IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 31, 2019, accepted February 10, 2020, date of publication March 2, 2020, date of current version March 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2977404

Rapid Re-Identification Risk Assessment for Anonymous Data Set in Mobile Multimedia Scene

ZHIGANG YANG[1,2,3,4], **(Member, IEEE), RUYAN WANG**[1,3,4], **(Member, IEEE), DAIZHONG LUO**[2], **AND YU XIONG**[2]

[1]School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2]School of Artificial Intelligence, Chongqing University of Arts and Sciences, Chongqing 402160, China
[3]Key Laboratory of Optical Communication and Networks, Chongqing 400065, China
[4]Key Laboratory of Ubiquitous Sensing and Networking, Chongqing 400065, China

Corresponding author: Zhigang Yang (ayzg163@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61771082, Grant 61871062, and Grant 61901078, in part by the Scientific and Technological Research Program of Chongqing Municipal Education Commission under Grant KJZD-K201901301, in part by the Natural Science Foundation of Chongqing of China under Grant cstc2013jcyjA40066, in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant KJQN201801316, and in part by the Industrial Technology Development Project of Chongqing Development and Reform Commission under Grant 2018148208.

ABSTRACT Ubiquitous mobile multimedia applications bring great convenience to users. However, when enjoying mobile multimedia services, users provide personal data to service platforms. Although the service platforms always claim that the collected personal data are de-identified, the risk of re-identifying users through linkage attacks still exists and is incalculable. This paper proposes a rapid prediction model for the overall re-identification risk based on the statistics of data sets (i.e., the number of individuals, number of attributes, distribution of attribute values, and attribute dependency). Our proposed model reveals the impact of statistics on the overall re-identification risk and adopts random sampling and semi-random sampling methods to predict the overall re-identification risk of data sets with and without strong dependency ordered attribute pairs. Experimental results show that for the data sets without strong dependency ordered attribute pairs, the random sampling method has a high prediction accuracy (the prediction error is less than 0.05). For the data sets with strong dependency ordered attribute pairs, the semi-random sampling method has a high prediction accuracy (the prediction error is less than 0.09). Exploiting our model, governments and individuals can quickly assess the privacy leakage risk of their data sets, given only the statistic of the data sets. Besides, this model can also evaluate the privacy risk of data collection schemes in advance according to historical statistics, and identify suspected services.

INDEX TERMS Multimedia, privacy, overall re-identification risk, attribute dependency.

I. INTRODUCTION

With the wide popularity of smart terminals and development of wireless communication technology, mobile multimedia applications become the indispensable tool for daily life and work [1]–[3]. Ubiquitous access, rich functions and good experience make mobile multimedia applications more and more popular. However, mobile multimedia service providers, in order to increase user viscosity, improve user experience, or reserve data resources, collect user personal information while providing services. While enjoying the convenience of mobile multimedia services, users must take on the risk of privacy disclosure. For example, the web browsing history will expose users' consumption habits, sexual orientation, political leanings and other private data. And trajectories of users will expose sensitive information such as home address and workplace. Although information collectors always claim that the purpose of collecting personal data is to provide better services to users, and personal information will be de-identified and properly preserved. But potential security problems remain, even if the information collectors are not malicious. Many incidents of service provider data breach, such as the Facebook data privacy scandal and the Equifax data breach, suggest that improper data sharing and ubiquitous hacking make data stored on servers highly vulnerable. Although the leaked data may not contain the user's identity, user's quasi-identifiers such as age, gender, and zip code in the anonymous data can be collected by many

The associate editor coordinating the review of this manuscript and approving it for publication was Dalei Wu.

multimedia application providers (e.g. Facebook and Twitter) through various smart terminals and IOT devices [4]–[6]. The combination of these quasi-identifiers is often used by attackers to re-identify the anonymous user. Famous attacks include re-identification of a Massachusetts hospital anonymous records by linking it to the public voter database [7] and de-anonymization of anonymous subscribers in large sparse data set (i.e., Netflix Prize data set) whose background knowledge (as few as 5-10 attributes) can be get from Internet Movie Database [8].

In order to protect citizens' privacy, many governments have promulgated privacy protection regulations or personal information protection laws, such as the General Data Protection Regulation (GDPR) in European Union and the Data Protection Act (DPA) in United Kingdom, considering that each person in data set should be anonymous. And GDPR define the higher standard for anonymization, personal data should not contain obvious identifiers and not be re-identifiable. However, the contradiction between data sharing and privacy protection still exists, and the scale of privacy protection is still difficult to define. Due to the lack of effective privacy risk assessment methods, how to strike a balance between privacy protection and data sharing is still a hard problem.

The re-identification risk of anonymous user, which is defined as the inverse of the number of records matching the user attribute group in data set, is the key indicator of privacy risk assessment. If only a unique record matches the user attribute in the data set, the probability of his re-identification risk is 1. If another 3 records match the same attribute group, the probability drops to 1/4. The famous privacy protection model k-anonymity requires each anonymous record in data set sharing the same attribute group with at least another $k - 1$ records [9]. But in the real world, the records of users in data set are highly unique. Rocher et al. find that 99.98% of Americans would be correctly re-identified by 15 demographic attributes [10]. The study shows that, even in a huge data set, almost all of users can be re-identified by enough attributes. Besides, from a qualitative perspective, number of individuals, distribution of attribute values, and attribute dependency may also affect the re-identification risk. But there is no simple method to briskly assess the re-identification risk based on the statistic of data set.

