Noname manuscript No.
(will be inserted by the editor)

Attacker Models with a Variety of Background Knowledge to De-identified Data

Satoshi Ito · Hiroaki Kikuchi · Hiroshi Nakagawa

Received: date / Accepted: date

Abstract De-identification is a process to prevent individuals from being identified from original transaction data by processing personal identification information. For de-identifying data, we should consider the risk of attackers that try to infer personal information from given de-identified datasets. However, the characteristics of attackers are not known. Attackers could find the weakest records to identify the user who owns them. In the case of big data, there are too many attributes to identify the weakest parts to be protected. To address these issues, we propose a new model of attackers that gain background knowledge from an attribute of the transaction data and estimate the risk of re-identification from the statistics of the dataset, such as the number of records and the background number of users. We present empirical results on evaluating the risks of actual transaction data using our proposed risk model.

Keywords Personal Identification Information · Privacy Risk Evaluation · De-identification · Transaction Data

Satoshi Ito
Meiji University Graduate School, Tokyo, 164-8525 Japan
E-mail: mmhm@meiji.ac.jp

Hiroaki Kikuch
Meiji University Graduate School, Tokyo, 164-8525 Japan
E-mail: kikn@meiji.ac.jp

Hiroshi Nakagawa
RIKEN Center for Advanced Intelligence Project, Tokyo, 103-0027, Japan
E-mail: hiroshi.nakagawa@riken.jp

1 Introduction

De-identification is a process to prevent individuals from being identified from original transaction data by processing personal identifying information (PII). Companies are required to assess the re-identification risks when employing big data extensively in their businesses. However, there is concern that attackers may try to re-identify individuals from available de-identified data using external background knowledge. For example, when an attacker tries to re-identify a user from purchase history data, he/she may use the background knowledges about attribute, e.g., the target user's age, sex, and ethnicity. It is unclear what kinds of background knowledge are available to these attackers. In addition, attackers may try to find the least protected records in published data in an effort at re-identification. For example, a unique record might be identified when it contains information about an unusual disease that a particular patient suffers. However, there are too many attributes and records to alter to reduce the risk of disclosure. It is not trivial to determine which attributes of the original data need to be processed to de-identify the data against attackers.

In order to determine the risky attributes to be processed, Domingo-Ferrer et al. proposed a model of a maximum-knowledge attacker who knows all original attribute values for all subjects (Domingo-Ferrer et al., 2015). In this attacker model, the attacker is supposed to know both the original and the de-identified datasets. So, it models a worst-case attacker who has all the background information he can possibly use. Therefore, this model has been adopted in much research, such as in the data anonymization competition, Privacy Workshop Cup (PWS CUP) (Kikuchi et al., 2016). In this model, the attacker can use all attributes as background knowledge to estimate the most likely linkages. However, this assumption is too strong to model a realistic environment and we relax the assumptions to make the attacker model more realistic.

El Emam studied the four realistic attacker models, "Intentional Attack", "Unintentional Attack", "Data Breach", and "Open Data Attack" (Emam and Arbuckle, 2013). These models need objective data, such as the probability of being attacked and the probability of a data breach occurring. However, it is well known that there is no universal attacker model because the impact differs from case to case. For example, the impact of disclosure of a cancer patient is more serious than that of a purchase from a convenience store.

In this paper, we study new model of risk of attackers depending on their background knowledge and estimate such risk for transaction data. Instead of assuming a *universal* model, we consider a *specific* attacker that gains background knowledge from an attribute of the transaction data with probability determined by statistics of the dataset, such as the number of records and the number of users who have the background knowledge. We consider not only the worst-case attacker, but also wide range of attackers with different background knowledge. In the worst-case of maximum knowledge attacker, the risk of disclosure is possibly overestimated since such a strong attacker is far from reality. We propose models to approximate the risk of re-identification of de-identified data as the expected value of the identification probability of the attacker with particular background knowledge, "the mean identification probability".

One of the drawbacks of the mean identification probability is its computational cost to evaluate. The processing time to examine all records of the dataset is proportional to the number of records, which could be huge. In addition, the cost depends of the number of attributes of the dataset. To address the cost issue, we propose three approximation methods, called the mean model, the low-cost model and the sampling model. There is a trade-off between the accuracy and the efficiency for the models and should be chosen based on the requirement of use case.

To show that our proposed method can be applied to a wide range of datasets, we present empirical results on evaluating the risks of the typical actual transaction data in our proposed risk model. In the experiments, we investigate four datasets, the purchase history, the hospitalization data for disease, the census income dataset, and the loan data, with distinct characteristics, e.g., the ratio of number of records and the number of users varies with the four datasets.

Our study makes three contributions: (1) we propose a new model of attacker and risk of data; (2) we propose three methods to approximate the measure of the risk of data; and (3) we make an empirical evaluation of actual transaction data by applying our risk model.

The remainder of the paper is organized as follows. In Section 2, we consider the models of datasets and attackers. In Section 3, we propose a theoretical risk model. Section 4 evaluates the risks of the actual transaction data by applying our risk model. Section 5 describes related work, and Section 6 concludes the paper.

2 The Dataset and Attacker Models

2.1 Dataset Model

We consider transaction data that consist of records (rows) and attributes (columns) and have an attribute that identifies individuals. We define our model as follows.

