Semantic Diversity: Privacy Considering Distance Between Values of Sensitive Attribute

Keiichiro Oishi[a,*], Yuichi Sei[a,b], Yasuyuki Tahara[a], Akihiko Ohsuga[a]

[a] *Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo, Japan*
[b] *JST, PRESTO, Kawaguchi, Saitama, Japan*

Abstract

A database that contains personal information and is collected by crowdsensing can be used for various purposes. Therefore, database holders may want to share their databases with other organizations. However, since a database contains information about individuals, database recipients must take privacy concerns into consideration. One of the mainstream privacy protection indicators, l-diversity, guarantees that the probability of identifying a sensitive attribute value of an individual in a database is less than $1/l$. However, when there are several semantically similar values in the sensitive attribute, there is a possibility that actual diversity is not satisfied, even if anonymization is performed to satisfy l-diversity. For example, an attacker may know that candidates of Alice's disease are a set of HIV-1(M), HIV-1(N), and HIV-2 if the anonymized database satisfies 3-diversity. In this case, the attacker can conclude that Alice has HIV, although the detailed type remains unknown. In this research, to solve how actual diversity cannot be taken into consideration with existing l-diversity, we proposed a novel privacy indicator, (l, d)-semantic diversity, and an algorithm that anonymizes a database to satisfy (l, d)-semantic diversity. We also proposed an analysis algorithm that is suitable for the proposed anonymizing algorithm

*Corresponding author
Email addresses: ohishi.keiichiro@ohsuga.lab.uec.ac.jp (Keiichiro Oishi), seiuny@uec.ac.jp (Yuichi Sei), tahara@uec.ac.jp (Yasuyuki Tahara), ohsuga@uec.ac.jp (Akihiko Ohsuga)

because the output of the anonymizing algorithm is difficult to understand. Our proposed algorithms were experimentally evaluated using synthetic and real datasets.
Keywords: Computer Security, Privacy Preserving Data Publishing, Anonymity, l-diversity

1. INTRODUCTION

In recent years, numerous organizations have possessed databases containing personal information that was obtained by crowdsensing and other sources for various purposes. A database containing personal information includes an identifier, quasi-identifiers (QIDs), sensitive attributes, and other attributes. Identifiers are attributes that can uniquely identify individuals (e.g., names and telephone numbers). QIDs are attributes that can identify individuals by combining individuals' information (e.g., gender, age, and ZIP code). Sensitive attributes are types of information that individuals do not want to be disclosed (e.g., annual income and health status); anonymization is used to protect sensitive attributes.

Many organizations want to analyze personal information while realizing the importance of personal privacy protection by anonymizing personal information according to existing indicators, such as k-anonymity [2], [6]. k-anonymity ensures that there are k or more records that have the same QID values. However, k-anonymity cannot protect against attribute disclosure because the k or more records might have the same sensitive attribute values.

The l-diversity [3] indicator extended k-anonymity, and it can protect against attribute disclosure. For example, in a case in which 3-diversity is satisfied, an attacker cannot recognize which one of {HIV-2, Influenza A, or cecum} is a true, sensitive attribute value of a certain person even if the attacker knows all of that person's QID values. However, when there are multiple similar sensitive attribute values in a database, there are cases where *actual* diversity is not satisfied even if the database satisfies l-diversity. Assume that database D

satisfies 3-diversity, and an attacker gets {HIV-2, HIV-1 (M), and HIV-1 (N)} as candidates for Alice's sensitive attribute value from this database. Although database D satisfies 3-diversity, it does not take into account the similarity between the sensitive attribute values; as a result, the attacker can conclude that Alice has HIV (although the type of HIV would remain unknown).

In this research, we propose (l, d)-semantic diversity, which is an indicator that can satisfy actual diversity, and an anonymization algorithm according to (l, d)-semantic diversity. In addition, as in many existing studies, we assume in this research that a database analyst wants statistical information about the database; therefore, we propose an analysis algorithm for obtaining statistical values in addition to an anonymization algorithm.

We will explain the assumed environment in Section 2 and present existing indicators and set tasks in Section 3. Section 4 introduces the new indicators used to solve the problem. Section 5 discusses our anonymization algorithm. Section 6 shows the experimental verification of analysis algorithms. The results of our simulations are presented in Section 7. Finally, Section 8 concludes the paper. This article is an extended version of [26].

2. ASSUMED ENVIRONMENT

We assume that a database contains personal information; therefore, the database needs privacy protection. Databases include identifiers, QIDs, sensitive attributes, and other attributes. Hereafter, the database described in this paper is a database containing personal information.

An organization holding such a database is called a "data holder" and the data publisher wants to share it with external "data users." However, external data users are not necessarily reliable. Data users can be attackers who are honest-but-curious; that is, they follow the proposed protocol but try to analyze individual information from published databases.

Therefore, data holders should anonymize databases to protect privacy from attackers. Anonymizing is a method that can preserve privacy while maintain-

ing usefulness of databases, and it processes databases according to privacy indicators, such as k-anonymity and l-diversity.

Data users want to analyze what kinds of sensitive attribute values a person with a specific QID holds. In other words, it is assumed that when data users receive a database containing personal information that they create multidimensional histograms (also called contingency tables or cross-tabulations), for example, and analyze them.

We assume that a sensitive attribute has distinct values because a data user's purpose in this paper is to generate multidimensional histograms. For example, numerical values, such as income, are expressed by \$20K-\$30K, \$30K-\$40K, and so on. In this research, it is assumed that a data holder collects data by crowdsensing. Fig. 1 shows a flow in which a data holder anonymizes data before collecting it and creates a database, then analyzes the data by creating a histogram from the anonymized database.

The attacker can see the created anonymized records or databases using our proposed algorithms. The attacker might have prior knowledge about a certain person's QID values, and may then try to identify the person's sensitive attribute value.

Anonymous communication systems (ACS) can protect communications between entities from traffic analysis by providing unidentifiability and unlinkability [23]. By using ACS such as [24] and [25], we assume that sender identity can be protected in this research.

3. RELATED WORK

3.1. k-*anonymity*

One of the privacy indicators widely used for database anonymization is k-anonymity [2]. A database containing certain personal information is defined as satisfying k-anonymity if there are at least k records with the same combination of QID values. As a result, even if an attacker knows all the QID values held

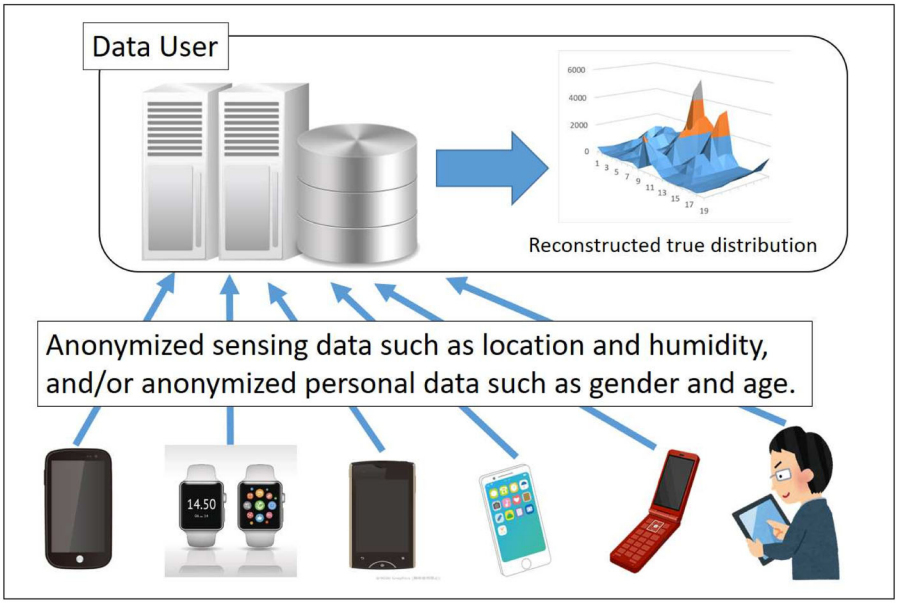


Figure 1: A scenario of anonymized data collection and analysis.

by a certain person, the individual cannot be identified since there are at least k records with the same combinations of QID values.

Generalization is commonly used as an anonymization method for realizing k-anonymity. The database represented by Table 2 is anonymized based on generalization from the disease database represented by Table 1. In Table 2, the QID values of A and B, the QID values of C and D, and the QID values of E, F, and G are the same, respectively. Because at least 2 records have the same QID values, Table 2 satisfies 2-anonymity.

However, there is a problem with k-anonymity. For example, the equivalence class of C and D in Table 2 have only "HIV-1(M)" as the sensitive attribute values.

k-Anonymity focuses on QIDs, and sensitive attributes are not considered. This problem can be particularly serious if there is a correlation between the QID and the sensitive attribute.

Table 1: Disease database.

Name	Gender	Age	ZIP code	Disease
Alice	Female	24	999-4565	Vocal folds polyp
Becky	Female	16	999-5636	HIV-2
Catriona	Female	21	999-4557	HIV-1(M)
Daisy	Female	22	999-4531	HIV-1(M)
Eddy	Male	34	999-1332	Tooth decay
Fred	Male	34	999-1335	Cecum
Gabriel	Male	48	999-1337	Gastric ulcer

3.2. l-*diversity*

One of the indicators that have extended k-anonymity is l-diversity [3]. A database containing personal information is defined as satisfying l-diversity when a certain equivalence class (a set of records that have the same QID values) holds at least l kinds of sensitive attribute values. By satisfying l-diversity, the diversity of sensitive attribute values held by a equivalence class is guaranteed, and it is possible to solve the aforementioned problem of k-anonymity. l-diversity has been widely studied in privacy-preserving data mining, such as [3], [15], and [16].

Table 3 is a database in which the database of Table 1 is anonymized to satisfy 2-diversity because every equivalence class has at least two kinds of sensitive attribute values. This makes it possible to disguise an individual's sensitive attribute value.

An attacker who knows Becky's QID values cannot decide whether Becky's disease is HIV-2 or HIV-1(M), according to Table 3. However, the attacker can conclude that Becky has HIV, although the specific type (HIV-2 or HIV-1(M)) is unknown, as described in Section 3.3.

Table 2: Disease database satisfying 2-anonymity.

Pseudonyms	Gender	Age	ZIP code	Disease
A	Female	$10\cdots29$	999-****	Vocal folds polyp
B	Female	$10\cdots29$	999-****	HIV-2
C	Female	2*	999-45**	HIV-1(M)
D	Female	2*	999-45**	HIV-1(M)
E	Male	$30\cdots49$	999-133*	Tooth decay
F	Male	$30\cdots49$	999-133*	Cecum
G	Male	$30\cdots49$	999-133*	Gastric ulcer

Table 3: Disease database satisfying 2-diversity.

Pseudonyms	Gender	Age	ZIP code	Disease
A	Female	2*	999-45**	Vocal folds polyp
B	Female	$10\cdots29$	999-****	HIV-2
C	Female	$10\cdots29$	999-****	HIV-1(M)
D	Female	2*	999-45**	HIV-1(M)
E	Male	$30\cdots49$	999-133*	Tooth decay
F	Male	$30\cdots49$	999-133*	Cecum
G	Male	$30\cdots49$	999-133*	Gastric ulcer

3.3. The problem of l*-diversity*

Although l-diversity solves the problem of k-diversity, l-diversity is still problematic in some cases. When there are several semantically similar values in a sensitive attribute, there is a possibility that actual diversity is not satisfied even if anonymization is performed to satisfy l-diversity.

