Safety Analysis of High-Dimensional Anonymized Data from Multiple Perspectives

Takaya Yamazoe and Kazumasa Omote

February 21, 2021

Safety analysis of high-dimensional anonymized data from multiple perspectives

Takaya Yamazoe[1] and Kazumasa Omote[1,2]

[1] Systems and Information Engineering, University of Tsukuba, Japan
s1920600@s.tsukuba.ac.jp
http://www.risk.tsukuba.ac.jp
[2] National Institute of Information and Communications Technology, Japan
omote@risk.tsukuba.ac.jp
https://www.nict.go.jp

Abstract. Recently, large-scale data collection has driven data utilization in the medical, financial, advertising, and several other fields. This increasing use of data necessitates privacy risk considerations. K-anonymization and other anonymization methods have been used to minimize data privacy risks, but they are unsuitable for large and high-dimensional datasets required in machine learning and other data mining techniques. Although subsequent methods such as matrix decomposition anonymization can anonymize high-dimensional data while maintaining a high level of utility, they do not clarify anonymized data safety or adequately analyze privacy risks.

Therefore, in this study, we performed a multi-perspective analysis on the privacy risks of datasets anonymized with some anonymization methods using various safety metrics. In addition, we propose a new technique for evaluating privacy risk for each attribute of anonymized data. Experimental results showed that our method effectively analyzed privacy risks of high-dimensional anonymized data. Furthermore, our evaluation of the resistance to data re-identification using existing techniques showed that anonymization methods have their suitable attack types, and it is important to assess data safety using various metrics before publishing.

Keywords: Anonymaization · Privacy · Safety metrics.

1 Introduction

Recently, data has become commonly utilized in all fields (e.g., medicine and finance), following the spread of web services and internet of things (IoT), as well as research developments in the machine learning and data mining fields. However, only a few institutions boast of both machine learning and data mining capabilities because the fields require advanced analytical techniques and huge amounts of data. For instance, assume a situation in which sensitive data has to be transferred to other institutions when a huge amount of data is required for machine learning or when a data analysis institution is asked to analyze the data. The transfer of data to other institutions may cause a violation of the sensitive information contained in the data. Examples of privacy violations from actual public data include the identification of a state legislator's personal information from the health care insurance data of a Massachusetts state legislator [1] and the unique identification of a user from the rating value of a user's movie released on Netflix [2]. These cases highlight the need for cautious processing and privacy protection when dealing with sensitive data. Anonymization is one of the techniques used to protect data privacy, and various data anonymization techniques have been proposed (e.g., k-anonymization [1] and noise addition). These methods have long been used to protect privacy when releasing datasets due to their use of intuitive data safety metrics [1][3][4].

However, conventional methods such as *k*-anonymization encounter difficulty in simultaneously maintaining a high level of utility and anonymizing large and high-dimensional datasets, which are used in techniques such as machine learning[7]. Thus, various techniques have been proposed to address this difficulty. One of the successful techniques combines matrix decomposition and k-anonymization or noise addition [5], but it does not sufficiently analyze the safety of the anonymized data. Although various evaluation metrics are required to analyze the safety of anonymized data [6], discussion is still lacking on evaluation metrics for the data anonymized by new techniques.

The aim of this study is to analyze the safety of high-dimensional datasets anonymized by matrix decomposition. Hence, we analyzed the safety of datasets anonymized by matrix decomposition, k-anonymization, and noise addition from multiple perspectives. To evaluate these risks, we propose a new technique to evaluate the vulnerability of each attribute of anonymized data, in addition to conventional techniques. Our proposed technique evaluates the vulnerability of each attribute, taking advantage of the fact that each attribute's distribution of values can only change slightly after anonymization. We futher discuss the safety of each anonymization method using various metrics. The major contributions of this study are as follows:

- We analyze anonymized data of comparable utility and provide some insight into the safety features of each anonymization method;
- We propose a new technique to evaluate the privacy risk of each attribute of anonymized data of using features that make the marginal distribution of each attribute unchanged after anonymization;
- We discuss the relationship between the safety features of each anonymization method and their resistance to malicious attacks on datasets, and the importance of using multiple metrics is demonstrated.

2 Related Works

2.1 *K*-anonymity

K-anonymization is a method that transforms data such that at least *k* records in the dataset have the same data within a quasi-identifier [1]. An intuitive metric whereby the number of data owners cannot be narrower than k is referred to as *k*-anonymity. When discussing *k*-anonymity, each attribute or combination is generally classified into attributes such as an identifier and a quasi-identifier. An identifier links the data owner to an individual (e.g., user id and username), whereas a quasi-identifier links the data owner to an individual by combining multiple data (e.g., age and gender). In the *k*-anonymization method based on *k*-anonymity, personal privacy is protected by deleting and processing quasi-identifiers such that the number of data owners cannot be narrowed down to *k* or fewer from a combination of quasi-identifiers. This metric is frequently used because it is intuitive and easy to understand. However, if the data to be handled is high-dimensional, the distance between the dataset records increase rapidly. Thus, k-anonymization does not efficiently anonymize high-dimensional data while retaining data utility [7]. Although l-diversity [3] and t-closeness [4] are other safety metrics that extend k-anonymity, neither can be used to evaluate data other than those anonymized by k-anonymization. The anonymization methods combined with matrix decomposition used in this study are not strict k-anonymization methods; thus, these metrics are unsuitable to ensure the safety of datasets anonymized by matrix decomposition anonymization.