We denote the average re-identification risk of all users in data set as overall re-identification risk (ORR). ORR is an important indicator for assessing the privacy disclosure risk of data set. For governments, it is an important tool to define the scale of privacy protection. For users, it is related to the security of sensitive personal information. For data collectors, it means the privacy risk of publishing anonymous data set. For attackers, it reveals the probability of successfully attacking. Although, the data collectors can easily calculate the ORR of the data set, they are unwilling to disclosure it, for commercial purposes or security considerations. Fortunately, for the purpose of data sharing, some data collectors publish incomplete information about the collected data, such as statistics or sampling data of the original complete data set. And the incomplete information may contain some knowledge about ORR. Therefore, how to predict the ORR of complete data sets when only partial information obtained is still an important and challenging problem in the field of privacy protection.

In summary, this paper proposes a rapid re-identification risk assessment (R3A) model for anonymous data set in mobile multimedia scene. The main contributions of this paper are as follows:

- Reveal the relationship between re-identification risk and statistics of data set, and first propose the rapid re-identification risk predicting method based on statistics.
- Propose information gain ratio and frequent tuple to describe the attribute dependency. Random sampling and semi-random sampling method are proposed for different degree of attribute dependency, achieving high prediction accuracy.

The rest of the paper is organized as follows. Related work is reviewed and summarized in Section II and Section III presents R3A model. Experiment result and analysis are given in section IV and 4 rules about entropy and ORR are discussed in Section V. Finally, we conclude the paper and propose some future work in Section VI.

II. RELATED WORK

Current research on privacy protection technology focuses on privacy protection means. Sweeney [9], proposed k-anonymity model, through generalization and concealment, each record at least share the same quasi-identifier attribute group with the other $k - 1$ records in the data set, thus the probability of successful linkage attack drops to $1/k$. Due to the values of sensitive attributes associated with quasi-identifier group may be similar, k-anonymity model does not preserve the privacy. Then, l-diversity and t-closeness and other more privacy protection model have been proposed [11], [12]. But these models always need to be refined for new types of attacks, and they are not suitable for modern high-dimensional data set. In 2006, Dwork [13], proposed the differential privacy, providing stringent mathematical underpinning and reliable privacy performance evaluation, can resist various attack considering the maximum background knowledge of attackers. Recently, the privacy protection models or technologies combined with artificial intelligence or blockchain technology become a new research hot spots [14]–[17]. However, the above privacy models focus on maximum protection against various attacks, and do not concern the re-identification risk of anonymous data set.

Studies on re-identification risk of data set are common in the fields of statistics and medicine. The re-identification risk of user is often defined as the product of membership probability and success linkage probability [18]. Member probability is the probability that the target user appears in the data set, which is decided by the attacker's background knowledge. The success linkage probability is determined by the number of records matching the target user's quasi-identifier attribute group in the data set. If there are k records in the data set matching the userİŕs quasi-identifier attribute group, the link success probability is $1/k$. Since the member probability is determined by the attacker's background knowledge, which is difficult to estimate, most researchers set the member probability as 1, then the re-identification risk of user equals the success linkage probability of user. Because the user with unique records, whose success linkage probability is 100%, is certain to be re-identified, some studies also equate unique probability with the re-identification risk [19].

El Emam *et al.* [20], emphasized that the uniqueness decreases with population size growth. They managed re-identification risk by controlling population size of data set. Sweeney and Golle et al. found that, in 1990 87% of U.S. population can be uniquely identified by birthdate, gender, and ZIP code [21], while in 2000 the ratio dropped to 63% [22]. Due to the uncertainty of the data collecting methods of above two studies, we do not know the real reason of the decline of the American population uniqueness. But we found the fact that, compared with 1990, America's population grew by 13% in 2000. And the growth of population size generally leads to the decrease of uniqueness. In study [10], Rocher et al. proposed a generative copula-based model to accurately predict the uniqueness of data records by random sampling from the complete data set. The study shows that, the uniqueness of data set can be predicted by partial information of complete data set (e.g., extremely incomplete sampling data sets). But the study did not concern the effect of statistics on uniqueness.

Against modern high-dimensional and sparse macro-data, Narayanan and Shmatikov [8] presents a new class of statistical de-anonymization (namely re-identification) attacks, which can easily identify anonymous subscribers by only 5-10 known attributes and uncover their potentially sensitive information. Merener [23] extended the study, established mathematical theory describing results on de-anonymization that can be achieved by an adversary under general and realistic assumptions. He also found the fact that when the auxiliary information including a rare attribute of D, the size of auxiliary information could be reduced in about 50%. The theory and algorithm applied on Joint Canada/United States Survey of Health 2004, which is less sparsity than Netflix database, getting a satisfactory success of empirical linkage attack.

The trajectory data set is a special data set, and trajectory uniqueness is a commonly highlighted research problem. Y. A. D. Montjoye et al. asserted in [24], that trajectories of 95% users can be uniquely determined by four spatio-temporal points, and in [25], that four spatio-temporal points can also uniquely fix the trajectories of 90% credit card holders. Both studies emphasized that, the trajectory uniqueness grows dramatically with the increasing time slots. Although two enormous data sets covering millions of users were analyzed by the above two studies, the trajectory uniqueness and its probability evaluation is scenario dependent and may be inapplicable to other trajectory data sets, encouraging vigorous discussions [19]. Tu *et al.* [26] proposed an attack system to recover user trajectories with an accuracy of 73%∼91%, from aggregated mobile data sets (i.e., the number of users covered by a cellular tower at a specific time stamp. Although the study did not reveal any statistical correlation between uniqueness and aggregated data, it hinted an association between them.

The above studies implied that, the statistic of data set, such as number of individuals, number of attributes, distribution of attribute values, may affect the uniqueness or re-identification risk of data set. However, no study had researched the deeper relationship between statistics and re-identification risk. To the best of our knowledge, we are very first to propose rapid re-identification risk assessment method based on statistics.