Definition 1 *Let T be a set of transaction data. Let m and n be the number of records and the number of users in transaction data T, respectively. Let D_X be the set of values for attribute X of T. Let R_x and U_x be the set of records containing a given $x \in D_X$ and the set of users that have x in attribute X, respectively. Let T' be a pseudonymised transaction data T such that replacing the identifiers (like user IDs) of T with unique characters chosen at random as pseudo IDs.*

Example 1 *As an example of T, Table 1 shows purchase transaction data T_{Example} of three users $(1, 2, 3)$ for three days $(2010/12/1 - 2010/12/3)$. For example, we can find that user 2 purchased three loaves of bread on 2010/12/1 from these data. In the case of T_{Example}, m is 10 and n is 3 and when X is* **date***, D_X is $\{2010/12/1, 2010/12/2, 2010/12/3\}$ and $|D_X|$ is 3. When x is $2010/12/1$, R_x is $\{1, 2, 3, 4\}$ and U_x is $\{1, 2\}$. Table 2 shows a processed purchase transaction data T'_{Example}. The values 1, 2, and 3 of* **user ID** *attributes of T are replaced with A, B, and C.*

2.2 Attacker Model

We consider an attacker that accidentally gains background knowledge x of attribute X, which helps to identify users (specified in T). In this paper, we assume that the probability of gaining background knowledge x for the attacker is proportional to the frequency of the records that contain x in transaction data. For example, when an attacker observe customer purchasing in front of supermarket, the records that frequently occur like "a customer purchased in 13:00" are likely to be available to attacker as background knowledge than the records that rarely occur like "a customer purchased in 2:00 midnight".

Definition 2 *The probability $Pr(x)$ of gaining background knowledge x for the attacker is proportional to the frequency of x in T, i.e., $Pr(x) = |R_x|/m$.*

The attacker that has processed data T' gains U_x as candidate users when he/she can access records that contain x of T as background knowledge. Therefore, we define the risk of re-identification denoted as 'idf' as a conditional probability of given x.

Definition 3 *An attacker identifies (*idf*) individual from side knowledge x of T, with a conditional probability given x, that is, $Pr(\text{idf}|x) = 1/|U_x|$.*

Based on the definitions, 2 and 3, a joint probability $Pr(\text{idf}, x)$ to gain background knowledge x for the attacker and to identify the individual is calculated as

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

Letting $\alpha_x = |R_x|/|U_x|$, we rewrite it as

$$Pr(\text{idf}, x) = \frac{\alpha_x}{m},$$

where α_x is the mean number of records for user that has x. Since α_x is very important for calculating and approximating risk, we define this as follows.

Definition 4 *Let α_x be the mean number of records for user about background knowledge x of attribute X. Let α_X be the mean of α_x for attribute X, i.e., $\alpha_X = \frac{1}{|D_X|}\sum_{x \in D_X} \alpha_x$.*

Example 2 *Table 3 shows $|R_x|$, $Pr(x)$, $|U_x|$, $Pr(\text{idf}|x)$, $Pr(\text{idf}, x)$ for the background knowledge x in three days in case of T_{Example} and X is* **date**. *In this case, D_X is $\{2010/12/1, 2010/12/2, 2010/12/3\}$. When x is $2010/12/3$, $Pr(x)$ is $3/10$ because R_x is $\{8, 9, 10\}$ and m is 10. The probability of gaining background knowledge is $3/m$ in all cases (before/after anonymization) because the attacker always gains background knowledge from the original transaction data. $Pr(\text{idf}|x)$ is $1/1$ because U_x is $\{3\}$. The attacker identifies a user u from x with the probability of*

$$Pr(\text{idf}, x) = Pr(x)Pr(\text{idf}|x) = 0.3 \cdot 1 = 0.3$$

Also, because $\alpha_x = 3/1 = 3$, simply we have

$$Pr(\text{idf}, x) = \frac{\alpha_x}{m} = \frac{3}{10} = 0.3.$$

Table 1 Example Purchase Data T_{Example}

ID	user ID	date	time	goods	price	number
1	1	2010/12/1	8:45	Bread	1.45	2
2	1	2010/12/1	8:45	Book	3.75	1
3	1	2010/12/1	20:10	Tea	0.85	2
4	2	2010/12/1	10:03	Bread	1.45	3
5	1	2010/12/2	15:07	Tea	0.85	3
6	3	2010/12/2	11:57	Bread	1.45	4
7	3	2010/12/2	11:57	Juice	1.25	4
8	3	2010/12/3	15:54	Book	3.75	1
9	3	2010/12/3	15:54	Tea	0.85	10
10	3	2010/12/3	15:54	Juice	1.45	10

Regardless of how the dataset is anonymized, our proposed methods allow to model how risk is reduced based on the number of risky records and chance for attacker to have the knowledge on the target attribute.

Definition 5 *(Mean Identification Probability) The mean identification probability is a probability of individual to be identified by the attacker who has all x of X as background knowledge, denoted by*

$$Pr(\text{idf}, X) = \sum_{x \in D_X} Pr(x)Pr(\text{idf}|x) = \sum_{x \in D_X} Pr(\text{idf}, x).$$

Example 3 *In case of $T_{Example}$, the identification probabilities of 4 records that contain "2010/12/1" are $Pr(\text{idf}|x) = 1/2$ and those of 3 records that contain "2010/12/2" are $Pr(\text{idf}|x) = 1/2$ and those of 3 records that contain "2010/12/3" are $Pr(\text{idf}|x) = 1$. So, the mean of $Pr(\text{idf}|x)$ is $Pr(\text{idf}, X) = 13/20$.*

Based on Definition 4 and 5, we have that

$$Pr(\text{idf}, X) = \sum_{x \in D_X} Pr(\text{idf}, x) = \sum_{x \in D_X} \frac{\alpha_x}{m}.$$

Example 4 *Given T_{Example} and X of* **date**, *we calculate the mean identification probability $Pr(\text{idf}, X)$ is calculated as*

$$Pr(\text{idf}, X) = \sum_{x \in D_X} \frac{\alpha_x}{m} = \frac{2 + 1.5 + 3}{10} = 0.65.$$

This means that the attacker who has background knowledge of the **date** *attribute identifies u from T_{Example} with the mean probability of 0.65.*

In this paper, we regard the mean identification probability, $Pr(\text{idf}, X)$, as the risk of an attribute X. We observe that the cost of calculation of risk is proportional to the number of records to calculate the risk.