2.2 Genome Privacy

Full sequencing of the human genome is now possible, and genomic data is rapidly being utilized in health care, research, and forensic science. Although genome data is invaluable in various fields, its use is highly likely to cause privacy invasion because genome sequences can uniquely identify individuals. Examples of privacy violations from genomic data range from patient disease condition leakage in the re-identification of anonymous participants in genome-wide association studies to genetic discrimination, such as using certain genetic predispositions to deny insurance. Consequently, privacy concerns necessitate considerable care when working with genomic data. Nevertheless, protection techniques and metrics for genomic privacy have not yet been established. Isabel proposed the use of various safety metrics to assess data privacy risks and investigated how an attacker can infer privacy information from genomic data [6]. Twenty-two metrics, including information entropy, mean-squared error, and Gini's coefficient, were used to analyze the potential risk of data privacy invasion. Analyzing the behavior of the 22 metrics showed that a single metric alone is insufficient to ensure data safety. This present study builds on the work of Isabel to analyze the safety of anonymized data from multiple perspectives using various metrics.

2.3 Maximum-knowledge attack

Record linkage is a method for re-identifying anonymized data. It is an attack that violates privacy by linking records in an anonymized dataset to those in an external dataset. Hence, it is important to consider record linkage risks when releasing anonymized data. However, in simulating record linkage, various items (e.g., the auxiliary information available to the attacker) are assumed. Furthermore, record linkage only focuses on record re-identification and does not include attacks where an attacker gains knowledge of specific sensitive attributes of an individual. Ferrer et al[10] proposed a technique for evaluating the risk of anonymized data disclosure. The technique solves the problem of record linkage described above. Specifically, a record linkage attack can be simulated without considering the background knowledge of the attacker and possible disclosure of attributes by assuming a maximum-knowledge attacker (i.e., one who has both the original and anonymized data). This scenario is described thus. The attacker possesses both the original dataset X and anonymized dataset Y. The attacker generates a dataset Y' from the anonymized dataset Y by permuting each attribute to remove the dependencies between the attributes. The attacker then computes the distance between X and Y and between X and Y', and the distributions of the distances are defined as $dist$ and $dist'$, respectively. Finally, the attacker compares the distributions. If both distributions are equal, then there is no evidence that X

contains any information that can be used to improve the linkage accuracy. This can be interpreted as X and Y being independent, indicating that the anonymization of Y is very strong. Thus, this method can be used to obtain the lower bound of achievable disclosure risk protection.

In addition to this method of evaluating the risk of record re-identification, they also propose a method for evaluating the vulnerability of each attribute of anonymized data. They divide the original dataset X with the number of m attributes into (x_b, x_m). Anonymized dataset Y is similarly divided into (y_b, y_m). x_b is a record that concatenates the attributes from x_1 to x_{m-1}. They then link the records using x_b and y_b and measure the distribution of distances x_m and y_m of the linked records. They evaluate the vulnerability of the attributes by comparing the distance distributions of the target attribute x_m and y_m, and the target attribute x_m and that of the permute dataset y'_m, in a similar way to the above assessment. Similar to the record re-identification risk described above, this assessment is also a technique to evaluate the lower bound of the vulnerability of each attribute.

2.4 Sensitive Attribute Disclosure

Recently, large-scale and high-dimensional data has been used for machine learning and data analysis. These data have many attributes, and it is difficult to identify the vulnerable attributes that should be protected among them. To solve these problems, Ito et al [11] proposed an attacker model to assess the vulnerability of attributes. The attacker model quantifies the probability that an attacker will gain background knowledge about an attribute by accident, based on the information about the values contained in the attribute. Let the dataset be T, and let m and n be the numbers of records and users in the dataset, respectively. Let D_x be the set of values for attribute X of T. Let R_x and U_x be the sets of records containing a given $x \in D_x$ and users that have x in attribute X, respectively. Then, the joint probability $Pr(idf, x)$ is represented by the following equation, using the probabilities $Pr(x)$ and $Pr(idf|x)$ that an attacker will gain background knowledge of an attribute x and a user with an attribute x, respectively.

$$Pr(idf, x) = Pr(x)Pr(idf|x) = \frac{|R_x|}{m}\frac{1}{|U_x|} \tag{1}$$

Ito et al showed that this risk model can be used to find the riskiest attributes in a dataset and guide the decision on which attributes to process or remove when anonymizing the data. As the risk model is intended to analyze the potential risk of attributes of the original dataset, it is not suitable for evaluating the privacy risk of anonymized datasets and cannot be used in this study.

3 Preliminary

In this section, we first describe the basis of our proposed method before detailing the method.

3.1 Matrix decomposition as anonymization

Matrix decomposition is a method of anonymizing high-dimensional data while maintaining a high level of utility. It decompose a matrix $M \in R^{n \times m}$ into two matrices, $U \in R^{n \times r}$ and $V \in R^{m \times r}$. Then, matrix $X = UV^T$ approximates M, and rank r is a parameter that specifies its accuracy. Mimoto et al [8][9] showed that combining matrix decomposition and k-anonymization or noise addition can anonymize high-dimensional data while keeping the utility of the data. Therefore, data anonymized by this technique is expected to be used for machine learning. Additionally, Mimoto showed experimentally that data anonymized using matrix decomposition are more useful than those anonymized using only k-anonymization or noise addition in training machine learning models. However, although the utility of anonymized data is well-established, the assessment of data safety is inadequate. As the combination of matrix decomposition and k-anonymization does not guarantee strict k-anonymity of the anonymized data, a careful analysis of the risk of information leakage from anonymized data is necessary. However, only simple record-matching tests and record links between anonymized data have been tested [8]. In this study, we analyze the privacy risks of datasets anonymized using matrix decomposition from multiple perspectives. Specifically, starting with the analysis of basic metrics (e.g., information entropy and distribution distance), we analyze the re-identification risk of anonymized data and the privacy risk of each attribute of the anonymized data.

