Privacy-Preserving Publishing Data with Full Functional Dependencies

Hui (Wendy) Wang, Ruilin Liu

Stevens Institute of Technology
Hoboken, NJ, USA
{hwang,rliu3@cs.stevens.edu}

Abstract. We study the privacy threat by publishing data that contains full functional dependencies (FFDs). We show that the cross-attribute correlations by FFDs can bring potential vulnerability to privacy. Unfortunately, none of the existing anonymization principles can effectively prevent against the FFD-based privacy attack. In this paper, we formalize the FFD-based privacy attack, define the privacy model (d, ℓ)-*inference* to combat the FFD-based attack, and design robust anonymization algorithm that achieves (d, ℓ)-*inference*. The efficiency and effectiveness of our approach are demonstrated by the empirical study.

1 Introduction

How to publish data that contains sensitive information of individuals has received considerable attention in recent years. It has been shown that simply removing explicit identifiers (IDs), e.g., name and SSN, from the released data is insufficient to protect privacy [9]. The existence of a set of non-ID attributes (called *quasi-identifiers($\mathcal{QI}$)*), e.g., the combination of zipcode, gender and date of birth, that can uniquely identify individuals, can be joined with information obtained from diverse external sources (e.g., public voting registration data) to re-identify the individuals in the released data. This is called the *record linkage attack*.

ID	QI		Sensitive	
Name	Sex	Zip	Phone	Disease
Alice	F	07921	1111111	Ovarian cancer
Bob	M	07920	2222222	Bronchitis
Calvin	M	07902	3333333	Diabetes
Doris	F	07901	1000001	Ovarian cancer
Eve	F	07902	3333333	Bronchitis
Flora	F	07903	2000001	Pneumonia

(a) The Original Microdata

QI		Sensitive	
Sex	Zip	Phone	Disease
*	079**	1111111	Ovarian cancer
*	079**	2222222	Bronchitis
*	079**	3333333	Diabetes
F	0790*	1000001	Ovarian cancer
F	0790*	3333333	Bronchitis
F	0790*	2000001	Pneumonia

(b) The 3-diversity table

QI		Sensitive	
Sex	Zip	Phone	Disease
*	079**	1111111	Ovarian cancer
*	079**	2222222	Bronchitis
*	0790*	3333333	Diabetes
F	0790*	1000001	Ovarian cancer
F	0790*	3333333	Bronchitis
F	0790*	2000001	Pneumonia

(c) The table after FD inference

Table 1. Anonymized microdata before&after FD inference

Various privacy principles have been proposed recently to defend against the record linkage attack (e.g., k-anonymity [7, 9] and ℓ-diversity [4]). However, by applying FDs on the released data that has met some aforementioned privacy principles, the attacker may be able to breach privacy. For example, assume the microdata in Table 1 (a) contains the functional dependency $F : Phone \rightarrow Zip$, which states that any two same phone numbers must correspond to the same zipcode. Assume that the attacker possesses the knowledge of F. Then by applying F on the 3-diversity table (i.e., every group that has the same QI-values contains at least three unique sensitive values) in Table 1 (b) , he/she can modify the zipcode value of the third tuple from "079**" to "0790*", since the second group contains the phone number 3333333 with zipcode "0790*". The anonymized table after the FD-based inference is shown in Table 1 (c). The third tuple only satisfies 1-diversity privacy guarantee.

The example shows that using FDs as adversary knowledge may bring privacy breach. Given the fact that it is not difficult for the attacker to obtain these functional dependencies from either the common sense or other sources, it is necessary to develop robust privacy criterion of publishing data when functional dependencies are available and are used as part of the adversary knowledge.

In this paper, we focus on full functional dependencies (FFDs). We have the following contributions.