Table 2 Processed Example Purchase Data T'_{Example}

ID	pseudo ID	date	time	goods	price	number
1	A	2010/12/1	8:45	Bread	1.45	2
2	A	2010/12/1	8:45	Book	3.75	1
3	A	2010/12/1	20:10	Tea	0.85	2
4	B	2010/12/1	10:03	Bread	1.45	3
5	A	2010/12/2	15:07	Tea	0.85	3
6	C	2010/12/2	11:57	Bread	1.45	4
7	C	2010/12/2	11:57	Juice	1.25	4
8	C	2010/12/3	15:54	Book	3.75	1
9	C	2010/12/3	15:54	Tea	0.85	10
10	C	2010/12/3	15:54	Juice	1.45	10

Table 3 Identification Probability about **date** Attribute of T_{Example}

x	$\|R_x\|$	$Pr(x)$	$\|U_x\|$	$Pr(\text{idf}\|x)$	$Pr(\text{idf}, x)$	α_x
2010/12/1	4	0.4	2	0.5	0.2	2
2010/12/2	3	0.3	2	0.5	0.15	1.5
2010/12/3	3	0.3	1	1	0.3	3
Sum	10	1.0			0.65	

3 Risk of the Attribute of Transaction Data

We have to calculate α_x for all x of X to calculate the mean identification probability based on Definition 5. However, examining all α_x for a big dataset is not efficient. So, we consider three alternative methods to approximate it as follows:

1. The mean model,
2. The low-cost model, and
3. The sampling model.

3.1 The Exact Solution

The risk of re-identification of T' depends on the background knowledge attribute X that the attacker has. Therefore, we define the risk of re-identification $R(X)$ as the mean identification probability of X, i.e., $R(X) = Pr(\text{idf}, X)$. We have to calculate α_x for all x of X to calculate the exact solution of $R(X)$ so the calculating cost is m, in this case.

3.2 The Mean Model

The mean model calculates a risk of attribute X with the mean of α_x.

Definition 6 *Let $R_{mean}(X)$ be the risk of attribute X in the mean model defined as*

$$R_{mean}(X) = \frac{\alpha_X |D_X|}{m}.$$

It is interesting that the risk calculated by the mean model gives the exact solution as follows.

The risk of X in the mean model is $R_{mean}(X) = Pr(\text{idf}, X)$. Based on Definition 6, $R_{mean}(X)$ is calculated as $R_{mean}(X) = \alpha_X |D_X|/m$. Based on Definition 4, α_X is calculated as $\alpha_X = \frac{1}{|D_X|} \sum_{x \in D_X} \alpha_x$. Therefore, $R_{mean}(X)$ is transformed as follows.

$$\begin{aligned} R_{mean}(X) &= \frac{\alpha_X |D_X|}{m} \\ &= \frac{|D_X|}{m} \sum_{x \in D_X} \frac{\alpha_x}{|D_X|} \\ &= \sum_{x \in D_X} \frac{\alpha_x}{m} \\ &= \sum_{x \in D_X} Pr(x) Pr(\text{idf}|x) \\ &= Pr(\text{idf}, X) \end{aligned}$$

Example 5 *When* $T = T_{\text{example}}$ *and* $X =$ **date**, *we have that*

$$Pr(\text{identify}, X) = \frac{\alpha_X |D_X|}{m} = \frac{\frac{13}{6} \cdot 3}{10} = 0.65$$

Unfortunately, we have to calculate α_x for all x of X to calculate α_X, so the calculating cost is m, in this model.

3.3 The Low-cost model

Based on our experiments in Section 4, we observed that the mean numbers of record per user (α_x) are close to 1.0 for many datasets. By assuming that the mean numbers of record per user, denoted by α_x, is 1.0 for all values x, we minimize the calculating cost to have an approximated risk for interested attributes.

Definition 7 *Let $R_{cost}(X)$ be the risk of X calculated in the low-cost model defined as*

$$R_{cost}(X) = \frac{|D_X|}{m}.$$

Example 6 *When* $T = T_{\text{example}}$ *and* $X =$ **date**, *we have that*

$$R_{cost}(X) = \frac{|D_X|}{m} = \frac{3}{10} = 0.3.$$

Theorem 1 *The error rate of the low-cost model is* $|1 - \frac{1}{\alpha_X}|$.

(Proof) The relative error rate of $R_{cost}(X)$ for the exact solution, according to Section 3.2, is

$$\begin{aligned} &= \frac{|R_{cost}(X) - Pr(\text{idf}, X)|}{Pr(\text{idf}, X)} \\ &= \frac{\left|\frac{|D_X|}{m} - \frac{\alpha_X |D_X|}{m}\right|}{\frac{\alpha_X |D_X|}{m}} = \left|\frac{1}{\alpha_X} - 1\right|. \end{aligned}$$

Thus, we have proved Theorem 1. (Q.E.D)

The number of records m of T and the number of kinds of X are given in this research. We do not have to use the records to approximate the risk in this model, that is, the cost for calculating risk is 0.

3.4 The Sampling Model

The sampling model is a model that approximates the risk of X by calculating the mean of α_x of a randomly selected subset of T. Note that we take a random sample of all records that contain a value x. For example, when "2010/12/1" is randomly chosen from attribute **date** of T_{Example}, we sample all records (four records) that contain "2010/12/1" from T_{Example}.

Definition 8 *Let $D'_X = \{x_1, \ldots, x_s\}$ be a subset of D_X of size s randomly sampled from D_X. Then, we define $\alpha_{x'} = \frac{1}{s}\sum_{i=1}^{s} \alpha_{x_i}$. Let $R_{sample}(X)$ be the risk of X calculated in the sampling model defined as*

$$R_{sample}(X) = \frac{\alpha_{x'}|D_X|}{m}.$$

Example 7 *When $T = T_{\text{example}}$ and $X =$ **date** and $s = 2$ and $D'_X = \{2010/12/1, 2010/12/3\}$, we have that*

$$R_{sample}(X) = \frac{\alpha_{x'}|D_X|}{m} = \frac{2.5 \cdot 3}{10} = 0.75$$

because $\alpha_{x_1} = 2$ and $\alpha_{x_2} = 3$.

Let σ_s be a standard deviation of s samples.

