A Sensitivity-adaptive ρ-uncertainty Model for Set-valued Data

Liuhua Chen, Shenghai Zhong, Li-e Wang, and Xianxian Li(✉)
{liuhuachengxnu@sina, wuhanzsh@gmail}.com,{wanglie, lixx}@gxnu.edu.cn

Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin 541004, China

Abstract. Set-valued data brings enormous opportunities to data mining tasks for various purposes. Many anonymous methods for set-valued data have been proposed to effectively protect individuals privacy against identify linkable attacks and item linkage attacks. In these methods, sensitive items are protected by a privacy threshold to limit the re-identified probability of sensitive items. However, lots of set-valued data have diverse sensitivity on data items. Then it leads to the over protection problem that these existing privacy-preserving methods are applied to process the data items with diverse sensitivity, and it reduces the utility of data. In this paper, we propose a sensitivity-adaptive ρ-uncertainty model to prevent over-generalization and over-suppression by using adaptive privacy thresholds. Thresholds, which accurately capture the hidden privacy features of the set-valued dataset, are defined by uneven distribution of different sensitive items. Under the model, we develop a fine-grained privacy preserving technique through Local Generalization and Partial Suppression, which optimizes a balance between privacy protection and data utility. Experiments show that our method effectively improves the utility of anonymous data.
Keywords: Set-valued Data; Anonymization; Privacy Preserving; Generalization and Suppression

1 Introduction

With the rapid development of information technology, the Internet produced a sea of data, such as web search query logs [1], electronic health records [2] and set-valued data [3–5], which can service for behavior prediction, commodity recommendation and information retrieval.

Set-valued data, where a set of sensitive and non-sensitive items are relevant to an individual, contains detailed individual information. For example, a transaction dataset is shown in table 1, where items, a, b, c and d, are non-sensitive and α, β are sensitive. Adversary *Hebe* knows his neighbor *Alice* bought item b and c, so that he is easy to infer that *Alice* has purchased sensitive items α and β. In a word, publishing non-mask dataset will reveal privacy of individuals. Thus, we have to sanitize the data for guaranteeing privacy before data publishing.

In order to resist against item linkage attack, the ρ-uncertainty model has been proposed in [6]. When an adversary knows any non-sensitive or sensitive

item, this model protect privacy information via ensuring that the probability of sensitive items inferred by known item set q is less than ρ.

The traditional approaches [6, 7] of ρ-uncertainty have a few drawbacks as follows:

1. "one size fits all" approach ignores the reality that different items have different data distributions.
2. Global suppression completely removes some items directly and incurs the loss of sensitive rules. Therefore, those removed items cannot be used by researchers.
3. Global generalization [8] brings a huge number of pseudo-association rules, so that the utility for data mining has severe distortion.

In this work, we propose a sensitivity-adaptive ρ-uncertainty model to address an important limitation of original ρ-uncertainty—that it provides only a uniform level of privacy protection for all items in a dataset. We propose a solution, LGPS, integrating local generation and partial suppression, to mask items varying in frequency and sensitivity.

In our LGPS solution, if the frequency of occurrence of a sensitive item is low, the sensitivity of the item is high and vice versa. LGPS divides records into separate groups and makes each sub-grouping satisfying the proposed model.

Table 1. Oiginal Data

Name	Id	Items
Chris	t1	a, d, α, β
Bob	t2	a, b, d
Alice	t3	b, c, α, β
Mary	t4	a, c, α
Dan	t5	a, b, α
Lucy	t6	a, α

Example 1. *Anonymization on set-valued datasets*

For original data as table 1, anonymous datasets, showing in table 2(a), table 2(b) and table 2(c), are masked by using original approach of TDControl(ρ=0.5) and our solution. In table 2(b), the uniform threshold is 0.5. And in table 2(c), according to data distribution of original dataset and formula 2 and 3 the thresholds ρ_α and ρ_β for privacy preserving are 0.7 and 0.3 when ρ, defined by users, is 0.3. In uneven datasets, flexible thresholds provide personal protection and retain more data utility. Comparing table 2(a) with table 2(b), the information loss of table 2(b) is less than table 2(a).