Table 3 is a database that satisfies 2-diversity. However, the sensitive attribute values of the equivalence class of B and C are "HIV-2" and "HIV-1(M)." It certainly has two types of sensitive attribute values, so it can be said that privacy is protected by l-diversity. However, a data user can conclude that the

people in the equivalence class have HIV with 100% probability, although the detailed type of HIV would remain unknown.

If we use l-diversity, the candidates of a person's sensitive attribute value have at least l values. However, there might be several l values that are semantically similar. In the worst case, all these values could be semantically similar. In this case, the data user can semantically identify the person's sensitive attribute value with very high probability. We define this case as a "situation where actual diversity is not satisfied," and the purpose of this research is to solve this problem. In other words, if the set of l values satisfies actual diversity, the values are not just different but semantically different.

If similar sensitive attribute values are present in the database, as shown in Table 3, it is impossible to satisfy actual diversity without considering the similarity of each sensitive attribute value in the database; unfortunately, l-diversity cannot accomplish this at all.

3.4. (l,e)*-diversity*

Haiyuan et al. [17] proposed that (l, e)-diversity is an indicator that extends l-diversity and is focused on actual diversity, as in our research. The parameter e of (l, e)-diversity controls the degree of the actual diversity. (l, e)-diversity assumes that the sensitive attribute values can be expressed as a tree structure, and the parameter e defines the depth that the values are considered to be semantically similar.

Example 1 (Tree structure of numerical attributes for (l, e)-diversity)**.** *Fig. 2 shows an example of a tree structure of sensitive attribute values. When we set e as three, the nodes where the depth is three are "Less than $20K," "$20K-$30K," "$30K-$50K," "$50K-$70K," "$70K-$80K," "$80K-$90K," "$90K-$100K," and "More than $100K." Therefore, we consider $20K and $25K to be semantically similar, but $20K and $40K are semantically different because $20K and $25K are grouped into the same category "$20K-$30K," but $20K and $40K are grouped into the different categories ("$20K-$30K" and "$30K-$40K," re-*

spectively.) (l, e)*-diversity ensures that every equivalence class has at least* l *semantically different values.*

For example, when we set e *as one, there are two categories: "Less than \$70K" and "More than \$70K." In this example, \$20K and \$40K are considered to be semantically similar values. Therefore, the smaller the parameter* e*, the better the privacy is protected.*

To satisfy (l, e)-diversity, the authors used Anatomy [12] as an anonymizing algorithm. Anatomy is a privacy-preserving method proposed by Xiao et al. that groups records so that the sensitive attribute values are l or more types, and it divides the QID table and the sensitive attribute table into different tables. The QID table and the sensitive attribute table have a common group ID so that it is possible to link both tables while satisfying l-diversity.

Tables 4 and 5 show examples in which Table 1 was processed by Anatomy to satisfy 2-diversity. Even if an attacker refers to two tables to estimate the sensitive attribute values of record A, since the sensitive attribute values corresponding to the group ID of A are vocal cord polyps and HIV-1(M), an attacker cannot create a unique estimate.

Table 4: Disease database satisfying 2-diversity (QID table).

Pseudonyms	Gender	Age	ZIP code	Group ID
A	Female	2*	999-45**	1
B	Female	$10\cdots29$	999-****	2
C	Female	$10\cdots29$	999-****	2
D	Female	2*	999-45**	1
E	Male	$30\cdots50$	999-133*	3
F	Male	$30\cdots50$	999-133*	3
G	Male	$30\cdots50$	999-133*	3

Haiyuan et al. consider actual diversity by using the unique parameter e when executing Anatomy.

Table 5: Disease database satisfying 2-diversity (sensitive attribute table).

Disease	Group ID
Vocal folds polyp	1
HIV-2	2
HIV-1(M)	2
HIV-1(M)	1
Tooth decay	3
Cecum	3
Gastric ulcer	3

However, (l, e)-diversity only supports sensitive attributes that can be categorized into tree structures. Also, the anonymizing algorithm of (l, e)-diversity only supports tree structures.

Therefore, if the distance of the sensitive attribute values is not suitable to be represented by a tree structure, (l, e)-diversity should not be used. We can use (l, e)-diversity for this tree structure, as shown in Example 1. However, this poses a privacy issue.

When we set e as one (i.e., the privacy protection level is at its maximum), (l, e)-diversity considers "\$60K-\$70K" and "\$70K-\$80K" to be semantically very different, although the difference between "\$60K-\$70K" and "\$70K-\$80K" is actually very small. We cannot avoid this problem as long as we use a tree structure for numerical attributes. Again, note that the anonymizing algorithm of (l, e)-diversity supports only tree structures. This means that we cannot ensure actual diversity by using (l, e)-diversity for numerical attributes.

3.5. t-*closeness*

In addition, t-closeness [4] exists as an extension of l-diversity. A database satisfying t-closeness means that the distance between the distribution of the attribute values of the entire database and the distribution of the attribute values within the equivalence class is less than t. By using this indicator, better pri-

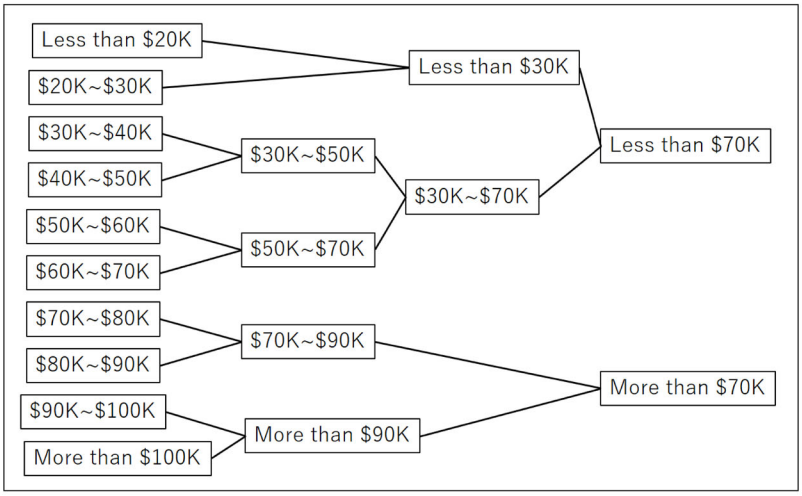


Figure 2: An example of classifying annual income into a tree structure.

vacy protection can be realized. However, t-closeness cannot solve the problem of l-diversity shown above. For example, if the database holds information for a fairly large number of HIV patients, then the above problem tends to occur. Conversely, in the case of a database in which few HIV patients are present, the equivalence class that contains even one HIV patient needs to be a large one because the distance between the distribution of attribute values of the entire database increases. Almost all of the information on an attribute value with a small frequency of appearance is lost, and the information on the record belonging to the same equivalence class is also greatly reduced.

3.6. Differential privacy

Differential privacy [8],[9],[10] exists as a privacy indicator that is different from k-anonymity, l-diversity, and t-closeness as they are described above. Differential privacy is defined by the assumption that data analyzers issue queries to a data holder and that the data holder anonymizes the answers.

This indicator is able to use crowdsensing as a data-collecting method, and the data holder does not share the anonymized database, but receives data

queries one by one while keeping the database. This assumption has merit, as the data analyzers can obtain only the desired information; however, a disadvantage exists in that there is a high cost to the holder, and the data analyzers cannot freely use the database. Considering the above disadvantages, it is assumed in this research that the data analyzers are requesting the anonymized data.

4. PROPOSED INDICATORS

In this section, we propose (l, d)-semantic diversity as an indicator that considers actual diversity that cannot be dealt with by existing indicators, such as l-diversity, t-closeness, and (l, e)-diversity.

4.1. Symbol definition

Let T be a database containing personal information, and let the i-th record of T be r_i; here $i = 1, \ldots, N$.

Let S be a domain of a sensitive attribute in T, and let F be the number of elements in S. Each value of S is expressed by $v_1, \ldots, v_F$. Furthermore, the sensitive attribute value of record r is expressed as $E(r)$.

For example, in Table 1 r_1 represents the first record, i.e., the record of [Alice, Female, 24, 999-4565, Vocal folds polyp], and r_2 represents the second record, i.e., the record of [Becky, Female, 16, 999-5636, HIV-2], and so on. The attribute Disease is a sensitive attribute in Table 1; therefore, S is the set of values of Disease without duplication, i.e., $S =$ {Volcal folds polyp, HIV-2, HIV-1(M), Tooth decay, Cecum, Gastric ulcer}. Because the number of elements of S is 6, $F = 6$. The symbols $v_1, \ldots, v_6$ represent Vocal folds polyp, HIV-2, HIV-1(M), Tooth decay, Cecum, Gastric ulcer, respectively. $E(r_1)$ is Vocal folds polyp, $E(r_2)$ is HIV-2, $E(r_3)$ is HIV-1(M), and so on.

Table 6 shows the symbols.

Table 6: Symbols.

Symbol	Description
T	Database containing personal information
N	Number of records in T
r_i	ith record of T
S	Domain of a sensitive attribute in T
F	Number of elements of S (i.e., $F = \|S\|$)
v_i	ith value in S
$E(r)$	Sensitive attribute value of record r in T

4.2. (l, d)-semantic diversity

Here, we show the definition of (l, d)-semantic diversity, which is a proposal indicator for problem-solving.

Definition 1 ((l, d)-semantic diversity):

Let T represent the database containing personal information.

For natural numbers l and d, if database T satisfies the following, then database T satisfies (l, d)-semantic diversity. Every equivalence class has at least l sensitive attribute values, and the minimum distance between the values is larger than or equal to d.

By utilizing (l, d)-semantic diversity as an anonymization indicator, it becomes possible to consider the distance between the sensitive attribute values, which could not have been considered in l-diversity.

The parameter d controls the degree of actual diversity. Thus, (l, d)-semantic diversity requires that every combination of distance between sensitive attribute values is defined. The distance does not need to be expressed by a tree structure, whereas (l, e)-diversity requires this be done. For example, if a sensitive attribute is represented by a numerical value, then the distance can be defined based on the absolute difference between values.

Example 2 (Absolute value-based distance of the numerical attributes for

(l, d)-semantic diversity). *Assume that the sensitive attribute is annual income, and the values are* $v_1, \ldots, v_{10}$ = *"Less than \$20K," "\$20K-\$30K," "\$30K-\$40K," "\$40K-\$50K," "\$50K-\$60K," "\$60K-\$70K," "\$70K-\$80K," "\$80K-\$90K," "\$90K-\$100K," and "More than \$100K," respectively.*

If we consider the representative values of $v_1, \ldots, v_{10}$ *to be 10K, 20K, ..., 90K, and 100K, then the semantic distance between* v_i *and* v_j *can be defined as* $|v_i - v_j|$. *In this case, the semantic distance between* v_i *and* v_{i+1} *is 10K.*

Note that the data user can define any distance as the semantic distance between v_i *and* v_j. *Therefore, the data user can also define the semantic distance* v_i *and* v_j *as* $|v_i - v_j|/10K$, *for example. This distance definition is the same as* $|i - j|$ *in this case.*

For another example, assume that the sensitive attribute is annual income, and the values are $v_1, \ldots, v_{10}$ = *"Less than \$10K," "\$20K," "\$25K," "\$30K," "\$40K," "\$45K," "\$60K," "\$80K," "\$95K," and "More than \$100K," respectively. In this case, the distance between* v_i *and* v_j *can be defined as* $|i - j|$, *for example.*

Note that in many existing studies, data users can freely define the distance functions, and this freeness is one of their advantages.

For example, t-closeness [4] is a famous privacy indicator that requires the sensitive attributes within each of the equivalence classs to have a similar distribution to their distribution in the entire database. The specific distance used between distributions is central to evaluating t-closeness, but the original definition does not advocate any specific distance [5]. Data users can use any distance definition.

