Re-identification risk prediction paradigm using incomplete statistical information and recursive hypergeometric distribution

Zhigang Yang (ayzg163@163.com)
Chongqing University of Posts and Telecommunications https://orcid.org/0000-0002-7268-5390
Ruyan Wang
Chongqing University of Posts and Telecommunications
Dapeng Wu
Chongqing University of Posts and Telecommunications
Boran Yang
Chongqing University of Posts and Telecommunications
Daizhong Luo
Chongqing University of Arts and Sciences

Article

Keywords: recursive hypergeometric distribution, statistical information, risk prediction paradigm

Posted Date: May 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-533949/v1

Re-identification risk prediction paradigm using incomplete statistical information and recursive hypergeometric distribution

Author Information

Affiliations

Contributions

Corresponding author

Abstract

The dataset anonymization has not eliminated the re-identification risk, the evaluation of which remains a huge challenge, especially given incomplete statistical information. The re-identification risk of individuals depends on their tuple frequency. The paper proposes the recursive hypergeometric (RH) distribution to accurately calculate the tuple frequency and leverages the binomial distribution to approximate the RH distribution and to efficiently predict the re-identification risk of individuals in both generated and real-world datasets. The experimental results show that our tuple frequency based re-identification risk (TFRR) prediction model has a superior performance (average AUC 0.86~0.98) for all types of datasets. Furthermore, we exploit the value dependence knowledge to rectify the prediction result for some subsets (average AUC 0.95~0.98). Our research reveals the general rule of the tuple frequency distribution and enables individuals and regulators to responsively predict the re-identification risk.

Introduction

Today we are living in an era of data explosion[1]. We have easier access to information services than at any time in history, but we also face unprecedented privacy risks because your service providers are extremely likely to know you better than you do[2,3,4]. Although service providers often allege that they have to collect as much personal data as possible to improve user experience, they fail to properly protect

user privacy[5,6]. In this regard, the government of many countries promulgated privacy protection laws, such as the General Data Protection Regulation[7] (GDPR) in Europe and Personal Information Security Specification (PISS) in China[8]. PISS emphasizes that all collected personal data should be immediately de-identified and stored separately from their profile data[9,10]. However, even after the de-identification, anonymized personal data still face re-identification risk and are vulnerable to linkage attacks launched by either honest but curious data collectors or malicious hackers[11,12]. Therefore, the re-identification risk of individual data not only reflects the privacy risk level of individuals but also supports regulators in formulating privacy protection policies. Beyond this, it is difficult for individuals and regulatory agencies to obtain the complete dataset maintained by service providers, and they can only infer the re-identification risk from the released incomplete statistical information.

The re-identification risk of an individual is closely related to her/his tuple frequency. The tuple frequency is defined as the count of a specific data value combination, where a high tuple frequency signifies a low re-identification risk. If an attacker has sufficient background knowledge for the linkage attack, individuals will be re-identified by her/his unique data records with 100% probability. Therefore, the uniqueness of individual data has attracted extensive research attention[13]. According to the 1990 and 2000 U.S. census data releases, it takes only three attributes, namely the date of birth, gender, and zip code, to uniquely identify 87% and 63% of the population[14,15]. Montjoye found that it takes only four spatiotemporal points in trajectory data to uniquely identify 95% of the individuals in the location dataset and 90% in the credit card dataset[16,17]. By exploiting the uniqueness contained in the sampled data records or statistical characteristics of datasets, a latent attacker can measure the uniqueness of individuals given incomplete statistical information[18] and even recover the original personal data[19]. However, using the uniqueness to describe the re-identification risk is sometimes inaccurate, because non-unique data records can still be exploited to re-identify individuals from anonymized datasets with a certain probability[20].

Results

The attribute dependence of experimental datasets.

Inspired by k-anonymity[21,22,23], we propose to leverage k-indistinguishability as an indicator to describe the re-identification risk of individuals. If the tuple frequency of an individual in an anonymized dataset is not less than k, then this individual is k-indistinguishable. If the probability of a specific individual being k-indistinguishable can be derived for $k = 2,3,\ldots$, one can have a relatively more comprehensive understanding of her/his re-identification risk. Unfortunately, given incomplete dataset information, the state-of-the-art privacy risk research cannot determine the probability of an individual being k-indistinguishable when $k > 2$. In light of this, the paper presents how to accurately predict the re-identification risk for a given individual with only the incomplete statistical information of the target dataset. Specifically, given some statistical information, the probability mass function (PMF) of the RH distribution can be used to estimate the frequency of the tuples containing strong dependent attribute pairs. In real-world applications, an approximate distribution of the RH distribution is employed to calculate the tuple frequency in an anonymized dataset for computational efficiency, and to further derive the probability of an individual being k-indistinguishable in the target dataset. Our experiments use random[24], demographic[25], medical[13], and educational[26] datasets, and the results show that for all involved datasets, the average AUC of our proposed TFRR is 0.86~0.98, suggesting a high prediction accuracy. For datasets containing strongly dependent attribute pairs, the value dependence knowledge is introduced to rectify the prediction results and the average AUC reaches 0.95~0.98. Our research reveals a general rule determining the distribution of the tuple frequency, which is applicable for all random datasets and most real-world datasets and provides a concise yet effective tool for the re-identification risk prediction of anonymized datasets. With the incomplete statistical information of the target dataset, both individuals and regulators can easily use this tool to predict the re-identification risk. Beyond this function, one can even predict the re-identification risk of submitting data to service providers according to their published data formats, statistical information, and privacy protection plans, and accordingly question whether they obey the existing privacy protection laws, so as to foresee and prevent privacy threats.

