Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-021-03369-5

ORIGINAL RESEARCH

Estimation of cost of *k*–anonymity in the number of dummy records

Satoshi Ito[1] · **Hiroaki Kikuchi**[1]

Received: 31 August 2020 / Accepted: 22 June 2021

Abstract
De-identification is a process to prevent individuals from being identified from original transaction data by processing personal identification information. *k*-anonymization, which processes data so that at least *k* users have the same records, is one of the representative methods of de-identification. One of the methods of *k*-anonymization is adding dummy records into the data to protect users who have unique histories. For this method, the cost for *k*-anonymization is the difference in the number of records between the original data and the processed data, and it can be calculated only after deciding the parameter *k* and processing data. However, we want to calculate the cost before processing and find the optimal value of *k* because processing the big data with various *k* is very costly. In this paper, we propose a new model of transaction data that gives us a probability distribution and an expected value of values in data under the assumption that all values occur independently with uniform probability. Applying our data model, it is possible to evaluate the cost of *k*-anonymized data even before processing.

Keywords Personal identification information · Privacy risk evaluation · De-identification · Transaction data

1 Introduction

Companies are required to assess the re-identification risks and to de-identify personally identifiable data before employing big data extensively in their businesses. De-identification is a process to prevent individuals from being identified from the original personally identifiable information. Technical Specification ISO/TS 20889 (ISO 2018) defines anonymization as "a process that removes the association between a set of identifying attributes and the data principal." The ISO definition classifies anonymization techniques into several techniques including statistical tools, cryptographic tools, suppression techniques such as masking, pseudonymization techniques, generalization techniques, and randomization techniques. In Japan, the Act on the Protection of Personal Information fully came into effect in 2015, in which a concept called "Anonymously Processed Information"[1] was introduced (ersonal Information Protection Commission Secretariat 2017).

In anonymizing data, we must evaluate the data from two perspectives: how accurate the characteristics of the original data are preserved by processing (utility), and how many individuals are not reidentified from the processed data (security). Anonymizing data requires secure and useful data. In Japan, the data competition "PWS Cup" (Kikuchi et al. 2016) has been held since 2015, and the anonymization method and the evaluation metrics for anonymized data were studied in this competition using purchase-history data and location-history data.

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 (Hundepool et al. 2012) (Duncan et al. 2011) (Torra 2017).

One representative method of de-identification is *k*-anonymity proposed by Latanya Sweeney and Pierangela Samaratiy (Samarati and Sweeney 1998; Sweeney 2006). *k*-anonymity is a privacy measurement model that ensures that for each identifier there is a corresponding equivalence class containing at least *k* records, and it is widely applied because of its simplicity. Many researchers have studied the modified methods to improve its privacy assurance.

✉ Satoshi Ito
mmhm@meiji.ac.jp

Hiroaki Kikuchi
kikn@meiji.ac.jp

[1] Meiji University Graduate School, Tokyo 164-8525, Japan

[1] Japanese version of de-identified information has slight differences to common anonymized data.

Published online: 21 March 2022

There are many issues in processing data to satisfy k-anonymity.

1.1 Complexity of *k*-anonymity

There are some ways for processing data to ensure k-anonymity. Either suppression or generalization or both are performed. The utility/privacy trade off makes it difficult to find optimal processing conditions. As more dummy records are added, the difference in the number of records between the original data and the processed data increases. Meyeson and Williams proved that the optimal k-anonymity is NP-hard in (Meyerson and Williams 2004). Therefore, the try-and-error method is attempted for optimizing the cost for k - anonymity. For example, the 2-anonymized data with 100 dummy records must be compared with that with 1,000 dummy records for cost for processing. The cost for processing depends on parameter k, and we wish to obtain the optimal value of k for given data. The optimal value of k is a very important value for a data processor that tries to process data so that it satisfies k-anonymity.

1.2 Empirical estimation of cost

Instead of the exact optimization, some heuristics have been applied to calculate the cost for processing data for any k. Examples include K-optimize (Bayardo and Agrawal 2005), Incognito (LeFevre et al. 2005), and Mondrian (LeFevre et al. 2006). In many studies, the costs were calculated empirically by varying k. For example, Basu et al. presented an empirical risk model for privacy based on k-anonymous data release (Basu et al. 2015). Their experiment used the car trajectory data gathered in the Italian cities of Pisa and Florence under a certified allowance of the empirical evaluation of the anonymized real-world data. They straightforwardly calculated processing costs defined by the probability of re-identification concerning several parameters k and the attacker's background knowledge. Regarding the problem of a dynamic data set, the processing costs are much worse. There are too many potential correlations between various processing methods and the conditions required to ensure k-anonymity. Xiao et al. proposed a new generalization principle m-invariance (Xiao and Tao 2007) that effectively limits the risk of privacy disclosure in republication. This method consists of generalization and adding counterfeit tuples that resemble those of other customers in other data sets for processing data so that it satisfies m-invariance (metrics such as k-anonymity).

Indeed, calculating the processing cost and the optimal value of k is very difficult because it greatly depends on the target data and use-case scenario. The processing cost and the optimal value of k are not estimated from the statistics of the given data without processing.

Our research aims to estimate processing costs and to find the optimal value of k without empirically investigating the data. We focus on the difference between the number of records between the original data and the processed data as the significant feature of data that allows us to estimate the optimal k without performing any heuristic algorithms. In this paper, we focus on a processing method by adding dummy records for k-anonymity and treat the number of dummy records as cost for k-anonymity. We will explain how to add dummy records to data for k-anonymity in Sect. 3.

We find that the cost of the difference of records is primarily determined by three values: (1) size of each cluster of individuals (s_i), (2) size of the range for each cluster ($|I(U_i)|$), and (3) size of the range for each user ($|I(u_i)|$). However, we are not able to obtain these values until we examine data because the number of clusters is one of necessary parameter of processing. Instead, we approximate these values (1), (2) and (3) with a mean (n/c), the expected value of the size of range ($E[y|m/c, \ell]$), and the expected value of the size of range ($E[y|m/n, \ell]$). We will explain how to estimate the cost of processing data by approximating these three values in Sect. 3.