Please note that this relationship cannot be expressed using a tree structure, although we can express this relationship using a semilattice structure where each node can have more than one parent (see Fig. 3).

Based on Fig. 3, we can consider that the distance between nodes under the same parent node (i.e., the distance between sibling nodes) is one. If nodes are not sibling nodes, but they are under the same grandparent node, the distance

between the nodes is two, and so on.

Please note that the distance definition is just an example in this paper. The data holder can use different distance definitions. (l, d)-semantic diversity can be used for arbitrary distance definitions, and this is one of the advantages of our proposed algorithm.

Again, (l, e)-diversity and its anonymizing algorithm can only be used for a tree structure.

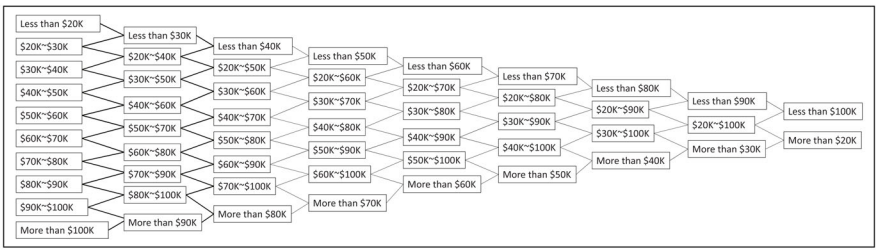


Figure 3: Semilattice structure for a numerical attribute.

In addition, if the categorization is a tree structure (e.g., disease), then it can also be defined by the depth of the tree of structure, such as with (l, e)-diversity.

Example 3 (Tree structure-based distance of the categorical attributes for (l, d)-semantic diversity)**.** *Fig. 2 shows an example of a tree structure of sensitive attribute values. The distance between "$30K-$40K" and "$40K-$50K" is one because they are sibling nodes, the distance between "$30K-$40K" and "$50K-$60K" is two, the distance between "$30K-$40K" and "$70K-$80K" is three, and so on.*

5. PROPOSED ALGORITHM

In this paper, we propose an analysis algorithm as well as two anonymization algorithms (a simple anonymization algorithm and a duplicate-processing anonymization algorithm). The duplicate-processing anonymization algorithm, which extends the simple anonymization algorithm, is specifically for numerical attributes.

Our proposed anonymization algorithm adds $l-1$ dummy records to each true record. Therefore, the anonymization algorithm outputs $N \times l$ records when there are N true records. We call the $N \times l$ records an anonymized database. Because the anonymization algorithm adds many random dummy records, data users cannot understand the anoymized database in a straightforward way. Therefore, we propose an analysis algorithm and the anonymization algorithm. The analysis algorithm's objective is to analyze the anonymized database generated by our proposed anonymization algorithm. The output of the analysis algorithm is an estimated histogram of the true N records. Note that the data users know the contents of the anonymized database but do not know the true N records.

5.1. Simple anonymization algorithm

Here, an anonymization algorithm that satisfies the (l, d)-semantic diversity defined above is shown. In this research, we do not use generalization, which is the basic method of anonymization, but instead use a method inspired by Sei et al. [1] to add several dummy records.

Table 7 is a database that has satisfied (2,4)-semantic diversity by adding a dummy record to each record in Table 1. Table 7 shows only one part of the anonymization result. Gender, age, and zip code are QIDs, and disease is a sensitive attribute value. To define the distance between sensitive attribute values, we use the tree structure used by the International Classification of Diseases (ICD[1]) and maintained by the World Health Organization.

We omit the full distance definitions of the disease values, as the distance of vocal fold polyps and HIV-2, the distance of HIV-2 and tooth decay, and the distance of HIV-1(M) and cecum are more than or equal to four according to ICD.

Adding dummy records makes it impossible to estimate individual records uniquely. Therefore, privacy can be claimed to be preserved because an attacker

[1] https://www.who.int/classifications/icd/en/

Table 7: Disease database satisfying (2,4)-semantic diversity

Pseudonyms	Gender	Age	ZIP code	Disease
A	Female	24	999-4565	Vocal folds polyp
A'	Female	24	999-4565	HIV-2
B	Female	16	999-5636	HIV-2
B'	Female	16	999-5636	Tooth decay
C	Female	21	999-4557	HIV-1(M)
C'	Female	21	999-4557	Cecum
⋮	⋮	⋮	⋮	⋮

wants to obtain one of the sensitive attribute values of the individual. Even if the attacker knows all of the QID values of A, he cannot know whether vocal fold polyps or HIV-2 is A's sensitive attribute value.

The proposed anonymization algorithm selects and adds $(l-1)$ sensitive attribute values that can satisfy (l, d)-semantic diversity. This anonymization algorithm is shown in Algorithm 1.

Algorithm 1 Anonymizing algorithm for database T

Input: Database T, domain of a sensitive attribute S, privacy parameters l and d
Output: Anonymized database T'
Create set T'
for $i = 1, \cdots, N$ **do**
  $Q \Leftarrow$ QID values of r_i
    /* Adds original sensitive attribute value */
  Create Set $R \Leftarrow \{E(r)\}$
    /* Adds dummy sensitive attribute values */
  for $j = 1, \cdots, (l-1)$ **do**
    $R \Leftarrow R \cup rand(extract(S, R, d))$
  end for
  $T' \Leftarrow T' \cup generate\_records(Q, R)$
end for
$shuffle(T')$
return T'

In Algorithm 1, the function $generate\_records(Q, R)$ generates $|R|$ records, where the QID values are Q, and the sensitive attribute value is one of R. For example, when Q is {Female, 24, 999-4565} and R is {Tooth decay, HIV-1(M), Cecum}, the function generates three records: [Female, 24, 999-4565, Tooth decay], [Female, 24, 999-4565, HIV-1(M)], and [Female, 24, 999-4565, Cecum].

The function $shuffle$ randomly shuffles the set. Dummy records thus consist of existing records' QID values and sensitive attribute values that are selected as dummies.

The function $extract(S, R, d)$ generates the set $\{e | e \in S \wedge v_j \in R \wedge dist(e, v_j) \geq d$ for all $j = 1, \ldots, |R|\}$, where $dist(e, v)$ represents the distance between sensitive attribute values e and v. The function $rand(B)$ randomly extracts an element from set B.

For example, suppose r_i represents the QID and sensitive attribute values of

the first record in Table 1, i.e., r_i is [Female, 24, 999-4565, Vocal folds polyp], and suppose $l = 2$. Q is [Female, 24, 999-4565] in line 5 in Algorithm 1. R is {vocal folds polyp} in line 7. In lines 9–11, we choose $l - 1$ random sensitive values so that the semantic distances between the chosen elements and the sensitive attribute value ("Vocal folds polyp" in this example) are greater than or equal to d. For example, Algorithm 1 chooses "HIV-2." In this case, R becomes {Vocal folds polyp, HIV-2}. Then, we have the set of records {[Female, 24, 999-4565, Vocal folds polyp], [Female, 24, 999-4565, HIV-2]} from the $generate\_records(Q, R)$ function. The set is added to T'. We repeat this process for all records. Finally, we shuffle all records of T'.

In this research, we assume that an attacker is honest but curious; that is, the attacker follows the proposed protocol but tries to analyze the individual information from published databases. This assumption is very common in research about privacy-preserving data mining. The attacker can see the anonymized records or the anonymized databases created by our proposed algorithms and might have prior knowledge about a certain person's QID values. The attacker thus tries to identify the person's sensitive attribute value. In existing studies for l-diversity, the attacker can conclude that the person has HIV (although the type of HIV would remain unknown); however, our proposed algorithms can even prevent this from happening.

5.2. Analysis algorithm

Let x_i and ω_i be the numbers of records that have v_i as their sensitive attribute values in the anonymized database and in the original database, respectively. The value of x_i is unknown for the data analyzers, and the purpose is to estimate the value of x_i with high accuracy.

By obtaining the x_i of Equation 1,

$$\omega_i = x_i + \sum_{k \neq i} q_{(k,i)} * x_k, \tag{1}$$

the total number of certain sensitive attribute values in the anonymized database is made up of the true values and the dummy values. The number of dummy

values is obtained as the expected value using Equation 1.

Here, $q_{(i,j)}$ is the probability of selecting v_j as a dummy when a certain record holds the sensitive attribute value v_i and is represented by Equation 2:

$$q_{(i,j)} = \begin{cases} \frac{l-1}{|extract(S,\{v_i\},d)|} & (v_j \in extract(S,\{v_i\},d)) \\ 0 & (otherwise). \end{cases} \tag{2}$$

The analysis algorithm is shown in Algorithm 2.

Algorithm 2 Analysis protocol

Input: Reported sets, domain size of sensitive attribute values F, database size N, parameters l and d

Output: Estimated distribution of sensitive attribute values

for $i = 1, \cdots, F$ **do**

  $F_i \Leftarrow$ size of $extract(S, \{v_i\}, d)$

  if $v_j \in dist(d, v_i)$ **then**

    $q_{(i,j)} \Leftarrow (l-1)/F_i$

  else

    $q_{(i,j)} \Leftarrow 0$

  end if

end for

$\hat{x_i} \leftarrow gauss(\omega_i = x_i + \sum_{k \neq j} q_{(k,i)} * x_k)$ for all i

  /* Calculate Simultaneous Equation */

return $\hat{x_i}(i = 1, \cdots, F)$

Here, the function $gauss(f(x))$ is a function for solving simultaneous equations.

For a simple discussion, we use a very simple example.

Let gender ("male" and "female") and age ("over 50" and "under 50") are the QIDs; obesity level is a sensitive attribute; and the values of the level are 1 (very low obesity), 2 (low obesity), 3 (normal), 4 (high obesity), or 5 (very high obesity). Therefore, $v_1, \ldots, v_5 = 1$ (very low obesity), ..., 5 (very high

obesity). In this example, the difference of their subscriptions ($|i - j|$) and the difference of the values ($|v_i - v_j|$) are the same, and we use these differences as the semantic distance. For example, the semantic distance between v_1 (obesity level 1) and v_3 (obesity level 3) is 2. Suppose that there are 1,000 participants. Table 8 shows the cross-tabulation of the true data. The elements of the cross-tabulation are the combination of {"male", "female"}, {"over 50", "under 50"}, and {1, 2, 3, 4, 5}. For example, 120 people are male, are aged over 50, and have obesity level 1.

When $l = 2$, each person reports two sensitive values from the basic anonymization algorithm and the duplicate-processing anonymization algorithm. One of the two sensitive values is the true value, and the other is a dummy. For example, if a person's true sensitive value is 1 (very low obesity), the person reports two values; one is 1 and the other is 2, 3, 4, or 5.

Table 9 shows the anonymized results of Table 8, where $l = 2$ and $d = 2$. For example, the reports from 143 people who are male and over 50 years old contain 1 (very low obesity). Note that the QID values are collected without change. Therefore, the total numbers of each row of Table 9 (460, 400, 760, and 380) are l times larger than the total numbers of each row of Table 8 (230, 200, 380, and 190).

We executed Algorithm 2 for each combination of QID values. In this example, we independently executed Algorithm 2 for ["male," "over 50"], ["male," "under 50"], ["female," "over 50"], and ["female," "under 50"].

Here, let us calculate the estimated value of each element for "male" and "over 50" of the histogram, that is, the estimated value of ["male," "over 50," "1" (very low obesity)], ["male," "over 50," "2" (low obesity)], ["male," "over 50," "3" (normal)], ["male," "over 50," "4" (high obesity)], and ["male," "over 50," "5" (very high obesity)].

