules. Apriori algorithm is used to find frequent itemsets and association rules from the dataset. Then using EMO algorithm, the transactions to be modified for the given set of sensitive rules are found. The sensitive itemsets are removed from these transactions to hide sensitive rules. Janakiramaiah Bonam et al. [21] proposed association rule hiding by data distortion technique using Particle Swarm Optimization (PSO) algorithm. The highest support sensitive item is chosen and the best transaction to remove this item is identified using PSO algorithm. The hiding is performed by modifying these transactions. Hiding each sensitive item requires multiple iterations. This process is repeated until all sensitive rules are hidden. Alaa K. Jumaah et al. [22] proposed a simple method of hiding sensitive rules. The support of the itemset in rule antecedent is increased, and the support of the itemset in rule consequent is decreased. The modification of support values of the itemsets is performed by removing or adding itemsets from the transactions. Peng Cheng et al. [23] proposed blocking-based association rule hiding method. The sensitive items are replaced with an unknown symbol such as "?" which makes the confidence of sensitive rules placed in a range instead of concrete value. Transactions for modification are selected using the border rules they contain. Yu-Chuan Tsai et al. [24] compared k-anonymization with the association rule hiding algorithm and found that the k-anonymization method to hide sensitive rules provides higher privacy gain.

Most of the methods studied in the literature hide sensitive itemsets or rules by modifying the transactions. Sensitive itemsets are removed or added to the transactions to change their support values. Few use frequent itemset lattice to

identify the transactions to remove the itemsets to reduce side effects optimally. These techniques produce sanitized database in the same format as original data which the user has to process again to generate frequent patterns or association rules. Also, these techniques address modifying only a few itemsets or rules and don't scale for preserving the whole database. These techniques also suffer from side effects and uncontrolled loss of information.

3. PROBLEM FORMULATION

Consider a database D representing a set of transactions T = $\{t_1, t_2, t_3,\ldots\}$ where t is a transaction. Each transaction t is a subset of I = $\{i_1, i_2, i_3,\ldots\}$ where I is a set of unique items i. An itemset is a set of items, and an itemset having k items is known as k-itemset. Itemsets A and B are said to exist in a transaction t if and only if $A \in t, B \in t, and A \cap B = \emptyset$.

3.1. Frequent Itemset

The support S of an itemset A is defined as the fraction of transactions in the database D having itemset A as given in Eq. 1. It is also called the frequency of the itemset. Minimum Support is the threshold for the support value where itemsets with support below this threshold are considered infrequent. Itemsets having support count more than or equal to this threshold are called Frequent Itemsets (FI).

$$Support(A) = \frac{No.of\ Transactions\ having\ Itemset\ A}{Total\ no.of\ Transaction\ in\ the\ Database\ D} \quad (1)$$

3.2. Closed Itemsets

Closed itemsets are the condensed set of itemsets generated by eliminating redundant itemsets. Frequent Itemsets and association rules can be mined from the closed itemsets for varying support and confidence threshold. An itemset is said to be closed if its support is not the same as any of its proper supersets. An itemset is called closed frequent itemset if it is closed itemset and its support is greater than the minimum support threshold in a given database D.

3.3. Privacy Preserved Frequent Itemset Mining

The frequent itemsets that the data owner does not want to disclose are called sensitive frequent itemsets. The privacy of the sensitive frequent itemsets is protected by removing them from the original data, by reducing their support count below a support threshold to hide them or distorting by adding noise to their support value to avoid disclosure of their original support value. Since we are considering the technique of distortion, the following issues need to be taken care;

- Side Effects: During the sanitization process, some non-sensitive infrequent itemsets can become frequent or frequent itemsets can become infrequent while protecting the sensitive itemsets.
- Information Loss: During the sanitization process, the support value of the sensitive frequent itemsets is altered. As a side effect, the support value of non-sensitive itemsets may also get altered. This deviation of support information from its original value is known as Information loss.

An optimal Privacy preserving frequent itemset mining techniques shall demonstrate minimum side effects with limited loss of information. In this paper, the information loss is measured as given in [25].

