TRANSACTIONS ON DATA PRIVACY 11 (2018) 219–237

Personalized Anonymization for Set-Valued Data by Partial Suppression

Takuma Nakagawa[1,2], **Hiromi Arai**[3,4], **Hiroshi Nakagawa**[3]

[1]The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan

[2]NS Solutions Corporation, 27-1, Shinkawa 2-chome, Chuo-ku, Tokyo, 104-0033, Japan

[3]RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan

[4]JST PRESTO, Gobancho, Chiyoda-ku, Tokyo, 102-0076, Japan

E-mail: takuma.nakagawa0725@gmail.com

Received 28 February 2018; received in revised form 5 May 2018; accepted 2 July 2018

Abstract. Set-valued data is comprised of records that are sets of items, such as goods purchased by each individual. Methods of publishing and widely utilizing set-valued data while protecting personal information have been extensively studied in the field of privacy-preserving data publishing. Until now, basic models such as k-anonymity or k^m-anonymity could not cope with attribute inference by an adversary with background knowledge of the records. On the other hand, the ρ-uncertainty model makes it possible to prevent attribute inference with a confidence value above a certain level in set-valued data. However, even in that case, there is the problem that items to be protected have to be designated as common to everyone. In this research, we propose a new model that can provide more suitable privacy protection for each individual by protecting different items designated for each record distinctively and build a heuristic algorithm to achieve this guarantee using partial suppression. In addition, considering the problem that the computational complexity of the algorithm increases combinatorially with increasing data size, we introduce the concept of probabilistic relaxation of privacy guarantee. Finally, we show the experimental results of evaluating the performance of the algorithms using real-world datasets.

Keywords. privacy-preserving, anonymization, set-valued data, attribute inference, association rule

1 Introduction

In recent years, the importance of utilizing big personal data has been emphasized. Among various types of data, set-valued data is useful as a data format that can express various kinds of information associated with individuals in a simple way. Set-valued data is data in which each record is represented as a set of items, such as goods purchased by an individual. Set-valued data is used in a growing number of applications including marketing research, recommendation systems, and biomedical studies. To utilize such data widely in the fields of business or research, there is a demand to give data to third parties or release data to the public. However, publishing these data can reveal private information of a data subject and should not be carried out directly.

To cope with this situation, techniques for performing data processing that preserve the privacy of the data subject while maintaining the utility of the data are widely studied in the field of privacy-preserving data publishing. k-anonymity [10] is one of the most basic data anonymization models, which aims to prevent *identity disclosure* (linking an individual to the corresponding record in the published data) in relational databases. Since k-anonymity cannot be applied to set-valued data that cannot distinguish between quasi identifiers (QIDs) and sensitive attributes, Terrovitis et al. [11] proposed a new model, k^m-anonymity. Although these models can prevent identity disclosure, there is the problem that they cannot guarantee that the private information contained in the record itself will not be disclosed, i.e., the risk of *attribute inference* or *sensitive information disclosure* remains. Then, ρ-uncertainty [3] was proposed to prevent attribute inference in set-valued data, in a similar way to ℓ-diversity [9] for relational data. It is based on the premise that items can be divided into sensitive attributes and non-sensitive attributes. The ρ-uncertainty model guarantees that no association rule whose consequence contains a sensitive attribute can be inferred with a confidence higher than parameter ρ.

However, this is not realistic because the same division of items is applied to all personal records. As has been pointed out [12], the information as to whether each item is sensitive can be strongly dependent on each individual. For example, in purchase history data, one person may want to hide the purchase of medicine, and another person may wish to hide the purchase of some entertainment supplies. Regarding privacy awareness, some individuals may agree to release information on the majority of their items to enhance the effectiveness of the research, and other individuals may feel that they want to protect information on all of their items as sensitive attributes. Cao et al. [3] stated that we can cope with this problem by deeming all items considered as sensitive by at least one person to be sensitive for all individuals, but this leads to an unnecessarily large information loss. Until now, to the best of our knowledge, no concrete countermeasure has been found. The only way to protect set-valued data from attribute inference was to suppose that some items are formally considered as sensitive attributes and can be distinguished from the others.

To tackle this problem, in this research, we propose a new model, *personalized ρ-uncertainty* that can formally handle the situation where each individual designates several items to be protected as sensitive attributes. Under this model, each individual can freely designate items that he/she wants to protect, and the data processor can perform anonymization without unnecessarily large utility loss under the conditions of this model. This is particularly useful when handling data such as purchase history and web search query logs that data subjects can select sensitive items.

Let us consider anonymization algorithms. Methods for processing set-valued data are roughly divided into two categories: *generalization* [1, 3, 4, 6, 8, 11, 12] and *suppression* [3, 4, 7, 18]. When considering attribute inference in a situation in which sensitive items and non-sensitive items are mixed, generalization methods are not suitable because the fact that we generalize items can give clues to an adversary [3]. For example, suppose that Bob has a sensitive item 'viagra' in his record and we generalize 'viagra' and 'aspirin' together as 'medicine'. If an adversary identifies Bob's record and knows that Bob has 'medicine' which represents 'viagra' or 'aspirin', he/she can infer that there must be some reason for this generalization. The expected reason is that his record includes medicine that he wants to hide and must be 'viagra'. That is, the adversary can infer that Bob has 'viagra' with high confidence because if Bob actually has 'aspirin', there must be no need to generalize it with 'viagra'. This way of thinking is related to the concept of minimality attack [15]. Therefore, in this study, we used the suppression method. With the *global suppression* method [3, 18], when an item is deleted, we delete it from all the records. This method has been preferred

User	Contents
Alice	milk, bread, *medicine*
Bob	apple
Carol	*milk*, *coffee*, *bread*
Dave	milk, medicine
Ellen	coffee, bread, apple
Frank	orange, *medicine*

(a) Original data

User	Contents
u_1	~~*milk*~~, bread, *medicine*
u_2	apple
u_3	*milk*, *coffee*, ~~*bread*~~
u_4	milk, medicine
u_5	coffee, bread, apple
u_6	orange, ~~*medicine*~~

(b) Processed data

User	Contents
u_1	bread, medicine
u_2	apple
u_3	milk, coffee
u_4	milk, medicine
u_5	coffee, bread, apple
u_6	orange

(c) Published data

Table 1: Example of anonymization of set-valued data

because it guarantees the truthfulness of the published data. However, since all the information on the items to be suppressed is lost, it adversely affects the utility. In recent years, instead of global suppression, methods using *partial suppression*, which allows respective items to be suppressed for each record, have been proposed [4, 7]. Jia et al. [7] confirmed that utility loss can be reduced considerably by realizing ρ-uncertainty by partial suppression instead of using global suppression. We built algorithms to achieve our proposed model by expanding the ρ-uncertainty method.

Table 1 shows an example of set-valued data representing purchase history. Table 1a shows the original data, Table 1b shows the processed data using the proposed algorithm with $\rho = 0.5$, and Table 1c shows the published data. An item written in italics means it was considered as sensitive by the user. In the personalized model, when a user uses a service, he/she can designate which product he/she wants to protect, for each genre or for each individual item (by entering this information into the system with checkboxes, for example). If an adversary can browse the original data and knows that Alice bought milk, he/she can infer that Alice also bought *medicine* with a confidence of $2/3$ (greater than ρ). On the other hand, if the adversary can only browse the published data, the fact that Alice bought *medicine* cannot be inferred with a confidence higher than ρ because the confidence of the association rule $\{\text{milk}\}\rightarrow\{\text{medicine}\}$ is 0.5. Similarly, all the privacy breaches that appeared in the original data are eliminated in the published data by suppressing some items. This makes it possible for many users to use the service and provide their personal data with peace of mind, which also leads to significant benefits for data publishers and data users.