In this case, $F = 5$ and $N = 230$. x_i represents the number of records that have v_i as their sensitive attribute values, "male" as their gender, and "over 50" as their age in N true records; and ω_i represents the numer of records that have v_i as their sensitive attribute values, "male" as their gender, and "over 50" as

their age in the anonymized database.

Table 8: Cross-tabulation of the true information, which is unknown.

	Obesity level					
Gender & Age	1	2	3	4	5	Total
Male & over 50	120	50	10	20	30	230
Female & over 50	80	60	30	10	20	200
Male & under 50	50	30	110	100	90	380
Female & under 50	20	20	40	40	70	190

Table 9: Cross-tabulation of the reported information, which is known to the data user.

	Obesity level					
Gender & Age	1	2	3	4	5	Total
Male & over 50	143	72	60	86	99	460
Female & over 50	108	70	65	64	93	400
Male & under 50	188	108	155	130	179	760
Female & under 50	83	62	73	57	105	380

Threrefore, the values of $x_1, \ldots, x_5$ are 120, 50, 10, 20, and 30, respectively, according to Table 8, because we are considering the combination of "male" and "over 50." Note that the data user does not know these values and that his or her aim is to estimate these values with high accuracy. The values of $\omega_1, \ldots, \omega_5$ are 143, 72, 60, 86, and 99, respectively, according to Table 9.

The set of $extract(S, v_i, 2)$ is $\{3, 4, 5\}$ for $i = 1$ because the distance between 1, and 3, 4, or 5 is greater than or equal to two, but the distance between 1 and 2 is only one. In the same way, $extract(S, v_i, 2)$ for $i = 2, \ldots, 5$ is $\{4, 5\}$, $\{1, 5\}$, $\{1, 2\}$, and $\{1, 2, 3\}$.

Therefore, the values of $q_{(2,4)}$, $q_{(2,5)}$, $q_{(3,1)}$, $q_{(3,5)}$, $q_{(4,1)}$, $q_{(4,2)}$ are $1/2$; the values of $q_{(1,3)}$, $q_{(1,4)}$, $q_{(1,5)}$, $q_{(5,1)}$, $q_{(5,2)}$, $q_{(5,3)}$ are $1/3$; and the $q_{(i,j)}$ values of

other combinations of i, j are 0.

The data user constructs the following five equations according to Algorithm 2:

$$\begin{cases} \omega_1 = x_1 + (q_{(2,1)} * x_2 + q_{(3,1)} * x_3 + q_{(4,1)} * x_4 + q_{(5,1)} * x_5) \\ \omega_2 = x_2 + (q_{(1,2)} * x_1 + q_{(3,2)} * x_3 + q_{(4,2)} * x_4 + q_{(5,2)} * x_5) \\ \omega_3 = x_3 + (q_{(1,3)} * x_1 + q_{(2,3)} * x_2 + q_{(4,3)} * x_4 + q_{(5,3)} * x_5) \\ \omega_4 = x_4 + (q_{(1,4)} * x_1 + q_{(2,4)} * x_2 + q_{(3,4)} * x_3 + q_{(5,4)} * x_5) \\ \omega_5 = x_5 + (q_{(1,5)} * x_1 + q_{(2,5)} * x_2 + q_{(3,5)} * x_3 + q_{(4,5)} * x_4). \end{cases} \tag{3}$$

In this example, Equation 3 can be written as the following equations based on Table 9 and each value of $q_{(i,j)}$:

$$\begin{cases} 143 = x_1 + (x_3/2 + x_4/2 + x_5/3) \\ 72 = x_2 + (x_4/2 + x_5/3) \\ 60 = x_3 + (x_1/3 + x_5/3) \\ 86 = x_4 + (x_1/3 + x_2/2) \\ 99 = x_5 + (x_1/3 + x_2/2 + x_3/2). \end{cases} \tag{4}$$

By solving Equation 4, the data user obtains $\hat{x_1}, \ldots, \hat{x_5} = 117.8, 52.5, 11.5, 20.5, 27.8$ where $\hat{x_i}$ represents the estimated value of x_i. In this example, we can say that the estimated values are close to the true values $(x_1, \ldots, x_5 = 120, 50, 10, 20, 30)$.

In the same way, by executing Algorithm 2 for other combinations of QID values (that is, [“female”, “over 50”], [“male”, “under 50”], and [“female”, “under 50”]), the data user finally gets the estimated cross-tabulation (Table 10). From Table 10, we can generate an equivalent histogram, if needed.

5.3. Duplicate-processing anonymization algorithm for numerical attributes

Next, we propose a duplicate-processing anonymization algorithm, which is an anonymizing algorithm that can obtain better results based on the simple anonymization algorithm described in Section 5.1. In this subsection, we assume that the sensitive attribute is a totally ordered set, such as the numerical attribute described in Section 4.2.

Table 10: Cross-tabulation estimated by the data user (each total value may not exactly agree with the sum of values of each row due to rounding.)

Gender & Age	Obesity level					Total
	1	2	3	4	5	
Male & over 50	117.8	52.5	11.5	20.5	27.8	230
Female & over 50	82.9	60.3	30.8	6.3	19.9	200
Male & under 50	55.1	27.8	105.3	97.8	94.1	380
Female & under 50	16.9	18.3	44.8	42.3	67.9	190

The main structure of the duplicate-processing anonymization algorithm is the same as Algorithm 1 of the simple anonymization algorithm. The only difference between it and the simple anonymization algorithm is the part of the function that is used for dummy selection.

If the sensitive attribute is a numeric type, it is possible that the probability of selecting a dummy represented by Equation 2 differs for each record when adding two or more dummy records using the simple anonymization algorithm. Thus, there are cases where good results cannot be obtained when applying the above analysis algorithm if we use the simple anonymization algorithm.

For example, consider a case where the sensitive attribute is annual income, $F = 9$, $S = \{\$30\text{K}, \cdots, \$110\text{K}\}$, $d = 2$, and $l = 3$, as shown in Fig. 4. In this example, consider that the semantic distance between v_i and v_j is $|v_i - v_j|/10K$ (which is the same as $|i - j|$ in this case). Assume that a true sensitive attribute of a certain record is \$50K. The simple anonymization algorithm selects two dummies from \$30K, \$70K, \$80K, \$90K, \$100K, and \$110K while keeping (3,2)-semantic diversity. This is because the distance between \$50K and these elements is greater than or equal to 2.

The probability that each value will be selected as a dummy is $2/6 = 1/3$.

This value is the same as the value calculated using Equation 2. [2]

In other words, if the true value is \$50K, then the simple anonymization algorithm will exclude \$40K, \$50K, and \$60K, which can be selected as dummies. More specifically, the simple anonymization algorithm excludes the true value and $2d$ elements which are not d away from the true value for selection of the first (or later) dummies. That is, the simple anonymization algorithm excludes $2d + 1$ elements for the selection of the first or later dummies based on the true value.

Suppose that \$80K is selected as the first dummy (Fig. 4(a)). Then, the simple anonymization algorithm selects a second dummy. The distance between the second dummy and \$50K (which is the true value) and the distance between the second dummy and \$80K (which is the first dummy) should be greater than or equal to 2 because we set $d = 2$. Only \$30K, \$100K, and \$110K satisfy this condition. In other words, if the first dummy is \$80K, the simple anonymization algorithm will exclude \$70K, \$80K, and \$90K from being selected for second (or later) dummies. More specifically, the simple anonymization algorithm excludes the first dummy and $2d$ elements that are not d away from the first dummy from being selected as second (or later) dummies. That is, the simple anonymization algorithm excludes $2d + 1$ elements from selection as second or later dummies based on the first dummy. The simple anonymization algorithm randomly selects one element from these three elements. That is, each element of these elements is selected with a 1/3 probability. This probability is the same as the probability calculated based on Equation 2.

On the other hand, assume that \$70K is selected as the first dummy (see Fig. 4(b)). In this case, the simple anonymization algorithm excludes \$40K,

[2]Here, let us check the value of Equation 2. We set $l = 3$, $d = 2$, S ={\$30K, ..., \$110K}, and the true value is \$50K in this example. The result of $extract(S$,{\$50K}, 2) is the set {\$30K, \$70K, \$80K, \$90K, \$100K, \$110K}. Therefore, the size of the set is 6. The probability that each element of the set will be selected as a dummy when the true value is \$50K is $(3 - 1)/6 = 1/3$.

$50K, and $60K from selection as the first or later dummies according to the true value of $50K, and excludes $60K, $70K, and $80K from selection as the second or later dummies according to the first dummy, $70K. Here, note that the element $60K appears in both the first exclusion list ($40K, $50K, and $60K) and the second exclusion list ($60K, $70K, and $80K). Therefore, the second dummy will not be selected from three elements but from four elements, which are $30K, $90K, $100K, and $110K, each with a 1/4 probability of being selected.

For another example, assume that $110K is selected as the first dummy (see Fig. 4(d)). In this case, the second dummy will be selected from $30K, $70K, $80K, and $90K, each with a 1/4 probability of being selected. In each case, the selection probability (1/4) is less than the value (1/3) calculated using Equation 2. As a result, the estimated values using the simple anonymization algorithm might be different and far from the true values.

Therefore, we propose a duplicate-processing anonymization algorithm, which modifies the simple anonymization algorithm. The simple anonymization algorithm excludes exactly $2d + 1$ elements that can be selected using the following steps. For example, assume that $50K is the true value and that $70K is selected as the first dummy. In this situation, the duplicate-processing anonymization anonymization algorithm excludes not only $40K, . . ., $80K, but $30K for selection of second or later dummies. As a result, the second dummy will be selected from $90K, $100K, and $110K, each with a 1/3 probability of being selected (see Fig. 4(c)). This value is the same as the value calculated using Equation 2.

As another example, assume that $110K is selected as the first dummy. The duplicate-processing anonymization algorithm excludes not only $40K, $50K, $60K, $100K, and $110K, but $30K for selection of second or later dummies (see Fig. 4(e)).

In Algorithm 1 of the simple anonymization algorithm, "*extract*" was used as a function to select a dummy, but in the duplicate-processing anonymization algorithm, a more complicated function is used. The details of the more complex

function are shown as Algorithm 3. The duplicate-processing anonymization algorithm replaces the "*extract*" of Algorithm 1 with Algorithm 3.

Algorithm 3 Duplicate-processing algorithm for anonymization algorithm

Input: Domain of a sensitive attribute S, Privacy level d, Set of the true value and the temporary selected dummies E

Output: A set of sensitive attribute values that can be selected

```
Create set R
Create value D
D ⇐ d
for e_i ∈ E do
  for j = 1, ··· , |S| do
    /* Confirm whether the distance between v_j and e_i is within d */
    if i − d ≤ j ≤ i + D OR i − d + |S| ≤ j OR j ≤ i + D − |S| then
      /*Confirm that a_j is already in R*/
      if a_j ∈ R then
          /* Update the value of D */
          D ⇐ (D + 1)
      else
        /* Adds values of sensitive attribute to be excluded */
          R ⇐ R ∪ a_j
      end if
    end if
  end for
end for
return S \ R
```

5.4. Mathematical Analysis and Implementation Technique

We analyze the mathematical property for a numerical-sensitive attribute, and we assume the attribute is a totally ordered set. The property for categorical-sensitive attributes depends heavily on the tree structures of the attribute values;

therefore, it is difficult to discuss the mathematical property in general.

First, we show that our proposed algorithms (both the basic anonymization algorithm and the duplicate-processing anonymization algorithm) always achieve (l, d)-semantic diversity (if and only if $l \cdot d \geq F$). Algorithm 1 is a logical algorithm, and we modify the algorithm in this subsection to always achieve (l, d)-semantic diversity for real implementation.