For each sensitive rule, such as $a \rightarrow \alpha$ or $\beta \rightarrow \alpha$, TDControl use global suppression and global generalization to hide sensitive rules. In the suppression process,

Table 2. Anonymization Dataset

(a) TDC(ρ=0.5)

Id	Items
t1	a,d
t2	a,b,d
t3	b,c
t4	a,c
t5	a,b
t6	a

(b) LGPS($\rho = 0.5$)

Id	Items
Gr1	a,b
	a,α
Gr2	a,d,β
	a,b,d
	b,β
	a,α

(c) LGPS(ρ_α=0.7,ρ_β=0.3)

Id	Items
Gr1	a,b
	a,α
Gr2	a,B,β
	a,b,B
	b,B,α
	a,B,α

the chosen item, which information loss of being suppressed is minimal, is removed in all records. At last, all sensitive items α and β are suppressed in table 2(a). Anonymous dataset masked by TDControl loses a lot of information and does not preserve enough utility for secondary analysis.

Differing from prior works [6, 7, 9, 10], our contributions can be summarized as follows:

1. We focus on diverse sensitivity on items and propose a sensitivity-adaptive ρ-uncertainty model for improving data utility.
2. Under the proposed model, we use the frequency of sensitive items to define flexible privacy thresholds and propose a sensitivity-adaptive ρ-uncertainty approach decreasing information loss. Flexible privacy threshold is non-trivial because it is adaptive to the distribution of data and it addresses the over-protection problem incurred by the unified threshold in the TDControl model.
3. Furthermore, we devise an effective algorithm called LGPS by using local generation and partial suppression to achieve anonymization. Not only does LGPS reserve some useful characteristics of those sensitive items, but also it introduces pseudo rules less. Experiments running on real datasets show that our approach is effective and information loss is lower than the implemented methods.

The rest of this paper is organized as follows. Section 2 describes related work on anonymization of set-valued dataset. Privacy model is introduced in section 3. The algorithm is shown in section 4. The result of experiments is demonstrated in section 5. Section 6 concludes this paper.

2 Related Work

Sweeney [11] raised the k-anonymous model in 2002, which aims to make each QID(quasi-identifier) including at least k matched records in sanitized dataset, so this model can effectively prevent identity linkage attacks. Yet, k-anonymity is insufficient to protect the privacy of set-valued data which is high-dimensional and sparse. So k^m-anonymity model, which makes an adversary at most know

m items, was proposed in [12]. Unfortunately, k-anonymity and k^m-anonymity are not able to prevent item linkage attacks. In 2007, Machanavajjhala [13] et al. proposed l-diversity model based on k-anonymity, which makes every sensitive attributes in each equivalence having at least l different attribute values, to prevent attribute linkage attacks.

Based on k-anonymous model, Sinhong [14] et al. proposed a new solution to anonymize set-valued data. This anonymous solution constructs a pseudo taxonomy tree based on utility metrics to instead the presetting taxonomy tree, so it can upgrade data utility and provide protection of individuals information. Yet, this solution is inadequate to prevent item linkage attacks.

Wang [15] et al. described high-dimensional sparse data using bipartite graphs with individuals attributes, and then the original graph is turned into an anonymous graph by clustering attributes in each node. Chen et al.[16] and Xiao et al.[17]propose two approaches, each of which satisfies differential privacy model, to protect high-dimensional datasets. The aim of the two approaches is mainly protecting individuals privacy information by adding noise. But these approaches will disclose privacy under item linkage attacks, if a small noise is added.

(h, k, p)-coherence ensures that any combination, in which at most h percent of records contain some sensitive items, has at least k records including p item. The work in [18] shows that the optimal solution of (h, k, p)-coherence is an NP-hard problem and gives a local optimization algorithm. However, the (h, k, p)-coherence criterion is insufficient to prevent an attacker who knows sensitive items of individuals.

PS-rule model proposed in [19], where P is a set of non-sensitive items and S is a set of sensitive items in dataset D, can simultaneously prevent identify linkable attacks and item linkage attacks. Given two item sets $I \in P$ and $J \in S$, the rule $I \to J$ is a PS-rule when $sup(I)$ is k or more and $conf(I \to J)$ is no more than c.