- Formally define the FFD-based attack.
- Define the (d, ℓ)-*inference* model to defend against the FFD-based attack.
- Propose novel *grouping stradegy* to archive (d, ℓ)-*inference* .
- Design an efficient anonymization algorithm to produce anonymized microdata.
- Demostrate the efficacy of our algorithm by an extensive set of experiments.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 introduces the preliminaries. Section 4 defines our privacy model. Section 5 proposes the intersection-grouping strategy. Section 6 presents the details of our anonymization algorithm. Section 7 presents the experimental results. We conclude the paper in Section 8.

2 Related Work

Privacy-preserving data publishing has received considerable attention in recent years. The k-anonymity model requires that in the published data, every individual is related with no less than *k* tuples [7,9]. The ℓ-diversity [4] model further requires that every QI-group contains at least ℓ sensitive values that have roughly the same frequency. Other variants of k-anonymity and ℓ-diversity, e.g., t-closeness [3] and (α, k)-anonymity [11], (c, k)-safety [5], are defined to address different privacy requirements. Unfortunately, none of them can defend against our defined FFD-based privacy attack.

Martin et al. [5] and Rastogi et al. [6] are the first to consider the adversary who knows arbitrary correlations between tuples. They show that if such correlations are available, then there exists privacy leakage on the published dataset that is of "meaningful" utility. Both of them focus on tuple correlations but do not consider FDs. Kifer [2] shows that the attacker may induce correlations from the sanitized dataset; such inferred correlations can enable potential vulnerabilities on the sanitized dataset. However, [2] does not provide any solution to defend against the FD-based attack. Tao et al. [10] study the correlation hiding problem of data publishing. They use i-masking operation to ensure the attributes of correlation which needs to hide are independent. Their work presents a different goal from ours.

3 Preliminaries

Functional Dependency. Given two attributes X and Y, a database instance D satisfies the functional dependency FD $F : X \rightarrow Y$ if for every two tuples $t_1, t_2 \in D$, if $t_1.X = t_2.X$, then $t_1.Y = t_2.Y$. We call X the *determinant* attributes and Y the *dependent* attributes, and their values the *determinant* values and *dependent* values. In this paper, *we only consider FFDs*, i.e., the FDs $X \rightarrow Y$ that hold for all values of X and Y.

Anonymization Framework. There are three types of attributes in the microdata: identifiers (ID), quasi-identifiers $\mathcal{QI}$, whose combination can play as the key and uniquely identify any individual, and sensitive attributes $\mathcal{S}$. There may exist multiple sensitive

attributes. We can consider them as a super attribute. For simplicity, we consider single sensitive attribute in the remaining of the paper.

In this paper, we consider *generalization* [8, 9], a popular anonymization technique. By generalization, numerical QI-values are recoded as an interval (e.g., age 20 is recorded as $[20, 40]$), while the categorical QI-values are replaced with higher level domain values in the taxonomy tree (e.g., city "Hoboken" is replaced with "New Jersey"). In this paper, we only consider numerical QI-values. The purpose of generalization is to hide each individual tuple into a group, which is called the *QI-group* where all tuples inside have the same QI-values after generalization.

It is important to measure the incurred information loss by generalization. In this paper, we consider the generalized loss metric [1], which measures the information loss as a ratio. The definition of the metric is given below.

Definition 1 (Information Loss). *For a data value v, if it is suppressed from the released dataset, its information loss equals $IL_v = 1$. Otherwise if it is generalized to an interval $[L_i, U_i]$, let Min and Max be the minimum and maximum values of the attribute A. The information loss of v is: $IL_v = (U_i - L_i)/(Max - Min)$.*
The information loss IL_t of a tuple t is defined as $IL_t = (\sum_{v_i \in t} IL_{v_i})/n$, where n is the number of non-ID attributes. The information loss of the microdata D is defined as $IL_D = (\sum_{t \in D} IL_t)/|D|$. □

4 Privacy Model

In this section, first, we define the FFD-based attack. Then we formally define the (d, ℓ)-*inference* model that combats the FFD-based attack.

4.1 FFD-based Attack

To analyze the impact of FFDs to privacy, we consider a popular privacy model, ℓ-diversity [4]. It requires that each QI-group must consist of at least ℓ "well-represented" distinct values that are of close frequency. We define d-closeness to address the requirement of close frequency. The definition of d-closeness is a simplified version of "well-represented" in [3].