Another problem in previous studies on anonymization of set-valued data is that the computational complexity increases combinatorially with increasing data size. As a countermeasure, for k^m-anonymity, a model that provides a probabilistic privacy guarantee using a sampling method was proposed [1]. In our research, we show that probabilistic relaxation by sampling method can also be applied to the proposed model, and we constructed an al-

gorithm based on this idea. As a relaxation model, we propose *personalized ρ^m-uncertainty* and *personalized (ε, δ)-ρ^m-uncertainty*.

The contributions of this paper are summarized as follows:

1. We propose a new model, personalized ρ-uncertainty, to prevent attribute inference in situations where individuals want to consider different items as sensitive and protect them. Also, to realize this, we constructed an algorithm using partial suppression.
2. We show that probabilistic relaxation using sampling can also be applied to the proposed model as a way to perform anonymization with a reasonable computation time even with large-scale data. We built an algorithm based on this idea.
3. Regarding the algorithms for realizing privacy guarantees by the proposed models, we evaluate their performances by numerical experiments using actual datasets. As a result, in realistic situations, we confirmed that anonymization can be performed with a much smaller utility loss when using the proposed model rather than the existing one.

The rest of this paper is organized as follows. In Section 2, we briefly review the existing anonymization models. In Section 3, we formally introduce the existing model, ρ-uncertainty. In Section 4, we formalize the proposed model, personalized ρ-uncertainty, and describe the algorithm for this model. Section 5 explains the relaxation model to cope with the difficulty of computational complexity. Section 6 shows the results of experimentally evaluating the performance of the proposed model, and Section 7 concludes the paper.

2 Related Work

In the researches on privacy-preserving data publishing, various anonymization models have been proposed so far. In this section, we briefly review these related works.

For relational databases that can distinguish between QIDs and sensitive attributes, k-anonymity [10], which prevents identity disclosure by ensuring that there are k or more records with exactly the same QID combination, is the most basic model. However, when anonymizing set-valued data, it is necessary to slightly change the way of thinking. He et al. [6] proposed an efficient algorithm to apply k-anonymity to set-valued data, but since there is no distinction as to which attribute is sensitive, this is not a practical privacy preservation method.

Terrovitis et al. [11, 12] proposed a new model, k^m-anonymity, which is a relaxation of k-anonymity for set-valued data. This guarantees that for any subset whose size is equal to or less than m of any record, there are $k-1$ or more different records containing it. By doing this, even if an adversary knows up to m of the items held by his/her target, it is expected that the corresponding record cannot be identified among k records, and more information is prevented from leaking. However, considering the case where all k records are the same, for example, even if the record cannot be specified, information on other items of the target can be leaked, i.e., the risk of attribute inference remains. Although they stated that we could think of the concept of ℓ^m-diversity [12] as a countermeasure for this, they also said that it cannot be practically applied, since we cannot identify which attributes are sensitive in set-valued data.

However, if we assume that items can be divided into QIDs and sensitive attributes, attribute inference in set-valued data can be prevented with simple models. Xu et al. [18] proposed (h, k, p)-coherency, Cao et al. [3] proposed ρ-uncertainty, and Loukides et al. [8]

proposed PS-rule model. Ghinita et al. [5] adopted a permutation approach for set-valued data, in a similar way to Anatomy [16] for relational data. In addition, disassociation by Terrovitis et al. [13] can also be applied to achieve ℓ-diversity.

There has been long-awaited demand to eliminate the assumption that the same items are sensitive for everyone and provide appropriate guarantees to each individual. For relational databases, there is research on personalized privacy preservation [17], but there are few studies dealing with personalized privacy of set-valued data. Wang et al. [14] proposed a novel model based on k-anonymity that provides a distinct privacy guarantee for each individual based on bipartite graphs. However, it is not sufficiently flexible because what can be specified in this model is only whether each individual is sensitive and whether each item is sensitive. In contrast, in our research, we propose a sufficiently flexible model that allows each individual to specify which items are sensitive.

When considering such a complicated privacy requirement, the methods using global generalization or global suppression lead to significant information loss. Jia et al. [7] proposed an algorithm to achieve ρ-uncertainty using partial suppression, preventing unnecessary information loss. In this research, we build algorithms based on this concept.

3 Preliminary

First, we formally define the existing model, ρ-uncertainty [3], which is the base of the proposed model.

Let D be a set-valued dataset, $\mathcal{I}$ be the universe of all items, and $\mathcal{U}$ be the set of all individuals contained in D. $D_u \subseteq \mathcal{I}$ refers to a record of an individual $u \in \mathcal{U}$. For any itemset $Q \subseteq \mathcal{I}$, the *support* of Q refers to the number of records containing Q in D and is denoted as $supp_D(Q)$. Under the ρ-uncertainty setting, it is assumed that sensitive items and non-sensitive items can be strictly distinguished, i.e., $\mathcal{I} = \mathcal{I}_S \cup \mathcal{I}_N$ and $\mathcal{I}_S \cap \mathcal{I}_N = \emptyset$ where $\mathcal{I}_S$ denotes the set of all sensitive items and $\mathcal{I}_N$ denotes the set of all non-sensitive items.

When an adversary knows that an individual u has a subset of items $Q \subset D_u$, if there is another itemset R containing some sensitive items (not included in Q) and the association rule $Q \to R$ has a high confidence, there is a privacy breach. Here, the confidence of an association rule $Q \to R$ is defined as follows:

$$conf(Q \to R) = \frac{supp_D(Q \cup R)}{supp_D(Q)}. \tag{1}$$

Here, Q and R are called the *antecedent* and *consequent* of the association rule, respectively. The condition of ρ-uncertainty is defined as follows.

Definition 1. A set-valued dataset D satisfies ρ-uncertainty if for any itemset $Q \subseteq \mathcal{I}$, one of the following conditions is satisfied:

1. for any sensitive item $e \in \mathcal{I}_S \backslash Q$, $conf(Q \to \{e\}) \leq \rho$.
2. $supp_D(Q) = 0$.

In the second condition of the above definition, we consider the case where an item contained in the adversary's background knowledge Q is suppressed by the anonymization process. In that case, the adversary cannot obtain any information about his/her target, which implies that there is no privacy breach.

4 Proposed Model

4.1 Personalized ρ-uncertainty

ρ-uncertainty can provide a practical privacy guarantee that prevents each individual from being presumed to have sensitive items with high probability. However, the information as to whether each item is sensitive can be strongly dependent on each individual, so it cannot be said that this condition determining sensitive items independently of individuals is realistic. In this research, we propose a new anonymization model that can be applied to realistic situations by allowing individuals to consider different items as sensitive and explicitly representing these conditions as *sensitive constraints*.

For each individual $u \in \mathcal{U}$, let E_u denote the set of items u considers as sensitive, and we call $E = \{(u, E_u) : u \in \mathcal{U}\}$ the sensitive constraint. For example, at e-commerce websites, we obtain E by having each customer register items that they want to protect when they use the service. An adversary defined as $adv = (u, Q)$ knows that the individual u (*target*) has the subset of items $Q \subset \mathcal{I}$ and estimates that u has his/her sensitive item $e \in E_u \backslash Q$ with a probability of $conf(Q \to \{e\})$. For the parameter ρ representing the predefined privacy intensity criterion, the conditions to be satisfied are defined as follows.