The ρ-uncertainty, a more sophisticated model to preserve sensitive information in set-valued data, was demonstrated in [6]. This criterion ensures that the probability of sensitive items inferred by an attacker is less than ρ and does not restrict background knowledge of attackers.

ρ-uncertainty approach based on partial suppression, presented in [7], can be adapted to either statistical analysis or association rules mining. However, this method does not take into account the difference of the sensitivity between sensitive items, in that overprotection of some low sensitivity items increase information loss. The first research result, published in [20], concerns the difference of sensitivity and proposes a solution to rank sensitive items by difference of items sensitivity.

3 Privacy Model

3.1 Privacy Concept

Item linkage attack refers that the probability, which a sensitive item associated to an individual is inferred by adversaries, is high. For example, an adversary

Table 3. Definition of Symbols

Symbol	Definition
D	The original dataset
$D^{'}$	Anonymous dataset
q	QID, a sub-group of items
$sup(q)$	Support of set q
e	A sensitive item
$F(e)$	Frequency of e
ρ_e	Privacy threshold of sensitive item e
$conf(q \to e)$	Confidence of the rule $q \to e$

knows an individual in Table 1 have bought non-sensitive commodities, b and c $(b, c \in q)$, and he/she can also infer that sensitive items α and β were purchased by the person. Therefore, users privacy is undermined while released data has not been masked.

Definition 1 (Sensitive association rule). Given a sensitive item e, the association rule $q \to e$ is a sensitive association rule.
Confidence of a rule [21]: if a rule $q \to e$, where e is sensitive and q is non sensitive, the confidence of this rule is a conditional probability and defined as follow:

$$conf(q \to e) = \frac{sup(q \cup e)}{sup(q)} \tag{1}$$

Here $sup(q \cup e)$ is the number of records including q and e, $sup(q)$ is the number of records including item set q.
ρ-uncertainty: If the confidence of any sensitive association rule in released dataset is less than ρ, the dataset achieves ρ-uncertainty ($\rho > 0$).

A unique threshold, for preventing item linkage attacks, is inappropriate since the distribution of different sensitive items is uneven. Furthermore, the view is natural that a sensitive item involved in most of records has a low sensitivity. And the item has a high sensitivity when it only appears in a handful of records. So we use diverse thresholds correlated with items sensitivity to mask data.

Definition 2 (Adaptive parameter δ_e). δ_e is an adaptive parameter to adjust the threshold ρ and is defined as follow:

$$\mathrm{F}(e) = sup(e)/\,|N| \tag{2}$$

$$\delta_e = \varepsilon_e \cdot \mathrm{F}(e) \tag{3}$$

Here e denotes a sensitive item, $sup(e)$ is the number of records including item e, and $|N|$ is the number of all records in dataset. $\varepsilon_e(\varepsilon_e > 0)$ is called a sensitive factor of e, and used to tune the value of δ_e combining with $F(e)$. If $F(e)$ is higher, a larger ε_e is chosen to prevent privacy disclosure. ε_e is adjusted by sensitivity of item e and user requirement of the protection strength.

Definition 3 (sensitivity-adaptive ρ-uncertainty). Let δ_e be an adaptive parameter as defined above, and $\rho_e = \rho + \delta_e(1 > \rho_e > 0)$. If the confidence of any

sensitive association rule in anonymous dataset is less than ρ_e, then the dataset achieves sensitivity-adaptive ρ-uncertainty.

Here ρ is the minimum privacy threshold, and ρ_e is the flexible threshold depended on the sensitivity of e and the distribution of e in the dataset.

Example 2. *Definition of flexible privacy thresholds*

Assuming that a dataset has 10 records, ρ=0.3, where 8 records contain item α, 6 records contain item β, and 2 records contain item Ω, the flexible threshold of ρ_α, ρ_β, ρ_Ω was defined as 0.7, 0.5, 0.3 by the frequency of occurrence as defined in Formula 2 and 3.

3.2 Information Loss Metric

Evaluation of the effectiveness is important for any anonymous algorithm. In this paper, Normalized Certainty Penalty [13, 22], an information loss metric for items in a generalization hierarchy, is used to evaluate the effectiveness of our algorithm. As shown in Fig. 1, a and b can be masked by A while c and d can be replaced with B. The root node, ALL, can represent each item in the dataset.