Theorem 2 *The maximum of error rate of the sampling model is*

$$\frac{\sigma_s m}{\sqrt{|s|}|D_X|\alpha_X}. \tag{1}$$

(Proof) According to Section 3.2, $Pr(\text{idf}, X) = \alpha_X$ $|D_X|/m$. The absolute error of $Pr(\text{idf}, X)$, $|R_{sample}(X) - Pr(\text{idf}, X)| < Var[Pr(\text{idf}, X)] = \sigma_s/\sqrt{s}$ when we suppose a confidence interval of 90 percent. Therefore, the relative error rate $R_{sample}(X)$ for the exact solution is

$$= \frac{|R_{sample}(X) - Pr(\text{idf}, X)|}{Pr(\text{idf}, X)}$$

$$< \frac{\frac{1}{\sqrt{s}}\sigma_s}{\frac{\alpha_X|D_X|}{m}} = \frac{\sigma_s m}{\sqrt{|s|}|D_X|\alpha_X}.$$

Thus, we have proved Theorem 2.

(Q.E.D)

We have to calculate $\alpha_{x'}$ to evaluate the risk. If the elements of D'_X are chosen with uniform probability $1/|D_X|$, the calculating cost is $sm/|D_X|$.

With Table 4, we show the summary of results for our three models. The calculating cost of the exact value and the mean model are maximum and that of low-cost model is minimum and that of sampling model depends on sampling size s. The error rate of the exact value and the mean model are minimum. Figure 1 illustrates a schematic view of our approximate models. The mean model exploits *all* transaction data. The sampling model exploits *a part* of transaction data. The low-cost model does use *no* transaction data for approximating the risk.

Table 4 Error rate and Calculating Cost of Our Approximate Models

Model	Risk	Error Rate	Cost
Exact Solution	$R(X)$	0	m
Mean	$R_{mean}(X)$	0	m
Low Cost	$R_{cost}(X)$	$\frac{1}{\alpha_X} - 1$	0
Sampling	$R_{sample}(X)$	Eq. (1)	$sm/\|D_X\|$

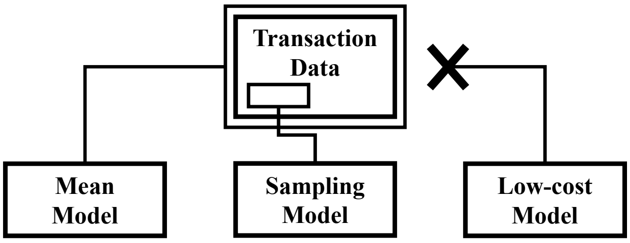

Fig. 1 Schematic Views of Our Approximate Models and Transaction Data for Calculation Risk

4 Risk Evaluation Experiment

4.1 Experiment Objective

In order to examine how universally our model works, we evaluate the risk of the actual datasets by the mean identification probability described in Section 3. For the risk evaluation experiment, we use the following three open datasets that are published in UCI Machine Learning Repository (UCI Machine Learning Repository, 2018c) and one dataset that is published in Lending Club (Lending Club, 2019b).

1. T_1: Online Retail Dataset, the purchase history data for one year in the UK. (UCI Machine Learning Repository, 2018d)
2. T_2: Diabetes Dataset, the hospitalization data of diabetics for 10 years. (UCI Machine Learning Repository, 2018b)
3. T_3: Adult Dataset, the census income dataset. (UCI Machine Learning Repository, 2018a)
4. T_4: LOAN DATA, loan data for all loans issued of 2007–2011. (Lending Club, 2019a)

Table 5 shows quantities m, n, the number of attributes of T_1, T_2, T_3 and T_4. We treat T_3 and T_4 as a transaction data, though these are not exactly transaction data.

4.2 Analysis of the Dataset

Table 5 shows the summary of attributes for T_1. These data consist of seven attributes and we adopt five attributes (**date**, **time**, **goods**, **price**, **number**) as the candidates for X. Figures 2–6 show the distribution of α_x for each attribute and Table 5 shows the statistics including α_X and $|D_X|$ for each attribute.

In the case of attributes **date** and **time**, the mean number of records α_x is generally high. For example, when X = **date** and $x = 2011/8/28$, $\alpha_x = 122$ (one user had 122 purchase records in one day). In our model, such x is evaluated to be risky because it is likely to be gained as the background knowledge and be used to identify an individual from T. On the other hand, in the case of the attributes, **goods**, **price**, and **number**, α_x is generally low and $\alpha_x = 1$ for most x.

Figures 7 and 8 show the scatter diagram for $|R_x|$ and $|U_x|$ when $T = T_1$ and $X = \textbf{date}, \textbf{price}$. The x-axis shows $|U_x|$ and the y-axis shows $|R_x|$. The red line shows the line of $|U_x| = \alpha_X |R_x|$ (the mean model) and the green line shows $|U_x| = |R_x|$ (the low-cost model).

Table 5 shows the summary of attributes for T_2 , T_3 and T_4. We take four attributes (**ethnicity**, **gender**, **age**, **time**) chosen out of 50 attributes as the candidates of X because these attributes are possible as background knowledge from T_2. T_3 consists of 17 attributes and we note four attributes (**age**, **martial**, **occupation**, **ethnicity**) as candidates of X as possible attributes as background knowledge.

Table 5 shows statistics α_X and $|D_X|$ about each attribute too. Figure 9 and 10 show the distribution of α_x for the attributes of **age** and **date** of T_2.

When $T = T_3$, $\alpha_X = 1$ for any x of any X because $m = n$ and $|R_x| = |U_x|$.