III. R3A MODEL

From a statistical perspective, we assume that data sets with the same statistics (i.e., number of individuals, distribution of attribute values, attribute dependency) have similar ORRs (this assumption will be verified by simulation in Chapter IV). Based on this assumption, we propose R3A model, in which the ORR of target data set can be predicted by the average ORR of random data sets with the same statistic. Considering that data owners may not disclose the attribute dependency, R3A model recommends two predicting methods, namely, full random sampling without the knowledge of attribute dependency and semi-random sampling considering attribute dependency.

A. DEFINITION

This section defines the terms used in the paper. We use the attribute value frequency matrix (AVFM) to describe the number of individuals, the number of attributes, and the distribution of attribute values. The attribute dependency are quantified by the information gain ratio.

Definition 1: AVFM

We consider a data set D containing n records. Each row is a user record with d quasi-attributes. The set of the j-th attribute values of all users is denoted as $q^{(j)}$, which containing l_j elements. The element d_{ij} represents the j-th attribute value of the user $x^{(i)}$. For example, d_{25} = male, representing the fifth attribute value of $x^{(2)}$ is male.

We denote the frequency of i-th element of $q^{(j)}$ as k_{ij} $(1 \leqslant j \leqslant d, 1 \leqslant i \leqslant l_j, i, j \in Z)$. Let $l = \max(l_j)(1 \leqslant j \leqslant d)$, the AVFM m of data set D can be defined as follows.

$$m = \begin{bmatrix} f_{11} & f_{12} & f_{13} & \cdots & f_{1d} \\ f_{21} & f_{23} & f_{24} & \cdots & f_{2d} \\ f_{31} & f_{32} & f_{33} & \cdots & f_{3d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{l1} & f_{l2} & f_{l3} & \cdots & f_{ld} \end{bmatrix} \tag{1}$$

And

$$f_{ij} = \begin{cases} k_{ij}, & 1 \leqslant i \leqslant l_j \\ 0, & l_j < i \leqslant l \end{cases} 1 \leqslant j \leqslant d \tag{2}$$

$$\sum_{i=1}^{l} f_{ij} = n(1 \leq j \leq d) \tag{3}$$

TABLE 1. An example of data set D.

Age	Gender	Smoking
20	Male	Yes
25	Male	Yes
25	Female	No
25	Female	No
35	Male	No

The generation of AVFM is described in detail below. As shown in Table 1, there are 5 records in data set D, each of which has three attributes. The attribute value ratios of the three attributes are 1:2:1, 3:2 and 2:3. The AVFM m of data set D is as follows.

$$m = \begin{bmatrix} 1 & 3 & 2 \\ 2 & 2 & 3 \\ 1 & 0 & 0 \end{bmatrix} \tag{4}$$

Obviously, if the AVFM of the data set D is known, we can easily calculate the number of individuals, the number of attributes and the distribution of attribute values in data set D.

Definition 2: ORR

If David's record $x^{(m)}$ is unique in data set D, then he is always correctly re-identified. If there are another two users sharing same record with David, then the re-identification probability of David is 1/3. We consider the number of potential false positives in the data set is $T \equiv \sum_{i=1}^{n}[x^i = x] - 1$. According to [10], the user with record x can be correctly re-identified with the probability of h_x, which is defined as follows.

$$\begin{aligned} h_x &\equiv \Pr(x\ correctly\ re - identified|\exists i, x^{(i)} = x) \\ &= \sum_{k=0}^{n-1} \frac{1}{k+1} \Pr(T = k) \end{aligned} \tag{5}$$

The ORR of data set D is equal to the average re-identification probability of every users in data set D, which is defined as follows.

$$orr \equiv \sum_{x \in D} h_x / n \tag{6}$$

Definition 3: Information gain ratio

We denote the set of all possible values of attribute A as l_A. Considering a is an element of l_A, f_a denotes the frequency of a in data set D. The entropy of attribute A is defined as follows.

$$H(A) = -\sum_{a \in l_A} \frac{f_a}{n} \log_2 \frac{f_a}{n} \tag{7}$$

We consider tuple (a, b) is an element of $l_A \times l_B$, and $f_{a\wedge b}$ denotes the frequency of (a, b) in data set D. The mutual information of attribute A and attribute B is defined as follows.

$$I(A, B) = \sum_{(a,b)\in l_A \times l_B} \frac{f_{a\wedge b}}{n} \log_2 \frac{f_{a\wedge b}/n}{(f_a/n)(f_b/n)} \tag{8}$$

The information gain ratio of B on A is defined as follows.

$$g(A, B) = \frac{I(A, B)}{H(A)} \tag{9}$$

Definition 4: Support and confidence

T and S are attribute groups of data set D, and $T \cap S = \Phi$. Tuple t is a value of T, and tuple s is a value of S. The support of t with respect to D is defined as the proportion of users in the data set which contains the item t.

$$\text{supp}(t) = \frac{|\{\, x | t \subseteq x\}|}{n} \tag{10}$$

The confidence value of a rule, $t \Rightarrow s$, with respect to a set of data set D, is the proportion of the users that contains t which also contains s.

Confidence $t \Rightarrow s$ is defined as:

$$\text{conf}(t \Rightarrow s) = \text{supp}(t \cup s)/\text{supp}(t) \tag{11}$$

and supp$(t \cup s)$ means the support of the union of the items t and s.

For example, the rule $\{smoking\} \Rightarrow \{male\}$ has a confidence of 1.0 in a data set, which means that for 100% of the smoker the rule is correct (100% of the smoker is male).