3.2 Evaluation of Utility

Various methods are used to evaluate the utility of anonymized data. For example, some methods use Hamming distance and cross-tabulation. In the Hamming distance methods, original datasets and anonymized ones are compared, and the ratio of different data records is calculated. In cross-tabulation methods, the tabulated values are obtained by the cross-tabulation of each original dataset and anonymized one, and the absolute error is calculated. Various methods can be used for the evaluation of utility, but it is necessary to consider the intended use of the dataset. In this study, we assumed that the anonymized data will be used for machine learning. Thus, we used the method proposed by Mimoto et al [8] for evaluation of the utility. Precisely, the F measure of the models trained using the original and anonymized datasets are set to F_{ori} and F_{ano}, respectively, and the utility of the anonymized dataset is evaluated by the following formula. We used logistic regression and random forests as machine learning algorithms to predict the test data and measure the F-measure.

$$Utility = \frac{F_{ano}}{F_{ori}} \tag{2}$$

4 Our Method

In the following section, we introduce our proposed method for evaluating anonymized data.

4.1 Attacker assumptions

In this study, we assumed that the attacker has the anonymized data. This situation corresponds to the case when the anonymized data are accessible to the public or when they are transferred to other institutions. In this case, the attacker may try to extract sensitive information about the original data from the anonymized data. Thus, we propose an attack method in which an attacker uses the anonymized data to estimate the original values of sensitive attributes. In addition, to assume an attacker with a vast background knowledge, we assumed an attacker in two levels.

Level1. Normal attacker The attacker has the anonymized data and has only background knowledge of the set of possible values for the target attributes.

Level2. Attacker with the distribution The attacker has the anonymized data. In addition to the set of possible values for the target attributes, the attacker also has background knowledge of the marginal distribution of target attributes in the original data.

4.2 Algorithm

We consider that when there is a high similarity between the distributions of identical attributes in the original and anonymized datasets, those attributes have a low level of anonymization. Therefore, we propose a method to evaluate the privacy risk of an attribute by using the distribution of that attribute in anonymized data. Let X be an attribute of the target dataset and D_x the set of values that can be taken by attribute X of the original dataset. The attacker calculates and ranks the distance of $x \in D_x$ from the value x_{ano} of the target attribute X_{ano} in the anonymized data. Finally, we use the calculated ranks to estimate the probability $Pr(x|x_{ano})$ that the original data is value x when the anonymized data is x_{ano}. The proposed method is shown in Algorithm-1. We also propose a second algorithm that assumes a level2 attacker who has a marginal distribution of the target attributes of the original data in addition to the assumptions of the Algorithm1 attacker. When the attacker of this assumption estimates the value of the original data from the anonymized data, the attacker adds weights to the computation of $Pr(x|x_{ano})$ using the marginal distribution $Pr(x)$. An attack by a level2 attacker is shown in Algorithm2.

5 Experiments

In this section, we describe the experiments performed to evaluate the utility and safety of the anonymized data using the proposed method. The data used in the experiments were processed by matrix-decomposition anonymization, using k-anonymization or noise addition. As the aim of this study is to analyze the safety of anonymized data, we first evaluated the utility of the anonymized data before analyzing the safety of the anonymized data found to have the same level of utility. An adult [12] dataset and a diabetes dataset [13] were used in the experiments. The adult dataset contained personal information (e.g., age, occupation, and gender) from the 1994 Census database by Barry Becker and had more than 100 attributes when one-hot encoded. The diabetes dataset contained over 50 features representing patient and hospital outcomes. It also had more than 100 attributes from one-hot encoding.

Algorithm 1 Algorithm for attacker of level1

```
Input: values of target attribute $X_{ano} = (x_{ano_1}, x_{ano_2}, ..., x_{ano_n})$, set of possible values $D_x$
Output: estimated original value $X_{pred} = (x_{pred_1}, x_{pred_2}, ..., x_{pred_n})$
for $x \in D_x$ do
    Initialize $d$ to $(d_1, d_2, ..., d_n)$
    Initialize $p$ to $(p_1, p_2, ..., p_n)$
    Compute $d_i = \frac{1}{dist(x, x_{ano_i})}$ for each $x_{ano_i} \in X_{ano}$
    Compute $p_i = \frac{Rank(d_i)}{n}$ for each $d_i \in d$
    Set $Pr(x|x_{ano_i})$ to $p_i$ for each $p_i \in p$
end for
for $i = 1$ to $n$ do
    Set $X_{pred_i}$ to $x$ which has maximum value $Pr(x|x_{ano_i}) \in D_x$
end for
return $X_{pred} = (x_{pred_1}, x_{pred_2}, ..., x_{pred_n})$
```

Algorithm 2 Algorithm for attacker of level2

```
Input: In addition to Algorithm1, marginal distributions $(Pr(x_1), Pr(x_2), ..., Pr(x_m))$
Output: estimated original value $X_{pred} = (x_{pred_1}, x_{pred_2}, ..., x_{pred_n})$
for $x \in D_x$ do
    Compute $p_i$ in the same way as Algorithm1
    Set $Pr(x|x_{ano_i}) \cdot Pr(x)$ to $p_i$ for each $p_i \in p$
end for
for $i = 1$ to $n$ do
    Set $X_{pred_i}$ to $x$ which has maximum value $Pr(x|x_{ano_i}) \in D_x$
end for
return $X_{pred} = (x_{pred_1}, x_{pred_2}, ..., x_{pred_n})$
```