3.4. Problem Statement

Most of the existing solutions are designed to hide the given sensitive itemsets by decreasing their support below a given threshold. Sanitization of the whole database for protecting original support values of itemsets needs to be studied. The sanitization is performed by selecting transactions containing the sensitive itemsets for modification. Then the sanitized database is released for mining. Instead, sanitization can be done on a model of the original data which can be queried for patterns after sanitization without again subjecting the data for data mining process. The existing techniques suffer from side effects and uncontrolled information loss. After the sanitization process, if a new set of sensitive itemsets needs to be protected, then the complete process has to be repeated on the original database.

4. PROPOSED METHOD

Our approach consists of two phases. In the first phase, original database is transformed into a closed itemsets lattice which provides the compact representation of the data, and in the second phase, data sanitization operation is performed on the closed itemsets to preserve the privacy of given sensitive itemsets.

4.1 Finding Safe Range for Perturbing Support of Sensitive Itemset

Consider a sample transactional database shown in Table 1. The closed itemsets of the sample database are given in Table 2. The lattice structure of the closed itemsets arranged based on their support levels as shown in Fig 1. Let us assume closed itemset (2, 3) as sensitive itemset whose support needs to be distorted. The actual support of the itemset in the database is 0.6. The actual support value shall be replaced with different new support value to perturb the sensitive itemset. This new support value must be in the safe range to avoid side effects, i.e., to avoid the altering of other non-sensitive itemsets support.

Table 1: Sample Transactional Database

Transaction ID	Items
T1	1,3,4
T2	2,3,5
T3	1,2,3,5
T4	2,5
T5	1,2,3,4
T6	3
T7	2,3,5
T8	1,3,4
T9	2,3,5
T10	1,2,3,4

The safe range for perturbation of an itemset is the difference of support value between its superset having the highest support (Floor Limit) and its subset having the lowest support (Ceiling Limit). For the sensitive itemset (2, 3), its supersets are (2,3,5) with support 0.4 and (1,2,3) with support 0.3. So, the highest support value among its supersets is 0.4 which is

its Floor Limit. Its subsets are (2) with support 0.7 and (3) with support 0.9. So, the lowest support value among its subsets is 0.7 which is its Ceiling Limit. So, the safe range to distort the support of the itemset (2,3) is between 0.4 and 0.7.

Table 2: Closed Itemsets of Sample Database

Closed Itemsets	Support
2	0.7
3	0.9
1,3	0.5
2,5	0.5
2,3	0.6
1,3,4	0.4
1,2,3	0.3
2,3,5	0.4
1,2,3,4	0.2
1,2,3,5	0.1

If the given sensitive itemset is not present in the closed itemset lattice, then it is added to the lattice with its support ranging from the highest support of its supersets from which it is derived and the lowest support of its subsets. The support of superset from which the itemsets is derived becomes the Floor Limit and support of the subset with minimum support among the subsets of the itemset becomes the Ceiling Limit.

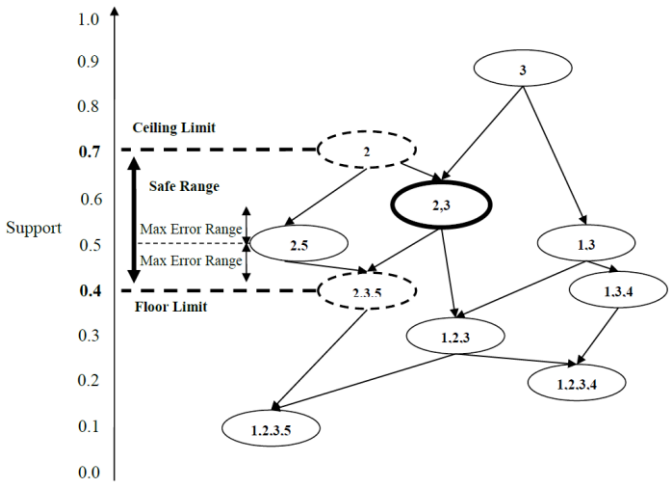

Figure 1: Closed Itemset lattice of sample database with Safe Range and Maximum Acceptable Error Range for distortion of support

After safe range for altering the support for sensitive itemset is found, the itemset can be perturbed in two ways:

1. Uncontrolled Randomized Perturbation (URP)
2. Controlled Randomized Perturbation (CRP)

4.2 Uncontrolled Randomized Perturbation

In the uncontrolled randomized perturbation [4], the support value of the sensitive itemset is allowed to vary randomly between the floor and ceiling limit. As the safe range can be different for different itemsets, the new support of a few sensitive itemsets can vary widely, and few may vary very little. In sparse data, the safe range can be sufficiently broad to perturb the sensitive itemset in a greater magnitude. Since only sensitive itemset is altered, there will not be side effects. The information loss is limited only to the loss of support of sensitive itemset. This technique can be used in perturbing only limited sensitive itemsets. The algorithm for uncontrolled random perturbation of the given sensitive itemset is given in Algorithm 1.