Definition 5: Attribute dependency

We consider tuple t is a value of attribute group T, and its frequency in data set D is f_t. The entropy of attribute group T is defined as follows.

$$H(T) = -\sum_{t \in T} \frac{f_t}{n} \log_2 \frac{f_t}{n} \tag{12}$$

The attribute group S is dependent on the attribute group T, if the knowledge of T can reduce the uncertainty of S (i.e., entropy). Obviously, the attribute dependency is asymmetric, we use the information gain ratio $g(S, T)$ to quantify the dependency of S on T. When $g(S, T) = 1$, S is completely dependent on T; when $g(S, T) = 0$, S is completely independent on T; when $0 < g(S, T) < 1$, S is partially dependent on T. We call S is weakly dependent on T when $0 < g(S, T) < 0.5$, and strongly dependent on T when $0.5 \leq g(S, T) < 1$. The relationship among dependency, information Gain Ratio and support is shown as Table 2.

Definition 6: Experience entropy

We consider data set D with n records consists of d attributes, which are denoted as A_1 to A_n. We define $\mathbb{M} = A_1 \times \ldots \times A_d$, tuple $m = (a_1, \ldots, a_d)$ is an element of $\mathbb{M}$, and the frequencies of a_1 to a_d in data set D are f_1 to f_d. The entropy of data set D is defined as follows.

$$H(D) = -\sum_{m \in D} \frac{f_m}{n} \log_2 \frac{f_m}{n} \tag{13}$$

And apparently, if every tuple in D is unique, the entropy is at its maximum.

$$max\ entropy = \log_2 n \tag{14}$$

TABLE 2. Dependency, information gain ratio and support.

Dependency		Information Gain Ratio	Support
Complete independency		$g(S,T)=0$	$\forall s \in S, \forall t \in T$, $\text{supp}(t \cup s) > 0 \Rightarrow \text{supp}(s \cup t) = \text{supp}(s)\text{supp}(t)$
Partial dependency	Weak dependency	$0 < g(S,T) < 0.5$	$\exists s \in S, \exists t \in T$, $\text{supp}(t \cup s) > 0$ and $\text{supp}(s \cup t) \neq \text{supp}(s)\text{supp}(t)$
	Strong dependency	$0.5 \leq g(S,T) < 1$	
Complete dependency		$g(S,T)=1$	$\forall s \in S, \forall t \in T$, $\text{supp}(t \cup s) > 0 \Rightarrow \text{supp}(s \cup t) = \text{supp}(t)$

We consider the probability of tuple m is $p_m = \frac{f_1 \times \ldots \times f_d}{n^d}$, and the experience entropy of data set D is defined as follows.

$$experience\ entropy = -\sum_{m \in \mathbb{M}} p_m \log_2 p_m \tag{15}$$

B. RANDOM SAMPLING METHOD

The AVFM of the data set implies all the statistical characteristics required by the random sampling method. We considered that the set of all data sets with the same AVFM m is $\mathbb{D}_m$. The set of overall re-identification risks of every data set in $\mathbb{D}_m$ is the population $\mathbb{R}_m$. Due to the extremely large capacity of $\mathbb{R}_m$, we adopted the random sampling method to analyze the statistical property of $\mathbb{R}_m$. We considered the capacity of each sample is 1, the method of sample selection is as follows: first, the standard data set is generated based on m. Second, each column element in the standard data set is randomly sorted to generate a new data set, and the over re-identification risk of the new data set is equivalent to a new sample which is randomly selected from $\mathbb{R}_m$. Then, repeat step 2 to get more random samples. Due to the capacity of $\mathbb{R}_m$ is extremely large, the sampling method is equivalent to sampling without replacement.

We considered the AVFM m of data set D is shown in formula 16, the standard data set based on m is D_m. The process of random sampling is described in detail below.

$$m = \begin{bmatrix} 2 & 3 & 2 \\ 3 & 3 & 2 \\ 1 & 0 & 2 \end{bmatrix} \tag{16}$$

The first column of D_m $[1\ 1\ 2\ 2\ 2\ 3]^{\mathrm{T}}$ is generated based on the first column of m. The elements in the standard data set do not represent the actual attribute value, but only the ordinal number of the attribute value in the corresponding attribute. Similarly, all columns of D_m are generated as follows.

$$D_m = \begin{bmatrix} 1 & 1 & 2 & 2 & 2 & 3 \\ 1 & 1 & 1 & 2 & 2 & 2 \\ 1 & 1 & 2 & 2 & 3 & 3 \end{bmatrix}^{\mathrm{T}} \tag{17}$$

Then D'_m is generated through randomly shuffling the order of elements in each column of D_m. For example,

$$D'_m = \begin{bmatrix} 2 & 1 & 1 & 2 & 3 & 2 \\ 2 & 1 & 1 & 2 & 1 & 2 \\ 3 & 2 & 3 & 1 & 2 & 1 \end{bmatrix}^{\mathrm{T}}$$

is a randomly generated data set, with two identical records and four unique records. The ORR of D'_m is $oor = \sum_{x \in D'_m} h_x/n = 5/6$. Obviously, D'_m is different from D_m, but having the same AVFM with D_m. So the *ORR* of D'_m is a new sample of $\mathbb{R}_m$.

We considered the AVFM m of target data set D is a $l \times d$ matrix, where the sum of each column is n. The algorithm of random sampling method is shown as Algorithm 1.