Let v_i be the ith element of S. We define

$$D(v_i, v_j) = \begin{cases} dist(v_i, v_j) & (i < j) \\ dist\_max(S) - dist(v_i, v_j) + 1 & (otherwise,) \end{cases} \tag{5}$$

where $dist\_max(S)$ represents the semantic distance between the maximum value and the minimum value of S.

For example, consider that we use annual income as the sensitive attribute described in Example 2, and consider that the semantic distance between v_i and v_j is $|i - j|$. In this case, Equation 5 is represented by

$$D(v_i, v_j) = \begin{cases} j - i & (i < j) \\ 10 - i + j & (otherwise.) \end{cases} \tag{6}$$

In this case, D(“\$80K-\$90K”, “\$90K-\$100K”) is one, and D(“\$90K-\$100K”, “Less than \$20K”) is two.

The maximum number of dummies that can be selected from S within values of v_i and v_j is represented by

$$A_{max}(v_i, v_j) = \left\lfloor \frac{D(v_i, v_j)}{d} - 1 \right\rfloor. \tag{7}$$

For example, consider that v_2 is “\$20K–\$30K,” v_8 is “\$80K–\$90K,” and $d = 2$. In this case,

$$A_{max}(v_2, v_8) = \left\lfloor \frac{6}{2} - 1 \right\rfloor = 2. \tag{8}$$

Therefore, we know that we can select two elements between “\$20K–\$30K” and “\$80K–\$90K” so that the semantic distance between each element is greater than or equal to 2. Actually, we can select “\$40K–\$50K” and “\$60K–\$70K”.

Let R be a set of the true sensitive attribute values and the previously selected dummies. Let e_i represent the ith smallest value of R where i starts from zero.

The maximum number of dummies that can be selected from S is represented by

$$A_{max}(R) = \sum_{i=1}^{|R|} A_{max}(e_i, e_{i+1 \pmod{|R|}}), \tag{9}$$

where $|R|$ represents the size of R.

For example, assume that R is $\{v_1, v_3, v_9\}$. In this case, $e_0, e_1, e_2 = v_1, v_3, v_9$. $A_{max}(R)$ is $A_{max}(e_0, e_1) + A_{max}(e_1, e_2) + A_{max}(e_2, e_0) = A_{max}(v_1, v_3) + A_{max}(v_3, v_9)$ $+ A_{max}(v_9, v_1) = 2$.

To select a dummy, after line 10 in Algorithm 1, we check the value of $A_{max}(R)$. After selecting the jth dummies, if the following inequation,

$$A_{max}(R) \geq l - j - 1, \tag{10}$$

is not satisfied, then we will execute line 10 again.

For example, assume that Algorithm 1 has already selected two dummies (i.e., $j = 2$ in line 10) and that $R = \{v_1, v_3, v_9\}$, $l = 5$ and $d = 2$. If Algorithm 1 then selects v_6 in $j = 3$ in line 10, R will become $R = \{v_1, v_3, v_6, v_9\}$. In this case, $A_{max}(R)$ is 0. Therefore, Algorithm 1 will roll back the selection of v_6, and reselect a dummy. If it selects v_5, then R will become $R = \{v_1, v_3, v_5, v_9\}$. In this case, $A_{max}(R)$ is 1, and the inequality $A_{max}(R) \geq$ 5-3-1 holds. Therefore, selecting v_5 is confirmed, and Algorithm 1 goes to the next loop ($j = 4$).

Theorem 5.1. *Algorithm 1 always achieves (l, d)-semantic diversity, if and only if $l \cdot d \geq F$.*

Proof. To achieve (l, d)-semantic diversity, each distance between arbitrary values in an anonymized set should be at least d. Because the size of the anonymized set is l, the total distance should be at least $l \cdot d$. Therefore, if and only if $l \cdot d \geq F$, the algorithm could achieve (l, d)-semantic diversity. □

We consider each distance between the sensitive attribute values. On the contrary, existing methods for l-diversity do not consider this distance. We show that the state-of-the-art method [1] for l-diversity does not achieve (l, d)-semantic diversity in most cases.

Theorem 5.2. *The lower boundary of the probability ρ of violation of (l, d)-semantic diversity per record from the algorithm proposed by [1] is*

$$\rho = 1 - \frac{{}_{F+l+d-dl-2}C_{l-1}}{{}_{m-1}C_{l-1}}, \tag{11}$$

and the lower boundary of the probability S of violation of (l, d)-semantic diversity of a database that has N records from the algorithm proposed by [1] is

$$S \geq 1 - (1 - \rho)^N. \tag{12}$$

Proof. The probability of violation of (l, d)-semantic diversity that uses the algorithm proposed by [1] depends on the true sensitive attribute value. If the true sensitive attribute value is v_1 or v_F, the probability is minimal because we can select $l - 1$ dummies from $F - d$ values only in this case. Here, we will show that the probability of a violation of (l, d)-semantic diversity uses the algorithm proposed by [1] in this situation. We assume that the true sensitive attribute value is v_1 without loss of generality.

Let u_0 represent the true value, and let $u_1, u_2, \ldots, u_{l-1}$ represent the selected dummies s.t., $dist(u_i, u_j) < dist(u_i, u_j)$ if $i < j$. Let E represent the set of $u_0, \ldots, u_{l-1}$ (see Fig. 5). Fig. 5 shows a situation where the number of elements F is 10 and l is 4. For example, $dist(u_0, u_1) = 2$ and $dist(u_0, u_3) = 7$.

To achieve (l, d)-semantic diversity, each distance between arbitrary values in E should be at least d. Because the number of dummies is $l - 1$, the total distance should be at least ld. Therefore, if and only if $ld \geq F$, the algorithm could achieve (l, d)-semantic diversity.

The algorithm proposed by [1] randomly chooses $l - 1$ elements from $F - 1$ elements. Let c_{all} and $c_{achieving}$ represent the number of all combinations of choosing $l - 1$ elements and the number of combinations that achieve (l, d)-semantic diversity, respectively.

The number of $(l-1)$-combinations from a set of $F-1$ elements is represented by

$$c_{all} = {}_{F-1}C_{l-1}. \tag{13}$$

We define $\alpha_i (i = 0, \ldots, l-1)$ as

$$\alpha_i = \begin{cases} dist(u_i, u_{i+1}) - d & (i \neq l-1) \\ dist(u_i, v_F) & (otherwise.) \end{cases} \tag{14}$$

To make E achieve (l, d)-semantic diversity, the following equation should be satisfied;

$$\begin{cases} \alpha_i \geq 0 \text{ for all } i \\ \sum_{i=0}^{l-1} \alpha_i = F - (l-1)d - 1. \end{cases} \tag{15}$$

The number of combinations of α_i $(i = 0, \ldots, l-1)$ that satisfies Equation 15 is represented by

$$c_{achieving} = {}_{F-(l-1)d-1+l-1}C_{l-1}. \tag{16}$$

Therefore, the lower boundary probability of violation of (l, d)-semantic diversity for each record is

$$1 - \frac{c_{achieving}}{c_{all}} = 1 - \frac{{}_{F+l+d-dl-2}C_{l-1}}{{}_{F-1}C_{l-1}}. \tag{17}$$

Because a database has N records, the probability is raised to the Nth power. □

Fig. 6 shows the results of the mathematical analysis when considering the probability of violation of (l, d)-semantic diversity with varying F, l, and d. We know from the figure that the probability of violation of (l, d)-semantic diversity of the proposed algorithm is always zero, whereas that of the baseline algorithm increases especially when d or l is large. This is because the number of combinations that satisfy Equation 16 decreases when d or l is large, with regard to the baseline algorithm.

6. EVALUATION EXPERIMENT

We evaluated the usefulness of the proposed anonymization algorithms (the simple algorithm (Algorithm 1) and the duplicate-processing algorithm (Algorithm 1 with Algorithm 3)) and the proposed analysis algorithm (Algorithm 2).

Usefulness was evaluated by the difference between the histogram generated by the analysis algorithm and the histogram from the original database.

6.1. Evaluation method

We used the mean squared error (MSE) of Equation 18 as an indicator to evaluate the usefulness of the analysis algorithm. The MSE shows the error between the correct data and the obtained data; the smaller the value, the more similar it is to the correct data:

$$MSE = Mean_i\Big(\frac{x_i}{N} - \frac{\hat{x_i}}{N}\Big)^2, \tag{18}$$

where the function $Mean_i$ calculates the arithmetic mean.

In Equation 18, x_i is the total number of sensitive attribute values v_i in the original database, and $\hat{x_i}$ is the total number of sensitive attribute values v_i guessed by the analysis algorithm. Equation 18 evaluates the difference between the true value x_i and the estimated value $\hat{x_i}$.

For example, assume that Table 8 represents the true information and Table 10 represents the estimated information. In this case, each x_i represents each value of Table 8 except for the total values, i.e., $x_i(i = 1, \ldots, 25)$ is 120, 50, 10, 20, 30, 80, ..., 20, 20, 40, 40, 70. Each $\hat{x_i}$ represents each value of Table 10 except for the total values, i.e., $\hat{x_i}(i = 1, \ldots, 25)$ is 117.8, 52.5, 11.5, 20.5, 27.8, 82.9, ..., 16.9, 18.3, 44.8, 42.3, 67.9. Therefore, the MSE is calculated by

$$\begin{aligned} MSE &= \frac{1}{25}\Big((\frac{120}{1000} - \frac{117.8}{1000})^2 + (\frac{50}{1000} - \frac{52.5}{1000})^2 + \ldots + (\frac{70}{1000} - \frac{67.9}{1000})^2\Big) \\ &\approx 8.06 \times 10^{-6}. \end{aligned} \tag{19}$$

To show the usefulness of the proposed analysis algorithm, it is possible to compare the estimated histograms using simple analysis and analysis algorithms in the existing research, specifically [1] and [17]. Estimation of the histogram using simple analysis is calculated by Equation 20:

$$\hat{x_i} = \frac{\omega_i}{l}. \tag{20}$$

Analysis algorithms in the existing research are shown in Equations 21 and 22:

$$\hat{x_i} = \frac{-qN + \omega_i}{1 - q}. \tag{21}$$

$$q = \frac{l - 1}{F - 1}. \tag{22}$$

Equation 22 is the probability of selecting a dummy, and the proposed analysis algorithm improved this equation to suit itself.

The estimated value obtained from Equations 20 and 21 and the estimated value obtained from the analysis algorithm were compared to the value of the MSE and subsequently evaluated.

6.2. Data used for the evaluation experiment

The data used for the experiment were a created dataset and two published datasets, so that there were three kinds of datasets in total. The created dataset was an income dataset where annual income was a sensitive attribute. This dataset had 100,000 records and included age and sex as QIDs.

In the income dataset, the sensitive attribute values were less than 1 million yen, less than 2 million yen, ..., and more than 25 million yen. The total number of sensitive attributes values was $F = 14$.

In addition, the sensitive attribute values in the generated dataset were not random; the medical condition refers to a patient survey published from the Ministry of Health, Labor and Welfare, and the annual income referred to a salary survey published by the National Tax Agency. Therefore, there is a

correlation between gender and age and each sensitive attribute; these are more realistic datasets.

As an open dataset, we used the adult dataset published by the UCI Machine Learning Repository. The adult dataset stored 14 kinds of QIDs, and there were about 30,000 records, which have also been utilized in a significant number of existing studies [11], [13]. In the evaluation experiment of this research, we used two kinds of adult datasets: Education (with educational background as the sensitive attribute) and fnlwgt (with fnlwgt as the sensitive attribute). The QIDs used 13 types, which were attributes other than a sensitive attribute.