Considering dataset D is a table with columns representing attributes and rows representing data records. Each cell in the table maintains the value of a particular attribute of a particular data record. A tuple is defined as an ordered list drawing one value per attribute, to enumerate all possible cases of data records in D, some of which may not appear in D. From the perspective of probability theory, the frequency of a specific tuple in a target dataset follows the RH distribution (see Methods). However, the dependence between the values in the tuple will affect the tuple frequency distribution. Therefore, we define value dependence (see Methods) to describe the dependence between the value pairs of a tuple, and use the value dependence knowledge of a particular tuple to rectify the prediction results. To grasp a general understanding of the dependence between an attribute pair, we define the attribute dependence and analyze the dependence between each attribute pair in experimental datasets (including random and real-world datasets).

The attribute dependence profiles an asymmetric relation between two attributes. The dependence of attribute B on attribute A can be calculated as follows,

$$\mathrm{D}(A \Rightarrow B) = \frac{\mathrm{I}(A;B)}{\mathrm{H}(B)}, \tag{1}$$

where $\mathrm{I}(A;B)$ is the mutual information of A and B, and $\mathrm{H}(B)$ is the entropy of B. (See Methods in Supplementary)

If the value of $\mathrm{D}(A \Rightarrow B)$ or $\mathrm{D}(B \Rightarrow A)$ reaches or exceeds a certain threshold, the attribute pair (A, B) is called as a strongly dependent attribute pair. The threshold is set to 0.5 in this paper, which is very consistent with the subsequent experimental results.

We use a parameter set (N, d, λ) to generate random datasets, where N is the size of the dataset, d is the number of attributes, λ is the indistinguishability indicator equal to the ratio of N to the size of sample space $\mathcal{X}$ (see Supplementary Note 1). For parameter combinations from $N \in \{30000, 150000, 750000, 3750000\}$, $d \in \{2,3,4,5\}$, $\lambda \in \{0.1, 0.5, 2.5, 12.5\}$, 64 different

parameter sets generate 64×50 random datasets, and the attribute dependence of these datasets are illustrated in Fig. 1.

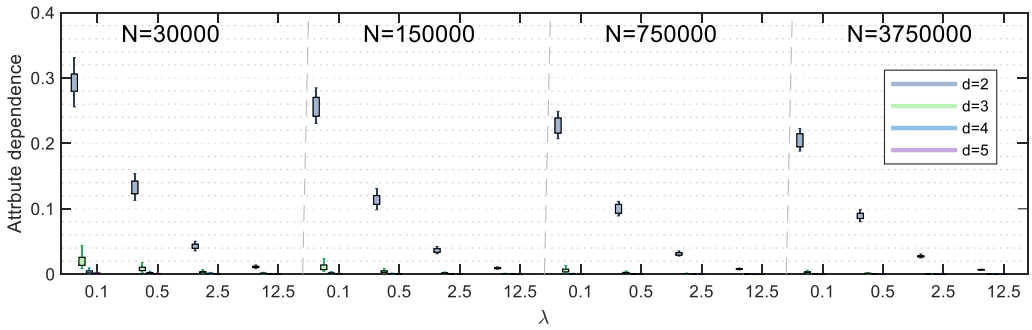

Fig.1 The attribute dependence of random datasets. 64 boxes represent the attribute dependence of the datasets randomly generated by the 64 parameter sets. The grey dotted vertical line divides the figure into 4 sectors, each of which corresponds to a different value of N, the X-axis labels λ, the gray, green, blue, and purple boxes indicate that $d=2,\ 3,\ 4$, and 5, respectively. With two of the parameters fixed, the attribute dependence decreases as the third parameter increases. The decreasing trend with d is the most obvious, followed by λ and N. When $d=2$ and $\lambda=0.1$, the average attribute dependence lies between 0.2 and 0.3. When $d\geq3$, the attribute dependence drops below 0.05, which signifies a very low attribute dependence for random datasets above three dimensions. When $d=2$ and $\lambda=0.1$, the high attribute dependence is due to the small λ and the limited dataset size (see Supplementary Note 2).