To address the optimal k problem in a utility/privacy trade off involved by a processing time-series data set such as payment history or trajectory data, the following problem must be solved. Our problem is finding how many unique y values (chosen out of ℓ values) are included when the transaction data of x records is given.

This problem is similar to a problem known as the "coupon collector problem" (Mitzenmacher and Upfal 2005). (Coupon Collector Problem) How many cereal boxes do we need to buy in order to collect all ℓ coupons when all coupons that are included in boxes occur independently with uniform probability? This problem can be solved in the expected value $E[X]$ of the number of boxes that we need to buy is $E[X] = \ell \ln \ell + \mathcal{O}(\ell)$ because the expected value is $E[X] = \sum_{i=1}^{\ell} \frac{\ell}{\ell - i + 1} = \ell \sum_{i=1}^{\ell} \frac{1}{i}$ and the harmonic number is $H(\ell) = \sum_{i=1}^{\ell} \frac{1}{i} = \ln \ell + \mathcal{O}(1)$. Unfortunately, we cannot apply the coupon collector's problem to our model because the goal is not exactly the same. We explain our problem in terms of the coupon collector problem. (Our Problem) How many unique y coupons (chosen out of ℓ coupons) are collected when we buy x cereal boxes? Table 1 shows the comparison between the coupon collector's problem and our model.

We propose a new model to solve our problem. Our model makes two theorems that values of records occur independently and identical distribution: (a) a distribution of conditional probability $Pr(y|x)$ of unique y values is chosen out of ℓ given data of x records and (b) an expected value $E[y|x, \ell]$ of y values. Our proposed model enables us

Table 1 The comparison between the coupon collector's problem and our model

	The coupon collector's problem	Our model
Assumption	Uniform probability of $1/\ell$	Uniform probability of $1/\ell$
Goal	The number of cereal boxes	The number of values are collected out of ℓ
Expected value	$\ell \ln \ell + \mathcal{O}(\ell)$	Theorem 2

to find the optimal value of k without suffering the brute-force processing cost of big data with various k.

Our contributions are as follows.

1. We propose a new model of time-series data that gives us a probability distribution and an expected value of values (e.g., purchased goods) in data under the assumption that all values occur independently and with uniform probability. We prove the probability distribution of a number of unique values (Theorems 1 and 2).
2. We identify the cost of arbitrary time-series data to ensure k-anonymity. Our proved theorems obtain the mean values of three values (size of each cluster of individual, size of the range for each cluster, and size of the range that the whole customers have) and calculate an expected value of the cost of k-anonymity ensured data without empirical evaluation (Theorem 3 and Corollary 1).
3. We show the comparison between the estimated cost and an actual cost of processed data that satisfies k-anonymity.

The remainder of the paper is organized as follows. In Sect. 2, we propose a new model of transaction data. In Sect. 3, we estimate the cost for k-anonymization of the transaction data using our model. In Sect. 4, we provide a conclusion to this paper.

2 The transaction data model

2.1 Preliminaries

Transaction data consists of records (rows) and attributes (columns) and has an attribute that identifies individuals. We define our model as follows.

Definition 1 Let T be transaction data consisting of a set of records. Let m and n be the numbers of records and users in transaction data T, respectively. T has an attribute that identifies individuals via user IDs and an attribute that ranges over ℓ values. Let $U = \{u_1, \ldots, u_n\}$ be a set of users, $I(U) = \{g_1, \ldots, g_\ell\}$ be a set of values for the attribute, and $I(u_i)$ be a set of values for a user u_i.

Example 1 Table (a) in Fig. 1 shows an example of purchase transaction data T_{ex} that has an attribute (user IDs) that identifies individuals and an attribute (Goods). In the example of T_{ex}, n is 3 because it contains three customers $U = \{Alice, Bob, Carol\}$, and m is 4 and ℓ is 2 because two goods $I(U) = \{Apple, Book\}$ have been purchased. For example, $I(Alice) = \{Apple\}$ because Alice has purchased an apple. We will explain about the Table (b–d) in Sect. 3.1.

Assumption 1 All ℓ values occur independently with uniform probability $1/\ell$. ($1/\ell$ assumption)

Definition 2 Let (x, y) be a state that transaction data has x records and y values out of ℓ. Let $Pr(Y = y|X = x)$ be the conditional probability of a random variable Y and y values given that random variable X has x records, i.e., the state (x, y). Let $E[y|x, \ell]$ be the expected value of the number of values chosen out of ℓ when the data has x records.

We regard the time-series events of purchase as a sequence of records that takes from ℓ values with uniform probability. The state evolves every time a new record of purchase arrives.

State (x, y) occurs in two ways. First, state $(x-1, y)$ changes into state (x, y) with a probability of y/ℓ when

(a) T_{ex}

User IDs	Goods
Alice	Apple
Bob	Apple
Bob	Book
Carol	Book

(b) T'_{ex}

Pseudonym	Goods	
1	Apple	
1	Book	*
2	Apple	
2	Book	
3	Book	
3	Apple	*

(c)

User IDs	$I(u_i)$
Alice	$\{Apple\}$
Bob	$\{Apple, Book\}$
Carol	$\{Book\}$

(d)