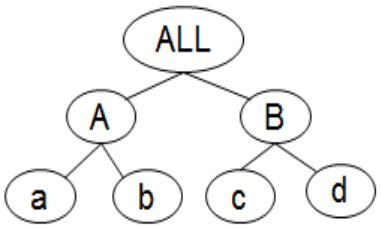


Fig. 1. Item Generalization Hierarchy H

NCP, a popular measurement of information loss for item generalization, is defined by in [22]. In our sensitivity-adaptive ρ-uncertainty model, NCP is redefined as below.

$$NCP_{(a)} = \begin{cases} 1 & \textit{if } a \textit{ is suppressed} \\ \frac{|u_m|}{|I|} & \textit{if } a \textit{ is generalized to node } m \in H \end{cases} \tag{4}$$

where a represents child node of m, $|u_m|$ is the total number of leaf nodes connected with m, and $|I|$ is the total number of non-sensitive items.

Example 3. *Information loss of items generalization*

When item a is generalized to A and c is replaced with B, the information loss is described as $IL_{(a)} = \frac{|u_A|}{|I|} = \frac{2}{4} = \frac{1}{2}, IL_{(c)} = \frac{|u_B|}{|I|} = \frac{2}{4} = \frac{1}{2}$.

If an item is suppressed such as in [6, 7], then the information loss is 1.

For a record t in dataset D and an item m in t, let C_t be the number of records in dataset. Then the information loss for the dataset is defined as

$$NCP_{(D)} = \frac{\sum_{t \in D} \sum_{m \in t} NCP_{(m)}}{\sum_{t \in D} C_t} \tag{5}$$

4 Anonymous Algorithm

Usually, there are many anonymous methods applied to mask datasets, such as kd-trees, SRT and AT[6], R-tree [23, 24]. In each top-down recursion, items are specialized in the hierarchy tree and records are assigned to different sub-groups. If the sub-group does not meet the sensitivity-adaptive ρ-uncertainty model, partial suppression was used to mask the group. Our anonymous method applies two algorithms in the whole anonymous process, so that information loss of anonymous data set is minimized.

LGPS: The LGPS first define the privacy constraints of sensitive items by calculating the frequency of sensitive items. Then, all non-sensitive items are initialized to *ALL* and information loss of the generalized item is 1. In next step, the algorithm checks validity of the generalized dataset. While the generalized dataset does not achieve sensitivity-adaptive ρ-uncertainty, partial suppression is used to mask this dataset. The algorithm terminates until every sub-group of records have been processed.

Algorithm 1 *LGPS(D)*

1: PrivacyThreshold(D);
2: **for** each $t \in D$ **do**
3: Initialized all no-sensitive item to ALL;
4: **end for**
5: PartialSuppressor(D,$(\rho_1, \rho_2, \ldots, \rho_h)$);
6: resultParts←Flexible_ρ_Uncertainty(D,H,$(\rho_1, \rho_2, \ldots, \rho_h)$);
7: **for** each subParts in resultParts **do**
8: specialData(subparts,$(\rho_1, \rho_2, \ldots, \rho_h)$;
9: **end for**

The Privacy Constraints define privacy thresholds of various sensitive items. In this part, the algorithm first calculates the support of sensitive items by traversing the entire dataset. In the traversing process, the support of the item is constantly updated while an item occurs again. Finally, the different ρ are defined according to the Formula 2 and 3.

Flexible_ρ_Uncertainty Anonymity: The FUA, which masks, in every top-down recursion, set-valued data by local generalization and partial suppression, is described as follow:

Line(1): The generalized item is added to set G.

Line(2-3): If G is empty and D is impossible to split down further, the D is

Algorithm 2 *PrivacyThreshold(D)*

initialize the sup of all sensitive item to 0;
for each $t \in D$ **do**
  update the sup of sensitive item;
  $n = n + 1$;
end for
let $F(e)$ be a set of sensitive item frequency;
for each sensitive item e **do**
  $F(e) = \frac{sup(e)}{n}$;
  according $\rho_e = \rho + \varepsilon_e \cdot F(e)$ define compute value ρ_e;
end for

released.

Line(5-8): Such as [22], the splitNode is replaced by its child node in generalized hierarchy tree. Then, according to different values of splitNodes child node, D is divided and various records are registered to disjoint subsets.