Algorithm 1: Uncontrolled Randomized Perturbation of Sensitive Itemsets

Input: Closed Itemset Lattice, Sensitive Itemsets
Output: Sanitized Closed Itemset Lattice with Perturbed Sensitive Itemsets without Side Effects

Step 1: Find the given Sensitive Itemset in the Closed Itemset Lattice at the level of size of the itemset. If found then jump to Step 3.
Step 2: If Sensitive Itemset not found (i.e., it is not closed itemset) then find its immediate supersets.
Step 3: Among its immediate supersets, find the superset having highest support. Set this support as Floor Limit. If supersets not available as in case of itemsets at the bottom of the lattice, select minimum support threshold or 0 as Ceiling Limit.
Step 4: Among its immediate subsets, find the subset having least support. Set this support as Ceiling Limit. If subsets not available as in case of itemsets at the top of the lattice, select the size of the database, i.e., the total number of transactions or 1 as Floor Limit.
Step 5: Generate a random number between the Floor Limit and Ceiling Limit, and assign it as support for the sensitive itemset

4.3 Controlled Randomized Perturbation

In this method, a maximum acceptable error is set while randomly perturbing the support of sensitive itemsets. The acceptable error is the percentage of original support of the sensitive itemset within which it can vary. The perturbed new support can be more or less than the original support but within the maximum acceptable error or safe range limit whichever is less. This technique can be mostly used to perturb the whole database or all itemsets within an acceptable error. The sanitized database generated using this technique protects the original information being disclosed, maintains the relationship between the itemsets and provides the maximum error the frequent itemset or association rule mining results contain. So, the original information is also protected, and the utility of the data is kept reasonable. The algorithm for controlled randomized perturbation of the database is given in Algorithm 2.

4.4 Treating Itemsets at Top or Bottom of the Lattice

For sensitive itemsets at the top of the lattice, the Ceiling Limit shall be 1 or database size in case of Uncontrolled Random Perturbation and case of Controlled Random Perturbation it is the maximum acceptable error support more than the original support of the itemset or its superset from which it was derived.

For itemsets at the bottom of the lattice, the Floor Limit shall be 0 or minimum support threshold in case of Uncontrolled

Random Perturbation and the maximum acceptable error support less than the original support.

Algorithm 2: Controlled Randomized Perturbation of Sensitive Itemsets

Input: Closed Itemset Lattice, Sensitive Itemsets
Output: Sanitized Closed Itemset Lattice with Perturbed Sensitive Itemsets within Given Range without Side Effects

Step 1: Find the given Sensitive Itemset in the Closed Itemset Lattice at the level of size of the itemset. If found then jump to Step 3.
Step 2: If Sensitive Itemset not found (i.e., it is not closed itemset) then find its immediate supersets.
Step 3: Among its immediate supersets, find the superset having highest support. Set this support as Floor Limit. If supersets not available as in case of itemsets at the bottom of the lattice, select minimum support threshold or 0 as Ceiling Limit.
Step 4: Among its immediate subsets, find the subset having least support. Set this support as Ceiling Limit. If subsets are not available as in case of itemsets at the top of the lattice, select the size of the database, i.e., the total number of transactions or 1 as Floor Limit.
Step 5: Find the Maximum Allowed Error Range (MAER) for the original support to vary based on the Maximum Acceptable Error % (MAE%) input i.e.

$$MAER = Itemset\ Original\ Support\ \times MAE\%$$

Step 6: Find the gap between the original support of the Itemset and Ceiling Limit.

$$Celinling\ Gap = Ceiling\ Limit - Original\ Support$$

If the distortion needs to be achieved by decreasing the original support, then find the gap between the original itemset support and Floor Limit.

$$Floor\ Gap = Original\ Support - Floor\ Limit$$

Step 7: If MAER is higher than the Ceiling Gap (or Floor Gap), then set the MAER to Ceiling Gap (or Floor Gap)
Step 8: Generate random number within MAER and add/subtract it to the original support of the Sensitive Itemset i.e.

$$Perturbed\ Support = Original\ Support \pm Random(MAER)$$

5. EXPERIMENTAL RESULTS

The proposed method has experimented on both real and synthetic datasets from [26]. The implementation of CHARM [27] closed itemset mining algorithm in SPMF Data Mining Library is extended to implement the proposed method. The experiments are performed on a computer system4 Core - Intel Corei5 processor and 8 GB RAM. The characteristics of the representative datasets are given Table 3.