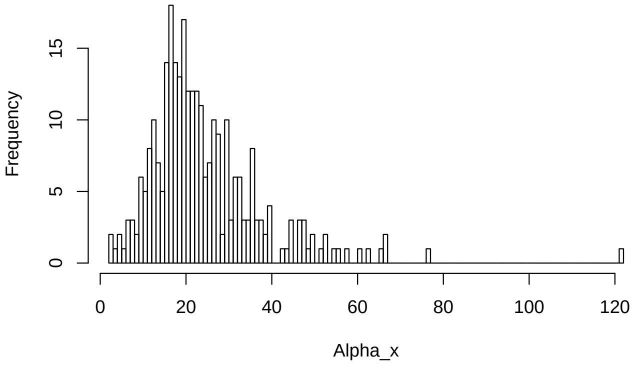

Fig. 2 Distribution of α_x when $T = T_1, X =$ **date**

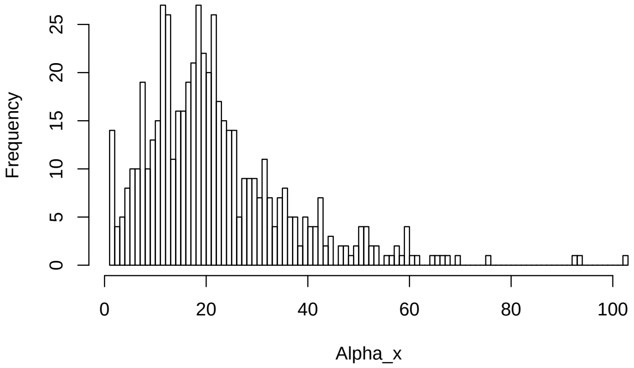

Fig. 3 Distribution of α_x when $T = T_1, X =$ **time**

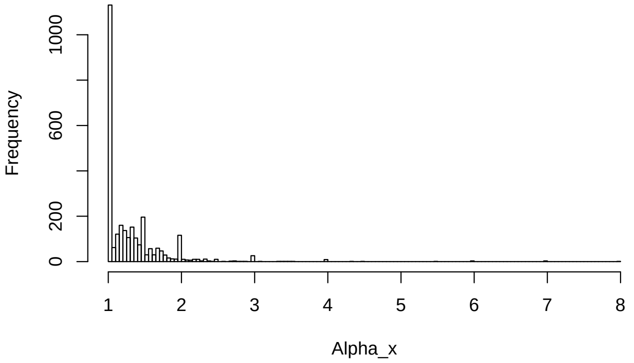

Fig. 4 Distribution of α_x when $T = T_1, X =$ **goods**

4.3 Result of Evaluation of Risk

We calculate the risk (exact solution) for each attribute involved. Table 5 shows $R(X)$ for each attribute of T_1, T_2, T_3, and T_4. For example, the risk of the attribute **date** of T_1 is 0.186. In the case of T_1, the attribute **time** is evaluated as the riskiest at 0.322. Hence, we should process attribute **time** by appropriate de-identification technique, e.g., rounding exact time (8:45) to hours (8:00), perturbation (8:42), and column suppression. This makes sense since attributes related **time** and **date** are often suppressed in de-identification use cases.

In the cases of T_2 and T_3, the risks are quite small because $|D_X|$ and α_X are small; the attribute **days** for T_2 and the attribute **age** for T_3 are evaluated as the riskiest for each dataset. Note that the mean number of records for user α_X is close to 1.0 in T_2, T_3, and T_4 and the ranks of mean identification probability $Pr(\text{idf}, X)$ are consistent with that of the unique values in attribute $|D_X|$. While, the risk of dataset with large α_X such as T_1 need to be carefully evaluated. Consequently, our model can be applied to arbitrary dataset.

Table 5 Details of T_1, T_2, T_3, T_4 and Risks of Each Attribute of These Data

T	m	n	#Attribute	X	Description	α_X	$\|D_X\|$	$Pr(\text{idf}, X)$
T_1	38,087	400	7	**time**	Purchase time (hh:mm)	22.23	551	0.322
				date	Purchase date (yyyy/mm/dd)	24.42	290	0.186
				goods	ID of purchased goods (number and character)	1.32	2,781	0.097
				price	Price of purchased goods (Pound sterling)	2.49	184	0.012
				number	Quantity of purchased goods (number)	3.15	97	0.008
T_2	101,766	71,518	50	**days**	Days in hospital (number)	1.05	14	$1.45 \cdot 10^{-4}$
				age	Age of patient (number)	1.35	10	$1.33 \cdot 10^{-4}$
				ethnicity	Ethnicity of patient (character)	1.31	6	$7.73 \cdot 10^{-5}$
				gender	Gender of patient (character)	1.28	3	$3.78 \cdot 10^{-5}$
T_3	32,561	32,561	16	**age**	Age of user (number)	1	73	$2.24 \cdot 10^{-3}$
				occupation	Occupation of user (character)	1	15	$4.61 \cdot 10^{-4}$
				martial	Marital status of user (character)	1	7	$2.15 \cdot 10^{-4}$
				ethnicity	Ethnicity of user (character)	1	5	$1.54 \cdot 10^{-4}$
T_4	42,538	42,538	145	**employment**	Employment of customers (character)	1	30,661	0.721
				income	Annual income of customers (number)	1	5,597	0.132
				amount	Amount of loan (number)	1	898	0.021
				grade	Grade of customers (character)	1	8	0.000