Line(9-12): For each subsets, some items are partial suppressed to hide sensitive rules by flexible threshold ρ_e, and some generalized items are instead in top-down recursion until G is empty or itself is impossible to further split down.

Algorithm 3 *Flexible_ρ_Uncertainty(D,H,($\rho_1, \rho_2, \ldots, \rho_h$))*

add generalized item in D to a set G;
if no further split down possible for D **then**
  Return and push D;
else
  splitNode←PickNode(D,G);
  for each data in D **do**
    add subParts in $D \leftarrow$ divideData(D,splitNode);
  end for
  for each subParts in D **do**
    PartialSuppressor(subParts,($\rho_1, \rho_2, \ldots, \rho_h$));
    Flexible_ρ_Uncertainty(subParts,H,($\rho_1, \rho_2, \ldots, \rho_h$));
  end for
end if

Partial Suppressor method[7]: In each subset, any rule whose confidence is more than its threshold ρ_Y, such as $conf(X \to Y) > \rho_Y$, is added to SR. Getting SR by flexible threshold is more suitable and effective for hiding sensitive rules. Until SR is empty, some item is masked by partial suppression. For selecting sensitive rule $X \to Y$, if any item appears in $X \cup Y$, until the confidence of this rule is less than threshold, each of them will be suppressed in records that have this rule. After hiding the selected sensitive rules, SR has to update for next suppression procession.

Table 4. A Transaction Dataset

(a) $(a, b, c, d) \rightarrow ALL$

Id	Items
t1	ALL, α, β
t2	ALL
t3	ALL, α, β
t4	ALL, α
t5	ALL, α
t6	ALL, α

(b) $ALL \rightarrow (A, B)$

Id	Items
Group1	A
	A,α
Group2	A,B,β
	A,B
	A,B,α
	A,B,α

(c) In group21 $A \rightarrow (a, b)$

Id	Items
$Group1'$	a,b
	a,α
Group21	a,B,β
	a,b,B
	b,B,α
	a,B,α

Example 4. *A top-down partitioning using LGPS*

First, uniform privacy threshold ρ be set to 0.3, In table 1, items, α and β, are sensitive, while all others are non-sensitive. In the process of finding flexible threshold, the ρ_α , ρ_β are defined as 0.7, 0.3, while the generalization hierarchy tree for non-sensitive items is constructed as Fig. 1. According to the generalization hierarchy tree, all non-sensitive item is initialized to top value ALL. And the dataset is masked as table 4(a). Due to flexible thresholds and confidence of sensitive rules, the SR is set as $\{ALL \rightarrow \alpha, ALL \rightarrow \beta, (ALL, \alpha) \rightarrow \beta, (ALL, \beta) \rightarrow \alpha\}$. While rules in SR disclose information to adversaries, the α in *t1* and the β in *t3* is removed to achieve sensitivity-adaptive ρ-uncertainty model. After this suppression, the anonymous dataset is described as $\{(ALL, \beta), ALL, (ALL, \alpha),($ ALL,$\alpha), (ALL, \alpha), (ALL, \alpha)\}$ and the *NCP(D)* is 14.

Anonymous data, in this example, is overgeneralization, and the data utility is insufficient for users. In addition, Candidate set of splitNode such as A and B is not empty. Therefore, item ALL is substituted with $\{A\}$, $\{B\}$, $\{A, B\}$ and all records are divided into different subgroups. While each subgroup is valid for the procession of dividing groups, partial suppression is used to mask each subgroup. As a result subgroup $\{A\}$ is not meet sensitivity-adaptive ρ-uncertainty, item α in *t5* is removed. In contrast, any item in group $\{A, B\}$ have not to be removed. The $NCP(D)^{'}$ in table 4(b), is the sum of each subgroups' *NCP* is 9. Because *NCP(D)* is more than $NCP(D^{'})$, this step satisfy dividing condition.