Pseudonym	$I(u_i)$
1	$\{Apple, Book\}$
2	$\{Apple, Book\}$
3	$\{Apple, Book\}$

Fig. 1 How to add dummy records to data

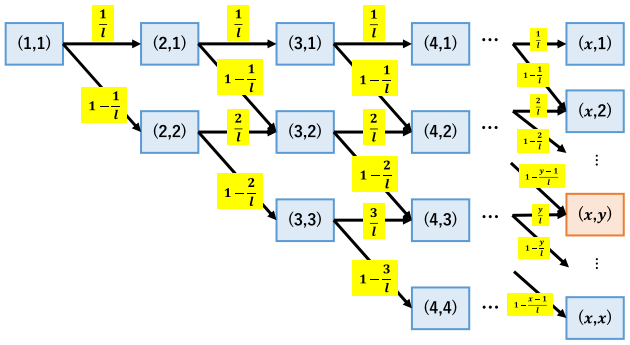

Fig. 2 A state transition diagram of (x, y)

an additional record has a value that has already existed in $(x-1)$ records. Second, state $(x-1, y-1)$ changes into state (x, y) with a probability of $1-(y-1)/\ell$ when an additional record has a value that has never existed in data. Therefore, the conditional probability $Pr(y|x)$ is calculated as follows.

$$Pr(y|x) = (1-\frac{y-1}{\ell})Pr(y-1|x-1) + \frac{y}{\ell}Pr(y|x-1) \quad (1)$$

Example 2 Figure 2 illustrates a state transition diagram of (x, y). In the case of purchase-history data, a state $(4, 2)$ means that data T has four records and contains two goods chosen out of ℓ. For example, $P(Y=2|X=4)$ is calculated as $Pr(Y=2|X=4) = ((1/\ell)(1/\ell)(1-1/\ell)) \cdot ((1/\ell)(1-1/\ell)(2/\ell)) \cdot ((1-1/\ell)(2/\ell)(2/\ell)) = 7(1-1/\ell)/\ell^2$, and $E[y|4,\ell]$ is $E[y|4,\ell] = \sum_{i=1}^{4} i \cdot Pr(i|4)$.

2.2 The transaction data model

In this paper, we propose a new model of transaction data that gives us a probability distribution $Pr(y|x)$ and an expected value $E[y|x,\ell]$ under the $1/\ell$ assumption.

Theorem 1 *The conditional probability $Pr(y|x)$ that states (x, y) occurs in the transaction data, that is, given x records y distinct values out of ℓ are chosen, is calculated as*:

$$Pr(y|x) = \prod_{j=0}^{y-1}(1-\frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_y=x-y} (\frac{1}{\ell})^{m_1} \cdots (\frac{y}{\ell})^{m_y}, \quad (2)$$

where $m_1, \ldots, m_y$ are positive integers such that the sum of these is $x-y$, and $x \geq y \geq 1$.

Proof We prove Eq. (2) by mathematical induction. When x is 1, y is 1 because x is greater than y which is greater than one, and $Pr(1|1)$ is calculated as $Pr(1|1) = (1-\frac{0}{\ell}) \cdot (\frac{1}{\ell})^0 \cdots (\frac{y}{\ell})^0 = 1$, so Eq. (2) holds. When x is $x'-1$ $(x' \geq 2)$, we suppose that Eq. (2) holds for any y $(1 \leq y \leq x'-1)$; that is, $Pr(y|x'-1)$ is $Pr(y|x'-1) = \prod_{j=0}^{y-1}(1-\frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_y=(x'-1)-y} (\frac{1}{\ell})^{m_1} \cdots (\frac{y}{\ell})^{m_y}$, and $Pr(y-1|x'-1)$ is $Pr(y-1|x'-1) = \prod_{j=0}^{y-2}(1-\frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_{y-1}=(x'-1)-(y-1)} (\frac{1}{\ell})^{m_1} \cdots (\frac{y-1}{\ell})^{m_{y-1}}$. Assigning these two formulas to Eq. (1), we obtain the theorem as follows.

$$\begin{aligned} Pr(y|x') &= (1-\frac{y-1}{\ell})Pr(y-1|x'-1) + (\frac{y}{\ell})Pr(y|x'-1) \\ &= \prod_{j=0}^{y-1}(1-\frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_{y-1}=x'-y} (\frac{1}{\ell})^{m_1} \cdots (\frac{y}{\ell})^{m_{y-1}} \\ &\quad + \prod_{j=0}^{y-1}(1-\frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_y=x'-y} (\frac{1}{\ell})^{m_1} \cdots (\frac{y}{\ell})^{m_y+1} \\ &= \prod_{j=0}^{y-1}(1-\frac{j}{\ell}) \cdot \sum_{m_1+\cdots+m_y=x'-y} (\frac{1}{\ell})^{m_1} \cdots (\frac{y}{\ell})^{m_y} \end{aligned}$$

Therefore, Eq. (2) holds with any $x (x \geq 1)$ because it holds with $x = x'$. □

Applying this theorem, we are able to calculate practically the probability distribution of state (x, y) that is hard to find from the state transaction diagram.

Theorem 2 *Under the $1/\ell$ assumption, an expected value of number of values Y chosen out of ℓ in transaction data T given T with x records is calculated as $E[y|x,\ell] = (-\ell)(1-\frac{1}{\ell})^x + \ell$.*

Proof Any particular value in ℓ occurs at least once in transaction data of x records with probability $1-(1-1/\ell)^x$ under the $1/\ell$ assumption. From the linearity of expectations, the expected value of sum of number of values ℓ values is given as $E[y|x,\ell] = (-\ell)(1-\frac{1}{\ell})^x + \ell$. □

We calculate the expected value of the number of dummy records applying this theorem in Sect. 3.