In this paper, because we assumed anonymization at the stage of data collection by crowdsensing, (l, e)-diversity cannot be used in the assumed environment in the first place, since (l, e)-diversity uses Anatomy as an anonymizing algorithm. In addition, since (l, e)-diversity assumes only datasets that can be categorized, it cannot be used for the income dataset and adult dataset (fnlwgt). However, it is clear that verification is indispensable from the similarity with this research.

Therefore, we did not assume that the adult dataset (Education) was collected by crowdsensing after anonymizing the data. The verification experiments with the adult dataset (Education) assumed a state where data were collected using a method that was not crowdsensing; the dataset was created and then anonymized.

Table 11: Datasets used for the experiments.

dataset	#Record	QID	Sensitive attribute	Reference data
Income dataset	100000	2	annual income	National Tax Agency
Adult dataset (fnlwgt)	30162	13	fnlwgt	adult dataset
Adult dataset (Education)	30162	13	educational background	adult dataset

Table 11 shows the information of the three datasets used in the experiments.

6.3. Distance definition

The categorization of educational background as the sensitive attribute and the distance definition between the sensitive attribute values in the anonymization algorithm were in accordance with the config.xml file that was published by the UCI Machine Learning Repository. This educational background information was a statistical classification of a tree structure of depth 4, and only the end nodes of depth 4 were used as the sensitive attribute values. Category classification followed the educational background data; the distance definitions between the sensitive attribute values were defined as distances to a common ancestor node. Fig. 7 shows this educational background data in a tree structure.

From Fig. 7, suppose that a certain record stores the sensitive attribute value "Doctorate." If the distance is 1, it refers to the descendant node of "Post grad," that is, "Masters." If the distance is 2, it refers to the descendant node of "University," that is, "Prof-school," and "Bachelors." Therefore, if anonymization were to satisfy $(3, 2)$-semantic diversity, then two dummies from "Assoc," "Some-college," "Secondary," or another category were selected and added.

7. EXPERIMENTAL RESULTS

Table 12 shows combinations of the anonymizing algorithm and the analysis algorithm used in the experiment. Since LE uses Anatomy as an anonymizing algorithm in (l, e)-diversity, the simple analysis was used as an analysis algorithm suitable for Anatomy.

7.1. The generated histograms

First, the generated histogram is shown. Histograms obtained from the income dataset, adult dataset (Education), and adult dataset (fnlwgt) are shown in Figs. 8, 9 and 10.

Table 12: Combinations of anonymizing algorithms and analysis algorithms.

Algorithm Name	Anonymizing Algorithm	Analysis Algorithm
N-Proposal	Duplicate-processing algorithm	Proposed analysis algorithm
N-Existing	Duplicate-processing algorithm	Existing analysis algorithm
N-Simple	Duplicate-processing algorithm	Simple analysis algorithm
N-Anonymity	Duplicate-processing algorithm	—-
Ex-Proposal	Simple algorithm	Proposed analysis algorithm
Ex-Existing	Simple algorithm	Existing analysis algorithm
Ex-Simple	Simple algorithm	Simple analysis algorithm
LE	Anatomy [12]	Simple analysis algorithm

Figs. 8, 9 and 10 show not only the histograms obtained from the combination of Table 12 but also the histograms obtained from the original data, and it is possible to compare how similar they are to the original.

The x and y axes in the histograms show the age, gender, and classification of the sensitive attribute, and the z axis shows the frequency. The sensitive attribute in Fig. 8 was categorized into 14 types; in Fig. 9, it was categorized into 16 types; and in Fig. 10, it was categorized into 148 types. The QID age was classified every 10 years.

We anonymized these histograms to satisfy (3,7)-semantic diversity for the income dataset, (2,3)-semantic diversity for the adult dataset (Education), and (3,10)-semantic diversity for the adult dataset (fnlwgt).

7.2. The MSE graph

Next, each of the MSE results of the estimates from the proposed analysis algorithm, the simple analysis, and the existing analysis algorithm are shown in Figs. 11, 12 and 13.

In Fig. 11, 12 and 13, the y axis represents the MSE value, and the x axis represents l or d. As shown in the figures, (a) is a graph of the MSE obtained with $l = 2$, where d is a variable, and graphs (b), (c), and (d) are of the MSE

obtained with $d = 2$, $l = 3$, and $d = 3$.

7.3. Safety Analysis

We confirmed that our proposed algorithms always achieved (l, d)-semantic diversity in the experiments. On the contrary, the method [1] failed many times. The probability of violating (l, d)-semantic diversity is shown in Figs. 14 and 15. These figures also show the mathematical results based on Equation 11.

The violation rate represents the rate at which the anonymization results do not satisfy (l, d)-semantic diversity. The anonymization results of the proposed simple anonymization algorithm and duplicate-processing anonymization algorithm always satisfy (l, d)-semantic diversity; that is, the violation rate of the proposed anonymized algorithms is always zero. Existing algorithms do not guarantee that the anonymization results will satisfy (l, d)-semantic diversity. Let n_g represent the number of times that the anonymization results satisfy (l, d)-semantic diversity, and let n_f represent the number of times that the anonymization results do not satisfy (l, d)-semantic diversity. The violation rate is defined as follows:

$$\text{Violation rate} = n_f/(n_g + n_f). \tag{23}$$

Based on the figure, we know that in any parameter settings, the mathematical results and simulation results are almost the same, and the mathematical results keep the lower boundary of the simulation results.

7.4. Discussion

Figs. 8, 9 and 10 show that the histograms estimated by the proposed analysis algorithm are the most similar to the original histograms. For example, the histograms estimated by the simple analysis algorithm are averaged overall, and the features are diminished. However, the histograms estimated by the proposed analysis algorithm clearly retain their characteristic unevenness.

Next, consider the graph of the MSE. With respect to each graph in Fig. 11, it is clear that the MSE of N-Proposal has a small value in all cases and that

Ex-Proposal is not less than N-Proposal as long as l and d are small, but the larger l or d is, the larger the MSE becomes. From this, the usefulness of the proposed analysis algorithm is shown.

The results of N-Proposal and Ex-Proposal show the usefulness of duplicate-processing algorithms. As described in Section 5.3, the probability of selecting a dummy changes as the values of l or d become larger, so the result worsens with the simple algorithm.

In addition, in Fig. 11(a), a histogram was created without using analysis algorithms from anonymizing data, and the MSE result was expressed as N-Anonymity. If the analysis algorithm is not used, since the number of records is l times the original database, the histogram is quite different from the histogram created from the original data. Since the anonymizing algorithm was assumed to be used with the analysis algorithm, this result is reasonable.

Each graph in Fig. 12 shows that the higher the values of l and d, the worse the MSE of Ex-Proposal becomes. The adult dataset (Education) has educational background as a sensitive attribute, and the duplicate-processing algorithm cannot be used; therefore, the result is poor, as mentioned above.

In addition, in Fig. 12(b), the MSE of Ex-Proposal is larger than the existing method (l, e)-diversity. Since (l, e)-diversity can be applied only when a sensitive attribute can be classified by a tree structure, it cannot be used for data collection via crowdsensing, and the proposed method cannot use the duplicate-processing algorithm. Therefore, it should be noted that (l, e)-diversity is not always better.

In this research, we assume that a record only has one sensitive attribute, as most existing studies assume the same. Treating multiple sensitive attributes will be a focus in our future work. Our previous work [1], which does not consider semantic diversity, can handle multiple sensitive attributes for realizing l-diversity. We believe that our proposed algorithms in the revised manuscript can treat multiple sensitive attributes when combined with our previous work [1].

Our proposed anonymization algorithm, which adds randomized records to

the original database, is relatively difficult for data users to understand. Moreover, it is difficult to obtain meaningful information from *each* record in the anonymized database. However, for statistical analysis, our proposed analysis algorithm can greatly reduce information loss, as compared with algorithms proposed in existing studies, as shown in Section 7. For example, as shown in Figure 8, the histogram generated by our proposed method (i.e., anonymized with our proposed anonymization algorithm and then reconstructed with our analysis algorithm) is very similar to the true histogram. Because histogram analysis (also called a cross-tabulation or contingency table analysis) is an important analysis method and our proposed method can generate a histogram very similar to the real one, we can say that the usability of the data is maintained for statistical analysis.

Note that generating anonymized histograms is a very important task in privacy-preserving data mining. Many existing studies have targeted the generation of anonymized histograms [9, 27, 28, 30, 29], although they cannot ensure semantic (l, d)-diversity.

8. CONCLUSION

In this research, a new indicator, (l, d)-semantic diversity, was proposed to solve the research problem, which is the possibility that actual diversity is not satisfied if a database holds similar sensitive attribute values, even if anonymization is performed to satisfy l-diversity.

By defining the distances between the sensitive attribute values from categorization and considering these distances with indicators, we solved the problem of existing indicators.

In addition, we proposed an anonymization algorithm using a dummy record addition method according to the proposed indicator and an analysis algorithm suitable for the proposed anonymization algorithm.

To confirm that the above proposals are correct, a verification experiment was conducted. Based on the results of the verification experiment, we confirmed

that the proposed analysis is satisfactory.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Numbers JP16K00419, JP16K12411, JP17H04705, JP18H03229, JP18H03340, JP18K19835. This work was supported by JST, PRESTO Grant Number JPMJPR1934.

References

[1] Y. Sei, H. Okumura, T. Takenouchi, A. Ohsuga, *Anonymization of Sensitive Quasi-Identifiers for l-diversity and t-closeness*, IEEE Transactions on Dependable and Secure Computing, 2019, vol.16, no.4, pp.580-593.

[2] K. LeFevre, D. DeWitt, and R. Ramakrishnan, *Mondrian Multidimensional K-Anonymity*, in Proc. IEEE ICDE, 2006, p.25.

[3] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, D. Kifer, and M. Venkitasubramaniam, *l-diversity: Privacy beyond K-Anonymity*, ACM TKDD, 2007, vol.1, no.1, p.3.

[4] N. Li, T. Li, and S. Venkatasubramanian, *t-closeness: Privacy beyond k-anonymity and l-diversity*, in Proc. IEEE ICDE, 2007, pp.106-115.

[5] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez. *t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation.* IEEE Transactions on Knowledge and Data Engineering, 27(11):3098–3110, 2015.

[6] Tu, Z., Zhao, K., Xu, F., Li, Y., Su, L., and Jin, D., *Protecting Trajectory From Semantic Attack Considering k-Anonymity, l-Diversity, and t-Closeness*, IEEE Transactions on Network and Service Management, 2019, vol.16, no.1, pp.264-278.

[7] B. Fung, K. Wang, R. Chen, and P. S. Yu, *Privacy-preserving data publishing: A survey of recent developments*, ACM Computing Surveys, 2010, vol.42, no.4, pp.14.

[8] C. Dwork, *Differential Privacy*, in Proc. ICALP, 2006, pp.1-12.

[9] G. Acs, C. Castelluccia, and R. Chen, *Differentially Private Histogram Publishing through Lossy Compression*, in Proc. IEEE ICDM, 2012, pp.1-10.

[10] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, *Enhancing data utility in differential privacy via microaggregation-based k-anonymity*, The VLDB Journal, vol.23, no.5, pp.771-794, 2014.

[11] K. LeFevre, D. DeWitt, and R. Ramakrishnan, *Incognito: Efficient full-domain k-anonymity*, in Proc. ACM SIGMOD, 2005, pp.49-60.