Algorithm 1 Random Sampling Method

```
Input:
AVFM of target data set $D$: $m$;
Number of samples: $nsamples$;
Sample capacity: $ncapacity$;
output:
Sample means;
Average sample mean;

Initialize: generate $n \times d$ standard data set $D_m$
for each $i \in [1, nsamples]$ do
    for each $j \in [1, ncapacity]$ do
        $D_m{}' = D_m$
        Shuffling all elements in each column of $D_m{}'$
        Calculate $ORR$ of $D_m{}'$
    end for
    Calculate the sample mean
end for
Calculate the average sample mean
```

C. SEMI-RANDOM SAMPLING METHOD

The dependency among the attributes in real-world data set would affect the predicting accuracy, so we need to use the attribute dependency background knowledge to correct the predicted results of real-world data set. Considering that it is difficult to obtain dependencies among three or more attributes, the experiment only considered dependencies between two attributes.

The semi-random sampling method is described in detail below. If attribute A is strongly dependent on B, the tuples with confidence or frequency exceeding a certain threshold in strong dependency ordered attribute pair (A, B) is considered as frequent tuples. For example, if (A, B) is a strong dependency ordered attribute pair and the confidence threshold is 0.8, then the tuple (a, b) with $conf(b \Rightarrow a) = 0.9$ is a frequent tuple. All frequent tuples and their confidences in (A, B) are called dependency background knowledge about (A, B). The semi-random sampling method is similar to the random sampling method, except that the semi-random sampling method can maintain attribute dependencies of target data set to some extent. For example, D'_m is a data set generated by semi-random sampling, (a, b) is a frequent tuple with $conf(b \Rightarrow a) = 0.9$ in D_m, then the $\text{conf}(b \Rightarrow a)$ in D'_m is 0.9.

We considered the confidence of frequent tuple (a, b) in target data set is b_a, the algorithm of semi-random sampling method is shown as Algorithm 2.

Algorithm 2 Semi-Random Sampling Method

Input:
AVFM of target data set D: m;
Number of samples: *nsamples*;
Sample capacity: *ncapacity*;
Confidence of each frequent tuple in D;
output:
Sample means;
Average sample mean;

```
Initialize: generate $n \times d$ standard data set $D_m$
for each $i \in [1, nsamples]$ do
    for each $j \in [1, ncapacity]$ do
        $D_m{}' = D_m$
        Shuffling all elements in each column of $D_m{}'$
        for each strong dependency ordered attribute pairs $(A, B)$ do
            for each frequent tuple $(a, b)$ do
                Switch elements of column $A$ to meet $\text{conf}(b \Rightarrow a) = b\_a$
            end for
        end for
        Calculate $ORR$ of $D_m{}'$
    end for
    Calculate the sample mean
end for
Calculate the average sample mean
```

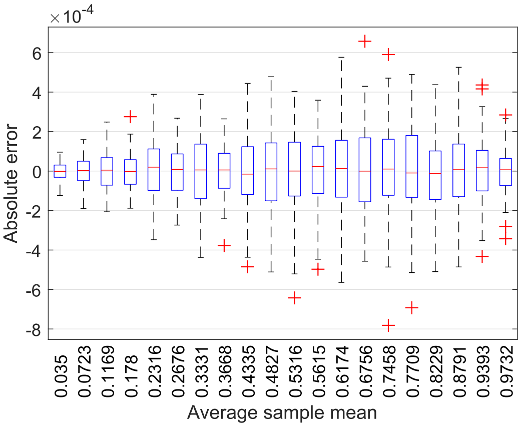

FIGURE 1. Absolute error of predicting ORR of random data set.

TABLE 3. Description of the 10 attributes considered in our study.

No.	Attribute	Type	Description
1	oshpd_id	Spatial	Hospital ID: 6-digit
2	age_yrs	Census	Age in years: from 0 to 85
3	sex	Census	Gender: 1-digit
4	ethncty	Census	Ethnicity: 1-digit
5	race	Census	Racial background: 1-digit
6	patzip	Census	ZIP code: 5-digit
7	patcnty	Census	County: 2-digit
8	los	Temporal	Length of stay
9	adm_qtr	Temporal	Admission quarter: 1-digit
10	charge	Confidential	Total charges for services

TABLE 4. Selected attributes of each target data set.

Data set No.	Selected attributes	Data set No.	Selected attributes
1	(2, 4, 5)	11	(3, 5, 6, 9)
2	(4, 7, 8)	12	(5, 6, 8)
3	(2, 4, 8)	13	(4, 6, 7)
4	(1, 8)	14	(5, 6, 7)
5	(2, 4, 5, 8)	15	(6, 7, 9)
6	(1, 7)	16	(1, 4, 5, 7, 9)
7	(1, 5, 8)	17	(1, 2, 3, 9)
8	(1, 8, 9)	18	(1, 2, 4, 8)
9	(2, 5, 7, 9)	19	(1, 2, 3, 5, 8)
10	(2, 3, 4, 7, 9)	20	(1, 2, 3, 4, 5, 7)

IV. SIMULATION RESULTS AND ANALYSIS

A. RANDOM SAMPLING METHOD

1) PREDICTING ORR OF RANDOM DATA SETS

We selected 20 representative AVFM with which the risks of the random data sets are approximately equally spaced distribution between 0 and 1. Then we randomly selected 100 samples from the ORR population $\mathbb{R}_m$ corresponding to each AVFM. The capacity of each sample was 50. We used the sample mean to predict the ORR of the target random data set. For simplicity, we made the target ORR equal to average sample mean. The absolute errors of predicting are shown in Figure 1. The x axis shows the average sample mean of each AVFM. The y axis shows the absolute error of predicting (i.e., the difference between the target ORR and the sample mean). As shown in Figure 1, the absolute error of each AVFM is close to zero. It means that the sample mean is centrally distributed around the average sample mean and occasionally some abnormal sample mean occurs, but the deviation between the outlier and the average sample mean is no more than 0.001. The results show that random data sets with the same AVFM have highly consistent ORR. That is, random data sets with the same statistics (i.e., number of individuals, distribution of attribute values) have similar ORRs. Our assumption was verified on the random data sets.