2.3 Analysis of our model

In this section, we analyze the conditional probability $Pr(y|x)$ and the expected value $E[y|x,\ell]$ given by our model introduced in Sect. 2.2. Figure 3 shows the probability distribution of a number of values y when ℓ is 100 and when the number of records x is 10, 25, 50, 75, and 100. For example, the blue line indicates $Pr(y|50)$, and it is a maximum value of 0.168 when y is 40. This implies that the probability is maximized at $Y = 40$ when the data T has 50 records, $\ell = 100$ values. Figure 4 shows the probability distribution of a number of records X when ℓ is 100

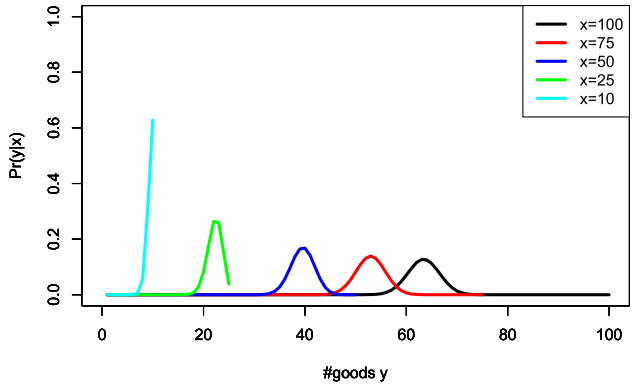

Fig. 3 The probability distribution of a number of values y when ℓ is 100

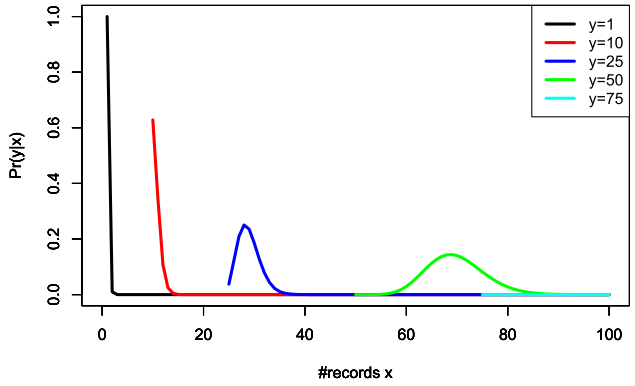

Fig. 4 The probability distribution of a number of records X when ℓ is 100

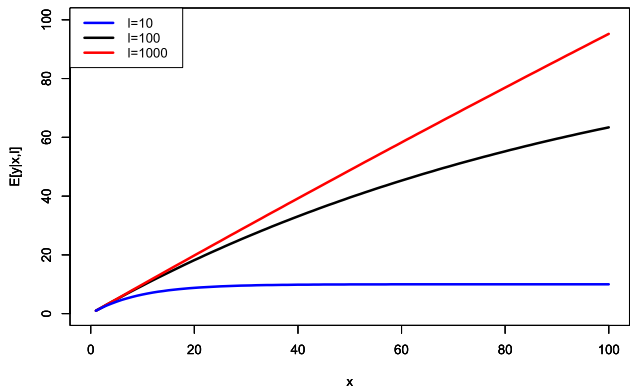

Fig. 5 The distribution of expected value $E[y|x, \ell]$ with regard to x

and when the number of values Y is 1, 10, 25, 50, and 75. For example, the blue line indicates $Pr(25|x)$, which has a maximum value 0.250 at $x = 28$. It means that the highest probability is at $X = 28$ when Y has 25 values. Figure 5 shows the distribution of expected value $E[y|x, \ell]$ with regard to x when ℓ are 10, 100, and 1000. For example, $E[y|x, 100]$ is represented by the black line and shows that the expected value is 63.40 when the data has 100 records.

3 The cost for *k*-anonymization

In this section, we estimate the cost for the k-anonymity of the transaction data using our model introduced in Sect. 2. We treat the cost for anonymization as the difference between the number of records of the original data and the processed data. We define some symbols in Sect. 3.1 and calculate the exact difference of records and then estimate the expected value of the difference of records in our data model.

3.1 Preliminaries

Definition 3 Let T' be a modified transaction data T by adding some dummy records. Let Δm be the number of dummy records, and let c be the number of clusters of users who share the same transaction history. Let $U_i = \{u_1^i, \ldots, u_{s_i}^i\}$ be the ith cluster, and let $s_i = |U_i|$ be the size (number of pseudonyms) of cluster U_i. U is divided into c equivalent classes, that is, $U = U_1 \cup \cdots \cup U_c$. Let $I(u)$ be a set of values that user u has purchased, and let $I(U_i) = \bigcup_{u \in U_i} I(u)$ be the union of sets $I(u)$ for all users u in cluster U_i.

Assumption 2 Identities of T' are pseudonymized and T' satisfies k-anonymity by adding some dummy records.

Example 3 Table (b) in Figure 1 shows data T'_{ex} that is processed from T_{ex}. In this data, three customers (Alice, Bob, and Carol) are pseudonymized with three pseudonyms (1, 2, and 3), and two dummy records (marked "*") are added so that three pseudonyms have the exact same purchase history ($c = 1, \Delta m = 2$). There is only one cluster $U_1 = \{1, 2, 3\}$ in this data, and s_1 is 3.

In this case, we mix up three customers Alice, Bob, and Carol by adding some dummy records such that three users have same set of goods. We detail the list of purchased goods for each customer of T_{ex} and T'_{ex} in Table (c) and Table (d) respectively, shown as $I(1) = I(2) = I(3) = I(Alice) \cup I(Bob) \cup I(Carol) = \{Apple, Book\}$. Because three pseudonyms have same set of goods as shown in Table (d), the processed data T'_{ex} satisfies 3-anonymity.

3.2 The exact solution of the difference of records

Proposition 1 *The number of dummy records Δm is calculated as*:

$$\begin{aligned}\Delta m &= \sum_{i=1}^{c}\sum_{j=1}^{s_i}(|I(U_i)| - |I(u_j^i)|)\\ &= \sum_{i=1}^{c}\left(s_i|I(U_i)| - \sum_{j=1}^{s_i}|I(u_j^i)|\right)\\ &= \sum_{i=1}^{c} s_i|I(U_i)| - \sum_{i=1}^{n}|I(u_i)|,\end{aligned}$$

where c is the number of clusters.