We use 4 real-world datasets collected from public sources: the populations of the U.S. (ADULT, Census), medical dataset (SPD), and education dataset (SCHOOL). The attribute dependence of each real-world dataset is shown in Fig. 2a. From each dataset, we create subsets by randomly selecting attributes (columns). The resulting 362 subsets cover a large range of attribute dependence values (0.000062-0.9886), numbers of attributes (2-28), and data records (44842-3985166 individuals).

The distribution of Max Attribute Dependence (MAD) of the 362 subsets is shown in Fig. 2b, where MAD is the maximum attribute dependence of all attribute pairs in a subset.

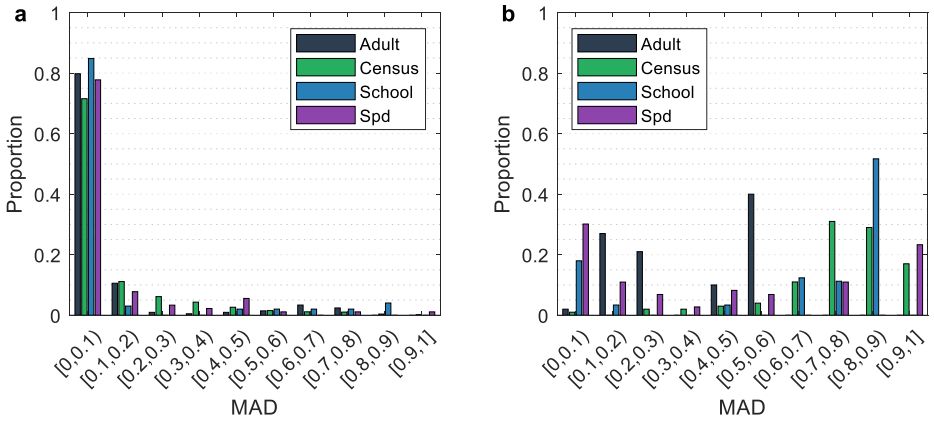

Fig.2 The attribute dependence of real-world datasets. **a** The attribute dependence of real-world datasets. The black, green, blue, and purple columns represent the attribute dependence of Adult, Census, School, and SPD, respectively. 70% of the attribute dependence are less than 0.1 and only 13% of the attribute dependence are greater than 0.3, which means that most of the attribute pairs in these four real-world datasets have a low dependence. **b** The MAD of subsets. 63.3% of the subsets have a $\mathrm{MAD} \geq 0.5$, Adult(40%), Census(92%), School(75.3%), SPD (41.1%), suggesting more than half of the subsets contain at least 1 pair of strongly dependent attributes.

The approximate RH distribution.

Because of the computational complexity of the RH distribution, we expect to find an approximate distribution to reduce the computational burden. According to the analysis in Methods, we find that when $N \to \infty$, the binomial distribution $B(n_d, P_{d-1})$ can be employed to approximate the RH distribution. To have a clearer understanding of the difference between them, we randomly select many tuples from the random datasets and use the PMFs of the two distributions to calculate the occurrence probability of these tuples. The maximum probability distance (MPD) is used to measure the difference between the binomial distribution and the RH distribution, defined as,

$$\mathrm{MPD} = \max_{\mathrm{k}} (\mathrm{f}_{\mathrm{B}}(k) - \mathrm{f}_{\mathrm{rH}}(k)), \tag{2}$$

where $f_B(k)$ and $f_{rH}(k)$ are the PMFs of $B(n_d, P_{d\text{-}1})$ and ${}_rH(N, d, n_1, \cdots, n_d)$.

We use the same 64 parameter sets as in the previous experiment to generate random datasets, and randomly select 1000 tuples from each dataset. The MPD between the binomial distribution and the RH distribution is shown in Fig. 3.

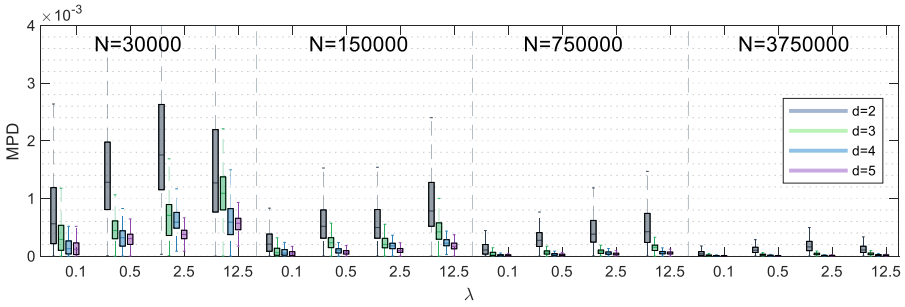

Fig. 3 The MPD between $B(n_d, P_{d\text{-}1})$ and ${}_rH(N, d, n_1, \cdots, n_d)$. The grey dotted vertical line divides the figure into 4 sectors, each of which corresponds to a different value of N, the X-axis labels λ, the gray, green, blue, and purple boxes indicate that $d = 2,\ 3,\ 4$, and 5, respectively. The MPD decreases as N or d increases if the other two parameters are fixed. When $N = 30000$, the MPD lies between 0-0.005. When $N \geq 750000$, the MPD drops to 0-0.001. This signifies that, if $N \to \infty$, the difference between $B(n_d, P_{d\text{-}1})$ and ${}_rH(N, d, n_1, \cdots, n_d)$ is negligibly small and $B(n_d, P_{d\text{-}1})$ can be exploited to approximate ${}_rH(N, d, n_1, \cdots, n_d)$ in calculating the tuple frequency.