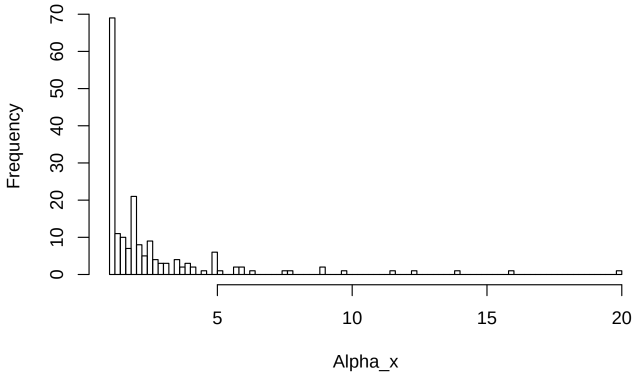

Fig. 5 Distribution of α_x when $T = T_1, X =$ **price**

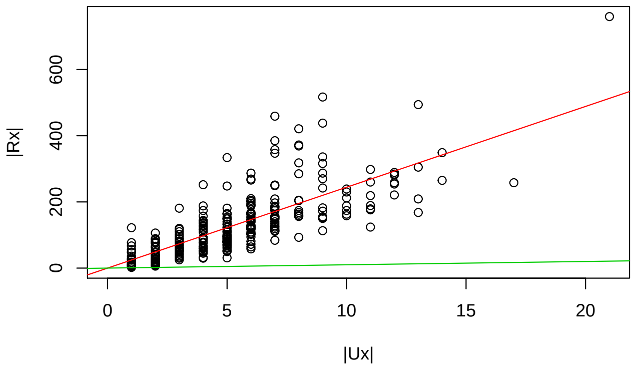

Fig. 7 Scatter diagram for $|R_x|$ and $|U_x|$ when $T = T_1, X =$ **date**

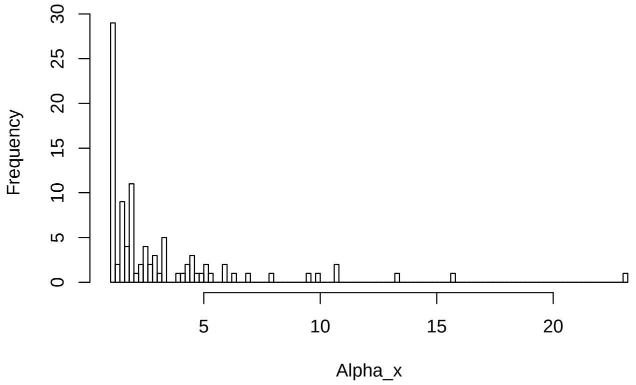

Fig. 6 Distribution of α_x when $T = T_1, X =$ **number**

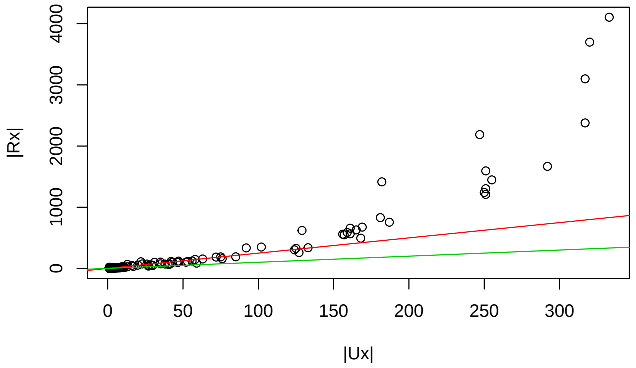

Fig. 8 Scatter diagram for $|R_x|$ and $|U_x|$ when $T = T_1, X =$ **price**

4.4 Accuracy and Cost of Our Models

We calculate the risk of each attribute for T_1, T_2, T_3, and T_4 efficiently with the mean model, the low-cost model, and the sampling model. Table 6 shows the empirical results for each model. Based on Section 3.2, the estimation value $R_{mean}(X)$ in the mean model is equal to $R(X)$ in Table 5. The evaluation value $R_{sample}(X)$ in the sampling model is provided with the 90 percent confidence interval ($\mu \pm \sigma$) when $s = 10$. The values marked * indicate the attribute that is evaluated as the highest risk for each data in our approximate methods. The risks in attributes are almost consistent in either model except the highest one in T_1. For example, the mean model (= the exact solution) identifies attribute **time** as the riskiest for T_1, while the low-cost model evaluates the **goods** attribute as the riskiest. In the

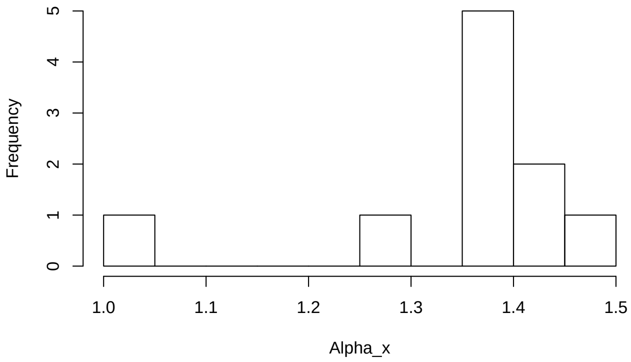

Fig. 9 Distribution of α_x when $T = T_2, X =$ **age**

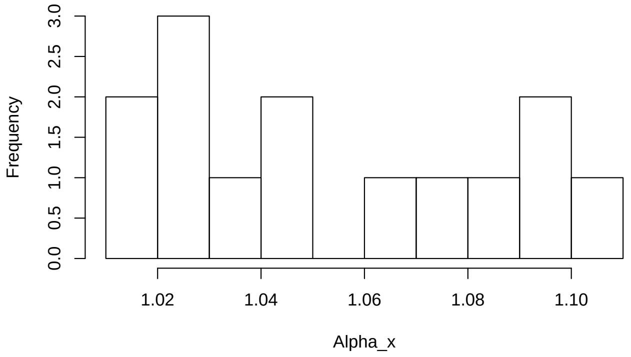

Fig. 10 Distribution of α_x when $T = T_2, X =$ **days**

case of the sampling model, the attribute becoming the maximum value in partial relation of confidence interval is **time**.

Figure 11 shows the scatter diagram for the error and the calculation cost for each model when $T = T_1$ and $X =$ **date**. The x-axis shows the logarithm of the calculation cost (the number of records) and the y-axis shows the error from the exact solution $Pr(\text{idf}, X)$.

The red points in Figure 11 show the results for these models. The gray points show the risks evaluated for all elements of D_X. We adopt the barycenter as the representative point for this model. Note that the sampling model evaluates the risk from s points that are randomly chosen from $|D_X|$ points. Table 7 shows the error and cost of each model.

Figure 12 shows the distribution of α_X when we sample 50 elements from D_X for 1,000 times randomly for $T = T_1, X =$ **date**. Table 8 shows the mean and the standard deviation of α_X for the number of groups sampled. Therefore, we are able to approximate α_X from α_x given x sampled randomly. The mean value rapidly converges, as the sampling size increase. Therefore, we evaluate the risk in the sampling model using $s = 10$.