Example 4 For example, suppose that transaction data T_{ex} is anonymized to T'_{ex} with $c = 1$. There is only one cluster $U_1 = \{1, 2, 3\}$ in data T, and s_1 is 3. The sets of goods for the three customers are $I(Alice) = \{Apple\}$, $I(Bob) = \{Apple, Bob\}$, and $I(Carol) = \{Book\}$, and the whole set of goods for cluster is $I(U_1) = \{Apple, Book\}$. In this case, the number of dummy records is $\Delta m = s_1|I(U_1)| - |I(u_1)| - |I(u_2)| - |I(u_3)| = 3 \cdot 2 - 1 - 2 - 1 = 2$.

The number of dummy records Δm is calculated from a size s_i of each cluster, size $|I(U_i)|$ of the set of values for ith cluster, and size $|I(u_i)|$ of the set of values that the whole customers have. Because we obtain s_i, $|I(U_i)|$, and $|I(u_i)|$ only after processing data with parameter c, the number of dummy records Δm is unknown before processing. Nevertheless, Δm is necessary for optimizing processing parameter c.

3.3 The expected value of the number of dummy records

Instead of computing the exact value, we estimate the approximate value of the number of dummy records Δm before processing with parameter c. We calculate an expected value of Δm before processing by approximating s_i, $|I(U_i)|$, and $|I(u_i)|$. In this paper, we introduce the following assumption.

Assumption 3 The sizes of all c clusters of n users are the same; thus, any cluster has n/c users (n/c assumption). All n users have the same number of records in T; thus, m/n records (m/n assumption).

The results show that all c clusters have m/c records under the two previous assumptions. From these assumptions ($1/\ell$, n/c, and m/n), the expected value of Δm is calculated as follows.

Theorem 3 *From three assumptions ($1/\ell$, n/c, and m/n), an expected value $E(\Delta m)$ of the number of dummy records for processing data is calculated as*

$$E(\Delta m) = n\ell\left((1 - \frac{1}{\ell})^{m/n} - (1 - \frac{1}{\ell})^{m/c}\right),$$

where n is the number of users, m is the number of records, and ℓ is the number of values.

Proof From Proposition 1, Δm is $\Delta m = \sum_{i=1}^{c} s_i|I(U_i)| - \sum_{i=1}^{n}|I(u_i)|$. Applying n/c, $E[y|m/c, \ell]$, and $E[y|m/n, \ell]$ to s_i, $|I(U_i)|$, and $|I(u_i)|$ in this formula, respectively, we obtain the following formula.

$$\begin{aligned}\Delta m &= \sum_{i=1}^{c} s_i|I(U_i)| - \sum_{i=1}^{n}|I(u_i)|\\ E(\Delta m) &= \sum_{i=1}^{c}\frac{n}{c}E[y|\frac{m}{c}, \ell] - \sum_{i=1}^{n}E[y|\frac{m}{n}, \ell]\\ &= n\{(-\ell)(1 - \frac{1}{\ell})^{m/c} + \ell\}\\ &\quad - n\{(-\ell)(1 - \frac{1}{\ell})^{m/n} + \ell\}\\ &= n\ell\{(1 - \frac{1}{\ell})^{m/n} - (1 - \frac{1}{\ell})^{m/c}\}\end{aligned}$$

Thus, the theorem is proved. □

Theorem 3 is related to the k-anonymization. When anonymized data T' with n customers and c clusters satisfies k-anonymity, every cluster in T' has at least k users. From the premise that all c clusters have n/c users in Theorem 3.1, k is less than n/c, and the number of dummy records for k-anonymization is calculated as follows.

Corollary 1 *The expected value $E'(\Delta m)$ of a number of dummy records for k-anonymization is calculated as*

$$E'(\Delta m) \geq n\ell\left((1 - \frac{1}{\ell})^{m/n} - (1 - \frac{1}{\ell})^{km/n}\right),$$

where n is the number of users, m is the number of records, and ℓ is the number of values.

3.4 Experimental evaluation

In this section, we analyze the expected value $E(\Delta m)$ of the number of dummy records.

First, Fig. 6 shows the distribution of $E(\Delta m)$ with respect to the number of users n when ℓ is 100 and c is 20, and m

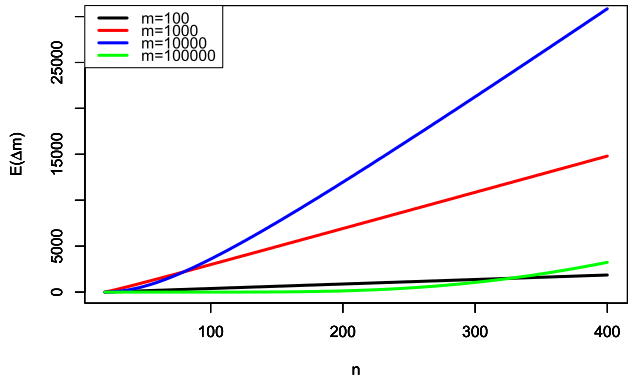

Fig. 6 The distribution of $E(\Delta m)$ with respect to the number of users n when ℓ is 100 and c is 20

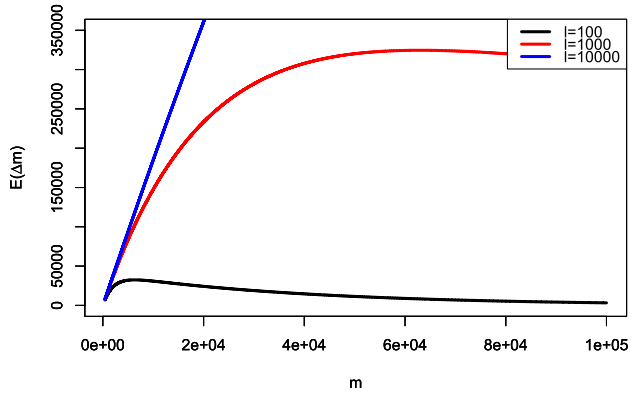

Fig. 7 The distribution of $E(\Delta m)$ with regard to the number of records m when n is 400 and c is 20