5.1 Evaluation of utility

It is important to keep anonymized data in such a way that they high utility. In this section, we describe the experiments conducted to evaluate the utility of anonymized data. We generated multiple matrix-decomposition anonymized datasets by adjusting the k values and noise levels. As we expected to use an anonymized dataset for machine learning, we trained a machine learning model with the generated anonymized dataset and obtained the F-measure. Furthermore, the utility of each anonymized dataset was calculated using the method introduced in Section 3.2. Tables 1 and 2 are the results of evaluating the utility of anonymized adult datasets, and Tables 3 and 4 are the results of evaluating the utility of anonymized diabetes datasets. The notations in the dataset columns are explained thus: k means using k-anonymization and its parameters, σ means using noise addition and its parameters, and d means using matrix decomposition and its parameters. The results in the tables show that k-anonymization is most useful when combined with matrix decomposition. More parameter tunings yield more useful results when the anonymization method is combined with matrix decomposition; however, we extracted a combination of parameters with similar utility in this experiment and included them in the table. Experimental results with the noise addition showed that the combination of matrix decomposition yields smaller σ values for anonymized data with the same utility than when noise addition is used alone. This shows that only a small amount of noise is needed to generate the anonymized datasets,

Table 1. Utility evaluation of anonymized adult datasets using matrix factorization and k-anonymization.

Score Dataset	F-measure	Utility
Ano(k=3)	0.833	0.988
Ano(k=20)	0.821	0.974
Ano(k=50)	0.818	0.97
Ano(k=70)	0.805	0.954
Ano(d=50, k=3)	0.841	0.998
Ano(d=50, k=20)	0.779	0.924
Ano(d=50, k=70)	0.775	0.919
Ano(d=30, k=30)	0.774	0.918
Ano(d=30, k=70)	0.778	0.922

Table 2. Utility evaluation of anonymized adult datasets using matrix factorization and noise addition.

Score Dataset	F-measure	Utility
Ano(σ=0.1)	0.821	0.974
Ano(σ=0.3)	0.819	0.971
Ano(σ=0.4)	0.776	0.92
Ano(σ=0.6)	0.774	0.918
Ano(d=80, σ=0.2)	0.774	0.918
Ano(d=50, σ=0.1)	0.833	0.988
Ano(d=50, σ=0.2)	0.781	0.926
Ano(d=30, σ=0.25)	0.78	0.925
Ano(d=30, σ=0.4)	0.779	0.923

Table 3. Utility evaluation of anonymized diabetes datasets using matrix factorization and k-anonymization.

Score Dataset	F-measure	Utility
Ano(k=3)	0.566	0.911
Ano(k=7)	0.553	0.89
Ano(k=10)	0.527	0.848
Ano(k=20)	0.521	0.838
Ano(d=80, k=3)	0.535	0.861
Ano(d=50, k=3)	0.571	0.919
Ano(d=50, k=10)	0.52	0.837
Ano(d=10, k=20)	0.517	0.832
Ano(d=10, k=30)	0.514	0.827

Table 4. Utility evaluation of anonymized diabetes datasets using matrix factorization and noise addition.

Score Dataset	F-measure	Utility
Ano(σ=0.1)	0.56	0.901
Ano(σ=0.2)	0.539	0.868
Ano(σ=0.3)	0.54	0.869
Ano(σ=0.45)	0.51	0.821
Ano(d=70, σ=0.01)	0.546	0.879
Ano(d=70, σ=0.03)	0.542	0.872
Ano(d=30, σ=0.01)	0.538	0.866
Ano(d=10, σ=0.05)	0.501	0.807
Ano(d=5, σ=0.01)	0.513	0.826

which is a positive result considering the utility of the data. Therefore, we performed the safety analyses using anonymized data with the same level of utility.

5.2 Evaluation of safety

In this section, we analyzed the safety of anonymized data using several safety metrics categorized as basic metrics such as entropy and distribution, robustness to record linkages, and attribute vulnerability.

Basic analysis First, we used techniques such as information entropy to perform simple analyses of anonymized data. The aim of this analysis was to consider the features and vulnerabilities of the anonymized data for each anonymization method. The metrics used were information entropy, distance between distributions measured by KL divergence, and the marginal distribution of each attribute's values.

The complexity of each attribute is measurable using information entropy. It can be considered safer when each attribute of the anonymized dataset is more complex than that of the original dataset, because it is more difficult to infer the original value

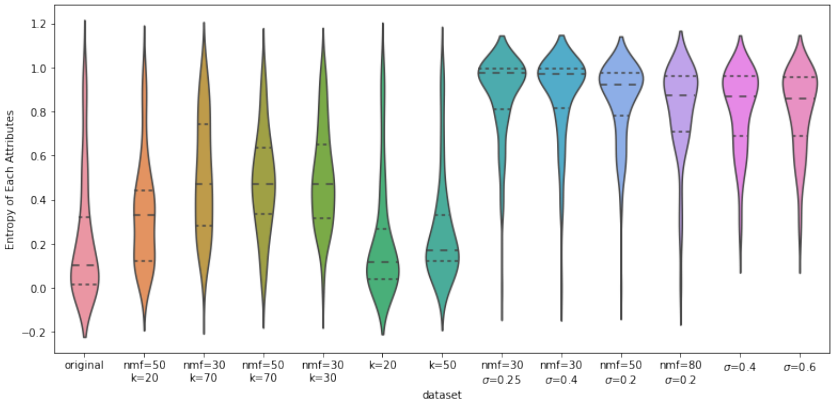

Fig. 1. Entropy details of each attribute in the adult dataset.