The first step in our proposed approach is to build closed itemset lattice which represents the original database. The closed itemset lattice will be referred as the model in this paper. Experimental results show that mining closed itemset is resource intensive compared to frequent itemset mining but, closed itemsets yields a high degree of compression of itemsets. The performance comparison of mining closed itemset, and frequent itemsets are given in Table 4.

Table 3: Dataset characteristics

Dataset	Number of Unique Items	Number of Records	Average Record Length
connect	129	67557	43
pumsb	2113	49046	74
chess	75	3196	37
accidents	468	340183	33
mushroom	119	8124	23
T10I4D100K	870	100000	10
T40I10D100K	942	100000	39
retail	16470	88162	10

Since the count of closed itemsets is an order of magnitude lesser for the medium and highly dense dataset, sanitization process takes lesser time compared to sanitization of itemsets database or frequent itemset lattice.

5.1 Side Effects and Information Loss in Uncontrolled Randomized Perturbation Method

No side effects are generated using this method while perturbing sensitive itemsets or all itemsets in the dataset. The frequent itemsets and association rules present in the original data are retained in the sanitized data except that their support and confidence values are distorted.

The information loss due to distortion of few sensitive frequent itemsets is limited to deviation in their support from original value in the sanitized model. The loss is negligible when the information content of the whole database is considered. The amount of information loss depends on the Safe Range of that particular sensitive itemset.

In case of distorting the whole database, i.e., all frequent itemsets, the information loss and the maximum error is minimal in highly dense data but very large for sparse datasets. Table 5 gives the details of information loss, and the maximum error occurred due to uncontrolled randomized perturbation of different datasets. The missed itemsets are those which doesn't get sufficient safe range to vary their support. These are the itemsets which have their superset and the subset support vary by just one support count.

5.2 Side Effects and Information Loss in Controlled Randomized Perturbation Method

There are no side effects in distorting limited sensitive itemsets or whole database using Controlled Randomized Perturbation method. The frequent itemsets and association rules available in the original data is retained in the sanitized data except that their support and confidence values are distorted within acceptable error range defined during sanitization.

Table 4: Reduction of Itemsets for Sanitization using Closed Itemsets for Different Minimum Support Threshold

Dataset	Minimum Support	Number of Closed Itemsets	Number of Frequent Itemsets	Reduction in number of Itemsets for Sanitization
connect	0.7	35875	4129839	11412 %
	0.75	24346	1585551	6413 %
	0.8	15107	533975	3435 %
	0.85	8252	142127	1622 %
	0.9	3486	27127	678 %
pumsb	0.75	101047	672390	565 %
	0.78	53417	272336	410 %
	0.81	25871	98194	280 %
	0.84	11442	30482	166 %
	0.87	4508	9262	105 %
chess	0.7	23892	48731	104 %
	0.75	11525	20993	82 %
	0.8	5083	8227	62 %
	0.85	1885	2669	42 %
	0.9	498	622	25 %
accidents	0.45	16123	16123	0 %
	0.5	8057	8057	0 %
	0.55	4051	4051	0 %
	0.6	2074	2074	0 %
	0.65	1093	1093	0 %
mushroom	0.04	16565	4360341	26223 %
	0.06	10217	1454179	14133 %
	0.08	6749	658107	9651 %
	0.1	4885	574431	11659 %
	0.12	3586	127053	3443 %
T10I4D100K	0.001	26806	27532	2.7 %
	0.002	13107	13255	1.2 %
	0.003	4509	4552	0.95 %
	0.004	1992	2001	0.45 %
	0.005	1073	1073	0 %
T40I10D100K	0.015	6539	6539	0 %
	0.02	2293	2293	0 %
	0.025	1221	1221	0 %
	0.03	793	793	0 %
	0.035	567	567	0 %
retail	0.0005	19114	19242	0.67 %
	0.001	7572	7589	0.22 %
	0.0015	4233	4237	0.09 %
	0.002	2691	2691	0 %
	0.0025	1882	1882	0 %

The information loss due to distortion of few sensitive frequent itemsets is negligible. It is limited to deviation in their support in the sanitized model, i.e., closed itemsets within the safe range or acceptable error range whichever is lesser. There is no additional, unexpected information loss as there are no side effects. The information loss due to sanitization of the whole database using Controlled Randomized Perturbation Method is given in Table 6. It can be noticed that the actual maximum error in all sanitized database is less the acceptable percentage error threshold input. The information loss is also minimal which increases the utility of the sanitized database.

