FF-Anonymity: When Quasi-Identifiers Are Missing

Ke Wang [#1], Yabo Xu [#1], Ada W.C. Fu [*2], Raymond C.W. Wong [*2]

[#]*Simon Fraser University*
[1]{*wangk,yxu*}*@cs.sfu.ca*
[*]*The Chinese University of Hong Kong*
[2]{adafu,cwwong}@cse.cuhk.edu.hk

***Abstract*— Existing approaches on privacy-preserving data publishing rely on the assumption that data can be divided into quasi-identifier attributes (QI) and sensitive attribute (SA). This assumption does not hold when an attribute has both sensitive values and identifying values, which is typically the case. In this paper, we study how such attributes would impact the privacy model and data anonymization. We identify a new form of attacks, called "freeform attacks", that occur on such data without explicit QI attributes and SA attributes. We present a framework for modeling identifying/sensitive information *at the value level*, define a problem to eliminate freeform attacks, and outline an efficient solution.**

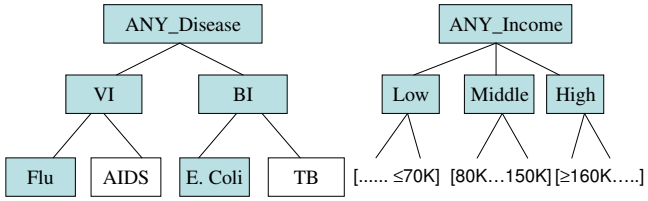
I. Introduction

Privacy-preserving data publishing (PPDP) focuses on publishing person-specific data (also called microdata) for the benefit of research. Previous works have considered microdata of the form $T(A_1, ..., A_m, SA)$. Each record corresponds to an individual. SA is the *sensitive attribute*. QI=$\{A_1, \cdots, A_m\}$, called the *quasi-identifier*, is a set of attributes that can be linked with an external source such as a voter list. In a linking attack, the adversary knows that some individual has a record in T, and observes the information qi on QI about the individual from an external source. The adversary's goal is to find the SA value of the individual. If the records in T that match qi are predominantly associated with a common SA value, the adversary could infer the individual's SA value with a high probability.

A. Abandoning the QI/SA paradigm

A fundamental assumption in previous works is that T can be split (vertically) into QI and the sensitive attribute SA. In this paper, the term *QI/SA paradigm* refers to this assumption and the works based on it. k-anonymity [1], l-diversity [2] and recent works all fall into this QI/SA paradigm. A common approach in the QI/SA paradigm is hiding the association between QI and SA. To this end, the generalization approach [1] partitions records into equivalence classes according to QI, and the bucketization approach [3] split records into sub-records according to the partition of QI and SA. Our insight is that determining QI and SA can be tricky and even undesirable. To explain this point, let us consider an example.

Example 1.1: Consider the table T1 on (Sex, Income, Disease) in Figure 1(a). Income and Disease have a hierarchical domain given by the taxonomies on the top of the figure. In our discussion, we use VI for "Viral Infectious", BI for "Bacterial Infectious", $E.Coli$ for "Escherichia coli", and TB for "Tubercle bacillus". Suppose that the following holds:



Sex	Income	Disease
F	190K	Flu
M	190K	AIDS
F	160K	AIDS
M	160K	AIDS
M	100K	E. Coli
M	80K	E. Coli
F	80K	TB
F	100K	TB

(a) T1: Raw table

Sex	Income	Disease
F	*	Flu
M	*	*
F	*	*
M	*	*
M	*	E. Coli
M	*	E. Coli
F	*	*
F	*	*

(b) T2: After suppressing sensitive values

Sex	Income	Disease
F	190K	VI
M	190K	VI
F	160K	VI
M	160K	VI
M	100K	E. Coli
M	80K	E. Coli
F	80K	TB
F	100K	TB

(c) T3: After generalizing Flu and AIDS

Fig. 1. A motivating example

A1 *Disease: Flu and $E.Coli$ are non-sensitive and can be observed on an individual; $AIDS$ and TB are sensitive and cannot be observed.*

A2 *Income: an exact income is sensitive and cannot be observed, but an income bracket such as $High$, $Middle$, and Low is non-sensitive and can be observed.*

A3 *Sex: F and M are non-sensitive and can be observed.*

We also assume that, if a value is non-sensitive, so are more generalized values. In a taxonomy, a non-sensitive value is denoted by a shaded node. The terms "non-sensitive", "observable" and "public" are interchangeable.