Through further research, we found that all the 100,000 random data sets do not contain strong dependency ordered attribute pairs. Compared with the total number of all random data sets with same AVFM, the number of random data sets used in simulation and the number of data sets containing strong dependency ordered attribute pairs are negligible. Due to the random distribution of the two in the population of the random data sets, the possibility of an intersection between the two is extremely tiny. Considering that there are usually strong dependency ordered attribute

TABLE 5. Attribute dependencies between any two non-sensitive attributes of SPD data set.

	oshpd_id	age_yrs	sex	ethncty	race	patzip	patnty	los	adm_qtr
oshpd_id	1	0.012	0.001	0.012	0.024	0.588	0.418	0.012	0.001
age_yrs	0.016	1	0.003	0.007	0.003	0.019	0.002	0.007	0
sex	0.007	0.017	1	0.001	0	0.008	0.001	0.005	0
ethncty	0.098	0.04	0.001	1	0.103	0.118	0.029	0.007	0
race	0.127	0.012	0	0.066	1	0.113	0.037	0.002	0
patzip	0.471	0.011	0.001	0.012	0.017	1	0.417	0.006	0.001
patnty	0.795	0.002	0	0.007	0.013	0.989	1	0.002	0
los	0.027	0.013	0.002	0.002	0.001	0.016	0.003	1	0
adm_qtr	0.003	0	0	0	0	0.003	0	0	1

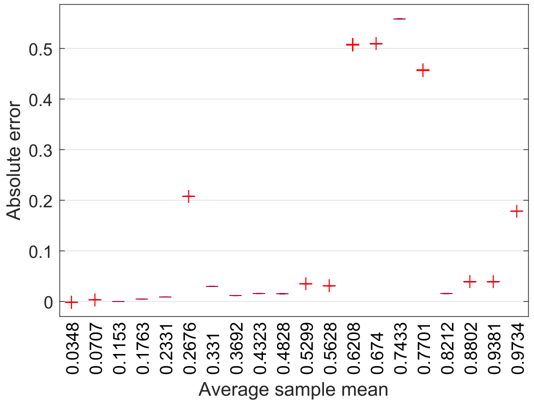

FIGURE 2. Absolute error of predicting ORR of real-world data set.

pairs in the real-world data sets, so we tested the predicting effect of the random sampling method on real data sets.

2) PREDICTING ORR OF REAL-WORLD DATA SETS

Since it was difficult to get a large number of real-world data sets, we generated 20 target data sets by selecting the intersecting positions of random rows and certain columns from a big real-world data set. The original real-world data set used in this study was the SPD data set with a capacity of 3985166 and 10 attributes [19]. Table 3 describes in detail the considered attributes, Table 4 shows the selected attributes of each target data set and Table 5 provides the attribute dependencies between any two non-sensitive attributes of SPD data set. As shown in Table 5, most ordered attribute pairs are weak dependencies, only three of them are strong dependencies. The three ordered attribute pairs are $(patnty, oshpd\_id)$, $(patnty, patzip)$ and $(oshpd\_id, patzip)$. It is understandable that in the real world, patient's hospital, county and ZIP code are highly correlated, and the dependencies among them are easily available from public information.

We used random sampling method to predict the ORRs of the real-world data sets, and the absolute errors of predicting are shown in the Figure 2. The x axis shows the average sample mean corresponding to the AVFM of each target data set. The y axis shows the absolute error of predicting (i.e., the difference between sample mean and target ORR). The predicting errors of groups 6, 13-16 and 20 were above 0.2, while the errors of other data sets were all below 0.05.

Through further research, we found that the attribute dependencies of the target data sets were close to the ones of the SPD data set. All target data sets with high predicting errors contained strong dependency ordered attribute pairs, while all data sets with low predicting errors did not contain strong dependency ordered attribute pairs. It shows that, the strong dependency ordered attribute pairs will heavily interfere the predicting accuracy of random sampling method.

B. SEMI-RANDOM SAMPLING METHOD

Compared with random sampling method, the background knowledge of attribute dependencies in target data set should be considered in semi-random sampling method. Considering that in reality the statistical characteristics of large population are easier to obtain than those of specific small population, we used knowledge of attribute dependencies in SPD data set to constrain the random data set. Due to the records of target data set is random sampling from SPD data set, the following two situations need to be considered: (1) The target data set do not contain some frequent tuples of SPD data set; (2) The confidence of frequent tuple of target data set is theoretically lower than the corresponding one of SPD data set. For situation one, we do not need to do anything. For situation two, we need to ensure that the confidence of frequent tuple of the random data set is equal the smallest one of the theoretical values of the target set and the background value of SPD data set. For example, if tuple (a, b) is a frequent tuple in SPD data set with $conf(b \Rightarrow a) = 0.9$, then in the random data set containing same frequent tuple, the confidence of (a, b) is equal to min(0.9, supp(a)/supp(b)), where the supp(a) and supp(b) are the supports of elements a and b in the target data set.