Table 5: Information Loss due to Uncontrolled Randomized Perturbation of the whole database

Dataset (Minimum Support)	Information Loss (%)	Max Error (%)	Itemsets Missed	Perturbation Time (Sec)
Connect (0.8)	0.54	14.9	10	16.25
Pumsb (0.81)	0.89	16.6	32	50.15
Chess (0.8)	0.74	15.8	119	1.35
Accidents (0.55)	3.91	47.1	3	0.78
Mushroom (0.08)	30.36	1417.8	4	3
T10I4D100K (0.003)	3230	15429	12	1
T40I10D100K (0.025)	768	2163	0	0.9
Retail (0.0015)	4720	21652	0	0.5

From the information presented in Table 4 to 6, it can be concluded that the proposed solution of using Closed Itemsets for sanitization instead of the database allows for simple and faster sanitization techniques. The algorithms presented are more efficient for sanitizing sparse database with no side effects and limited information loss. Also, the relation between the itemsets are kept intact, and only their magnitude is varied. These features increase the utility of the data in the sanitized database compared to other perturbation techniques described in the literature.

6. CONCLUSION

In this paper, a controlled randomized perturbation of closed itemsets for preserving privacy of sensitive itemsets is proposed. This work is an improvement of our previous work on uncontrolled perturbation of closed itemsets which suffer from a spurt of large errors introduced and huge information loss. The novelty of the proposed method is in limiting the randomized distortion of the support information of sensitive closed itemsets within a specified threshold which is within the safe perturbation range. Experimental results show that the proposed method is effective in limiting the errors introduced for perturbation, avoiding side effects, minimizing information loss and preserve original information from disclosing. The utility of the data after sanitizing using CRP is more compared to URP due to the decrease in error and information loss. The future work includes exploring Closed Itemset as the model for sanitizing database for other privacy-preserving data mining tasks such as anonymization, classification, and clustering. Also, improve the sanitization algorithms towards satisfying personalized privacy requirements.

Table 6: Information Loss due to Controlled Randomized Perturbation of the whole database

Dataset (Minimum Support)	No. of Closed Itemsets	[Input] Maximum Acceptable Error Threshold (%)	Information Loss (%)	[Output] Actual Maximum Error (%)	No. Of Itemsets Missed	Perturbation Time
Connect (0.8)	15107	5	0.598	4.96	0	17.1 sec
		10	0.611	9.78	0	
		15	0.603	8.38	0	
		20	0.608	8.79	0	
		25	0.608	8.79	0	
		50	0.608	8.79	0	
		100	0.608	8.79	0	
Pumsb (0.81)	25871	5	0.873	4.93	0	46.15 sec
		10	0.938	9.94	0	
		15	0.971	14.67	0	
		20	1.046	19.47	0	
		25	1.089	23.80	0	
		50	1.11	41.97	0	
		100	1.149	92.28	0	
Chess (0.8)	5083	5	0.439	4.62	340	1 sec
		10	0.616	9.45	280	
		15	0.639	14.28	292	
		20	0.642	14.87	288	
		25	0.637	14.87	287	
		50	0.637	14.87	287	
		100	0.637	14.87	287	
Accidents (0.55)	4051	20	2.15	19.91	0	0.82 sec
		40	3.11	38.88	0	
		60	3.52	58.08	0	
		80	3.49	52.02	0	
		100	3.49	52.02	0	
Mushroom (0.08)	6749	20	6.43	19.73	3	2.3 sec
		40	11.38	39.57	1	
		60	14.58	59.76	2	
		80	19.03	79.49	1	
		100	20.78	99.60	1	
T10I4D100K (0.003)	4509	20	7.5	19.76	10	0.8 sec
		40	14.4	39.53	11	
		60	22.4	59.65	9	
		80	28.5	79.55	10	
		100	36.9	99.73	15	
T40I10D100K (0.025)	1221	20	10	19.94	0	0.08 sec
		40	21	39.77	0	
		60	31	59.84	0	
		80	40	79.56	0	
		100	50	99.89	0	
Retail (0.0015)	4233	20	8	19.71	1	0.6 sec
		40	15	39.41	1	
		60	24	59.40	1	
		80	33	79.36	2	
		100	41	98.86	0	

References

[1] Charu C. Aggarwal and Philip S. Yu, "Privacy-Preserving Data Mining: Models and Algorithms", New York: Springer, 2008.