For *Group1*, the item A in every record is instead by $\{a\}$, $\{b\}$, $\{a, b\}$. Record *t5* is assigned to *group*$\{a, b\}$, and another is added into the *group*$\{a\}$. The number of records is less than $1/\rho_\alpha$, so group $\{A\}$ is not valid. As a result $Group1^{'}$ information loss $NCP(Group1^{'})$ is less than *NCP(Group1)*. This specialization process is valid. In addition, records in $Group1^{'}$ have not generalized item and achieve sensitivity-adaptive ρ-uncertainty, so $Group1^{'}$ is released as $\{(a, b), (a, \alpha)\}$.

In *Group2*, generalized items, A and B, appear. According to A or B, the divided subgroups for all records is not valid. So item A and B are specialized by its child, and the record in this subgroup is described as *Group21* and *Group22*. Because *NCP(Group21)* is less than *NCP(Group22)*, the best specialization $A \rightarrow (a, b)$ is chosen to specialize records in *Group2*. Then records in this subgroup are released for users.

When this specialization is legal and *NCP(Group21)* is less than *NCP(Group2)*, this subgroup is described as $\{(a, B, \alpha), (a, b, B), (a, B, \alpha), (a, B, \alpha)\}$. But a generalized item B exits in G, the specialization $B \rightarrow (c, d)$ is executed. Then 4 sensitive rules, $\{d \rightarrow \beta, (a, d) \rightarrow \beta, c \rightarrow \alpha$ and $(a, c) \rightarrow \alpha\}$, are added to *SR*. For anonymizing this group, items d in *t1* and α in *t3* are suppressed. But the *NCP(Group211)* is greater than *NCP(Group21)*. Therefore, the specialization for item B is illegal and this group is released as $\{(a, B, \alpha), (a, b, B), (a, B, \alpha), (a, B, \alpha)\}$. All records in original dataset are published as Table 4(c).

5 Experimental Study

5.1 Dataset and parameters

Our experiments run on three real-world datasets introduced in [6, 7], *BMS-POS*, *BMS-WebView-1* and *BMS-WebView-2*. *BMS-POS* is a transaction log from several years of sales and an electronics retailer. And *BMS-WebView-1* and *BMS-WebView-2* are click-stream data from two e-commerce web sites. All of those are widely used as benchmark datasets in the knowledge discovery community. Information about the three datasets is listed in table 5.

Table 5. Characteristics of the three datasets

Datasets	#Trans	#Distinct items	# Max.trans.size	# Avg.trans.size
BMS-WV1	59,602	497	267	2.5
BMS-WV2	77,512	3340	161	5.0
BMS-POS	551,597	1657	164	6.5

Specifically, we measure execution time and data utility to compare our LGPS algorithms, with TDControl[6] and Dist[7]. All algorithms were implemented in *C++* and ran on an *Intel(R) Core(TM) i3-2100 cpu* machine with 4GB RAM running the Linux operating system.

5.2 Data Utility

We evaluate our algorithms with three performance factors: a) information loss(*KL*-divergence[7]), b)the difference of the number of mined rules, c)the size of datasets .

Kullback-Leibler divergence measures the distance between two probability distributions and determines the similarity between the original data and the anonymous data. When D is original data and $D^{'}$ is anonymous data, *KL*-divergence define as follows:

$$KL(D^{'} \parallel D) = \sum_{i} D(i) \log \frac{D(i)}{D^{'}(i)} \tag{6}$$

Where $D(i)$ and $D^{'}(i)$ is the occurrence probability of item i in original dataset D and the anonymized dataset $D^{'}$.

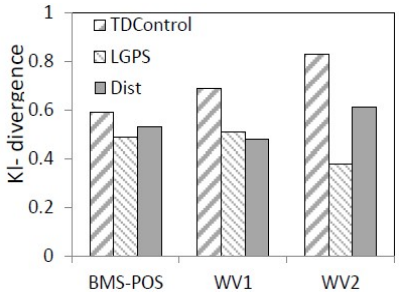
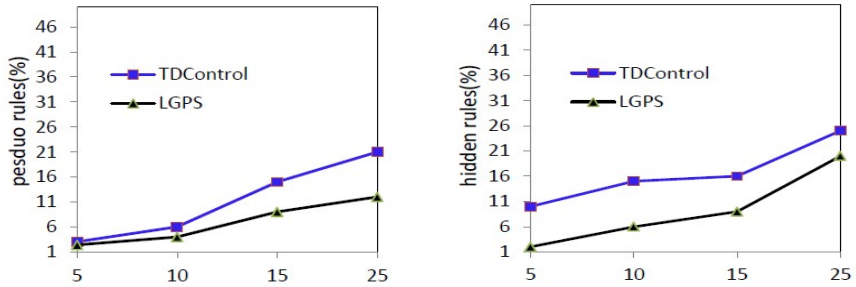