We considered the tuple of strong dependency ordered attribute pairs, with confidence greater than 0.9, or with confidence between 0.5 and 0.9, and frequency exceeding 398 were frequent tuples. The absolute error of semi-random sampling method is shown as Figure 3. With the background knowledge of attribute dependency, the ORR prediction accuracy was greatly improved, and the absolute predicting error was limited to 0.09. Obviously, the background knowledge we considered was incomplete, if more background knowledge was obtained, the absolute prediction error would be further reduced. The results show that data sets with same statistic (i.e., number of individuals, number of attributes, distribution of attribute values, attribute dependency) have

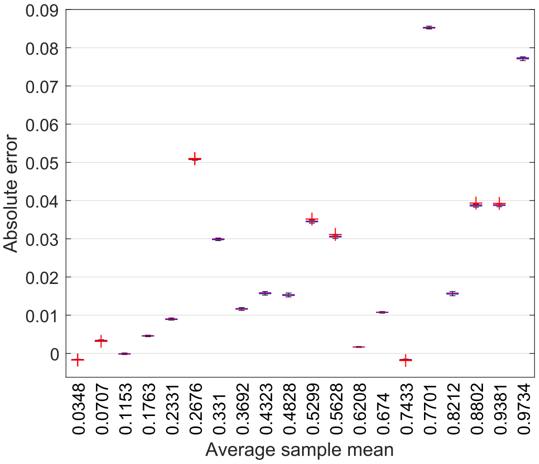

FIGURE 3. **Absolute error of predicting ORR of real-world data sets.**

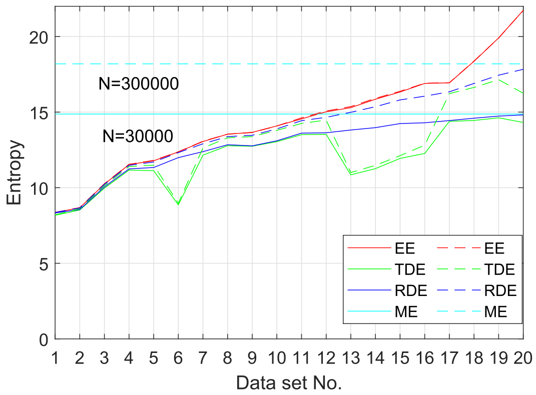

FIGURE 4. **Entropies of test data sets.**

highly consistent overall re-identification risk. It means that, for real-world data sets, our assumptions are also correct.

V. DISCUSSION

Here, we discuss the relationship between entropy and ORR. We obtained 40 testing data sets by randomly selecting 30000 and 300,000 records from the SPD data set according to the attribute combination shown in table 4. The Max entropy (ME), the experience entropy (EE), the entropy of random data set with the same AVFM (RDE), the entropy of target data set (TDE), of the forty data sets are shown as Figure 4. The solid line represents the data set capacity of 30000, and the dashed line represents the data set capacity of 30000. Based on information theory, statistics knowledge and experimental results, we have summarized the following four rules.

Rule 1: The ME is absolutely determined by the capacity of the data set, and the larger the capacity, the greater the maximum entropy. If the TDE is equal to the ME, which means that each tuple of the target data set is unique, then the ORR of the target data set is 1. EE must be greater than or equal to TDE and RDE.

Rule 2: When selecting the same combination of attributes, the EEs of the sampling data sets with 30000 records and 300,000 records are very close, because they are all from SPD data set, having the close proportion of attribute values of each attribute. If the capacity of data set changes, but the combination of attributes and the proportion of each attribute value remain, EE will be greater than ME when the capacity of the data set is small enough. This is because too small data set capacity will make the number of tuples in the data set far lower than the capacity of tuple space $\mathbb{T}$, resulting in the ME of data set will be lower than the EE calculated based on probability distribution. For example, we consider a data set with 1000 records and 4 attributes, and each attribute has 10 attribute values, the frequency of each attribute value is 100. Then ME of data set is 9.9658, which is lower than the EE 13.2877 of data set.

Rule 3: RDE is less than or equal to EE. When the capacity of tuple space $\mathbb{T}$ remain, RDE is close to EE if the data set capacity is large enough. The larger the RDE, the larger the ORR. For data sets with same AVFM, the larger the volume, the larger the RDE, but the lower the ORR.

Rule 4: TDE is less than or equal to RDE, because the attribute dependencies in the real-world data set will weaken the uncertainty of the data, and random data sets destroy the attribute dependencies and maximize the entropy of the data set. When there are no strong dependency ordered attribute pairs in the target data set, the TDE is very close to the RDE, and the ORR of the target data set is very close to the one of the random data set. When there are strong dependency ordered attribute pairs in the target data set, TDE will deviate from RDE greatly, and the ORR of the target data set will be much lower than that of the random data set. In general, for data sets with same capacity, the ORR of the data set with significantly larger TDE is greater than the one of the data set with smaller TDE.

In short, when the data set capacity is large enough, there is ME $\geqslant$ EE $\geqslant$ RDE $\geqslant$ TDE. The ORR of the data set is highly correlated with the TDE, and the dependencies among the attributes of the data set will make the TDE deviate from the RDE, and the ORR of the target data set will be much lower than that of the random data set. It means that two data sets with same AVFM, the ORR of the one with stronger attribute dependencies is lower than the other. And the attribute dependencies (e.g, the dependency of beer on diaper) are exactly the value of the data set. This suggests that data privacy and value are not always contradictory. Differential privacy technology preserves user privacy by adding random noise, but random noise will destroy attribute dependency and reduce data availability. If we can maintain attribute dependency while adding noise, the availability of the data will be preserved without privacy risk increasing.

For modern sparse data sets with thousands of attributes, i.e., each user includes for fewer non-null attributes, R3A is not suitable. However, because of recording too many user attributes, modern high-dimensional sparse data sets have high privacy risk. From the perspective of privacy protection, high-dimensional data sets should be divided into lowdimensional data sets for which R3A model has a good prediction effect. From the perspective of privacy attack prevention, attackers can only acquire the knowledge of a few attributes, and R3A model is capable of predicting the ORR of data set composed of these attributes.

VI. CONCLUSION

In this paper, we propose R3A model to rapidly predict the ORR of data set. Our model has high prediction accuracy, when considering the background knowledge of attribute dependency (i.e., the confidence of all frequent tuples). Fortunately, in the real-world, the background knowledge can be easily obtained through public data. That provides a wide space for the application of our model. For example, R3A model can be used to rapidly assess the privacy disclosure risk, providing references for government policy making and personal privacy estimation.