[2] Agrawal. R and Srikant. R, "Privacy-preserving data mining". In proceedings of the 2000 ACM-SIGMOD International Conference on Management of Data (SIGMOD 2000), pp 439–450. Dallas, Texas, USA.

[3] Atallah M, Bertino E, Elmagarmid A, Ibrahim M and Verykios V, "Disclosure limitation of sensitive rules". In proceedings of the Knowledge and Data Engineering Exchange (KDEX'99), pp 45–52. Chicago, IL, USA, 1999.

[4] Surendra H., Mohan H.S, "Distortion-Based Privacy-Preserved Association Rules Mining Without Side Effects Using Closed Itemsets". In: Abraham A., Dutta P., Mandal J., Bhattacharya A., Dutta S. (eds) Emerging Technologies in Data Mining and Information Security. Advances in Intelligent Systems and Computing, vol 813. pp 591-601, Springer, Singapore, 2019.

[5] S. Rizvi and J. Harita, "Maintaining data privacy in association rule mining". In proceedings of 28th International Conference on Very Large Databases (VLDB), 2002.

[6] Agrawal S., Krishnan V., Haritsa J.R, "On Addressing Efficiency Concerns in Privacy-Preserving Mining". In: Lee Y., Li J., Whang KY., Lee D. (eds) Database Systems for Advanced Applications. DASFAA 2004. Lecture Notes in Computer Science, vol 2973. Springer, Berlin, Heidelberg

[7] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim and V. S. Verykios, "Disclosure limitation of sensitive rules". In proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), pp 45-52. Chicago, IL, USA.

[8] Dasseni E., Verykios V.S., Elmagarmid A.K., Bertino E, "Hiding Association Rules by Using Confidence and Support". In: Moskowitz I.S. (eds) Information Hiding. IH 2001. Lecture Notes in Computer Science, vol 2137. Springer, Berlin, Heidelberg.

[9] Y.-H. Wu, C.-M. Chiang, and A.L.P. Chen, "Hiding Sensitive association rules with limited side effects". IEEE Transactions on Knowledge and Data Engineering, pp 29-42, 2007.

[10] V.S. Verykios, A. K. Emagarmid, E. Bertion, Y. Saygin and E. Dasseni, "Association rule hiding". IEEE Transactions on Knowledge and Data Engineering, 16(4), pp. 434-447, 2004.

[11] S.R.M Oliveira and O.R. Zaiane, "Privacy preserving frequent itemset mining". In proceedings of the 2002 IEEE International Conference on Privacy, Security and Data Mining (CRPITS 2002), pp 43-54.

[12] S.R.M Oliveira and O.R. Zaiane, "Protecting sensitive knowledge by data sanitization". In proceedings of the third IEEE International Conference on Data Mining (ICDM 2003), pp 211-218.

[13] Lin CW., Hong TP., Hsu HC, "Hiding Sensitive Itemsets with Minimal Side Effects in Privacy Preserving Data Mining". In: Pan JS., Snasel V., Corchado E., Abraham A., Wang SL. (eds) Intelligent Data analysis and its Applications, Volume I. Advances in Intelligent Systems and Computing, vol 297. Springer, Cham, 2014.

[14] Gregory Caiola and Jerome P. Reiter, "Random Forests for Generating Partially Synthetic, Categorical Data". Transaction Data Privacy, pp. 27-42, 2010.

[15] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava and Xiaokui Xiao, "PrivBayes: private data release via Bayesian Networks". In proceedings of ACM SIGMOD International Conference on Management of Data, pp. 1423-1434, 2014.

[16] Tzung-Pei Hong, Chun-Wei Lin, Kuo-Tung Yang and Shyue-Liang Wang, "Using TF-IDF to hide sensitive itemsets". Applied Intelligence, Springer, vol. 38, issue 4, pp 502 510, 2013.