Definition 2 (d-closeness). *Given two sensitive values s_1 and s_2, let f_1 and f_2 be their frequency, then s_1 and s_2 are considered as d-*close *if $|f_1 - f_2| \leq d$. Given a QI-group G that consists of a set of distinct sensitive values, G is d-close if $\forall$ sensitive values $s_i, s_j \in G$, s_i and s_j are d-close.* □

Based on the definition of d-closeness, we give a simplified version of ℓ-diversity.

Definition 3 (ℓ-diversity). *Given a microdata D, let D^* be its anonymized version. Then D^* is ℓ-*diverse *if $\forall$ sensitive attribute S of D, each QI-group $G \in D^*$ consists of at least ℓ distinct sensitive values on S that are d-close.* □

Next, we explain the details of FFD-based attack. In the anonymized dataset, it is possible that different QI-groups share the same sensitive or QI- values. This enables the possibility of the FFD-based attack. Intuitively, assume both QI-groups G_1 and G_2 include the value a, where a is a determinant value of the FD $F : \mathcal{A} \rightarrow \mathcal{B}$. Let b be the corresponding dependent value of a in the original microdata. Even though b can be generalized to different values b_1^* and b_2^* in G_1 and G_2, due to the presence of FFDs,

the attacker still can infer that b_1^* and b_2^* must correspond to the same original value. Thus he/she can "intersect" b_1^* and b_2^*. It is such intersection that enables the FFD-based attack. In the next, first, we formally define $b_1^* \cap b_2^*$, the *intersection* operation on generalized values. To distinguish from the conventional intersection operation $\cap$, we use $\cap^*$ to denote the intersection operation on generalized data values.

Definition 4 (Intersection of Generalized values). *Given two generalized values b_1^* and b_2^*, which are two intervals [l_1, u_1] and [l_2, u_2], if these two intervals overlap, then $b_1^* \cap^* b_2^* = [max(l_1, l_2), min(u_1, u_2)]$, otherwise $b_1^* \cap^* b_2^* = NULL$.* □

For example, given two generalized *Age* values $b_1^* = [20, 40]$ and $b_2^* = [30, 50]$, $b_1^* \cap^* b_2^* = [30, 40]$.

As aforementioned, due to FFDs, the attacker can conclude that generalized dependent values b_1^* and b_2^* in the tuples in G_1 and G_2 that contain the same determinant values indeed correspond to the same original value. Then he/she can replace both b_1^* and b_2^* in these tuples with $b_1^* \cap^* b_2^*$. Such replacement separates the tuples in QI-group G_1 (G_2, resp.) into two sets, the one of the values b_1^* (b_2^*, resp.), and the one of the values $b_1^* \cap^* b_2^*$. We formally define these two sets below. To distinguish from the conventional set intersection/difference($\cap/-$) operation, we use $G_1 \cap_F G_2$ and $G_1 -_F G_2$ to denote the set intersection/difference operations of G_1 and G_2 based on the reasoning of FFD F. We use $t[\mathcal{A}]$ to denote the values of attributes $\mathcal{A}$ of the tuple t.

Definition 5 (FFD-based Intersection/Difference of QI-groups). *Given the FFD F : $\mathcal{A} \rightarrow \mathcal{B}$ of the microdata D and two QI-groups G_1, G_2, let $G_{12} = \{t | t \in G_1, \exists t' \in G_2$ s.t. $t[\mathcal{A}]$=$t'[\mathcal{A}]\}$. Then $G_1 \cap_F G_2$ is a set of tuples J s.t. $\forall t \in G_{12}$, $\exists t' \in J$ s.t.*
(1) $\forall$ attribute $A \in \mathcal{A}$, $t'[A] = t[A]$,
(2) $\forall$ attribute $B \in \mathcal{B}$,

$$t'[B] = \begin{cases} t[B] & \text{if } t[B] \text{ is not generalized} \\ G_1[B] \cap^* G_2[B] & \text{if } t[B] \text{ is generalized} \end{cases}$$

Furthermore, $G_1 -_F G_2 = G_1 - G_{12}$. □

For example, for the two QI-groups G_1 and G_2 in Table 1 (a) with FFD F : $Phone \rightarrow Zip$, $G_1 \cap_F G_2$ ={*, 0790*, 3333333, Diabetes}, and $G_1 -_F G_2$ returns the first two tuples in G_1.