REFERENCES

[1] D. Wu, Q. Liu, H. Wang, Q. Yang, and R. Wang, "Cache less for more: Exploiting cooperative video caching and delivery in D2D communications," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1788–1798, Jul. 2019.

[2] Z. Li, Y. Jiang, Y. Gao, L. Sang, and D. Yang, "On buffer-constrained throughput of a wireless-powered communication system," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 2, pp. 283–297, Feb. 2019.

[3] C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Vallati, "Reducing power consumption with QoS constraints in IEEE 802.16e wireless networks," *IEEE Trans. Mobile Comput.*, vol. 9, no. 7, pp. 1008–1021, Jul. 2010.

[4] D. Wu, Z. Zhang, S. Wu, J. Yang, and R. Wang, "Biologically inspired resource allocation for network slices in 5G-enabled Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9266–9279, Dec. 2019.

[5] Z. Zhang and L. Wang, "Social tie-driven content priority scheme for D2D communications," *Inf. Sci.*, vol. 480, pp. 160–173, Apr. 2019.

[6] D. Wu, H. Shi, H. Wang, R. Wang, and H. Fang, "A feature-based learning system for Internet of Things applications," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1928–1937, Apr. 2019.

[7] L. Sweeney, "Weaving technology and policy together to maintain confidentiality," *J. Law, Med. Ethics*, vol. 25, nos. 2–3, pp. 98–110, 1997.

[8] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2008, pp. 111–125.

[9] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, May 2012.

[10] L. Rocher, J. M. Hendrickx, and Y.-A. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature Commun.*, vol. 10, no. 1, p. 3069, Jul. 2019.

[11] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "*L*-diversity: Privacy beyond *k*-anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, p. 3-es, Mar. 2007.

[12] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115.

[13] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds. Berlin, Heidelberg: Springer, 2006, pp. 1–12.

[14] J. Xiong, M. Zhao, M. Bhuiyan, L. Chen, and Y. Tian, "An AI-enabled three-party game framework for guaranteed data privacy in mobile edge crowdsensing of IoT," *IEEE Trans Ind. Informat.*, to be published.

[15] J. Xiong, X. Chen, Q. Yang, L. Chen., and Z. Yao, "A task-oriented user selection incentive mechanism in edge-aided mobile crowdsensing," *IEEE Trans. Netw. Sci. Eng.*, to be published.

[16] R. Wang, H. Liu, H. Wang, Q. Yang, and D. Wu, "Distributed security architecture based on blockchain for connected health: Architecture, challenges, and approaches," *IEEE Wireless Commun.*, vol. 26, no. 6, pp. 30–36, Dec. 2019.

[17] A. Alshammari and D. B. Rawat, "Intelligent multi-camera video surveillance system for smart city applications," in *Proc. IEEE 9th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2019, pp. 0317–0323.

[18] K. Kuhn, F. Prasser, and F. Kohlmayer, "The importance of context: Risk-based de-identification of biomedical data," *Methods Inf. Med.*, vol. 55, no. 04, pp. 347–355, Jan. 2018.

[19] Y.-A. de Montjoye and A. S. Pentland, "Response to comment on 'Unique in the shopping mall: On the reidentifiability of credit card metadata,"' *Science*, vol. 351, no. 6279, p. 1274, Mar. 2016.

[20] K. El Emam, A. Brown, and P. AbdelMalik, "Evaluating predictors of geographic area population size cut-offs to manage re-identification risk," *J. Amer. Med. Inform. Assoc.*, vol. 16, no. 2, pp. 256–266, Mar. 2009.

[21] L. Sweeney, "Uniqueness of simple demographics in the U.S. Population," Tech. Rep. LIDAP-WP4, 2000.

[22] P. Golle, "Revisiting the uniqueness of simple demographics in the US population," in *Proc. 5th ACM Workshop Privacy Electron. Soc. (WPES)*. New York, NY, USA: ACM, 2006, pp. 77–80.

[23] M. M. Merener, "Theoretical results on de-anonymization via linkage attacks," *Trans. Data Privacy*, vol. 5, no. 2, pp. 377–402, Aug. 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2423651.2423652

[24] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, vol. 3, no. 1, p. 1376, Mar. 2013.

[25] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, Jan. 2015.

[26] Z. Tu, F. Xu, Y. Li, P. Zhang, and D. Jin, "A new privacy breach: User trajectory recovery from aggregated mobility data," *IEEE/ACM Trans. Netw.*, vol. 26, no. 3, pp. 1446–1459, Jun. 2018.

ZHIGANG YANG (Member, IEEE) received the M.S. degree from the Chongqing University of Posts and Telecommunications, in 2006, where he is currently pursuing the Ph.D. degree. He is currently an Associate Professor with the Chongqing University of Arts and Sciences. His research interests include edge computing, network security, and privacy.

RUYAN WANG (Member, IEEE) received the Ph.D. degree from the University of Electronic and Science Technology of China, in 2007. He is currently the Dean of the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications. His research interests include network performance analysis and multimedia information processing. He was a recipient of the Danian Huang Team from the Ministry of Education of the People's Republic of China.

DAIZHONG LUO received the M.S. degree from Chongqing University, in 2005. He is currently a Professor with the Chongqing University of Arts and Sciences. His research interests include software reuse, software architecture, and software product line.

YU XIONG received the B.S. degree from Southwest University, in 2003. He is currently a Lecturer at the Chongqing University of Arts and Sciences. His research interests include wireless networks and network security.