The possibility of an individual being k-indistinguishable in random datasets.

We randomly select the data records of 1000 individuals from 64 random datasets and use Eq. 11 to estimate the possibility of these individuals being k-indistinguishable (see Methods for details). The result of binary classification is shown in Fig. 4.

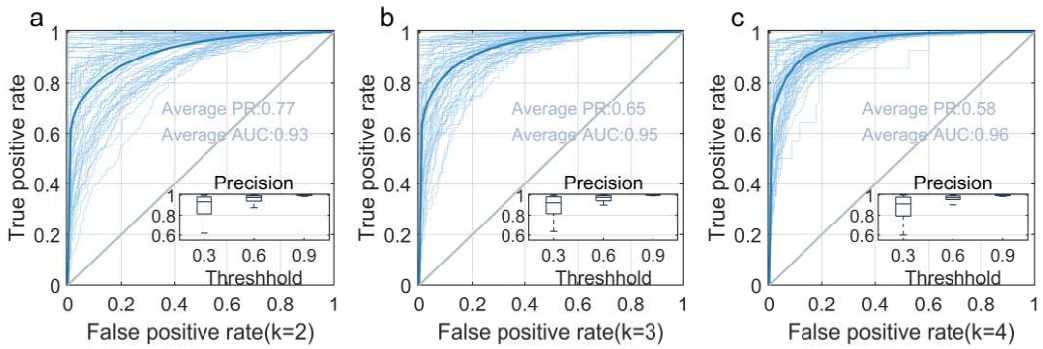

Fig. 4 Receiver operating characteristic (ROC) curves for random datasets (light blue for each dataset and dark blue for the average). **a**, **b,** and **c** identify the impact of different values of k on prediction performance. The AUC of our proposed TFRR prediction is over 0.93, showing a high prediction performance for random datasets. The average positive rate declines sharply as k grows, suggesting fewer individuals will be k-indistinguishable in random datasets if k increases. When the classification threshold is set to 0.6, the precision of our proposed TFRR exceeds 0.9.

The possibility of an individual being k-indistinguishable in real-world datasets.

We randomly select the data records of 1000 individuals from each subset and then calculate the possibility of these individuals being k-indistinguishable. If all the data samples from a subset are either positive or negative, no binary prediction will be performed on this subset (See Supplementary Methods). For readability, we present only the results for subsets from Spd in Fig. 5. Other prediction results are available in Supplementary.

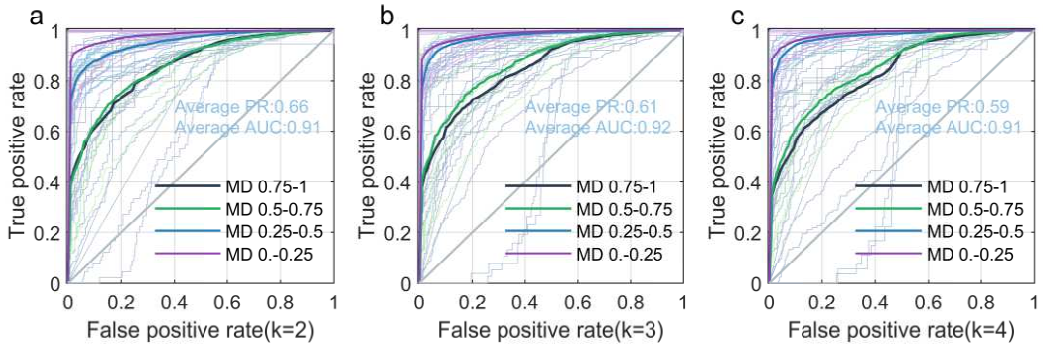

Fig. 5 ROC curves for real-world datasets. **a**, **b**, and **c** identify the impact of different values of k on prediction performance. Gray (black), light green (green), light blue (blue), and light purple (purple) represent the ROC (average ROC) curves when the MADs are [0.75-1), [0.5-0.75), [0.25-0.5), and [0-0.25), respectively. For subsets whose MAD<0.5, the average AUC lies between 0.83~0.87. The average Positive Rate (PR) and average AUC of each subfigure are **a** average PR=0.66 average AUC=0.91. **b** average PR=0.61 average AUC=0.92 **c** average PR=0.59 average AUC=0.91. See Fig. 1 in Supplementary for more information.

Exploiting value dependence knowledge to rectify prediction results.