Now we are ready to define *FFD-based privacy attack*.

Definition 6 (FFD-based Privacy Attack). *Given a microdata D, let D^* be its generalized version that satisfies ℓ-diversity. Then the full functional dependency $F : A \rightarrow B$ ($A, B \subseteq \mathcal{QI} \cup \mathcal{S}$) enables the* FFD-based privacy attack *if there exist two QI-groups $G_1, G_2 \in D^*$ such that at least one of followings is non-empty and does not satisfy ℓ-diversity: (1) $G_1 \cap_F G_2$, (2) $G_2 \cap_F G_1$, (3) $G_1 -_F G_2$, and (4) $G_2 -_F G_1$. Otherwise, we say D^* is* safe at the presence of F. □

We have shown an example of FFD-based privacy attack in Section 1.

Although FFDs may threat privacy, not all FFDs can enable the FFD-based attack. Based on this, we define *safe* and *unsafe* FFDs.

Definition 7 (Safe/Unsafe FFDs). *A functional dependency F is* safe *if for any microdata D that satisfies F, all of its possible generalized versions D^* are safe at the presence of F. Otherwise, we say F is* unsafe. □

Based on the definition, we distinguish the "safe" FFDs from the "unsafe" ones. We have:

Theorem 1 ((Un)safe FFDs). *Given the microdata D that contains the QI attributes $\mathcal{QI}$ and sensitive attributes $\mathcal{S}$, let $F : \mathcal{A} \rightarrow \mathcal{B}$ ($\mathcal{A}, \mathcal{B} \subseteq \mathcal{QI} \cup \mathcal{S}$) be one of its FFDs. then F is* safe *iff $\mathcal{A} \subseteq \mathcal{QI}$. Otherwise, F is* unsafe. □

Due to space limit, we skip the proof. Next, we define the (d, ℓ)-*inference* model to defend against the privacy attack by unsafe FDs.

Definition 8 ((d, ℓ)-inference). *Given microdata D, let D^* be its anonymized version that consists of the QI-groups $\mathcal{G}\{G_1, \ldots, G_n\}$. Let S_i be the set of distinct sensitive values of the QI-group G_i ($1 \leq i \leq n$). Then D^* satisfies* (d, ℓ)-inference *if*
*(1) d-**close**: $\forall G \in \mathcal{G}$, all sensitive value sets in G are d-close,*
*(2) ℓ-**overlapping**: $\forall G_i, \ldots, G_j \in \mathcal{G}$, if $|S_i \cap \cdots \cap S_j| \neq 0$, then $|S_i \cap \cdots \cap S_j| \geq \ell$, i.e., there are at least ℓ shared distinct values in $S_i \cap \cdots \cap S_j$,*
*(3) ℓ-**non-overlapping**: $\forall G_i, G_j \in \mathcal{G}$, $|S_i - S_j| \geq \ell$, i.e., there are at least ℓ non-overlapping distinct values in $S_i - S_j$.* □

Both ℓ-overlapping and ℓ-non-overlapping conditions consider the worst case, i.e., the "smallest" results of intersection and set difference. Thus ℓ-overlapping considers interactions of multiple QI-groups (maybe more than 2), while ℓ-non-overlapping considers the difference of two QI-groups. Note that the (d, ℓ)-inference model is not based on the reasoning of the generalized datasets anymore. Instead, it only reasons on the sensitive values and can be applied on the original microdata directly.