Definition 2. A set-valued dataset D is safe w.r.t. $adv = (u, Q)$ if one of the following conditions is satisfied:

1. for any sensitive item $e \in E_u \backslash Q$, $conf(Q \to \{e\}) \leq \rho$.
2. $supp_D(Q) = 0$.

This condition guarantees that an adversary who knows a subset of his/her target's record cannot speculate on any sensitive information of the target with a confidence higher than ρ or cannot find a record in the data that can correspond to the target. Based on this condition, the privacy requirement of the whole data is defined as follows.

Definition 3. A set-valued dataset D satisfies personalized ρ-uncertainty if D is safe w.r.t. any adversary $adv = (u, Q)$.

This requirement is equivalent to Definition 1 if the sensitive items for all individuals are equal, so it can be said that this is a more versatile model as an extension of the existing model. Note that each individual can designate items that they do not actually have as sensitive items, as is the case with ρ-uncertainty. By doing this, it is possible to prevent misunderstanding about items that the person does not actually have.

The validity of considering only association rules whose consequents contain only one sensitive item is guaranteed by the following lemma. This is the straightforward adaptation of the lemma by Cao et al. [3] to our setting.

Lemma 4. *If an association rule whose antecedent is $Q \subseteq D_u$ and consequent is $\{e\}$ $(e \in E_u)$ satisfies $conf(Q \to \{e\}) \leq \rho$, $conf(Q \to R) \leq \rho$ holds for any itemset $R \subseteq \mathcal{I}$ satisfying $e \in R$.*

Proof.

$$conf(Q \to R) = \frac{supp(Q \cup R)}{supp(Q)} \tag{2}$$

$$\leq \frac{supp(Q \cup \{e\})}{supp(Q)} \tag{3}$$

$$= conf(Q \to \{e\}) \leq \rho. \tag{4}$$

□

We call an association rule $Q \to \{e\}$ a *sensitive association rule* if there is a user $u \in \mathcal{U}$ satisfying $Q \subseteq D_u$ and $e \in E_u$.

4.2 Algorithm

As mentioned in Section 1, methods for processing set-valued data are roughly divided into generalization and suppression. If generalization is used in a situation where sensitive items and non-sensitive items are mixed, there is a risk that the fact itself that we generalize items will give clues to an adversary [3]. Therefore, this time we adopt suppression as the processing method, and, in particular, we use partial suppression. There is a method by Jia et al. [7] using partial suppression for ρ-uncertainty. In this paper, we extend it so that it can be applied to the proposed model.

When constructing an anonymization algorithm, it is necessary to first decide what kind of indicators to use to analyze the utility of the data. If we know the tasks in which we use anonymized data, we could design an algorithm that maximizes a utility metric to measure the performance of one of the tasks while satisfying privacy constraints. Here, we focus on the case where how to use data is not specifically defined. Hence, as a general-purpose criterion to some extent, we aim to maintain the amount of information and the statistical properties of the whole data as much as possible and define the utility metrics as follows. Let D be the original dataset, D' be the dataset being processed, and $D(i)$ be the normalized frequency of item i in D.

1. Ratio of items suppressed:

$$util_{\mathrm{info}}(D, D') = \frac{\sum_{i \in \mathcal{I}} (supp_D(i) - supp_{D'}(i))}{\sum_{i \in \mathcal{I}} supp_D(i)}. \tag{5}$$

2. Difference in frequency distribution of each item (measured by KL-divergence):

$$util_{\mathrm{dist}}(D, D') = \sum_{i \in \mathcal{I}} D'(i) \log \frac{D'(i)}{D(i)}. \tag{6}$$

In most anonymization models, including the proposed one, the problem of minimizing the utility loss while satisfying the privacy constraint is NP-hard. Another difficulty is that the proposed model does not satisfy *monotonicity*. For example, in the case of preventing identity disclosure by generalization, the privacy risk decreases monotonically as generalization proceeds (Property 1 [11]). However, when considering the risk of attribute inference, such a property does not hold, i.e., new risks may be created by processing. It takes huge amounts of computation to perform this processing while preventing such risks completely, and it is difficult to achieve by an efficient algorithm. Thus, to be realistic, here, we adopt a simple policy that we iterate suppression focusing on each adversary that is not safe and remove the risk until all risks are eliminated.

For an association rule $Q \to \{e\}$ that has a confidence higher than ρ, define a *suppression number* N_s of an item $i \in Q \cup \{e\}$ as follows:

$$N_s(i,\ Q \to \{e\}) = \begin{cases} \lceil supp_D(Q \cup \{e\}) - supp_D(Q) \cdot \rho \rceil & \text{if } i = e, \\ \left\lceil \frac{supp_D(Q \cup \{e\}) - supp_D(Q) \cdot \rho}{1 - \rho} \right\rceil & \text{if } i \in Q. \end{cases} \tag{7}$$

User	Contents
u_1	$x, \boldsymbol{y}$
u_2	x, y
u_3	x, y
u_4	x

(a) $conf(\{x\} \to \{y\}) = 0.75$

User	Contents
u_1	$x, \boldsymbol{y}$
u_2	$x, \not{y}$
u_3	x, y
u_4	x

(b) $conf(\{x\} \to \{y\}) = 0.5$

User	Contents
u_1	$\not{x}, \boldsymbol{y}$
u_2	$\not{x}, y$
u_3	x, y
u_4	x

(c) $conf(\{x\} \to \{y\}) = 0.5$

Table 2: Example of suppression

This definition represents the number of suppressions needed to satisfy $conf(Q \to \{e\}) \leq \rho$ when we select one item $i \in Q \cup \{e\}$ and suppress it from the records containing $Q \cup \{e\}$. For example, suppose that user u_1 has items $x \notin E_{u_1}$ and $y \in E_{u_1}$. Let $supp(\{x\}) = 4$, $supp(\{y\}) = 3$, and $\rho = 0.5$ (see Table 2a). If we choose y as an item to be suppressed, picking $3 - 4 \times 0.5 = 1$ record containing $\{x, y\}$ and suppressing y from it results in $conf(\{x\} \to \{y\}) = 2/4 \leq \rho$ (Table 2b). If we choose x instead, we should suppress it from $\frac{3-4\times0.5}{1-0.5} = 2$ records, and we get $conf(\{x\} \to \{y\}) = 1/2 \leq \rho$ (Table 2c). Both of these cases of suppression satisfy the condition between $conf(\{x\} \to \{y\})$ and ρ.

The overall view of the algorithm, SUPPRESSOR, is shown in Algorithm 1. In this algorithm, for each ℓ and $u \in \mathcal{U}$, we assume adversaries whose target is u and whose background knowledge contains ℓ items. The notation $\mathcal{R}_u^\ell$ represents the set of sensitive association rules that can lead to privacy breach regarding such adversaries. If no rule is contained in $\mathcal{R}_u^\ell$ for all ℓ and u, it can be said that the dataset satisfies Definition 3. Otherwise, we need to suppress some items to eliminate the risks. For each rule $Q \to \{e\}$ whose confidence exceeds ρ in D', we select item d and suppress it in N_s records. When choosing d, we maximize $F(d, e, Q, D, D')$ defined as follows:

$$F(d, e, Q, D, D') = \frac{D'(d) \log \frac{D'(d)}{D(d)}}{N_s(d,\ Q \to \{e\})}. \tag{8}$$

This is a function designed by combining two objectives: (1) to reduce the number of suppressed items and (2) reduce the difference in the frequency distribution of items between D and D'. These are based on the two utility metrics described previously. Finally, it outputs processed data that is safe w.r.t. all adversaries.