For the subsets containing strongly dependent attribute pairs, approximation distribution $\mathrm{B}(n_i p_{ij}, p')$ (see Methods for details) is exploited to recalculate the prediction scores of the true positive samples whose prediction scores are less than 0.5. Only the samples of the subsets from Spd are used to recalculate, because most of the subsets from the other real-world datasets contain at least two strongly dependent attribute pairs, meaning that the samples in these datasets may contain two or more strongly dependent value pairs. But $\mathrm{B}(n_i p_{ij}, p')$ can only rectify the negative impact of 1 strongly dependent value pair on the prediction performance. The improved ROC curves are shown in Fig. 6.

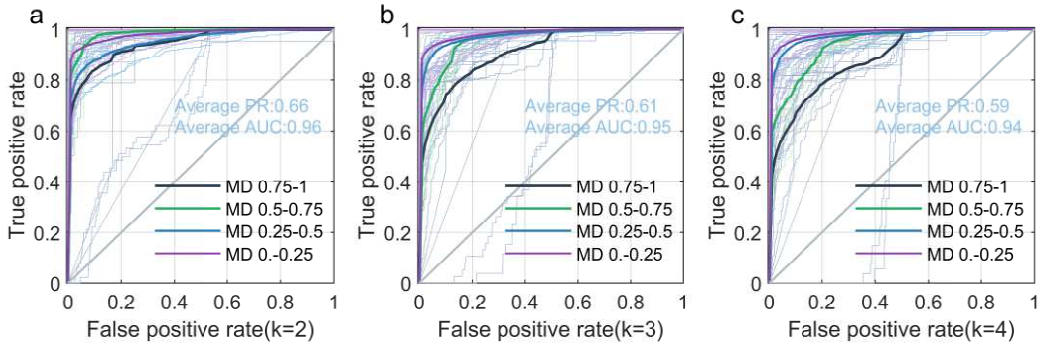

Fig. 6 Rectified ROC curves. This experiment only considers the value dependence knowledge of attribute pair (ZIP code, county), (ZIP code, hospital ID), and (hospital ID, county). The complex dependence knowledge involving more than 3 attribute values is ignored (see Methods). The average PR

and average AUC of each subfigure are **a** average PR=0.66 average AUC=0.96. **b** average PR=0.61 average AUC=0.95 **c** average PR=0.59 average AUC=0.94. The AUC curves of all the tested subsets are rectified by using the value dependence knowledge, except for three subsets with extremely low positive or negative sample ratio (AUC: 0.54~0.80), and the AUC of the other subsets are relatively high (AUC: 0.86~1). This shows that the value dependence knowledge can be leveraged to efficiently improve the prediction performance (see Note 3 in Supplementary).

Discussion

The tuple frequency follows the RH distribution. This rule enables the re-identification risk prediction of individuals before even obtaining the complete dataset. Due to the computational complexity of the RH distribution, the binomial approximation can help predict the privacy risk of individuals.

Both attribute dependence and value dependence are employed to measure the relation between attributes and values. The dependence between values will break the randomness of the value distribution, and the prediction error will be unacceptable when using the binomial approximation to predict the tuple frequency of strongly dependent value pairs. To estimate the frequency of such tuples, we need to not only know the parameters including $N, d, n_1, \cdots, n_d$, but also obtain the attribute dependence knowledge, which can be obtained from:

1. Statistical information published by statistical agencies or data service providers[27,28.]

2. Data sampled from complete datasets. Since the value proportion in the sample is an unbiased estimate of the value proportion in the subset, if the sample and the size of the subset are obtained, the attribute dependence knowledge of the subset can be derived using statistical inference. This enables the estimation of the frequency distribution of the tuple matching a particular individual using the sampled data from the complete dataset.

The knowledge of attribute dependence and value dependence can reveal the internal relation between different attributes and values of data records, which are important indicators for the value distribution of

data[29]. Existing privacy protection methods, such as differential privacy[30,31,32], can hide the original data while ensuring their availability by adding random noise regularly. Although we can obtain useful information, such as the tuple frequency distribution and top-k data from the de-identified dataset[33], the value and utilization of the de-identified dataset are substantially reduced due to the impact of the added noise on the attribute dependence and value dependence[34]. Therefore, we plan to study how to customize the differential privacy budgets and noise generation methods according to the predicted re-identification risk for specific individuals, and how to maximize the preservation of the attribute dependence and value dependence information. In addition, trajectory datasets typically contain temporal and sequential location data with strong attribute dependence[35], which makes the binomial approximation ineffective in privacy risk prediction. This also points out a new research direction for future work.

Methods

The d-RH distribution.