5 Intersection-grouping (IG)

The key to achieve (d, ℓ)-*inference* is to appropriately group the sensitive values so that all these groups meet the three conditions of (d, ℓ)-*inference*. To address this, we propose the *intersection-grouping* strategy. It puts the sensitive values into groups that intersect (but not contain) in a chain. To avoid considering intersection of arbitrary number of groups, we do not allow the intersection of more than two groups. Furthermore, we require that all overlapped groups construct a chain, i.e., given a set of groups $G_1, \ldots, G_m$, G_i only intersects with G_{i+1} and G_{i-1}, but not the others.

The intersection-grouping approach consists of two steps: (1) *bucket construction* to construct d-close, ℓ-diverse, disjoint buckets and (2)*intersected groups construction* to construct intersected groups from the buckets. Due to the space limit, we omit the details. Instead, we use an example to show how our grouping stradegy works.

Example 1. Given the sensitive values $\{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9\}$ of frequency (1, 7, 9, 10, 26, 30, 40, 45, 50), assume $d = 3$ and $\ell = 2$. By Step 1, we construct the buckets $B_1\{s_2, s_3, s_4\}$ of frequency (7, 9, 10), $B_2\{s_5, s_6\}$ of frequency (26, 29), and $B_3\{s_7, s_8, s_9\}$ of frequency (40, 43, 43). There are 1 tuple containing s_1, 1 tuple containing s_6, 2 tuples containing s_8, and 7 tuples containing s_9 that will be removed. In Step 2, we construct the following groups: G_1 (s_2, s_3, s_4, s_5, s_6) of frequency (7, 9, 10, 10, 10), G_2: (s_5, s_6, s_7, s_8, s_9) of frequency (16, 19, 19, 19, 19), and G_3: (s_7, s_8, s_9) of frequency (21, 24, 24). There are 1+ 1 + 2 + 7 = 11 tuples in total that are removed. □

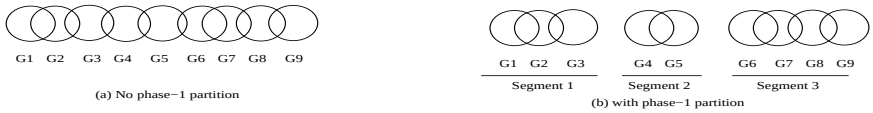


Fig. 1. Without & with Phase-1 Partition

6 Anonymization Algorithm

In this section, we explain the details of the algorithm that constructs the QI-groups for anonymization. The purpose of the QI-group is two-fold: (1) achieve (d, ℓ)-*inference* privacy guarantee, and (2) minimize information loss, including both by tuple suppression and by generalization. Our anonymization algorithm has two phases, phase-1 that minimizes the information loss by tuple suppression, and phase-2 that minimizes the information loss by tuple generalization. In particular, *phase-1* partitions the tuples by their sensitive values, so that each group satisfies (d, ℓ)-*inference*, while the phase-2 constructs QI-groups of minimized information loss by generalization from the constructed phase-1 partitions.

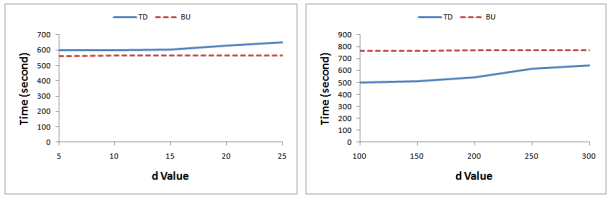
Phase-1 partition. We split the sensitive values into smaller disjoint segments, and apply IG on these segments. Figure 1 illustrates the effect of the partition. We prove that finding an optimal parition is a NP-hard problem, and propose two heuristic partitioning approaches, namely *top-down* and *bottom-up*. Both heuristics are designed in a greedy fashion. Due to the limit space, we omit the details.

Phase-2 QI-Group Construction. To further reduce the information loss, we split each partition into smaller groups of same sensitive values. Since the partition P satisfies (d, ℓ)-inference, it is straightforward all QI-groups from P must satisfy (d, ℓ)-inference.