Note that in the proposed model, a divide-and-conquer framework [7] cannot be applied. In the existing ρ-uncertainty model, if a large dataset is divided horizontally into several parts and the safety is guaranteed by executing the algorithm independently for each part, the combined dataset is also safe. However, when considering personalized ρ-uncertainty, even if each part is processed independently, the safety of the whole dataset is not guaranteed because items specified as sensitive attributes by people included in other parts cannot be considered. Therefore, we did not divide the dataset this time but processed the whole data at once.

We describe the time complexity of Algorithm 1. Let n be the number of users included in D, i.e., $n = |\mathcal{U}|$, M_D be the maximum record size, i.e., $M_D = \max_{u\in\mathcal{U}} |D_u|$, and M_E be the maximum number of sensitive items assigned by a single user, i.e., $M_E = \max_{u\in\mathcal{U}} |E_u|$. For each ℓ and u, we enumerate sensitive association rules $\mathcal{R}_u^\ell$. For an antecedent $Q \subseteq D_u$, the cost of computing $supp_{D'}(Q)$ is $\mathrm{O}(n|Q|)$ and the cost of computing $conf(Q \to \{e\})$ regarding all $e \in E_u$ is $\mathrm{O}(nM_E)$. Since the number of patterns of Q whose size is ℓ is $\mathrm{O}(\binom{M_D}{\ell})$, enumerating $\mathcal{R}_u^\ell$ takes $\mathrm{O}(n\binom{M_D}{\ell}(M_D + M_E))$ time. In each iteration regard-

Algorithm 1 SUPPRESSOR

Input: dataset D, sensitive constraint E, parameter ρ
Output: processed dataset D' satisfying personalized ρ-uncertainty

1: $D' \leftarrow D$
2: $safe \leftarrow$ **false**
3: **while not** $safe$ **do**
4: $safe \leftarrow$ **true**
5: **for** $\ell = 1$ to $\max_u(|D_u|)$ **do**
6: **for** u in $\mathcal{U}$ **do**
7: $\mathcal{R}_u^\ell \leftarrow \{Q \to \{e\} \mid Q \subseteq D_u, |Q| = \ell, e \in E_u, conf(Q \to \{e\}) > \rho\}$
8: **if** $\mathcal{R}_u^\ell \neq \emptyset$ **then**
9: $safe \leftarrow$ **false**
10: **for** $Q \to \{e\}$ in $\mathcal{R}_u^\ell$ **do**
11: **if** $conf(Q \to \{e\}) > \rho$ **then**
12: $d \leftarrow \arg\max_{d \in Q \cup \{e\}} F(d, e, Q, D, D')$
13: $N \leftarrow N_s(d,\ Q \to \{e\})$
14: select N records randomly from records in D' containing $Q \cup \{e\}$, and suppress d from them
15: **return** D'

ing one rule in $\mathcal{R}_u^\ell$ (Steps 11–14), the bottleneck is computing $conf(Q \to \{e\})$, and this cost is $\mathrm{O}(n|Q|)$. Since $|\mathcal{R}_u^\ell| = \mathrm{O}(\binom{M_D}{\ell} M_E)$, the cost of Steps 7–14 is $\mathrm{O}(n\binom{M_D}{\ell} M_D M_E)$. Summarizing this with respect to $\ell = 1, \ldots, M_D$ and $u \in \mathcal{U}$, the cost of Steps 4–14 is $\mathrm{O}(\sum_{\ell=1}^{M_D} n^2 \binom{M_D}{\ell} M_D M_E) = \mathrm{O}(n^2 \cdot 2^{M_D} \cdot M_D M_E)$. In the algorithm, this process will be repeated until all risks are eliminated. Because it is not guaranteed that the number of risks decreases monotonically, estimating the number of iterations of the **while** loop is difficult. At least, if a risk is found in one iteration, one or more items must be suppressed. If all the items are suppressed, the dataset obviously meets the condition of Definition 3. Therefore, it can be said that the number of iterations in the worst case is $\mathrm{O}(nM_D)$. Assuming now that the number of iterations is T, the computational complexity of the entire algorithm is $\mathrm{O}(Tn^2 \cdot 2^{M_D} \cdot M_D M_E)$. When $M_D \ll M_E$ and we can consider M_D as a constant, the cost is $\mathrm{O}(Tn^2 M_E)$.

5 Model Relaxation

To guarantee personalized ρ-uncertainty, we must consider adversaries who have any possible subset as background knowledge and investigate the safety of the data against them. Therefore, it is difficult to directly apply this algorithm to real large-scale data since the number of patterns to be considered as adversaries' knowledge increases combinatorially with the maximum record length of data. A similar problem also occurs in the existing ρ-uncertainty model, and in prior researches limiting its application target to data with a very small maximum record length was inevitable.

In this section, to cope with this problem, we propose new models with relaxed conditions that can anonymize large-scale data appropriately.

5.1 Personalized ρ^m-uncertainty

First, as a simple approach, we define the next model that limits the size of the adversary's background knowledge to less than a certain size.

Definition 5. A set-valued dataset D satisfies personalized ρ^m-uncertainty if D is safe w.r.t. any adversary $adv = (u, Q)$ satisfying $|Q| \leq m$.

This model relaxes the condition by limiting the strength of the adversary assumed in Definition 3 in a similar way to the relaxation of k-anonymity to k^m-anonymity. However, in this case, too, since the number of patterns to be checked increases combinatorially with m, it is difficult to apply in some cases. Next, as a further contrivance, we introduce a method to reduce the computational cost by probabilistically relaxing the model.

5.2 Personalized (ε, δ)-ρ^m-uncertainty

The basic idea of probabilistic relaxation is to assume that an adversary appears according to a certain probability distribution and guarantee the safety of the data against such an adversary with a certain probability using a sampling method. The model described here involves applying the idea of Gergely et al. [1], which proposed to probabilistically relax k^m-anonymity.

Let $\mathcal{A}^\ell$ be a set of adversaries who have a subset of size ℓ of the records held by the target as background knowledge, i.e., $\mathcal{A}^\ell = \{(u, Q) : u \in \mathcal{U},\ Q \subseteq D_u,\ |Q| = \ell\}$, and let α_ℓ be a random variable taking a value on $\mathcal{A}^\ell$. Also, let H_ℓ be the probability that D is not safe w.r.t. α_ℓ. After these preparations, we define a model that is probabilistic relaxation of personalized ρ^m-uncertainty.

Definition 6. A set-valued dataset D satisfies personalized (ε, δ)-ρ^m-uncertainty if $Pr[H_\ell < \varepsilon] \geq 1 - \delta$ for all $\ell \leq m$.

When this definition is satisfied for sufficiently small $\varepsilon, \delta > 0$, it is highly probable that D is safe w.r.t. most adversaries.

5.3 Algorithm with Sampling

We used sampling to check if the data meets the condition of Definition 6. The following theorem gives a sufficient requirement for the samples necessary for this purpose.