The d-dimensional recursive hypergeometric (d-RH) distribution describes the frequency of a specific tuple in a d-dimensional dataset, namely the tuple frequency. For a d-dimensional dataset D of size N, $\mathcal{X}$ is the universe of all possible tuples in D. Considering $A_1, A_2, \ldots, A_d$ are the value sets of d attributes, $x^i = \{x_1^i, x_2^i, \ldots, x_d^i\} (1 \le i \le |\mathcal{X}|)$ denotes a possible tuple in D, where $x_j^i \in A_j (1 \le j \le d)$ is the j-th value of x^i. Let the frequencies of $x_1^i, x_2^i, \ldots, x_d^i$ in corresponding columns (attributes) be $n_1, n_2, \ldots, n_d$. Let the frequency of x^i in D be X_d, such that $X_d \sim_{\mathrm{r}} \mathrm{H}(N, d, n_1, \cdots, n_{\mathrm{d}})$, then ${}_{\mathrm{r}}\mathrm{H}(N, d, n_1, \cdots, n_d)$ is the d-RH distribution.

Considering $X_j (1 \le j \le d)$ satisfies the j-RH distribution characterized by $N, j, n_1, \cdots, n_j$.

When $j = 1$,

$$\mathrm{P}(X_j = k) = \begin{cases} 1 & k = n_1 \\ 0 & else \end{cases} \tag{3}$$

When $j \geq 2$, the PMF of the j-RH distribution can be obtained recursively as follows,

$$\begin{aligned} \mathrm{P}(X_j = k) &= \sum_{m=k}^{n_{j-1}} \mathrm{P}(X_{j-1} = m)\,\mathrm{P}(X_j = k / X_{j-1} = m) \\ &= \sum_{m=k}^{n_{j-1}} \mathrm{P}(X_{j-1} = m) \cdot \frac{C_m^k C_{N-m}^{n_j - k}}{C_N^{n_j}} \end{aligned} \tag{4}$$

Equation 4 can be interpreted as that, given that sub-tuple $x^i(j-1) = \{x_1^i, \ldots, x_{j-1}^i\} (1 \leq i \leq |\mathcal{X}|, 2 \leq j \leq d)$ appears m times in the first $j-1$ columns of D, the times that $x^i(j)$ appears in the first j columns follow the hypergeometric distribution $\mathrm{H}(N, m, n_j)$.

Therefore, the PMF of ${}_{\mathrm{r}}\mathrm{H}(N, d, n_1, \cdots, n_d)$ can be obtained similarly by the recursive calculation of Eq. 4.

The PMF of 2-RH distribution ${}_{\mathrm{r}}\mathrm{H}(N, 2, n_1, n_2)$ is,

$$\mathrm{P}(X_2 = k) = \frac{C_{n_1}^k C_{N-n_1}^{n_2 - k}}{C_N^{n_2}}. \tag{5}$$

The 2-RH distribution ${}_{\mathrm{r}}\mathrm{H}(N, 2, n_1, n_2)$ and the hypergeometric distribution $\mathrm{H}(N, n_1, n_2)$ are identical. Therefore, we consider that the hypergeometric distribution is only a special case of d-RH distributions with $d = 2$.

The binomial approximation of the d-RH distribution.

Since the frequency of the d-th value of x^i is n_d, the frequency of x^i must be less than or equal to n_d, namely $X_d \leq n_d$. Considering r is a record in D whose d-th value is x_d^i and a n_d round experiment is designed. In each round, we randomly draw one value per attribute in the first $d-1$

columns without replacement and combine them with the fixed x_d^i to obtain r. The count of rounds in which record r matching tuple x^i is equal to X_d, such that $X_d \sim_{\mathrm{r}} \mathrm{H}(N, d, n_1, \cdots, n_{\mathrm{d}})$.

Because we draw values without replacement in each round, the probability of record r matching tuple x^i varies[36]. But when $N \to \infty$, this probability stays almost the same. Therefore, the count of rounds satisfies the binomial distribution. The probability of record r matching tuple x^i is as follows,

$$P_{d-1} = n_1 \times \cdots \times n_{d-1} / N^{d-1} \tag{6}$$

Then the probability of record r matching tuple x^i for k rounds is as follows,

$$\mathrm{P}(X_d = k) = C_{n_d}^k P_{d-1}^k (1 - P_{d-1})^{n_d - k}. \tag{7}$$

In conclusion, when $N \to \infty$, the binomial distribution $\mathrm{B}(n_d, P_{d-1})$ approximates the d-RH distribution ${}_{\mathrm{r}}\mathrm{H}(N, d, n_1, \cdots, n_d)$.

Value dependence.

There is a dependence between each pair of values in a tuple. For example, in a shopping dataset, someone who bought a long dress is highly likely to be female. The value long dress determines to a large extent the value female. The confidence and lift[37] in association rule learning are employed as indicators to describe the dependence between a pair of values.