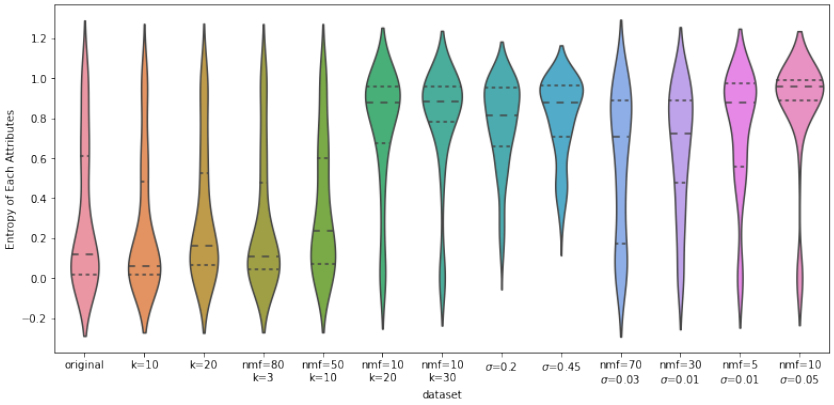

Fig. 2. Entropy details of each attribute in the diabetes dataset.

from the anonymized value. Figures 1 and 2 are plots of the information entropy of each attribute of the original and anonymized datasets on a violin graph. In the data using noise addition, the noise increased the complexity of the values. In the adult dataset, it can be seen that a combination with matrix decomposition has the same level of complexity as using noise addition alone. This positive result allows the same level of complexity to be achieved a smaller amount of noise. Noise is not added to anonymized datasets using k-anonymization; hence, the complexity of the values is at the same level as that of the original dataset. However, combining matrix decomposition tends to increase the complexity of the values. Thus, matrix decomposition apparently increases the complexity of the values and improves the anonymity of the data. However, as in the diabetes dataset, anonymizing a dataset using matrix decomposition does not increase the complexity of the values. Therefore parameters should be carefully chosen. When the complexity of the values of each attribute in the anonymized data is not high, there is likely a risk of a higher success rate of attacks against the vulnerability of the attribute.

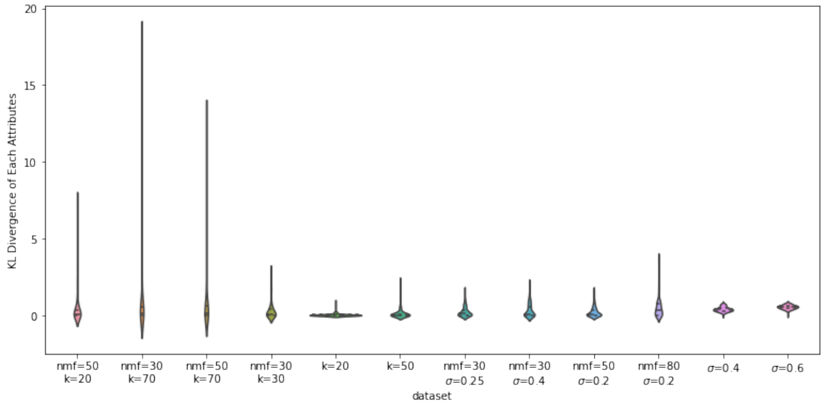

Fig. 3. KL divergence of each attribute in the adult dataset.

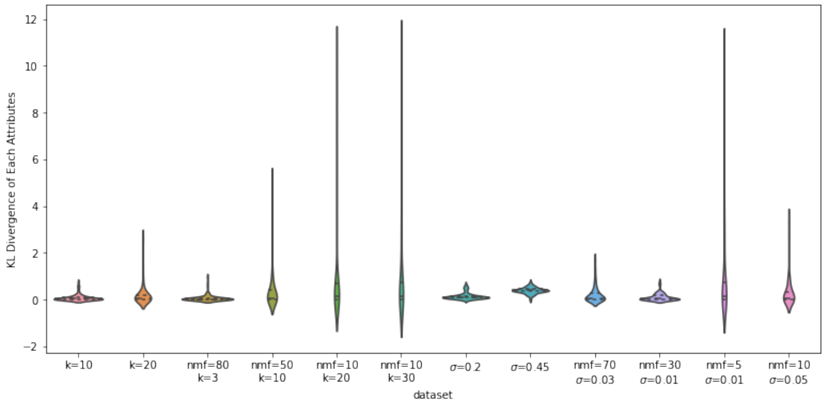

Fig. 4. KL divergence of each attribute in the diabetes dataset.

Figures 3 and 4 illustrate the distances between the distributions of each attribute of the anonymized and original datasets, measured by KL divergence. A larger distribution distance from the original data generally means a stronger level of privacy protection. When only k-anonymization is used and the value of k is small, we can confirm that there are many attributes with small distribution distances between the original and anonymized data. The anonymized data combined with matrix decomposition has attributes with larger distribution distances, which indicates stronger anonymization. This tendency can also be seen in the anonymized dataset of noise addition. When there are many attributes with small distribution distances between the original and anonymized data, we can predict that the risk is higher for attacks that take advantage of attribute vulnerability and record linkage.

Figure 5 illustrates plots of the distribution of some adult dataset attributes. Each graph represents the original data and some anonymized data. The title of the graph indicates the attributes of the target, and the legend indicates the target dataset. The attributes of the graphs are as follows: ”husband” means having a husband, ”never-

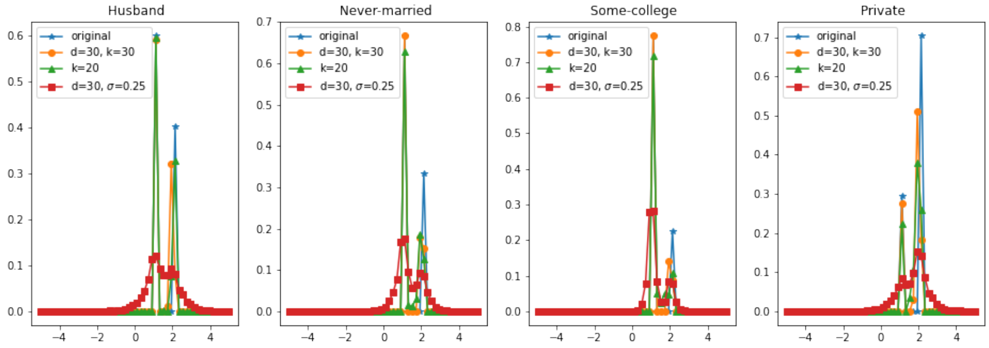

Fig. 5. Distribution of some attributes of the adult dataset.