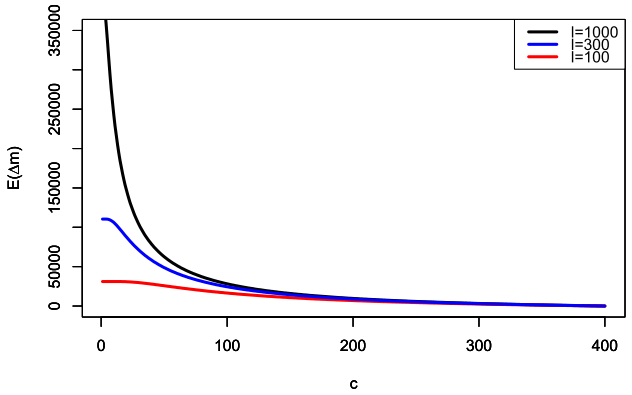

Fig. 8 The distribution of $E(\Delta m)$ with regard to the number of clusters c when n is 400 and m is 10,000

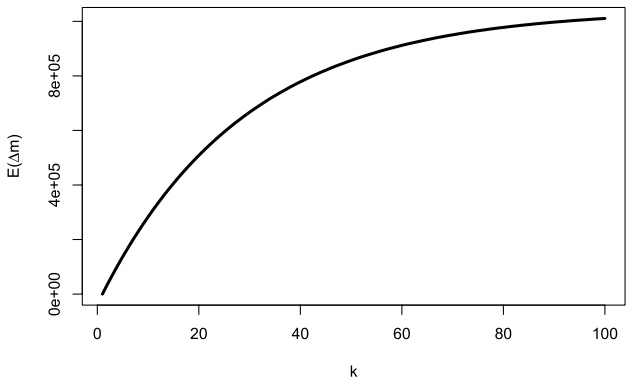

Fig. 9 The distribution of the expected value $E'(\Delta m)$ with regard to k

is 100, 1,000, 10,000, and 100,000. For example, the blue line indicates the expected number of dummy records for data with 10,000 records. We find $E(\Delta m)$ is 30,850 when n is 400. As the number of users n increases, the expected value $E(\Delta m)$ of the number of dummy records also increases.

Second, Fig. 7 shows the distribution of $E(\Delta m)$ with regard to the number of records m when n is 400 and c is 20, and ℓ is 100, 1,000, and 10,000. For example, the red line indicates the expected number of dummy records for data when ℓ is 1,000. It is a maximum of 324,570 records when m is 63,037. This result shows that the number of dummy records decreases when the number of records of data is too large.

Finally, Fig. 8 shows the distribution of $E(\Delta m)$ with regard to the number of clusters c when n is 400 and m is 10,000, and ℓ is 100, 300, and 1,000. This result shows that as c decreases, $E(\Delta m)$ decreases. Figure 9 shows the distribution of the expected value $E'(\Delta m)$ with regard to k of k-anonymization when n is 400 and m is 38,000, and ℓ is 2,700. As k increases, $E'(\Delta m)$ increases, and the utility of the data accordingly decreases.

3.5 Comparison between the estimated costs and the actual costs

In our past research (Ito et al. 2020), we revealed that about 160,000 dummy records are required to 4-anonymize purchase-history data containing 400 customers and 38,087 records. Table 2 shows the comparison of experimental result Δm of (Ito et al. 2020) and the estimated cost $E(\Delta m)$ calculated in this paper. The reason why the estimates are not very close to the experimental result is that the number of clusters c is fixed ($c = 50$) at any k.

In (Ito et al. 2020), we also evaluated the de-identified data comprehensively based on the metrics $\alpha E(\Delta m) + E(Reid)$ referring to the metrics (utility + security)/2 used in PWS Cup 2016 (Let α be a coefficient to normalize $E(\Delta m)$ to the range of $0 \leq \alpha E(\Delta m) \leq 1$). We show the relationship between c and the comprehensive

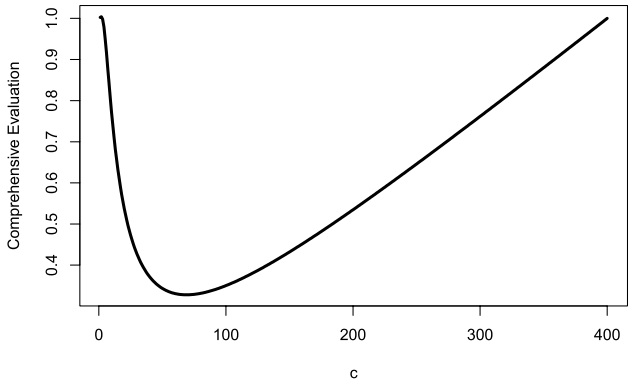

Fig. 10 Relationship between the number of clusters c and comprehensive evaluation of de-identified data

evaluation of de-identified data in Fig. 10 when $n = 400$, $m = 38,000$, $\ell = 2,700$, $\alpha = 1/1,042,653$. In this case, the comprehensive evaluation value is the smallest when $c = 69$, or in other words, the de-identified data processed based on our method will be the best when $c = 69$. In this case, the optimum value of k is 5 because k is less than $n/c = 400/69 = 5.80$, as mentioned in Sect. 3.3.

3.6 Influences of assumptions

We examine some influences of two assumptions, that is, all ℓ values occur independently and with uniform probability $1/\ell$ ($1/\ell$ assumption) and all users have the same number of records in T (m/n assumption). These assumptions are not always satisfied in big data.