Let $x_1 \in A_1$ and $x_2 \in A_1$ be two values, E_x and E_y be the events that record r contains x_1 and x_2, then the confidence and lift of the association rule that x_1 implies x_2 are as follows.

$$\mathrm{conf}(x_1 \Rightarrow x_2) = \frac{|\{r \in D; \{x_1, x_2\} \subseteq r\}|}{|\{r \in D; \{x_1\} \subseteq r\}|} = \mathrm{P}(E_{x_2} \mid E_{x_1}) \tag{8}$$

$$\mathrm{lift}(x_1 \Rightarrow x_2) = \frac{|\{r \in D; \{x_1, x_2\} \subseteq r\}|}{|\{r \in D; \{x_1\} \subseteq r\}| \times |\{r \in D; \{x_2\} \subseteq r\}|} = \frac{\mathrm{P}(E_{x_1} \cap E_{x_2})}{\mathrm{P}(E_{x_1}) P(E_{x_2})} \tag{9}$$

If $\mathrm{conf}(x_1 \Rightarrow x_2)$ or $\mathrm{conf}(x_2 \Rightarrow x_1)$ reaches or exceeds a certain threshold, the value pair (x_1, x_2) is called as a strongly dependent value pair, and the threshold is set to 0.5 in this paper.

The frequency distribution of tuples with strongly dependent value pairs.

The physical significance of d-RH distribution can be summarized as follows. Let $x = (x_1, \cdots, x_d)$ be a possible tuple of a d-dimensional dataset D with size N, X be the frequency of tuple x in D. Let the frequencies of $x_1, \cdots, x_d$ in the corresponding column be $n_1, n_2, \ldots, n_d$. Considering $\mathcal{D}^i$ is the set of all datasets in which the frequency of x is i, and $\mathcal{D} = \mathcal{D}^0 \bigcup \ldots \bigcup \mathcal{D}^\infty$ is the set of all datasets that may contain x. The d-RH distribution is based on classical probability theory. From the perspective of classical probability theory[38], each dataset in $\mathcal{D}$ has the same odds to be D, and the probability of tuple x appearing k times in D can be obtained by $P(X = k) = \left|\mathcal{D}^k\right| / \left|\mathcal{D}\right|$. When $n_1, n_2, \ldots, n_d \ll N$, in most datasets of $\mathcal{D}$, x does not contain strongly dependent value pairs. Therefore, for the tuples without strongly dependent value pairs, each dataset in $\mathcal{D}$ has almost the same odds to be D and the d-RH distribution is suitable for predicting the frequency of such tuples. For the tuples with strongly dependent value pairs, the occurrence probability of each dataset is unequal, and therefore the d-RH distribution will fail in prediction.

Fortunately, introducing the background knowledge of value dependence can improve the prediction result. Assume that all the values of the tuple x are divided into groups A and B, and all values in A are the members of the strongly dependent value pair, whereas all values in B are not the members. The times of group A appearing in D is denoted by m, the number of members in B is s, ordered by $i_1, \ldots i_s$, then the frequency of tuple x can be denoted by an $(s+1)$-RH distribution with parameter $N, s+1, n_{i_1}, \cdots, n_{i_s}, m$.

When the number of values in A is greater than 2, the dependence between members becomes complex. Therefore, we only consider in this paper the cases that group A has only 2 values, and give the binomial approximation for these cases. Considering group A only has two values x_i and x_j, where $1 \le i, j \le d$ and $\text{conf}(x_i, x_j) = p_{ij} > \text{conf}(x_j, x_i)$, then

$$n' = \begin{cases} n_i p_{ij} & d > 2 \\ n_i & d = 2 \end{cases} \tag{10}$$

$$p' = \begin{cases} \left(\prod_{k=1, k \neq i, j}^{d} n_k \right) \Big/ N^{d-2} & d > 2 \\ p_{ij} & d = 2 \end{cases} \tag{11}$$

The frequency distribution of x in D can be approximated by $\text{B}(n', p')$.

The probability of a specific individual being k-indistinguishable.

Considering $r = (r_1, \cdots, r_d)$ is the record of an individual p in the d-dimensional dataset D, X_d is the frequency of the tuple matching r in D, such that X_d follows the d-RH distribution.

Then the probability of p being k-indistinguishable when $k \ge 2$ can be calculated as follows

$$\begin{aligned} &\text{P}(p \text{ is } k-\text{indistinguishable}) \\ &= \text{P}(X_d \ge k \mid r \text{ in } D) \\ &= \text{P}(X_d \ge k) / \text{P}(r \text{ in } D) \\ &= \{1 - \textstyle\sum_{i=0}^{k-1} \text{P}(X_d = i)\} / \{1 - \text{P}(X_d = 0)\} \end{aligned} \tag{12}$$

Data Availability

All data information related to the article is provided in the Supplementary Methods.

Code Availability

All simulations were implemented in Matlab. The source code to reproduce the experiments will be is deposited in Code ocean or Github.