Theorem 7. *Consider a dataset D and a sensitive constraint E. Let S denote a set of samples picked from a certain probability distribution on $\mathcal{A}^\ell$ independently, and suppose $|S| \geq \frac{\log(1/\delta)}{2\varepsilon^2}$ regarding parameters $\varepsilon, \delta > 0$. Then, if D is safe w.r.t. any adversary $\alpha_\ell \in S$, $Pr[H_\ell < \varepsilon] \geq 1 - \delta$ holds.*

Proof. Let $n = |S|$ and X_i $(i = 1, \ldots, n)$ be random variables that take 0 if D is safe w.r.t. the sample $adv_i = (u_i, Q_i) \in S$ and 1 if not. Let $\hat{H}_\ell$ be the rate of adversaries in S for which D is not safe, i.e., $\hat{H}_\ell = \sum_i X_i/n$. Since S is the set of samples independently picked from $\mathcal{A}^\ell$, $E[\hat{H}_\ell] = H_\ell$ holds. Following Hoeffding's inequality, we obtain

$$\forall \varepsilon > 0,\ Pr[H_\ell - \hat{H}_\ell \geq \varepsilon] \leq e^{-2|S|\varepsilon^2}. \tag{9}$$

$\hat{H}_\ell = 0$ if D is safe w.r.t. any adversary included in S, and

$$Pr[H_\ell < \varepsilon] \geq 1 - e^{-2|S|\varepsilon^2}. \tag{10}$$

Algorithm 2 SAMPLESUPPRESSOR

Input: dataset D, sensitive constraint E, parameters $\rho,\ m,\ \varepsilon,\ \delta$
Output: processed dataset D' satisfying personalized (ε, δ)-ρ^m-uncertainty

$D' \leftarrow D$
$n_S \leftarrow \left\lceil \frac{\log(1/\delta)}{2\varepsilon^2} \right\rceil$
$safe \leftarrow$ **false**
while not $safe$ **do**
  $safe \leftarrow$ **true**
  for $\ell = 1$ to m **do**
    $S \leftarrow n_S$ samples of adversary α_ℓ sampled from $\mathcal{A}^\ell$ independently
    $\mathcal{R}_S \leftarrow \{Q \to \{e\} \mid \exists (u, Q) \in S, e \in E_u, conf(Q \to \{e\}) > \rho\}$
    if $\mathcal{R}_S \neq \emptyset$ **then**
      $safe \leftarrow$ **false**
    for $Q \to \{e\}$ in $\mathcal{R}_S$ **do**
      if $conf(Q \to \{e\}) > \rho$ **then**
        $d \leftarrow \arg\max_{d \in Q \cup \{e\}} F(d, e, Q, D, D')$
        $N \leftarrow N_s(d,\ Q \to \{e\})$
        select N records randomly from records in D' containing $Q \cup \{e\}$, and suppress d from them
return D'

Here, given parameters $\varepsilon, \delta > 0$, we get $Pr[H_\ell < \varepsilon] \geq 1 - \delta$ if D is safe w.r.t. any sample in S satisfying

$$|S| \geq \frac{\log(1/\delta)}{2\varepsilon^2} \tag{11}$$

and this concludes the proof. □

According to this theorem, we should set $|S| \geq 116$ when $\varepsilon = \delta = 0.1$, $|S| \geq 600$ when $\varepsilon = \delta = 0.05$, and $|S| \geq 23{,}026$ when $\varepsilon = \delta = 0.01$. This value does not depend on the maximum record length or m, so using this condition prevents the computational cost from rapidly increasing with the values of the parameters and makes it possible to handle large-scale data at practical costs.

Algorithm 2, SAMPLESUPPRESSOR, shows the operation to proceed with suppression while checking whether the dataset satisfies this condition. For each ℓ, we select $n_S = \left\lceil \frac{\log(1/\delta)}{2\varepsilon^2} \right\rceil$ samples from $\mathcal{A}^\ell$ independently and call the set of samples S. The notation $\mathcal{R}_S$ represents the set of sensitive association rules that can lead to privacy breach regarding sampled adversaries. If no rule is contained in $\mathcal{R}_S$ for all ℓ, it can be said that the data meet the condition of Definition 6. Otherwise, we suppress some items to eliminate the risks in the same manner as in Algorithm 1. Finally, if any unsafe adversary does not appear at all for all $\ell \leq m$, the condition is satisfied, and the algorithm outputs D' at that point.

To perform sampling, it is necessary to determine the probability distribution on $\mathcal{A}^\ell$ that an adversary α_ℓ follows. In [1], Gergely et al. proposed a method of sampling from a uniform distribution by using Markov Chain Monte Carlo method, taking into consideration the difference in appearance frequency of itemsets. In our case, however, since we must also consider who is the target of the sampled adversary, we use a more simple way as

follows: first, select one user with ℓ or more items from all users uniformly at random, then pick a subset of ℓ items uniformly at random from that user's record.

Let us consider the time complexity of Algorithm 2. For each adversary $(u, Q) \in S$, the cost of computing $supp_{D'}(Q)$ is $\mathrm{O}(n|Q|)$ and the cost of computing $conf(Q \to \{e\})$ regarding all $e \in E_u$ is $\mathrm{O}(nM_E)$. Thus, enumerating $\mathcal{R}_S$ takes $\mathrm{O}(n_S n(|Q| + M_E)) = \mathrm{O}(n_S n(m + M_E))$ time. Since the cost of suppressing items regarding one rule $Q \to \{e\}$ is $\mathrm{O}(n|Q|)$, as discussed in Section 4.2, and $|\mathcal{R}_S| = \mathrm{O}(n_S M_E)$, the cost of suppressing items regarding $\mathcal{R}_S$ (Steps 11–15) is $\mathrm{O}(n_S n m M_E)$. Summarizing the above for $\ell = 1, \ldots, m$, one iteration of the **while** loop is $\mathrm{O}(n_S n m^2 M_E)$. When the number of this iteration is T, the time complexity of the entire algorithm is $\mathrm{O}(T n_S n m^2 M_E)$. Since it is difficult to estimate T theoretically, the computation time for realistic data is experimentally investigated.

6 Experimental Evaluation

6.1 Settings

We conducted experiments on three real-world datasets, BMS-WebView-1, BMS-WebView-2 and BMS-POS (introduced by Zheng et al. [19]), which are widely used as benchmarks in the field of data mining. BMS-WebView-1 and BMS-WebView-2 contain several months worth of click-stream data from e-commerce websites and BMS-POS contains several years worth of point-of-sale data from an electronics retailer. As mentioned above, it is difficult to apply Algorithm 1 directly to these datasets because the computational cost increases combinatorially when the maximum record length of data is large. Thus, we extract only records whose sizes are 5 or less from the original data and use them as experimental data following previous studies [3, 4, 7]. We call these extracted datasets BMS-WebView-1', BMS-WebView-2' and BMS-POS' respectively. The characteristics of the datasets are shown in Table 3. For Algorithm 2, original datasets were used.

In these datasets, there is no distinction between sensitive items and non-sensitive items. In the previous studies, a certain percentage of all the items were randomly selected as sensitive items, and the others were defined as non-sensitive items. However, since this is the first study to consider the situation where each individual can select different sensitive items in set-valued data, it is necessary to investigate how this situation affects the results of anonymization. Here, as Experiment 1, we conduct experiments on the following two cases. One involves randomly selecting a certain percentage of items to be sensitive for everyone (fixed) and the other involves randomly selecting a certain percentage of items to be sensitive for each individual distinctively (personalized). Because these cases provide different privacy protection criteria, there is no point in directly comparing the values of

Dataset	$\lvert D\rvert$	$\lvert\mathcal{I}\rvert$	Max. record length	Avg. record length
BMS-WebView-1	59,602	497	267	2.5
BMS-WebView-1'	54,737	472	5	1.7
BMS-WebView-2	77,512	3,340	161	4.6
BMS-WebView-2'	58,044	3,207	5	2.2
BMS-POS	551,596	1,657	164	6.5
BMS-POS'	306,982	1,177	5	2.7

Table 3: Characteristics of datasets

the utility metrics, but we can observe the effect of allowing individuals to choose different sensitive items separately.