Table 6 Approximated Evaluation Values for Each Model

T	X	$R_{mean}(X)$	$R_{cost}(X)$	$R_{sample}(X)(s=10)$
T_1	**time**	*0.3217	0.0145	*[0.1411, 0.5998]
	date	0.1860	0.0076	[0.1267, 0.2786]
	goods	0.0965	*0.0730	[0.0718, 0.0982]
	price	0.0121	0.0048	[0.0036, 0.0132]
	number	0.0080	0.0025	[0.0017, 0.0152]
T_2	**days**	*1.45E-04	*1.38E-04	*[1.46E-04, 1.52E-04]
	age	1.33E-04	9.83E-05	[1.21E-04 , 1.42E-04]
	ethnicity	7.73E-05	5.90E-05	[6.92E-05, 8.31E-05]
	gender	3.78E-05	2.95E-05	[3.08E-05 , 4.30E-05]
T_3	**age**	*2.24E-03	*2.24E-03	*[2.24E-03, 2.24E-03]
	occupation	4.61E-04	4.61E-04	[4.61E-04, 4.61E-04]
	martial	2.15E-04	2.15E-04	[2.15E-04, 2.15E-04]
	ethnicity	1.54E-04	1.54E-04	[1.54E-04, 1.54E-04]
T_4	**employment**	*0.7208	*0.7208	*[0.7208, 0.7208]
	income	0.1316	0.1316	[0.1316, 0.1316]
	amount	0.0211	0.0211	[0.0211, 0.0211]
	grade	0.0002	0.0002	[0.0002, 0.0002]

Table 7 Error and Calculation Cost for Proposed Model

Model	Cost	Error
Mean	38087	0
Sample	131.3	0.073
Cost	0	0.178

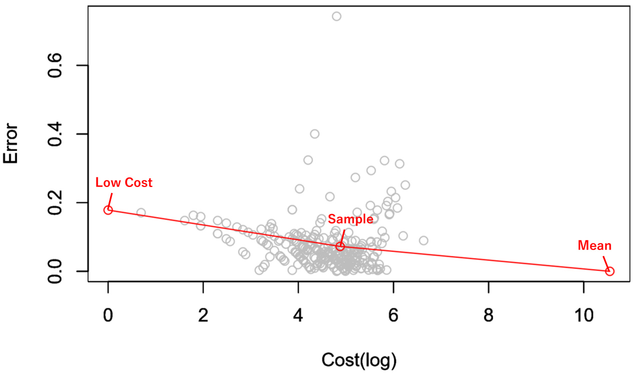

Fig. 11 Scatter Diagram for the Error and Calculation Cost for Each Model When $T = T_1$ and $X =$ **date**

Table 8 Mean and Standard Deviation when some x are Sampled from D_X and $T = T_1$ and $X =$ **date**

#Sample	mean	σ
1	24.03	13.33
50	24.47	1.75
100	24.39	1.09
150	24.41	0.78
200	24.41	0.53
250	24.42	0.33
$\lvert D_X \rvert$	24.42	0

4.5 Practical Way to identify knowledge of attacker.

It is getting difficult to know which attribute can be known to attackers since we are now in the era of big data where a large quantity of information is available to everyone including attackers. However, we believe that there are some objective ways to assume the background knowledge; 1. Based on open statistics, we as-

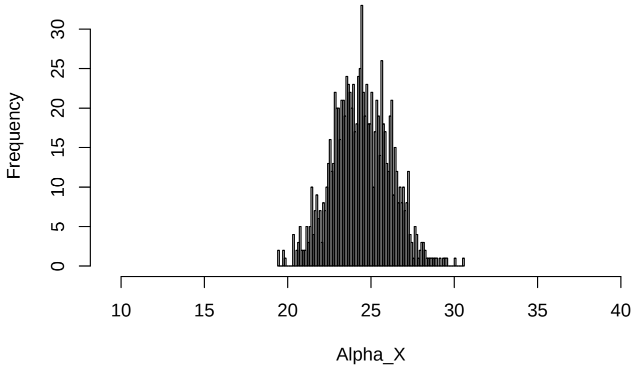

Fig. 12 Distribution of α_X when 50 x are sampled from D_X and $T = T_1$ and $X =$ **date**

sume the knowledge of attributes to be given to attacker. 2. Based on some criteria, we assume some class of attackers with distinct knowledges. 3. Based on the known cyber incidents so far, we accumulate the quantity of attributes and estimate the approximation of quantity of knowledge. Some of these may require future study.

5 Related Works

There are two representative studies to evaluate the privacy level of data, k-anonymity (Sweeny, 2006) and differential privacy (Dwork, 2006). k-anonymity was proposed by Sweeney in 2006. It evaluates the privacy level of data according to whether the data have at least k indistinguishable records in terms of quasi-identifiers. Differential privacy was introduced by Dwork in 2006. It evaluates the privacy level of data according to whether the possibility of restoring personal data from differences in analysis results of the data is high.

There are many works to extend the fundamental models in de-identification. Tamir et. al proposed k-concealment(Tassa et al., 2018) that is an alternative model of k-anonymity. Domingo-Ferrer et al. showed that the ε-differential privacy and t-closeness that is an extension of k-anonymity are strongly related to one another (Domingo-Ferrer and Soria-Comas, 2015). Stokes defined (k, l)-anonymity (Stokes, 2012) that is relaxation of k-anonymity and n-confusion (Stokes and Torra, 2012) which is a generalization of k-anonymity. These schemes help to reduce the risk of re-identification, but the risk still remains. Our approximate methods are able to evaluate the remaining risk of data that was already de-identified by either of these works including the k-anonymization method.

Technical Specification ISO/TS 25237 (ISO, 2008) defines ano-nymization as “a process that removes the association between the identifying data and the data subject”. Many anonymization algorithms have been proposed to preserve privacy while retaining the utility of the data that have been *anonymized.* That is, the data are made less specific so that a particular individual cannot be identified. Anonymization algorithms employ various operations, including *suppression* of attributes or records, *generalization* of values, replacing values with *pseudonyms*, *perturbation* with random noise, sampling, rounding, swapping, top/bottom coding, and micro aggregation (ICO, 2012) (Aggarwal and Yu, 2008).