Fig. 2. varying datasets

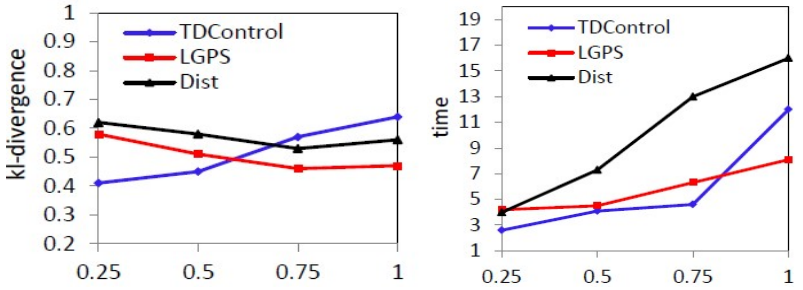
Due to the limitation of hardware, we have to make length less than or equal to 5 when we find sensitive association rules. In Fig. 2, TDControl and Dist are basic approaches to mask data by ρ-uncertainty model while ρ is the mean of all flexible ρ_e in our approach. Although, for the dataset *WV1*, information loss of LGPS is larger than information loss of original approaches, LGPS provides a stronger protection and gets an anonymous dataset whose data distribution is more similar to original dataset.



(a) Sensitive items in the domain(%) (b) Sensitive items in the domain(%)

Fig. 3. the difference of the number of mined rules

The smaller the difference of the number of mined rules between the anonymous dataset and the original dataset is, the higher the data utility is. So the number of the hidden rules and the pseudo rules is used to evaluate data utility. In Fig. 3(a) and Fig. 3(b), LGPS hides fewer rules and introduces fewer pseudo rules than TDControl when the proportion of sensitive items in original dataset is increasing.



(a) *KL*-divergence (b) Execution time on *BMS-POS*

Fig. 4. Varying the size of dataset

Fig. 4(a) shows the experiments results of *KL*-divergence and execution time. LGPS is less efficient than TDControl and Dist when the size of dataset is small. With the increased size of the dataset, LGPS becomes more effective. When we do experiments on the whole *BMS-POS*, as shown in Fig. 4(b), our approach may take more time than TDControl. However, mostly anonymous process applied in data publish is offline, so it is reasonable that we spend more time and get more utility of anonymous data.

6 Conclusion

We proposed a sensitivity-adaptive ρ-uncertainty model to mask dataset when the distribution of sensitive items has great vary widely. Flexible threshold for different sensitive items is defined by the occurrence frequency of itself. In order to reduce information loss and increase utility of anonymous data, the partial suppression and local generalization are used in top-down recursion. Comparing to previous methods[6, 7], this approach only delete or generalize fewer items to satisfy the advanced privacy model.

In the future work, we focus on influence for privacy when adversary infers individuals information by the semantic of a combination of different items. And we pay more attention to the influence of background known by adversary. For example, when adversary knows an item is not containing in the record, most of implemented models are inadequate for protecting individuals information. Encountering various backgrounds known by adversary, we should find some new model and construct some new algorithm to conceal sensitive information.

Acknowledgments. The research is supported by the National Science Foundation of China (Nos. 61272535, 61363009, 61365009, 61502111), Guangxi Bagui Scholar Teams for Innovation and Research Project, Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing, Guangxi Natural Science Foundation (Nos. 2015GXNSFBA139246, 2013GXNSFBA019263, 2014GXNSFBA118288), Science and Technology Research Projects

of Guangxi Higher Education (Nos. 2013YB029, 2015YB032), the Guangxi Science Research and Technology Development Project (No.14124004-4-11) ,Youth Scientific Research Foundation of Guangxi Normal University and Innovation Project of Guangxi Graduate Education (No. YCSZ2015104).