In practice, however, even when individuals can select sensitive items separately, it is hard to say that the situation where each individual selects totally different items is realistic. As Experiment 2, therefore, we assume a more realistic situation as follows. First, a certain percentage of items are defined as sensitive items (to be precise, items that tend to be sensitive). In addition to that, we reproduce a situation in which each individual specifies whether to protect each item according to the need by adding 1% of noise (personalized+). In other words, each person considers each item that tends not to be sensitive as a sensitive attribute with a probability of 1% and each item that tends to be sensitive as a sensitive attribute with a probability of 99%. To cope with such a situation with the existing models, the only way is to designate all items that are sensitive to at least one person as sensitive to everyone. We also prepare the data in this manner (fixed+). Note that, when sensitive items are common to everyone, Algorithm 1 is almost the same as the *Dist* [7] algorithm (without divide-and-conquer framework). Thus, fixed+ can be regarded as a baseline corresponding to the current algorithm.

In general, the performance of the anonymization method depends greatly on the characteristics of the dataset to be handled. Moreover, since we randomly select sensitive items at this time, the results can also be affected by this way of selecting. It is necessary to consider the effect of these factors. Hence, here, we first randomly extract 10% of the records from the whole dataset and randomly select sensitive items. We perform anonymization of the data obtained in this way ten times and examine the average value and standard deviation of the utility metrics of the results.

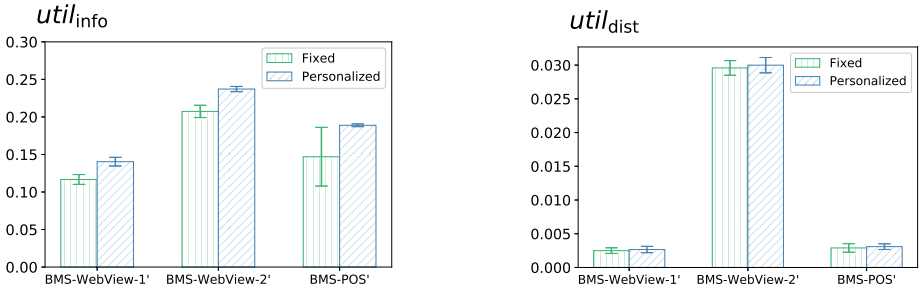
Evaluation is performed mainly by analyzing the ratio of items suppressed ($util_{\mathrm{info}}$) and KL-divergence ($util_{\mathrm{dist}}$). In addition, to consider more practical utility, we also introduce the results of investigating how the results of association rule mining, which is a typical application of set-valued data, changes through the anonymization process (see the result of Experiment 2 for details). For any metric, the smaller the value, the higher the utility of the processed data is.

All programs were coded in Python3 and run on AWS Linux instance with 61-GiB memory and eight 2.3-GHz CPUs.

6.2 Results

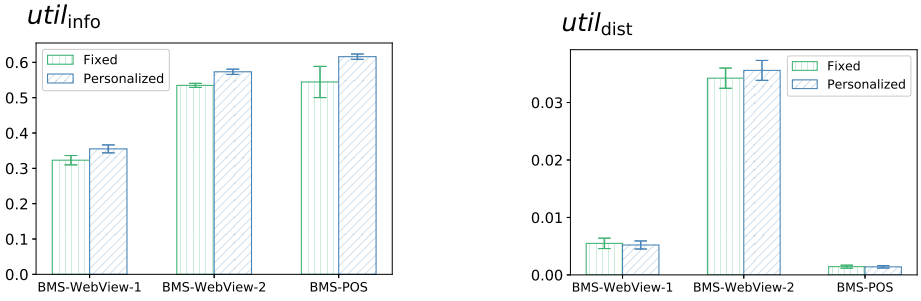
6.2.1 Experiment 1 (fixed vs. personalized)

First, we investigated the effect of selecting different sensitive items separately by comparing fixed and personalized. Figure 1 shows the results of applying Algorithm 1 to the three datasets with a maximum record length of 5. The value of ρ was set to 0.5 and the percentage of sensitive items was set to 40% following the previous studies [3, 7]. The graphs represent the average values and standard deviations in ten trials. Since the privacy requirement is more complicated when sensitive items are different depending on individuals (personalized), the loss of utility is consistently larger with respect to the ratio of suppressed items ($util_{\mathrm{info}}$). Nevertheless, 75 to 90% of the items are left without being suppressed, and in these cases as well, this is considered to be sufficient for practical use. As for KL-divergence, the difference between fixed and personalized was slight, and this tendency was similar in other experimental results. When comparing the three datasets, the utility loss in BMS-WebView-2' was the largest, which is particularly noticeable from the values of KL-divergence (Figure 1b). This is thought to be due to the facts that BMS-



(a) (b)

Figure 1: Results of Algorithm 1



(a) (b)

Figure 2: Results of Algorithm 2

WebView-2' has the largest number of items and it is difficult to maintain the frequency distribution by balancing items to be suppressed in this dataset.

Next, Algorithm 2 is applied to the original datasets, and the utility loss is examined. Figure 2 shows the results of experiments with $\rho = 0.5$, $m = 5$, $\varepsilon = \delta = 0.05$. As a whole, more items are suppressed than the cases when the maximum record length is limited to 5. This is probably due to the fact that the number of combinations of items to be considered is greatly increased by long records, and the privacy risks to be eliminated are also increased. The difference between fixed and personalized is smaller than in the result of Algorithm 1 in most cases, and the value of KL-divergence may be smaller in personalized than in fixed.

Through the results of Experiment 1, it can be observed that in the case of personalized, the utility loss is greater than in the case of fixed, but the increase is not so large, so it is practically acceptable.

6.2.2 Experiment 2 (fixed+ vs. personalized+)

For more realistic situations, we investigated with personalized+ and fixed+ settings using BMS-WebView-2, which has the largest number of item types. To consider more practical utility, we introduce the results of investigating how the results of association rule mining, which is a typical application of set-valued data, change through the anonymization process. Specifically, we investigated how much the set of association rules obtained with the Apriori algorithm [2] changes by applying our algorithms to the dataset. Apriori is a well-known algorithm to efficiently enumerate all association rules whose support and

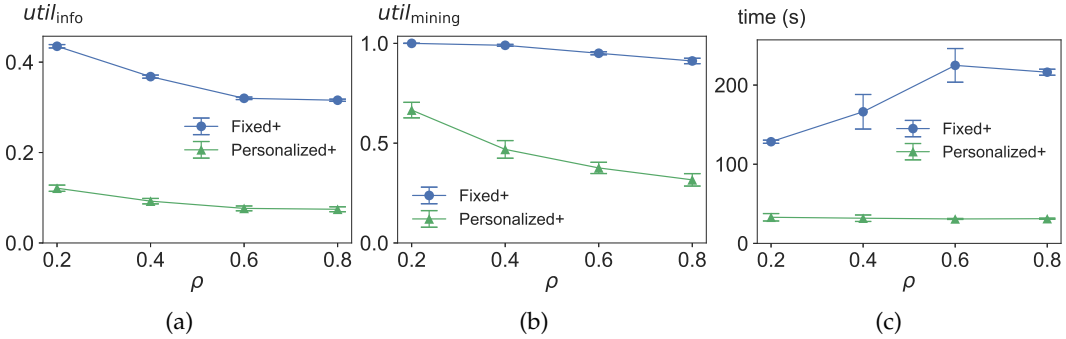


Figure 3: Results of Algorithm 1 while varying ρ (BMS-WebView-2′)