Koot et al. proposed a method to quantify anonymity via an approximation of the uniqueness probability using a measure of the Kullback–Leibler distance (Koot et al., 2011). Monreale et al. proposed a framework for the anonymization of semantic trajectory data, called c-safety (Monreale et al., 2011). Based on this framework, Basu et al. presented an empirical risk model for privacy based on k-anonymous data release (Basu et al., 2015). Their experiment using car trajectory data gathered in the Italian cities of Pisa and Florence allowed the empirical evaluation of the protection of anonymization of real-world data.

In 2017, Torra presented a general introduction to data privacy (Torra, 2017). Li and Lai proposed a definition of a new δ-privacy model that requires that no adversary could improve more than δ privacy degree (Li and Lai, 2017). Tomoaki et al. consider low-rank matrix decomposition as one of the anonymization methods and evaluate its efficiency for time-sequence data (Mimoto et al., 2018).

6 Conclusions

In this paper, we studied the attacker who gains background knowledge from an attribute of the transaction data and proposed three approximate models that evaluate the risk of a dataset by approximating the mean identification probability of the attacker.

We applied our model to four actual datasets (the purchase history data for one year in the UK, the hospitalization data of diabetics for 10 years, the census income dataset, and the loan data) and evaluated the risks of these data. Our experiment reveals that our risk model (mean identification probability) is able to find the riskiest attribute from dataset and serves as a guide to decide the attributes to process or delete when we de-identify data. For example, the attribute **time** is the riskiest for five attributes in purchase history dataset T_1 and should be processed first.

The mean model provides the exact risk in the cost of examining all records. On the contrary, the low cost model gives the approximation of risk in the minimized cost, but with relatively large error. The sampling model approximates the risk with intermediate cost between the two extreme models. We have clarified mathematical properties of three models in terms of the accuracy and the cost. The riskiest attribute of dataset was found by these three approximate models in a high accuracy with less calculating cost.

Our future studies include the extension of our model to the aggregated risk of multiple attributes known to attacker, the extension of model so that we consider the quality of background knowledge rather than the quantity of knowledge. This study noted the risk of attributes ‘before’ de-identification in this paper. We will study the risk of attribute ‘after’ de-identification and the method to choose appropriate method to de-identify riskily records.

References

Aggarwal C, Yu P (2008) A general survey of privacy-preserving data mining, models and algorithms. *Privacy-preserving Data Mining* pp 11–52

Basu A, Monreale A, Trasarti R, Corena JC, Giannotti F, Pedreschi D, Kiyomoto S, Miyake Y, Yanagihara T (2015) A risk model for privacy in trajectory data. Journal of Trust Management pp 2–9

Domingo-Ferrer J, Soria-Comas J (2015) From t-closeness to differential privacy and vice versa in data anonymization. Journal Knowledge-Based Systems Volume 74 Issue 1:151–158

Domingo-Ferrer J, Ricci S, Soria-Comas J (2015) Disclosure risk assessment via record linkage by a maximum-knowledge attacker. 2015 Thirteenth Annual Conference on Privacy, Security and Trust (PST) *IEEE*:28–35

Dwork C (2006) Differential privacy. Proceedings of ICALP 2006 LNCS vol. 4052:1–12

Emam KE, Arbuckle L (2013) Anonymizing health data case studies and methods to get you started *O’Reilly*

ICO (2012) Anonymisation: managing data protection risk code of practice

ISO (2008) Health informatics – pseudonymization ISO Technical Specification ISO/TS 25237

Kikuchi H, Yamaguchi T, Hamada K, Yamaoka Y, Oguri H, Sakuma J (2016) What is the best anonymization method? – a study from the data anonymization competition pwscup 2015. Data Privacy Management Security Assurance (DPM2016) LNCS 9963:230–237

Koot MR, Mandjes MRH, van’t Noordende GJ, de Laat CTAM (2011) Efficient probabilistic estimation of quasi-identifier uniqueness. In Proceedings of ICT OPEN 2011 14–15:119–126

Lending Club (2019a) LOAN DATA [online]. `https://www.lendingclub.com/info/download-data.action`, [Accessed 15 Apr 2019]

Lending Club (2019b) [online]. `https://www.lendingclub.com/`, [Accessed 15 Apr 2019]

Li Z, Lai TH (2017) δ-privacy: Bounding privacy leaks in privacy preserving data mining. DPM/CBT 2017 Springer, LNCS 10436:124–142

Mimoto T, Kiyomoto S, Hidano S, Basu A, Miyaji A (2018) The possibility of matrix decomposition as anonymization and evaluation for time-sequence data. 2018 16th Annual Conference on Privacy, Security and Trust (PST), IEEE pp 139–145

Monreale A, Trasarti R, Pedreschi D, Renso C, Bogorny V (2011) c-safety: a framework for the anonymization of semantic trajectories. Transactions on Data Privacy 4(2):73–101

Stokes K (2012) On computational anonymity. Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2012 pp 336–347

Stokes K, Torra V (2012) n-confusion: a generalization of k-anonymity. EDBT/ ICDT Workshops 2012 pp 211–215

Sweeny L (2006) k-anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5):557–570

Tassa T, Mazza A, Gionis A (2018) k-concealment: An alternative model of k-type anonymity. TRANSACTIONS ON DATA PRIVACY 5:189–222

Torra V (2017) Data privacy: Foundations, new developments and the big data challenge. Studies in Big Data 28, Springer

UCI Machine Learning Repository (2018a) Adult Data Set [online]. `https://archive.ics.uci.edu/ml/datasets/adult`, [Accessed 17 Dec 2018]

UCI Machine Learning Repository (2018b) 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`, [Accessed 17 Dec 2018]

UCI Machine Learning Repository (2018c) [online]. `http://archive.ics.uci.edu/ml/index.php`, [Accessed 17 Dec 2018]

UCI Machine Learning Repository (2018d) Online Retail Data Set [online]. `https://archive.ics.uci.edu/ml/datasets/online+retail`, [Accessed 17 Dec 2018]