[12] X. Xiao, and Y. Tao, *Anatomy: Simple and effective privacy preservation*, in Proc. the 32nd international conference on VLDB Endowment, 2006, pp.139-150.

[13] V. Iyengar, *Transforming data to satisfy privacy constraints*, in Proc. ACM SIGKDD, 2002, pp.279-288.

[14] N. Mohammed, B. Fung, P.C.K. Hung, and C.K. Lee, *Centralized and Distributed Anonymization for High-Dimensional Healthcare Data*, ACM TKDD, 2013, vol.4, no.4, p.18.

[15] M. Yuan, L. Chen, S. Y. Philip, and T. Yu, *Protecting sensitive labels in social network data anonymization*, IEEE Transactions on Knowledge and Data Engineering, 2013, vol.25, no.3, pp.633-647.

[16] K. Mancuhan, and C. Clifton, *Statistical Learning Theory Approach for Data Classification with l-diversity*, in Proc. SIAM International Conference on Data Mining, 2017, pp.651-659.

[17] Haiyuan Wang, Jianmin Han, Jiyi Wang, Lixia Wang, *(l,e)-Diversity — A Privacy Preserving Model to Resist Semantic Similarity Attack*, JCP, 2014, Vol.9, no.1, pp.59-64.

[18] Z. Huang, and W. Du, *OptRR: Optimizing Randomized Response Schemes for Privacy-Preserving Data Mining*, in Proc. IEEE ICDE, 2008, pp.705-714.

[19] R. Agrawal, and R. Srikant, *Privacy-Preserving Data Mining*, in Proc. ACM SIGMOD, 2000, pp.439-450.

[20] X. Sun, M. Li, and H. Wang, *A family of enhanced (L, α)-diversity models for privacy preserving data publishing*, Future Generation Computer Systems, 2011, vol.27, no.3, pp.348-356.

[21] H. Jianmin, Y. Juan, Y. Huiqun, and J. Jiong, *A multi-level l-diversity model for numerical sensitive attributes*, Journal of Computer Research and Development, 2011, 1, pp.147-158.

[22] N. Mohammed, B. Fung, P. C. Hung, and C. K. Lee, *Anonymizing healthcare data: a case study on the blood transfusion service*, in Proc. ACM SIGKDD, 2009, pp.1285-1294.

[23] M. Alidoost Nia and A. Ruiz-Martinez, *Systematic literature review on the state of the art and future research work in anonymous communications systems*, Comput. Electr. Eng., 2018, vol.69, pp.497-520.

[24] K. Emura, A. Kanaoka, S. Ohta, and T. Takahashi, *A KEM/DEM-based construction for secure and anonymous communication*, in Proc. IEEE 39th Annual Computer Software and Applications Conference, 2015, pp.680-681.

[25] B. Wu, B. J. Shastri, P. Mittal, A. N. Tait, and P. R. Prucnal, *Optical signal processing and stealth transmission for privacy*, IEEE Journal of Selected Topics in Signal Processing, 2015, vol.9, no.7, pp.1185-1194.

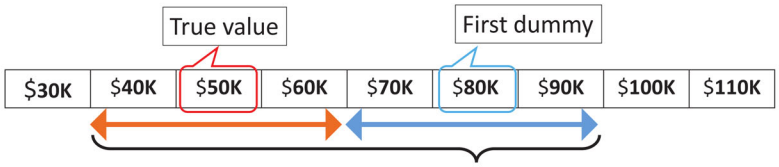
[26] K. Oishi, Y. Sei, Y. Tahara, and A. Ohsuga, *Proposal of l-Diversity Algorithm Considering Distance between Sensitive Attribute Values*, in Proc. IEEE CIDM, 2017, pp.2065-2072.

[27] R. Chen, Q. Xiao, Y. Zhang, and J. Xu. *Differentially Private High-Dimensional Data Publication via Sampling-Based Inference*, In Proc. ACM KDD, pages 129–138, 2015.

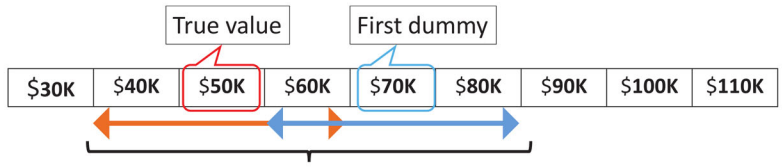
[28] W. Qardaji, W. Yang, and N. Li. *PriView: practical differentially private release of marginal contingency tables*, In Proc. ACM SIGMOD, pages 1435–1446, 2014.

[29] J. Soria-Comas, J. Domingo-Ferrer, D. S´ anchez, and S. Mart´ınez. *Enhancing data utility in differential privacy via microaggregation-based k-anonymity*, The VLDB Journal, 23(5):771–794, 2014.

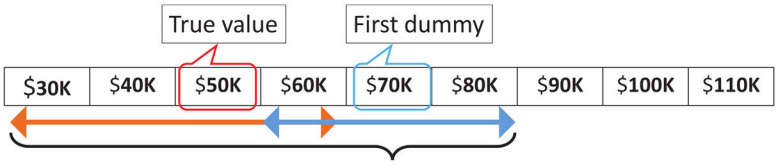
[30] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez. *Improving the Utility of Differentially Private Data Releases via k-Anonymity*, In Proc. IEEE TrustCom, pages 372–379, 2013



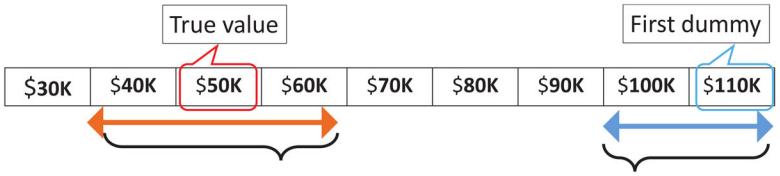
(a) The distance between the true and the first dummy values is greater than or equal to $2d$.



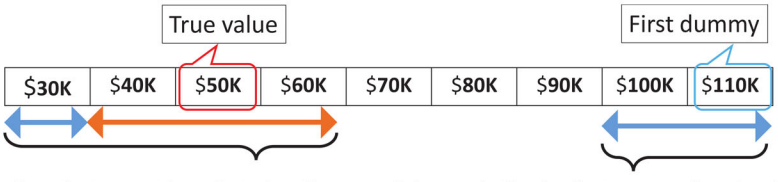
(b) The distance between the two values is less than $2d$ in the simple algorithm.



(c) The distance between the two values is less than $2d$ in the duplicate-processing algorithm.



(d) The first dummy is near the side in the simple algorithm.



(e) The first dummy is near the side in the duplicate-processing algorithm.

Figure 4: A specific example of a duplicate-processing algorithm.

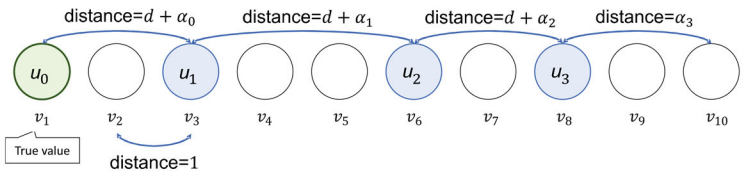
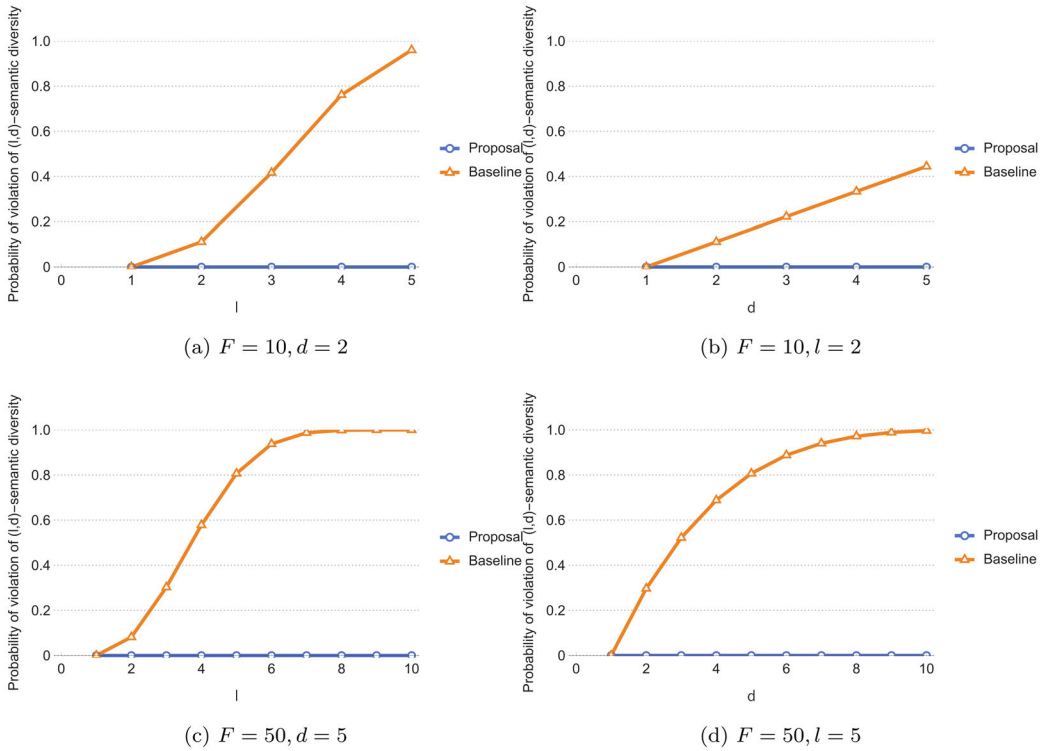


Figure 5: An example situation where the number of elements F is 10 and l is 4.



(a) $F = 10, d = 2$ (b) $F = 10, l = 2$

(c) $F = 50, d = 5$ (d) $F = 50, l = 5$

Figure 6: Probability of violation of (l, d)-semantic diversity for each record.

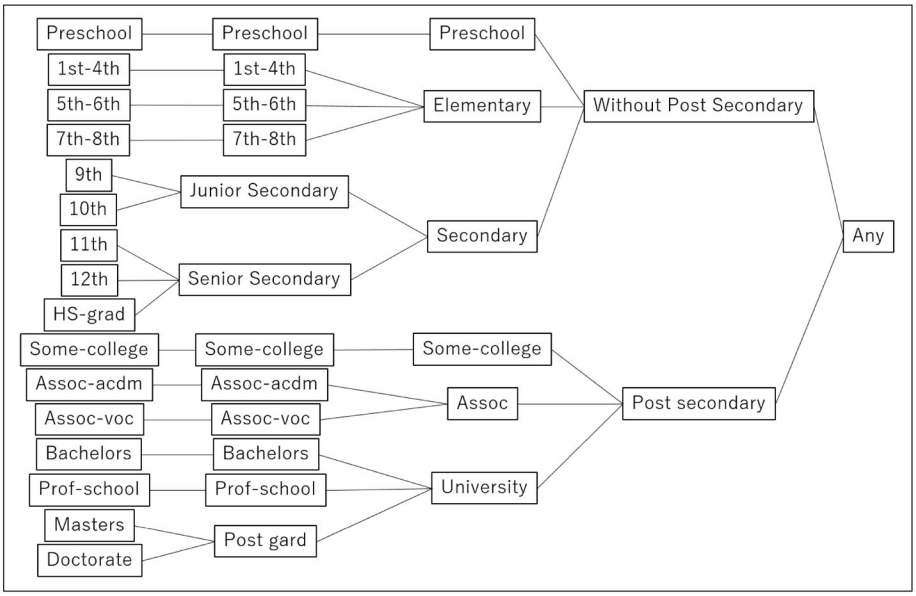


Figure 7: Classification of educational background.