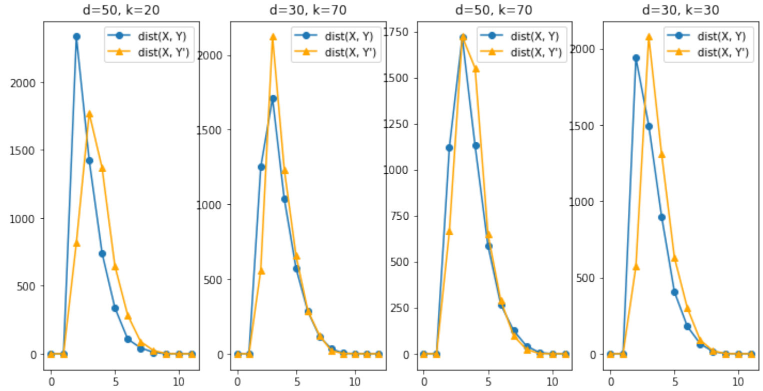

Fig. 6. Experimental results of maximum knowledge attack on an adult dataset using k-anonymization.

married" means, "some-college" means having attended college, and "private" means having a private occupation. From the figure, it can be seen that the distribution of k-anonymized data is similar to that of the original data, even when combined with matrix decomposition. When noise addition is used for anonymization, the distribution is far from the original data. Matrix decomposition did not seem to affect the distribution of attribute values, but noise addition affected the distribution of attribute values. This tendency was also observed for other attributes and parameters. If the distribution of attribute values is not different from that of the original data, the original values are more likely to be inferred even if the unusual values are anonymized. In other words, it is easy to infer the original values of the anonymized data by simply ranking the attribute values by size.
Based on the results of the above basic metric analyses, we evaluated the actual risk of information leakage in attacks on anonymized data by the following attacks.

Maximum knowledge attack We evaluated the resistance of the anonymized dataset to record linkage. For the evaluation method, we used the maximum knowledge attack

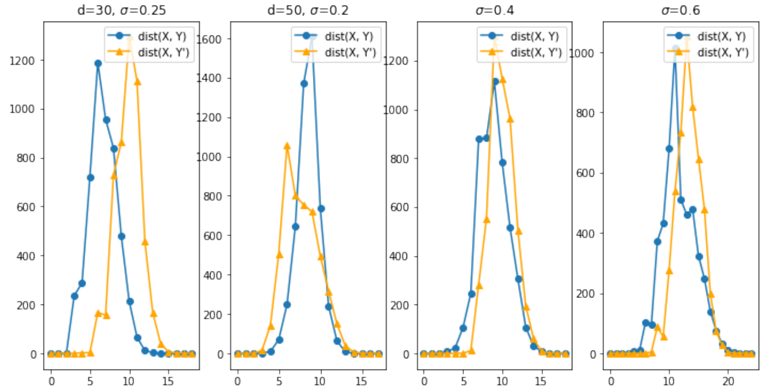

Fig. 7. Experimental results of maximum knowledge attack on an adult dataset using noise addition.

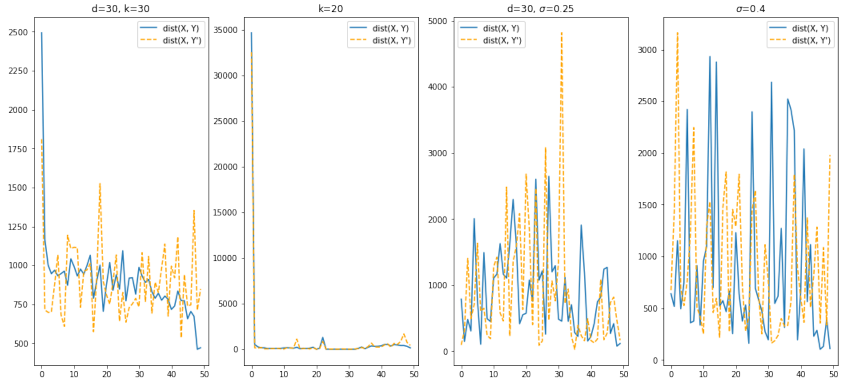

Fig. 8. Experimental results of maximum knowledge attack on an adult dataset attribute ”Workclass”.

of Ferrer et al [10]. This technique makes it is possible to evaluate the lower bound of the information leakage risk of the original data from the anonymized data. Following Ferrer’s method, we generated a permuted dataset Y' based on the anonymized data Y by removing the dependency between attributes. The distances between X and Y, and between X and Y' were then calculated before comparing the two distributions of distance. KL divergence was used to compute the distance between the two distributions. The closer the two distributions are, the less the information in X that an attacker can use to improve the accuracy of the record linkage, meaning the anonymized dataset Y shows that strong anonymization is applied.

Evaluation results of the maximum knowledge attack for adult datasets are shown in Figs 6 and 7. The title of each graph indicates the parameters of the anonymization method used. Each graph illustrates the distribution of the distance between the original data X and the anonymized data Y, and the distance between the original data X and the permuted data Y'. In Tables 5 and 6, the distance between the two distributions is calculated by KL divergence. In the evaluation of the maximum knowl-

Table 5. Distribution distance between $dist(X, Y)$ and $dist(X, Y')$ of anonymized datasets by k-anonymization.