Figure 11 shows the distribution of frequencies of goods of the subset of Online Retail Data Set (UCI Machine Learning Repository 2020) that has 400 customers and 2781 goods. Note that we use the subset that used in PWS Cup and Online Retail Data Set contains more than 400 customers. The most frequent good occurs more than 1000 times, and many goods occur only one time; that is, this distribution is extremely skewed. Figure 12 shows the number of records of users of the data. This distribution is similarly skewed. Unfortunately, the $1/\ell$ and m/n condition are not exactly satisfied in the Online Retail Data Set.

Therefore, in this section, we try to investigate the cost estimate if we remove these two assumptions ($1/\ell$ and m/n) from our model as follows.

Definition 4 Let p_j be an occurrence probability of j-th value out of ℓ values. Let b_i be a number of records of i-th user out of n users.

Table 2 Relationship between k and Δm

	Δm(Ito et al. 2020)	$E(\Delta m)$	$E_2(\Delta m)$	$E_3(\Delta m)$	Jaccard *Reid*	Random *Reid*
$k = 2$	183,902	36,188	31,868	36,213	0.1729	0.1223
$k = 3$	175,449	71,158	60,968	65,312	0.1726	0.1222
$k = 4$	162,474	104,950	87,768	92,113	0.1723	0.1218
$k = 8$	125,798	229,122	177,815	182,159	0.1681	0.1218

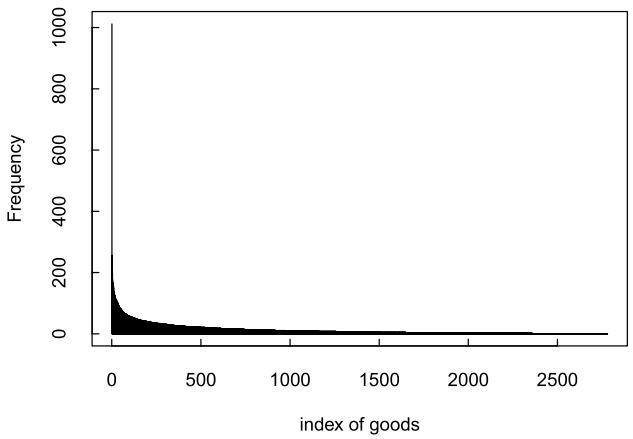

Fig. 11 The distribution of goods of a subset of the Online Retail Data Set

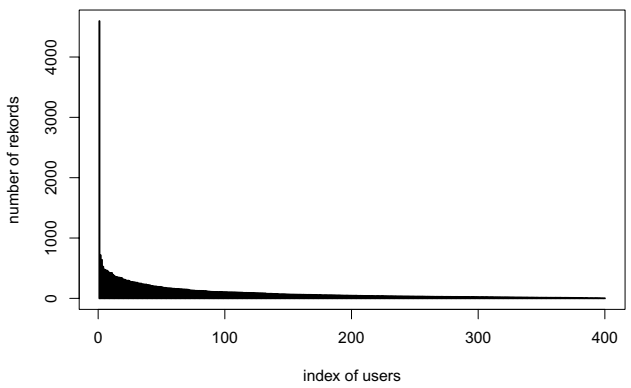

Fig. 12 The number of records of users of the Online Retail Data Set

Under the $1/\ell$ assumption, p_j is equal to $1/\ell$ in any value. Under the m/n assumption, b_i is equal to m/n in any user.

Corollary 2 *Under the n/c assumption and the m/n assumption, an expected value $E_2(\Delta m)$ of the number of dummy records for processing data is*

$$E_2(\Delta m) = n \sum_{j=1}^{\ell} \{(1-p_j)^{m/n} - (1-p_j)^{m/c}\},$$

where n is the number of users, m is the number of records, and ℓ is the number of values.

Proof From Proposition 1, Δm is $\Delta m = \sum_{i=1}^{c} s_i |I(U_i)| - \sum_{i=1}^{n} |I(u_i)|$. Applying n/c, $E[y|m/c, \ell]$, and $E[y|m/n, \ell]$ to s_i, $|I(U_i)|$, and $|I(u_i)|$ in this formula, respectively, we obtain the formula of the corollary 2 as follows.

$$\begin{aligned}
E_2(\Delta m) &= \sum_{i=1}^{c} \frac{n}{c} E[y|\frac{m}{c}, \ell] - \sum_{i=1}^{n} E[y|\frac{m}{n}, \ell] \\
&= n \sum_{j=1}^{\ell} \{1-(1-p_j)^{m/c}\} \\
&\quad - \sum_{i=1}^{n} \sum_{j=1}^{\ell} \{1-(1-p_j)^{m/n}\} \\
&= n \sum_{j=1}^{\ell} \{1-(1-p_j)^{m/c}\} \\
&\quad - n \sum_{j=1}^{\ell} \{1-(1-p_j)^{m/n}\} \\
&= n \sum_{j=1}^{\ell} \{(1-p_j)^{m/n} - (1-p_j)^{m/c}\}
\end{aligned}$$

□

Corollary 3 *Under the n/c assumption, an expected value $E_3(\Delta m)$ of the number of dummy records for processing data is*

$$E_3(\Delta m) = n \sum_{j=1}^{\ell} \{1-(1-p_j)^{m/c}\} - \sum_{i=1}^{n} \sum_{j=1}^{\ell} \{1-(1-p_j)^{b_i}\},$$

where n is the number of users, m is the number of records, and ℓ is the number of values.