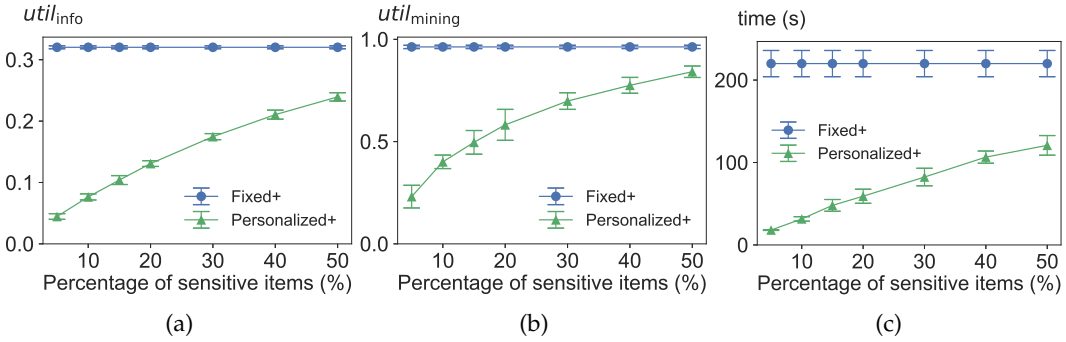


Figure 4: Results of Algorithm 1 while varying percentage of sensitive items (BMS-WebView-2′)

confidence are above certain values, $minsup$ and $minconf$. Here, $minsup$ was set to 0.05% of the number of the records and $minconf$ was set to 30%. These values reflect a practical scenario [7]. We measured utility loss by the Jaccard distance between the sets of rules mined from the data before and after processing. That is, when the set of rules in the original data is R_D and that in the processed data is $R_{D'}$, the utility metric is defined as

$$util_{\text{mining}}(D, D') = 1 - \frac{|R_D \cap R_{D'}|}{|R_D \cup R_{D'}|}. \tag{12}$$

In the following experiments, for BMS-WebView-2′, the utility loss is measured using two metrics, $util_{\text{info}}$ and $util_{\text{mining}}$, instead of $util_{\text{dist}}$. For BMS-WebView-2, since we cannot compute $util_{\text{mining}}$ for the original dataset containing large records due to the computational complexity of the Apriori algorithm, only the values of $util_{\text{info}}$ are shown.

In the previous section, the percentage of sensitive items was set to 40% following the previous studies [3, 7]. In realistic situations, the percentage of items that are sensitive to many people (e.g., drugs, sexual items, etc.) is considered not to be so large, so we set this value to 10% if not stated otherwise. For other parameters, $\rho = 0.5$, $m = 5$, and $\varepsilon = \delta = 0.05$ unless otherwise specified, and the effect of changes in these parameters on the results was examined in detail. We also show the computation time of the algorithms.

Figures 3 and 4 show the results of applying Algorithm 1 to BMS-WebView-2′ with $util_{\text{info}}$, $util_{\text{mining}}$ and runtime. We conducted the experiment while varying ρ (Figure 3) and the

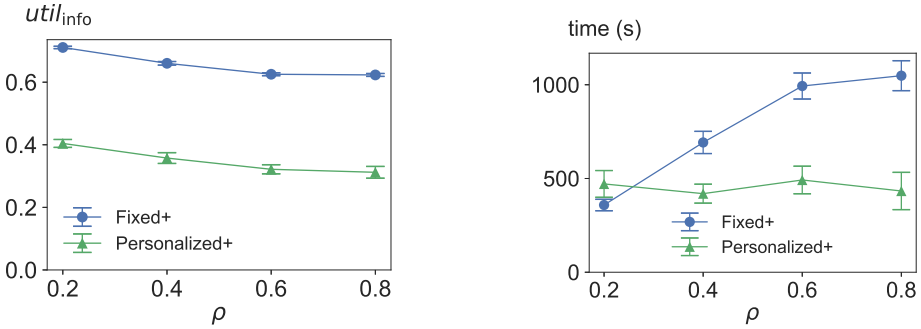


Figure 5: Results of Algorithm 2 while varying ρ (BMS-WebView-2)

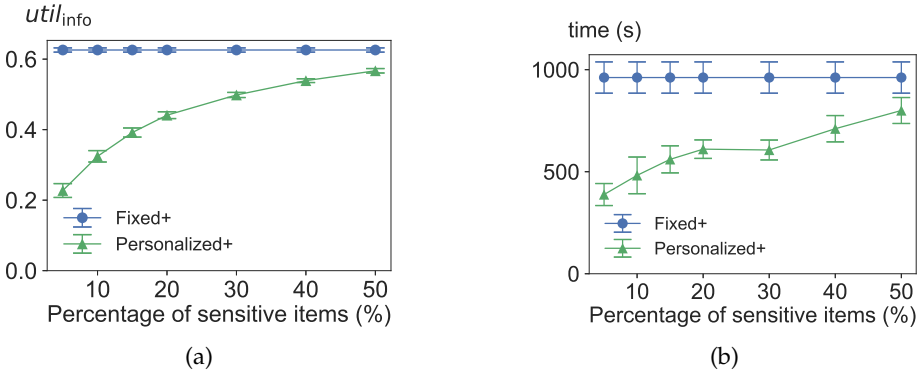


Figure 6: Results of Algorithm 2 while varying percentage of sensitive items (BMS-WebView-2)

percentage of sensitive items (Figure 4). Fixed+ in this experiment is the baseline corresponding to the current model. In this case of fixed+, it is assumed that virtually all items must be considered as sensitive for everyone, and the information loss increases greatly due to unnecessary suppressions. Specifically, the sets of association rules mined from the dataset before and after processing are almost completely different ($util_{\text{mining}} \approx 1$). In contrast, the utility loss in the proposed model (personalized+) is much smaller than that in the current model, and the runtime is also smaller. Regarding ρ (Figure 3), the smaller the value, the more severe the privacy protection requirement, resulting in greater utility loss. Regarding the percentage of sensitive items (Figure 4), as the number of sensitive items increases, the number of sensitive association rules to be considered becomes larger, so utility loss and necessary computation time also increase. The difference between fixed+ and personalized+ is prominent when the percentage of sensitive items is small, and these results indicate the superiority of performing anonymization with the proposed model in realistic situations. The runtime in personalized+ is hardly affected by ρ, but as the percentage of sensitive items increases, it almost linearly increases, which is consistent with the analysis in Section 4.2.