Datasets	KL Divergence
Ano(d=50, k=20)	0.265
Ano(d=30, k=70)	0.079
Ano(d=50, k=70)	0.043
Ano(d=30, k=30)	0.248

Table 6. Distribution distance between $dist(X, Y)$ and $dist(X, Y')$ of anonymized datasets by noise addition.

Datasets	KL Divergence
Ano(d=30, σ=0.25)	2.535
Ano(d=50, σ=0.2)	0.323
Ano(σ=0.4)	0.653
Ano(σ=0.6)	0.731

edge attack of the k-anonymized dataset, it can be seen that the distribution distance between $dist(X, Y)$ and $dist(X, Y')$ is small when the parameter is $(d = 30, k = 70)$ or $(d = 50, k = 70)$ and large when the parameter is $(d = 50, k = 20)$ or $(d = 30, k = 30)$. In comparison with the results of the basic analysis, the risk of record linkage tends to increase as the number of attributes with KL divergence close to 0 increases. In other words, anonymized data with parameters $(d = 30, k = 70)$ or $(d = 50, k = 70)$, which have attributes with high KL divergence, have a lower risk of being affected by record linkage. Similarly, in the case of anonymization using noise summation, the risk of being affected by record linkage tends to be higher because the KL divergence between attributes is smaller for all parameters. When using noise addition for anonymization, the risk of record linkage will not be reduced if the noise is not increased, as stated in Ferrer et al.

In addition to the risk of record re-identification, we evaluated the vulnerability of each attribute of the anonymized data using the method of maximum knowledge attack. As introduced in Section 2.5, we linked X and Y using records other than the target's attribute x_b and y_b and compared the distance distributions of x_m and y_m with the distance distributions of x_m and y'_m for a record. The results of evaluating the vulnerability of attribute "Workclass" of the adult dataset are shown in Fig 8. From the figure, it can be seen that the two distributions of k-anonymized data are similar, and that the attributes of the anonymized data are difficult to identify. On the contrary, the two distributions are far apart in the anonymized data with noise addition, but this can be mitigated using matrix decomposition. This trend was also observed in the evaluation of other attributes. Thus, it is possible to evaluate the vulnerability of each attribute of anonymized data using Ferrer's method, but the evaluation results are not intuitive. Moreover, the level of anonymization to be achieved is unclear.

Data invasion Finally, we used our proposed data invasion attack to evaluate the privacy risk of anonymized data attributes. Our method estimates the original value from the attribute values of the anonymized dataset. Using this proposed method, we can evaluate the privacy risk of each anonymized data attribute rather than the risk of re-identification as in record linkage. The privacy risk per attribute was evaluated, and this can be used in use cases such as applying further anonymization to high-risk attributes only.

Algorithm-1 and Algorithm-2 were introduced for each assumed level of the attacker. The results of Algorithm 1 are shown in Table 7 and Algorithm 2 in Table 8. The baseline is the percentage of the most common value x in the target's attributes. This baseline refers to the highest accuracy when an attacker can guess the original value at

Table 7. Results of the data invasion attack on adult datasets. The attacker is Level-1 and does not have a marginal distribution of the attribute values of the original data.

Attribute Dataset	Workclass	Education	Marital	Occupation	Relationship	Race	Sex	Native
BaseLine	0.704	0.325	0.456	0.129	0.401	0.853	0.666	0.900
Ano(k=20)	0.543	0.476	0.715	0.696	0.805	0.429	0.985	0.311
Ano(k=50)	0.443	0.434	0.659	0.548	0.769	0.401	0.981	0.092
Ano(d=30, k=30)	0.290	0.409	0.550	0.586	0.772	0.385	0.856	0.104
Ano(σ=0.4)	0.286	0.284	0.454	0.420	0.604	0.352	0.857	0.023
Ano(d=30, σ=0.25)	0.222	0.289	0.368	0.289	0.500	0.245	0.671	0.046

Table 8. Results of the data invasion attack on adult datasets. The attacker is Level-2 and has a marginal distribution of the attribute values of the original data.

Attribute Dataset	Workclass	Education	Marital	Occupation	Relationship	Race	Sex	Native
BaseLine	0.704	0.325	0.456	0.129	0.401	0.853	0.666	0.900
Ano(k=20)	0.841	0.740	0.846	0.766	0.936	0.891	0.994	0.902
Ano(k=50)	0.842	0.600	0.826	0.596	0.916	0.888	0.992	0.900
Ano(d=30, k=30)	0.704	0.538	0.737	0.652	0.720	0.853	0.965	0.900
Ano(σ=0.4)	0.704	0.393	0.682	0.544	0.695	0.853	0.895	0.900
Ano(d=30, σ=0.25)	0.704	0.428	0.634	0.425	0.617	0.853	0.736	0.900

random. If the guess accuracy of the data invasion attack is higher than this baseline, then some information has been leaked from the anonymized dataset. Conversely, if it is smaller than the baseline, it indicates a higher level of anonymization. From Table 7, it can be seen that when the attacker does not own the distribution of the original data (Algorithm1 case), they can still estimate the value of the original data with a higher accuracy than the baseline. Particularly, the k-anonymized data showed that the estimates were highly accurate. The reason for this is supported by the analysis of the distribution of values performed in the basic analyses (Fig 5). As the distribution of the values after anonymization has not changed much, the original values could easily be estimated in this attack using value ordering. Previous experiments have shown that k-anonymization is resistant to record re-identification. Thus, the risk of being identified is low. Nevertheless, care is needed in combining it with matrix decomposition because the resistance is also weak. The data invasion attack was less accurate when noise-addition anonymization was used; this result is also consistent with the results of the basic analysis. However, looking at the results of Algorithm-2, it can be seen in Table 8 that the probability of the original value being estimated is high regardless of which anonymization method is used. That is, an attacker has a marginal distribution of the original data, they can easily infer the original data from the anonymized data. From this result, we consider that when publishing anonymized data, it is necessary to keep the marginal distribution of the original data as confidential information or distort the distribution of attribute values significantly.