Proof From Proposition 1, Δm is $\Delta m = \sum_{i=1}^{c} s_i |I(U_i)| - \sum_{i=1}^{n} |I(u_i)|$. Applying n/c, $E[y|m/c, \ell]$, and $E[y|b_i, \ell]$ to s_i, $|I(U_i)|$, and $|I(u_i)|$ in this formula, respectively, we obtain the formula of the corollary 3 as follows.

$$\begin{aligned}
E_3(\Delta m) &= \sum_{i=1}^{c} \frac{n}{c} E[y|\frac{m}{c}, \ell] - \sum_{i=1}^{n} E[y|b_i, \ell] \\
&= nE[y|\frac{m}{c}, \ell] - \sum_{i=1}^{n} \sum_{j=1}^{\ell} \{1-(1-p_j)^{b_i}\} \\
&= n \sum_{j=1}^{\ell} \{1-(1-p_j)^{m/c}\} \\
&\quad - \sum_{i=1}^{n} \sum_{j=1}^{\ell} \{1-(1-p_j)^{b_i}\}
\end{aligned}$$

□

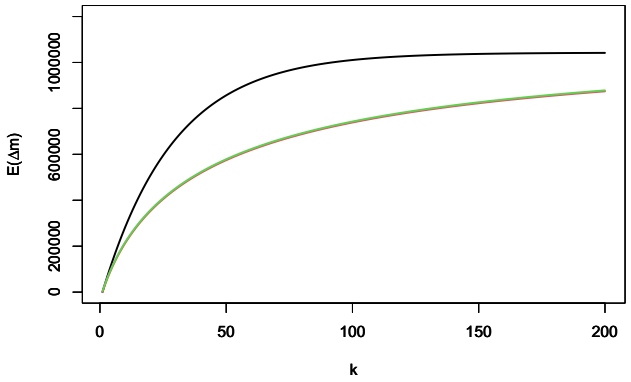

Fig. 13 The comparison of cost for k-anonymity of 3 models

We calculate $E_2(\Delta m)$ and $E_3(\Delta m)$ by finding all p_j (2,781 goods) and all b_i (400 users) from the Online Retail Dataset and substituting n/k for c in these equations like Corollary 1. Figure 13 shows a comparison of the expected value of the number of dummy records for k-anonymity of 3 models ($E(\Delta m)$, $E_2(\Delta m)$, and $E_3(\Delta m)$). The red and green lines indicate $E_2(\Delta m)$ and $E_3(\Delta m)$. These two lines almost overlap and are lower than the black line ($E(\Delta m)$) in any k ($1 \leq k \leq 200$). Table 2 shows the comparison of these models ($E(\Delta m)$, $E_2(\Delta m)$, and $E_3(\Delta m)$) for $k = 2, 3, 4, 8$. Even though some assumptions are removed from $E_2(\Delta m)$ and $E_3(\Delta m)$, the behavior of three models are almost same. From these results, we find that the two assumptions work to increase the estimated value of cost for k-anonymity.

It is hard for $E_2(\Delta m)$ and $E_3(\Delta m)$ to be calculated because these equation include p_j and b_i that are not able to be obtained without analyzing the data. So, when we want to obtain the optimum value of the processing parameter (k, c), the equation of $E(\Delta m)$ is more useful than that of $E_2(\Delta m)$ and $E_3(\Delta m)$. Note that we do not study the parameters except m to calculate an estimated cost in this paper

As shown in Table 2, the behaviors of the estimated values ($E(\Delta m)$, $E_2(\Delta m)$, and $E_3(\Delta m)$) and the actual value (Δm) are very different. The more the value of k, the more the estimated values increase while the actual value decrease. From our experimental results, we find that these differences are not based on the assumptions and we estimate that the reason of these differences is that the compared processing method (Ito et al. 2020) adjust some parameters to minimize the processing cost.

4 Conclusions

In this paper, we proposed a new model of transaction data that allows us to estimate a probability distribution and an expected value of utility of anonymized data under the assumptions that all values in data occur independently and have a uniform probability. Our data model calculates the expected value of the number of values y out of ℓ when the data has x records. We calculated the reduction of utility incurred by anonymization by using the number of dummy records for k-anonymizing data from the original data and parameter k. Applying our model, it is possible to evaluate the quality of de-identified data even before processing.

Our future studies will strive to improve our model so that the $1/\ell$ assumption is unnecessary and that other useful properties of de-identified data can be calculated without processing.

References

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. J Trust Manag 2:9

Bayardo RJ, Agrawal R (2005) Data privacy through optimal k -anonymization. ICDE 05:217–228

Duncan G, Elliot M, Salazar J (2011) Statistical confidentially. Springer, New York

Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Nordholt E, Spicer K, Wolf P (2012) Statistical disclosure control. Wiley, New York

ISO (2018) Privacy enhancing data de-identification terminology and classification of techniques ISO Technical Specification ISO/TS 20889

Ito S, Harada R, Kikuchi H (2020) De-identification for transaction data secure against re-identification risk based on payment records. to be published in JIP

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 Priv Manag Secur Assur (DPM2016) LNCS 9963:230–237

LeFevre K, DeWitt DJ, Ramakrishnan R (2005) Incognito: efficient full-domain k-anonymity. SIGMOD 05:49–60

LeFevre K, DeWitt DJ, Ramakrishnan R (2006) Mondrian multidimensional k-anonymity. ICDE 06:1–11

Meyerson A, Williams R (2004) On the complexity of optimal k-anonymity. In: Proceedings of ACM PODS, pp 223–228

Mitzenmacher M, Upfal E (2005) Probability and computing: randomized algorithms and probabilistic analysis. Cambridge University Press, United Kingdom, pp 32–34 (**Section 2.4.1**)

Personal Information Protection Commission Secretariat (2017) Report by the personal information protection commission secretariat: anonymously processed information–towords balanced promotion of personal data utilization and consumer trust

Samarati P, Sweeney L (1998) Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Computer Science Laboratory, SRI International Technical Report SRI-CSL-98-04

Sweeney L (2006) k-anonymity: a model for protecting privacy. Int J Uncertain Fuzziness Knowl-Based Syst 10(5):557–570

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

UCI Machine learning repository (2020) Online retail data set [online]. https://archive.ics.uci.edu/ml/datasets/online+retail. [Accessed 3 Dec 2020]

Xiao X, Tao Y (2007) m-invariance: toward privacy preserving republication of dynamic datasets. Proc SIGMOD 07:689–700