References

1. Marx, V. (2013). The big challenges of big data. Nature, 498(7453), 255-260.
2. Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073-1076.
3. Thompson, S. A., & Warzel, C.. How to Track President Trump. The New York Times, 12, 20 (2019).
4. Raskar, R., Schunemann, I., Barbar, R., Vilcans, K., Gray, J., Vepakomma, P., ... & Werner, J. (2020). Apps gone rogue: Maintaining personal privacy in an epidemic. arXiv preprint arXiv:2003.08567.
5. Kozlowska, I. (2018). Facebook and data privacy in the age of Cambridge Analytica. Seattle, WA: The University of Washington. Retrieved August, 1, 2019.
6. Ablon, L., & Libicki, M. (2015). Hacker's bazaar: The markets for cybercrime tools and stolen data. Def. Counsel J., 82, 143.
7. Voigt, P., & Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing, 10, 3152676.
8. Wilmer Hale. China Issues New Personal Information Security Specification.3, 24 (2020). https://www.wilmerhale.com/en/insights/client-alerts/20200324-china-issues-new-personal-information-security-specification
9. Shi, M. L., Sacks, S., Chen, Q. H., & Webster, G. (2019). Translation: China's personal information security specification. New America. https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinas-personal-information-security-specification/

10. Hong, Y. et al. Information security technology—Personal information security specification. https://www.tc260.org.cn/upload/2020-09-18/1600432872689070371.pdf (2020)

11. Clauβ, S., Kesdogan, D., & Kölsch, T. (2005, November). Privacy enhancing identity management: protection against re-identification and profiling. In Proceedings of the 2005 workshop on Digital identity management (pp. 84-93).

12. Henriksen-Bulmer, J., & Jeary, S. (2016). Re-identification attacks—A systematic literature review. International Journal of Information Management, 36(6), 1184-1192.

13. Sánchez, D., Martnez, S. & Domingo-Ferrer, J. Comment on unique in the shopping mall: on the reidentifiability of credit card metadata. Science 351, 1274 (2016).

14. Sweeney, L. (2000). Uniqueness of simple demographics in the US population. LIDAP-WP4, 2000.

15. Golle, P. (2006, October). Revisiting the uniqueness of simple demographics in the US population. In Proceedings of the 5th ACM Workshop on Privacy in Electronic Society (pp. 77-80).

16. De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. Scientific reports, 3(1), 1-5.

17. De Montjoye, Y. A., Radaelli, L., & Singh, V. K. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. Science, 347(6221), 536-539.

18. Rocher, L., Hendrickx, J. M., & De Montjoye, Y. A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications, 10(1), 1-9.

19. Tu, Z., Xu, F., Li, Y., Zhang, P., & Jin, D. (2018). A new privacy breach: User trajectory recovery from aggregated mobility data. IEEE/ACM Transactions on Networking, 26(3), 1446-1459.

20. Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. Journal of the American Medical Informatics Association, 17(2), 169-177.

21. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

22. Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 571-588.

23. El Emam, K., & Dankar, F. K. (2008). Protecting privacy using k-anonymity. Journal of the American Medical Informatics Association, 15(5), 627-637.

24. Yang, Z., Wang, R., Luo, D., & Xiong, Y. (2020). Rapid Re-Identification Risk Assessment for Anonymous Data Set in Mobile Multimedia Scene. IEEE Access, 8, 41557-41565.

25. Asuncion, A., & Newman, D. (2007). UCI machine learning repository.

26. Romania government. (2014). Primary school enrollment 2014.

27. Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys (Csur), 42(4), 1-53.

28. Abella, A., Ortiz-de-Urbina-Criado, M., & De-Pablos-Heredero, C. (2019). The process of open data publication and reuse. Journal of the Association for Information Science and Technology, 70(3), 296-300.

29. Zheng, Y., Zhang, L., Xie, X., & Ma, W. Y. (2009, November). Mining correlation between locations using human location history. In Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems (pp. 472-475).

30. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 211-407.

31. Friedman, A., & Schuster, A. (2010, July). Data mining with differential privacy. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 493-502).

32. Xiong, J., Ma, R., Chen, L., Tian, Y., Li, Q., Liu, X., & Yao, Z. (2019). A personalized privacy protection framework for mobile crowdsensing in IIoT. IEEE Transactions on Industrial Informatics, 16(6), 4231-4241.

33. Fanti, G., Pihur, V., & Erlingsson, Ú. (2015). Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries. arXiv preprint arXiv:1503.01214.

34. Ren, X., Yu, C. M., Yu, W., Yang, S., Yang, X., McCann, J. A., & Philip, S. Y. (2018). LoPub: High-Dimensional Crowdsourced Data Publication with Local Differential Privacy. IEEE Transactions on Information Forensics and Security, 13(9), 2151-2166.

35. Yang, Z., Wang, R., Wu, D., & Luo, D. (2020). UTM: A trajectory privacy evaluating model for online health monitoring. Digital Communications and Networks.

36. Berkopec, A. (2007). HyperQuick algorithm for discrete hypergeometric distribution. Journal of Discrete Algorithms, 5(2), 341-347.

37. Hornik, K., Grün, B., & Hahsler, M. (2005). arules-A computational environment for mining association rules and frequent item sets. Journal of statistical software, 14(15), 1-25.

38. Jaynes, E. T. (2003). Probability theory: The logic of science. Cambridge university press.

Acknowledgments (optional)

This work was supported.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.pdf