6 Discussion

6.1 Comparison of our method with conventional methods

Our method shows that an attacker can infer an original value of target attribute from the anonymized data when the attacker knows anonymized dataset and the possible values of the original data. This technique is an attack based on the results of basic analysis, which shows that the distribution of each attribute value in the anonymized data is not so different from the original data. This attack gives an intuitive indication of the extent to which the value of the target attribute is likely to be inferred. Ferrer's method [10] can also evaluate the vulnerability of each attribute of anonymized data, but there are two problems: the assumption of attacker is too strong, and it is unclear how much of the data should be anonymized in practice. We have assumed a realistic attacker and can check the level of anonymization by comparing the baseline with the evaluation results. We show in Table 9 that our attack more strongly reflects the difference in distribution distance between the original data and the anonymized data. This table shows the correlation coefficient between distribution distance between the anonymized data and the original data for an attribute and each evaluation method. From the table, we can see that the accuracy of the proposed method increases as the distribution distance between the original data and the anonymized data gets closer, which confirms that our proposed method can provide intuitive indicators that the distribution of data is sensitive to privacy.

6.2 Matrix Decomposition and Privacy

For record re-identification attacks, when k-anonymization is insufficient, we may reduce the risk of record re-identification attacks by adjusting the parameters with matrix-decomposition anonymization. When adjusting parameters, the distribution distance between the original and anonymized data should be large to reduce the risk of record re-identification. When using noise addition, the noise should be increased rather than the parameters adjusted because risk reduction results were not observed from matrix decomposition.

While matrix decomposition may provide a small reduction in privacy risk for attacks that take advantage of attribute vulnerabilities, Table 7 shows that an attacker can estimate the original value with greater accuracy than the baseline even when matrix decomposition is used. Furthermore, no anonymization method can be effective if the attacker has a marginal distribution of the original data.

We confirmed that the risk of certain attacks can be reduced by using matrix decomposition as described above. However, matrix decomposition may still increase privacy risk; therefore, it is necessary to consider the scenario in which the data are attacked and to analyze them sufficiently in advance.

7 Conclusion

In this study, we conducted a multifaceted safety analysis of anonymization techniques proposed to anonymize large-scale and high-dimensional data. Analyzing the anonymized data using basic metrics revealed that the distribution of the data for each attribute did not change significantly after anonymization, and we proposed a method

Table 9. Correlation coefficients between the evaluated value of anonymized data attributes and the distance distribution of attribute values of the original and anonymized data.

Method Attribute	Ferrer's[10]	Proposed1	Proposed2
Workclass	0.503	-0.942	-0.991
Education	0.249	-0.509	-0.722
Marital	0.912	-0.962	-0.966
Occupation	0.323	-0.529	-0.655

for estimating the original data using this feature. Experimental results further showed that our attack can be used to estimate the value of the original data with high accuracy when the attacker knows the distribution of the original data. Our evaluation of the resistance to data re-identification using existing techniques established that each anonymization method has its own suitable attack and it is crucial to assess the safety of the data using various metrics before the data are published.

References

1. S.Latanya, k-anonymity: a model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, Volume 10, Issue 5, pp.557-570, 2002
2. A.Narayanan, V.Shmatikov, Robust de-anonymization of large sparse datasets, In Proceedings of 2008 IEEE Symposium on Security and Privacy(SP), 2008
3. A.Machanavajjhala, D.Kifer, J.Gehrke, l-Diversity: Privacy Beyond k-Anonymity, ACM Transactions on Knowledge Discovery from Data, 2007
4. L.Ninghui, L.Tiancheng, V.Suresh, t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, IEEE 23rd International Conference on Data Enginering, 2007
5. T.Mimoto, S.Kiyomoto, S.Hidano, A.Basu, A.Miyaji, The Possibility of Matrix Decomposition as Anonymization and Evaluation for Time-sequence Data, 16th Annual Conference on Privacy, Security and Trust (PST), 2018
6. W.Isabel, Genomic Privacy Metrics: A Systematic Comparison, IEEE Security and Privacy Workshops (SPW), 2015
7. C.Aggawal, On k-anonymity and the curse of dimensionality, In Proceedings of the 31st international conference on Very learge data bases, pp.901-909, 2005
8. T.Mimoto, S.Kiyomoto, S.Hidano, A.Basu, A.Miyaji, The Possibility of Matrix Decomposition as Anonymization and Evaluation for Time sequence Data, Annual Conference on Privacy, Security, 2018
9. T.Mimoto, S.Kiyomoto, S.Hidano, A.Miyaji, Anonymization Technique Based on SGD Matrix Factorization, IEICE TRANSACTIONS on Information and Systems, 2020
10. J.Ferrer, S.Ricci, J.Comas, Disclosure Risk Assessment via Record Linkage by a Maximum-Knowledge Attacker, Annual Conference on Privacy, Security and Trust (PST), 2015
11. S.Ito, H.Kikuchi, H.Nakagawa, Attacker models with a variety of background knowledge to de-identified data, Journal of Ambient Intelligence and Humanized Computing, 2019
12. UCI machine learning repository (2018) Adult data set. https://archive.ics.uci.edu/ml/datasets/adult.
13. UCI Machine Learning Repository (2018) Diabetes 130-US hospitals for years 1999–2008 Data Set [online]. https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008