7 Experiments

We have done an extensive set of experiments to evaluate both the effectiveness and efficiency of our anonymization algorithm. Due to space limit, in this section, we briefly describe our experiment design and results.

We use a workstation machine of 2.4GHz Intel core and 3GB of RAM. We implement the algorithms in C++. We use both synthetic datasets in the experiments. To measure the impact of frequency distribution of sensitive values to anonymization, we generated two types of synthetic datasets, the one of sensitive values of close-to-uniform, and the one of sensitive values of skewed distribution. We call these two datasets the U-*dis* dataset and S-*dis* dataset. We designed the functional dependency *salary-class* $\rightarrow$ *work-class* for the synthetic dataset.



(a) U-dis set. $\ell = 5$ (b) S-dis set. $\ell = 5$

Fig. 2. Time Performance Comparison (TD: Top-down, BU: Bottom-up)

Time Performance. First, we focus on the top-down and bottom-up approaches and measure the impact of the d value to the performance of anonymization. Figure 2 (a)

& (b) show that for both U-dis and S-dis datasets, the performance of the bottom-up approach is relatively stable with changing d value. However, the time performance of top-down gets worse with larger d, as larger d results in more partitions. We also measure the performance of the phase-1 partition of both top-down and bottom-up approaches. The experiment result shows that it is around 0.016 second for both U-dis and S-dis datasets. Thus it is negligible compared with the total time of anonymization (shown in Figure 2 (a) & (b)). Due to space limit, we omit the result.

Information Loss. We use the metric in Section 3 to measure the information loss. Intuitively, the smaller the ratio is, the better is the data utility. From our experiment results, we observe that our information loss is at most 0.31. This proves the effectiveness of our approach.

8 Conclusion

In this paper, we studied the problem of privacy-preserving publishing of data that contains full functional dependencies. We formally defined the privacy model, (d, ℓ)-*inference*, and developed robust and efficient algorithms that anonymize the data with minimized information loss. Our empirical studies using both synthetic and real datasets demonstrated the efficiency and effectiveness of our algorithm.

For the future work, we will consider multiple FFDs. Furthermore, we will move to CFDs, i.e., the FDs $X \rightarrow Y$ that do not hold for all values of X and Y. It turns out that the CFD-based analysis is far more intriguing than FFDs, and the (d, ℓ)-*inference* model fails to effectively defend against its consequent privacy attack. Thus in the future, we plan to strengthen the (d, ℓ)-*inference* model to defend against the CFD-based attack.

References

1. V.S. Iyengar, “Transforming Data to Satisfy Privacy Constraints”, SIGKDD, 2002.
2. D. Kifer, “Attacks on Privacy and deFinetti’s Theorem”, SIGMOD 2009.
3. N. Li, T. Li, “t-Closeness: Privacy Beyond K-anonymity and l-diversity”, ICDE 2007. Datasets”, SIGMOD 2006.
4. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. “l-Diversity: Privacy Beyond k-Anonymity”, ICDE 2006.
5. D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke and J. Y. Halpern, “Worst-Case Background Knowledge for Privacy-Preserving Data Publishing”, ICDE 2007.
6. V. Rastogi, D. Suciu, and S. Hong, “The boundary between privacy and utility in data publishing”, VLDB 2007.
7. P. Samarati and L. Sweeney, “Generalizing Data to Provide Anonymity when Disclosing Information”, PODS, 1998.
8. P. Samarati. “Protecting Respondents’ Identities in Microdata Release”, TKDE, 2001.
9. L. Sweeney, “K-anonymity: A Model for Protecting Privacy”, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557570, 2002.
10. Y. Tao, J. Pei, J. Li, X. Xiao, K. Yi, and Z. Xing “Hiding Correlation by Independence Masking”, ICDE 2010.
11. R. C. Wong, J. Li, A. W. Fu, and K. Wang, “(α, k)-Anonymity: An Enhanced k-Anonymity Model for Privacy-Preserving Data Publishing”, SIGKDD, 2006.