References

1. Saygin, Y., Verykios, V.S., Elmagarmid, A.K.: Privacy preserving association rule mining. In: Research Issues in Data Engineering: Engineering E-Commerce/E-Business Systems, 2002. RIDE-2EC 2002. Proceedings. Twelfth International Workshop on, IEEE (2002) 151–158
2. Han, J., Luo, F., Lu, J., Peng, H.: Sloms: A privacy preserving data publishing method for multiple sensitive attributes microdata. Journal of Software **8**(12) (2013) 3096–3104
3. Xiao, X., Tao, Y.: M-invariance: towards privacy preserving re-publication of dynamic datasets. In: Proceedings of the 2007 ACM SIGMOD international conference on Management of data, ACM (2007) 689–700
4. Ghinita, G., Tao, Y., Kalnis, P.: On the anonymization of sparse high-dimensional data. In: Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on, IEEE (2008) 715–724
5. Liu, J.Q.: Publishing set-valued data against realistic adversaries. Journal of Computer Science and Technology **27**(1) (2012) 24–36
6. Cao, J., Karras, P., Raïssi, C., Tan, K.L.: ρ-uncertainty: inference-proof transaction anonymization. Proceedings of the VLDB Endowment **3**(1-2) (2010) 1033–1044
7. Jia, X., Pan, C., Xu, X., Zhu, K.Q., Lo, E.: ρ-uncertainty anonymization by partial suppression. In: Database Systems for Advanced Applications, Springer (2014) 188–202
8. Tripathy, B., Reddy, A.J., Manusha, G., Mohisin, G.: Improved algorithms for anonymization of set-valued data. In: Advances in Computing and Information Technology. Springer (2013) 581–594
9. Vaidya, J., Clifton, C.: Privacy preserving association rule mining in vertically partitioned data. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2002) 639–644
10. Fung, B., Wang, K., Yu, P.S.: Top-down specialization for information and privacy preservation. In: Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on, IEEE (2005) 205–216
11. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **10**(05) (2002) 557–570
12. Terrovitis, M., Mamoulis, N., Kalnis, P.: Privacy-preserving anonymization of set-valued data. Proceedings of the VLDB Endowment **1**(1) (2008) 115–125
13. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) **1**(1) (2007) 3
14. Lin, S., Liao, M., : Towards publishing set-valued data with high utility. (2014)
15. Wang, L.e., Li, X.: A clustering-based bipartite graph privacy-preserving approach for sharing high-dimensional data. International Journal of Software Engineering and Knowledge Engineering **24**(07) (2014) 1091–1111
16. Chen, R., Mohammed, N., Fung, B.C., Desai, B.C., Xiong, L.: Publishing set-valued data via differential privacy. Proceedings of the VLDB Endowment **4**(11) (2011) 1087–1098

17. Xiao, X.: Differentially private data release: Improving utility with wavelets and bayesian networks. In: Web Technologies and Applications. Springer (2014) 25–35
18. Xu, Y., Wang, K., Fu, A.W.C., Yu, P.S.: Anonymizing transaction databases for publication. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2008) 767–775
19. Loukides, G., Gkoulalas-Divanis, A., Shao, J.: Anonymizing transaction data to eliminate sensitive inferences. In: Database and Expert Systems Applications, Springer (2010) 400–415
20. Ye, Y., Liu, Y., Wang, C., Lv, D., Feng, J.: Decomposition: Privacy preservation for multiple sensitive attributes. In: Database Systems for Advanced Applications, Springer (2009) 486–490
21. Verykios, V.S., Elmagarmid, A.K., Bertino, E., Saygin, Y., Dasseni, E.: Association rule hiding. Knowledge and Data Engineering, IEEE Transactions on **16**(4) (2004) 434–447
22. He, Y., Naughton, J.F.: Anonymization of set-valued data via top-down, local generalization. Proceedings of the VLDB Endowment **2**(1) (2009) 934–945
23. Wang, S.L., Tsai, Y.C., Kao, H.Y., Hong, T.P.: On anonymizing transactions with sensitive items. Applied Intelligence **41**(4) (2014) 1043–1058
24. Gkoulalas-Divanis, A., Loukides, G.: Pcta: privacy-constrained clustering-based transaction data anonymization. In: Proceedings of the 4th International Workshop on Privacy and Anonymity in the Information Society, ACM (2011) 5