Next, Figures 5 and 6 show the results of applying Algorithm 2 to BMS-WebView-2 while varying ρ (Figure 5) and the percentage of sensitive items (Figure 6). Since we cannot

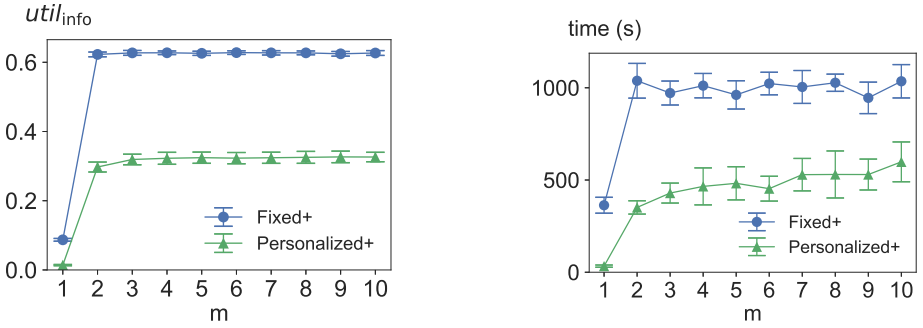


Figure 7: Results of Algorithm 2 while varying m (BMS-WebView-2)

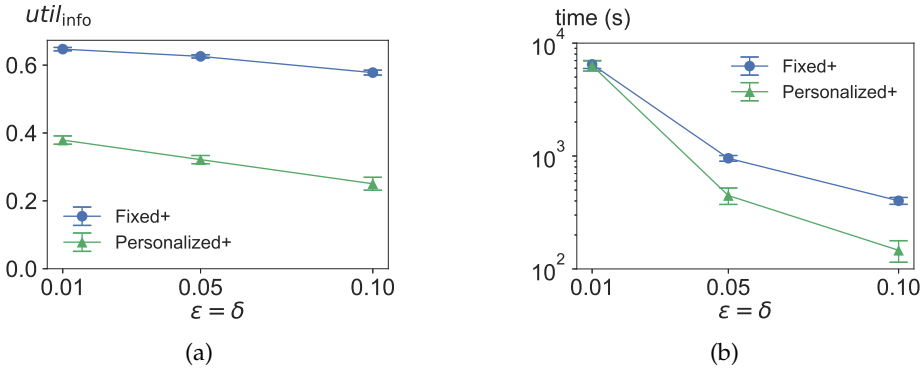


Figure 8: Results of Algorithm 2 while varying ε, δ (BMS-WebView-2)

apply the Apriori algorithm to this dataset, only $util_{\text{info}}$ and runtime are shown. The effect of the change in ρ and the percentage of sensitive items has the same tendency as the result of Algorithm 1. Although the utility loss is greater than in the case of BMS-WebView-2', the proportion of suppressed items is not so large in personalized+, especially when the percentage of sensitive items is small. This result suggests that sufficiently useful data can be obtained using the probabilistic relaxation model even for data including long records that cannot be handled with current models.

Finally, we examine the effect of parameters specific to the probabilistic relaxation model. Figures 7 and 8 show the results of applying Algorithm 2 to BMS-WebView-2 while varying m (Figure 7) and ε, δ (Figure 8). As m becomes larger, the conditions for privacy protection become stricter, considering adversaries with more background knowledge. Figure 7 shows that, when m is about 3 or more, the utility loss hardly changes. This is thought due to the facts that there are relatively few records including many items and that many individuals can be sufficiently specified by the background knowledge about 3 in size. As ε and δ become smaller, that is, as the number of samples needed increases according to (11), the number of items suppressed increases slightly. This is an expected result because the privacy requirement becomes stricter when the number of samples for which we must confirm the safety of the data increases. As ε and δ decrease, the computational complexity of one iteration increases in proportion to (11), and the termination condition of the algorithm

becomes more severe; hence, the total computation time becomes significantly large. This means that the safety of the data also increases accordingly. As computing power permits, this parameter should be set small, and safe anonymization should be done.

7 Conclusion

In this paper, we firstly reviewed the existing models concerning anonymization of set-valued data. Then we discussed the problem that it is necessary to assume that only some items are sensitive attributes for everyone to prevent attribute inference in particular. To deal with this problem, we proposed a new model, personalized ρ-uncertainty, that allows individuals to designate different items as sensitive attributes and provides privacy protection criteria appropriate for each. Furthermore, we also proposed other models, personalized ρ^m-uncertainty and personalized (ε, δ)-ρ^m-uncertainty, to cope with another big problem, the increase of the computational complexity. We designed algorithms to achieve these guarantees and conducted experimental evaluation using the real-world datasets. As a result, we confirmed that anonymization can be performed with a much smaller utility loss when using the proposed model rather than the existing one. We also showed that the probabilistic relaxation model can deal with large-scale data including long records that cannot be handled with current models and obtain sufficiently useful processed data.

Future tasks include development of better performance algorithms and improving our theoretical understanding of the performance of the algorithms.

Acknowledgements

This work was supported by JSPS Grant-in-Aid for Scientific Research (B), Grant Number 15H02700.

References

[1] Gergely Acs, Jagdish Prasad Achara, and Claude Castelluccia. Probabilistic km-anonymity efficient anonymization of large set-valued datasets. In *Proceedings of the 2015 IEEE International Conference on Big Data*, pages 1164–1173, 2015.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, 1994.

[3] Jianneng Cao, Panagiotis Karras, Chedy Raïssi, and Kian-Lee Tan. ρ-uncertainty: inference-proof transaction anonymization. In *Proceedings of the VLDB Endowment*, volume 3, pages 1033–1044, 2010.

[4] Liuhua Chen, Shenghai Zhong, Li-e Wang, and Xianxian Li. A sensitivity-adaptive ρ-uncertainty model for set-valued data. In *International Conference on Financial Cryptography and Data Security*, pages 460–473, 2017.

[5] Gabriel Ghinita, Yufei Tao, and Panos Kalnis. On the anonymization of sparse high-dimensional data. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 715–724, 2008.

[6] Yeye He and Jeffrey F. Naughton. Anonymization of set-valued data via top-down, local generalization. In *Proceedings of the VLDB Endowment*, volume 2, pages 934–945, 2009.

[7] Xiao Jia, Chao Pan, Xinhui Xu, Kenny Q. Zhu, and Eric Lo. ρ-uncertainty anonymization by partial suppression. In *Database Systems for Advanced Applications*, pages 188–202, 2014.

[8] Grigorios Loukides, Aris Gkoulalas-Divanis, and Jianhua Shao. Efficient and flexible anonymization of transaction data. *Knowledge and Information Systems*, 36(1):153–210, 2013.

[9] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[10] Latanya Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[11] Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis. Privacy-preserving anonymization of set-valued data. In *Proceedings of the VLDB Endowment*, volume 1, pages 115–125, 2008.

[12] Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis. Local and global recoding methods for anonymizing set-valued data. *The VLDB Journal*, 20(1):83–106, 2011.

[13] Manolis Terrovitis, Nikos Mamoulis, John Liagouris, and Spiros Skiadopoulos. Privacy preservation by disassociation. In *Proceedings of the VLDB Endowment*, volume 5, pages 944–955, 2012.

[14] Li-e. Wang and Xianxian Li. Personalized privacy protection for transactional data. In *Proceedings of the 10th International Conference on Advanced Data Mining and Applications*, pages 253–266, 2014.

[15] Raymond Chi-Wing Wong, Ada Wai-Chee Fu, Ke Wang, and Jian Pei. Minimality attack in privacy preserving data publishing. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, pages 543–554, 2007.

[16] Xiaokui Xiao and Yufei Tao. Anatomy: simple and effective privacy preservation. In *Proceedings of the 32nd International Conference on Very Large Data Bases*, pages 139–150, 2006.

[17] Xiaokui Xiao and Yufei Tao. Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, pages 229–240, 2006.

[18] Yabu Xu, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. Anonymizing transaction databases for publication. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–775, 2008.

[19] Zijian Zheng, Ron Kohavi, and Llew Mason. Real world performance of association rule algorithms. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 401–406, 2001.