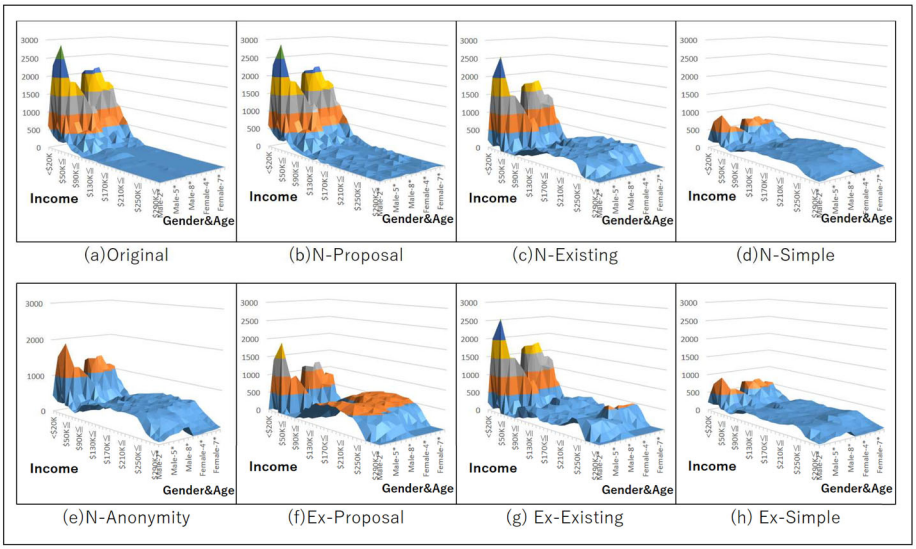


Figure 8: Histograms of the income dataset.

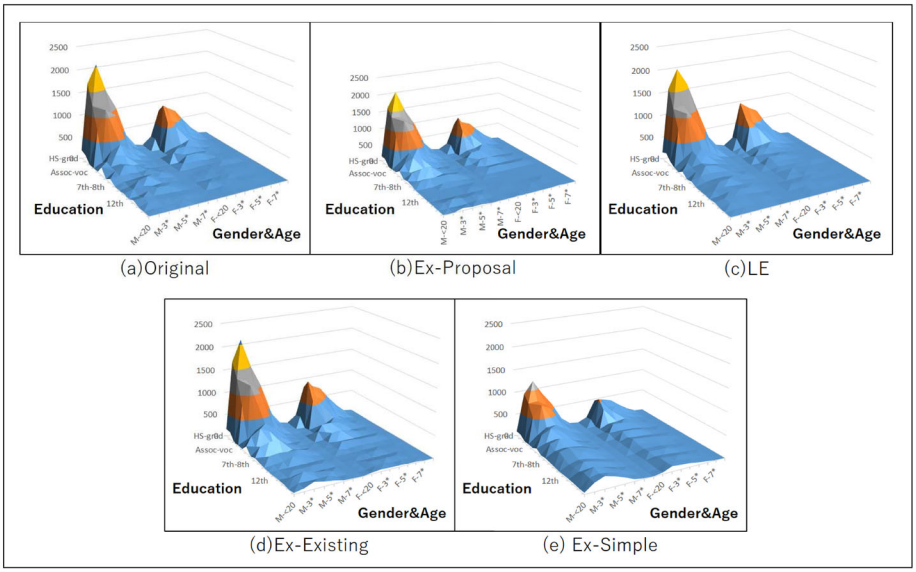


Figure 9: Histograms of the adult dataset (Education).

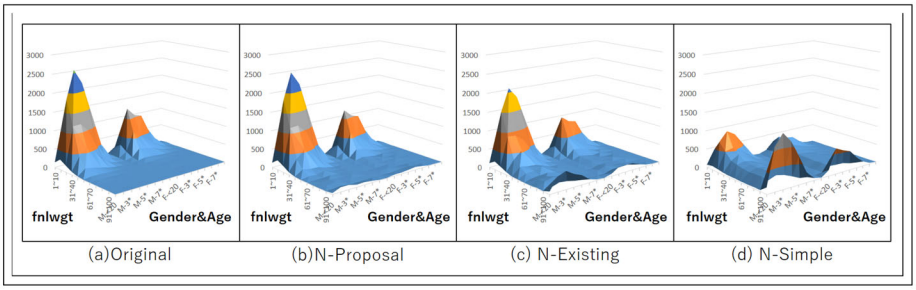


Figure 10: Histograms of the adult dataset (fnlwgt).

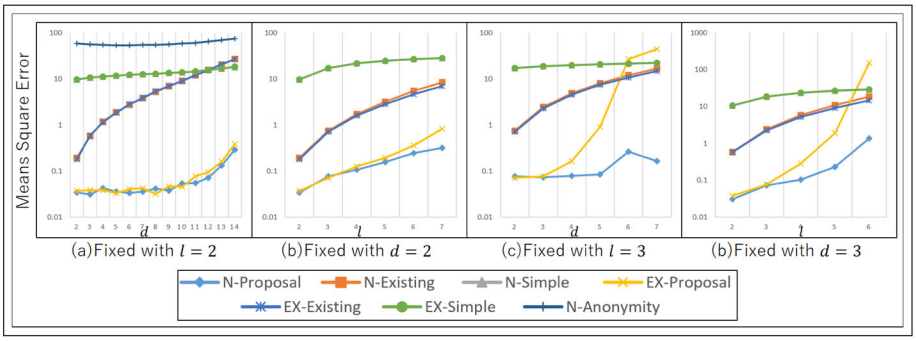


Figure 11: The MSE for the income dataset.

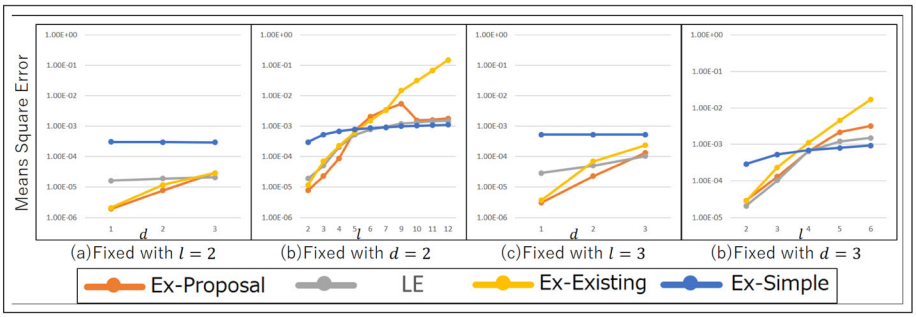


Figure 12: The MSE for the adult dataset (Education).

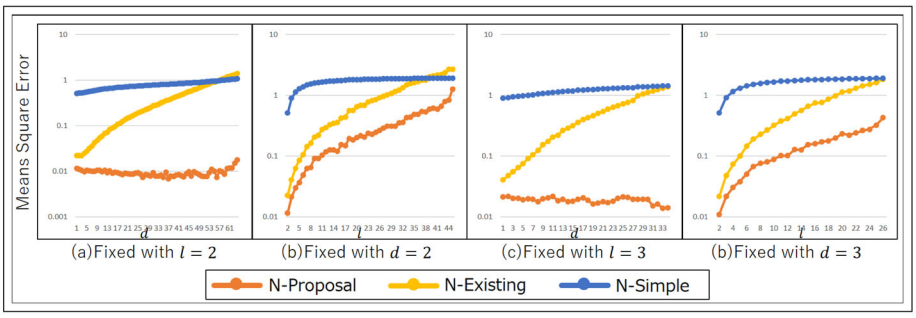


Figure 13: The MSE for the adult dataset (fnlwgt).

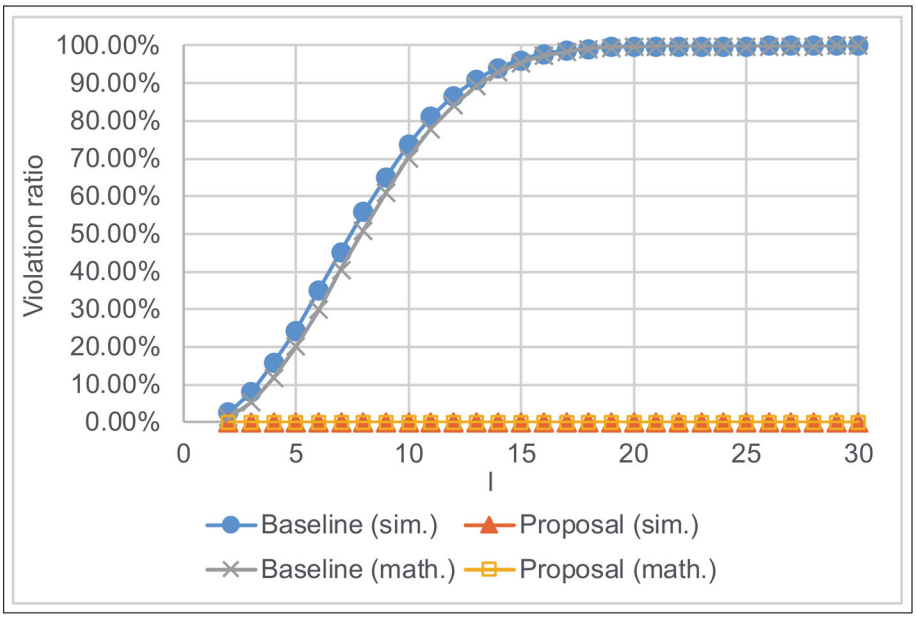


Figure 14: Violation ratio of (l, d)-semantic diversity ($d = 3$).

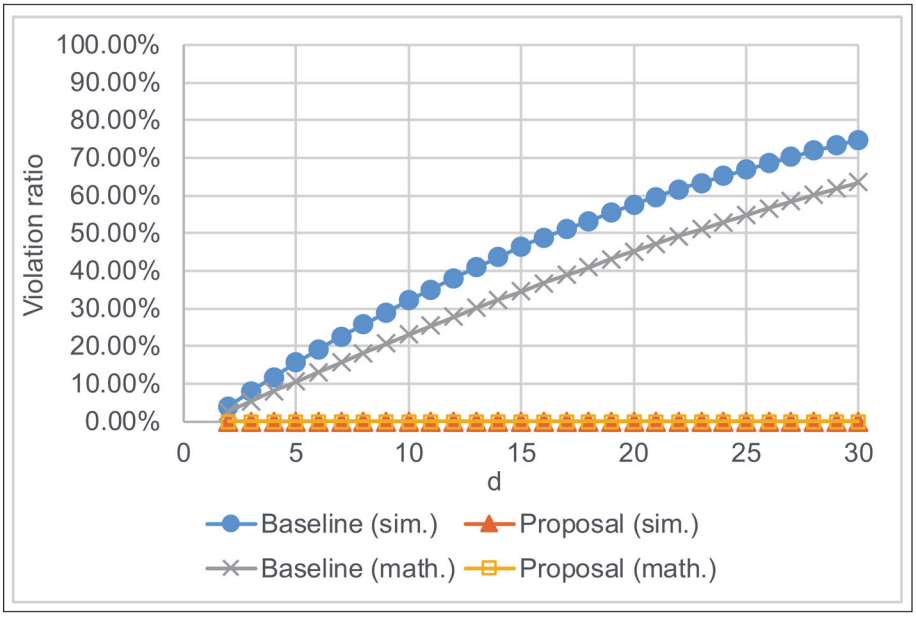


Figure 15: Violation ratio of (l, d)-semantic diversity ($l = 3$).