[17] Chirag N. Modi, Udai Pratap Rao and Dhiren R. Patel, "Maintaining Privacy and Data Quality in Privacy Preserving Association Rule Mining", In proceedings of IEEE International Conference on Computing Communication and Network Technologies, 2010.

[18] Hai Quoc Le, Somjit Arch-int, Huy Xuan Nguyen and Ngamnij Arch-int, "Association rule hiding in risk management for retail supply chain collaboration", Computers in Industry, Elsevier, vol. 64, issue 7, pp 776-784, 2013

[19] Amirhosain Shahsavari and Shahram Hosseinzadeh, "CISA and FISA: Efficient Algorithms for Hiding Association Rules Based on Consequent and Full Item Sensitivities". In proceedings of IEEE International Symposium on Telecommunications, pp 977-982, 2014.

[20] Peng Cheng, Jeng-Shyang Pan and Chun-Wei Lin Harbin, "Use EMO to Protect Sensitive Knowledge in Association Rule Mining by Removing Items". In proceedings of IEEE Congress on Evolutionary Computing, pp 1108-1115, 2014.

[21] Janakiramaiah Bonam, A. Ramamohan Reddy and G. Kalyani, "Privacy Preserving in Association Rule Mining by Data Distortion Using PSO, ICT and Critical Infrastructure". In proceedings of the 48th Annual Convention of Computer Society of India - vol II, Advances in Intelligent Systems and Computing, vol 249, pp 551-558, 2014.

[22] Alaa K. Jumaah, Sufyan Al-Janabi and Nazar Abedlqader Ali, "An Enhanced Algorithm for Hiding Sensitive Association Rules Based on ISL and DSR Algorithms", International Journal of Computing and Network Technology, teach 3, no. 3, pp 83-89, 2015.

[23] Peng Cheng, Ivan Lee, Li, Kuo-Kun Tseng and Jeng-Shyang Pan, "BRBA: A Blocking Based Association Rule Hiding Method". In proceedings of the 13th AAAI Conference on Artificial Intelligence, ACM, pp 4200-4201, 2016.

[24] Yu-Chuan Tsai, Shyue-Liang Wang, Cheng-Yu Song and I-Hsien Ting, "Privacy and Utility of k-anonymity on Association Rule Hiding". In proceedings of the 3rd Multi-disciplinary International Social Networks

Conference on Social Informatics, Data Science, Article no. 42, pp 42:1-42:6, 2016.

[25] Kagklis V, Verykios VS, Tzimas G, Tsakalidis AK, "An integer linear programming scheme to sanitize sensitive frequent itemsets", In proceedings of IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2014). Pp 771-775.

[26] Frequent Itemset Mining Implementations Repository (2018), [Online]. Available: http://fimi.ua.ac.be/data/

[27] Mohammed J. Zaki and Ching-Jui Hsiao, "CHARM: An Efficient Algorithm for Closed Itemset Mining". In proceedings of International Conference on Data Mining, pp. 457-473, 2002.

Authors Profile

Mr. Surendra H received the B.E degree in electronics and communications engineering, and M.Tech degree in computer science and engineering from Visveswaraya Technological University, Belgaum, India in the year 2004 and 2013 respectively. He is currently pursuing Ph.D. in the Department of Information Science and Engineering of SJB Institute of Technology, Bengaluru, India. He was a software engineer with Ingersoll Rand Engineering and Technology Center, Bengaluru. His interests are data science, big data, and information privacy.

Dr. Mohan H S received the Bachelor's degree in computer science and engineering from Malnad College of Engineering, Hassan, India in the year 1999, M.Tech in computer science and engineering from Jawaharlal Nehru National College of Engineering, Shimoga, India in the year 2004 and Ph. D. in computer science & engineering from Dr. MGR University, Chennai, India. He is working as a Professor and Head in the Department of Information Science and Engineering at SJB Institute of Technology, Bengaluru, India. He is having a total of 19 years of teaching experience. His area of interests is Networks Security, Image processing, Data Structures, Computer Graphics, Finite Automata, and Formal Languages and Compiler Design. He has obtained the Best Teacher award for his teaching during the year 2008 at SJB Institute of Technology, Bengaluru, India. He has published and presented papers in journals, international and national conferences.