Within the QI/SA paradigm, since Income and Disease contain sensitive values, they can only be modeled as the sensitive attribute SA. The following are two possible ways of modeling:

- *Case 1: QI=$\{Sex\}$ and SA=Disease, and*
- *Case 2: QI=$\{Sex\}$ and SA=Income.*

Suppose that we consider 2-diversity [2] as our privacy goal. It is easy to see that T1 satisfies 2-diversity in both Case 1 and Case 2: no single SA value occurs in more than 1/2 percent

of the records sharing a common value on QI. Therefore, the probability of linking an individual to SA via QI is no more than 50%.

However, publishing T1 is not *safe under our assumptions A1, A2 and A3. Consider an adversary who tries to find the individual's sensitive value. First, with M and $High$ being observable (i.e., A2 and A3), the adversary can link an individual P_1 via M and $High$ to the two records $(M, 190K, AIDS)$ and $(M, 160K, AIDS)$ and infer $AIDS$ with 100% certainty. Note that this linking does not require observing the exact income of the individual. Let us denote this attack by*

$$\{M, High\} \rightarrow AIDS.$$

Second, with F and Flu being observable (i.e., A1 and A3), the adversary can link an individual P_2 to the first record via F and Flu, thus, to the exact income 190K, with 100% certainty. Let us denote this attack by

$$\{F, Flu\} \rightarrow 190K.$$

These attacks exist even though T1 is 2-diverse in both Case 1 and Case 2. Note that in these attacks, Disease acts as both an identifying attribute (in $\{F, Flu\} \rightarrow 190K$) and SA (in $\{M, High\} \rightarrow AIDS$). To prevent such attacks, new solutions are required.

One solution is suppressing the sensitive $AIDS$, TB and all exact incomes, as in T2 in Figure 1(b). This solution loses too much information since sensitive values usually are important for data analysis. A preferred solution is generalizing Flu and $AIDS$ to VI, as in Figure 1(c), and the result table T3 is safe for release. For example, $\{F, VI\} \rightarrow 190K$ has only 50% certainty and $\{M, High\} \rightarrow VI$ is no longer a threat because VI is non-sensitive. ■

This example illustrates several interesting points.

First, information on sensitive attributes may be observed on an individual. The QI/SA paradigm assumes that no information on SA can be observed on an individual. In the above example, Flu for Disease and $High$ for Income can be observed. Both Disease and Income are considered sensitive attributes. This example shows that, even for such attributes, a less sensitive value or a higher level value can be easily observed.

Second, "public" and "sensitive" information often is distinguished at the value level. The QI/SA paradigm assumes that for each attribute, either all values are public or all values are sensitive. In the above example, Income and Disease have *both* types of values. The QI/SA paradigm is handicapped in dealing with such attributes: treating them as SA would *under-protect* the data because observable values (such as Flu and $High$)) are mistakenly treated as non-observable, whereas treating them as QI attributes would *over-protect* because some sensitive values (such as 190K and $AIDS$) may not be easily observed.

Third, a linking attack may be through observing a different set of attributes for a different individual. The attack $\{M, High\} \rightarrow AIDS$ on the individual P_1 is by observing values on Sex and Income, whereas the attack $\{F, Flu\} \rightarrow$ 190K on the individual P_2 is by observing values on Sex and Disease. The QI/SA paradigm essentially assumes that all individuals have the same set of observable attributes and the same sensitive attribute. This is not reasonable because the adversary will not confine herself to any pre-determined QI.

B. Challenges and contributions

We consider a table $T(A_1, ..., A_m)$ where each attribute A_i has *both* observable values *and* sensitive values. To our knowledge, this is the first work that addresses the linking attacks under this setting. The following summarizes the challenges and our contributions.

Challenge/Contribution 1: The first challenge is modeling observability and sensitivity "at the value level". Requiring the publisher to specify this information for all domain and generalized values is neither feasible nor scalable. We present a framework for these specifications, with an focus on general principles so that new instantiations can be adapted.

Challenge/Contribution 2: We identify a class of *freeform attacks* of the form $X \rightarrow a$, where a and the values in X can be at *any* level of *any* attributes. $X \rightarrow a$ is a privacy breach if X is observable, a is sensitive, and X is associated with a. As shown in Example 1.1, a unique challenge posed by freeform attacks is that they may occur at a general level without occurring at a special level. We propose the notion of *FF*-anonymity (*F*ree*F*orm-anonymity) to eliminate freeform attacks.

Challenge/Contribution 3: We show that finding an optimal FF-anonymization is NP-hard and present an efficient solution for finding a "minimally generalized", not necessarily optimal, FF-anonymization.

II. Modeling Sensitivity and Observability

A. Terminology

Consider a set of attributes $U = \{A_1, ..., A_m\}$. Each A_j has a taxonomy (tree) arranged from general values at high levels to specific values at low levels. "Nodes" and "values" are interchangeable. ANY_j denotes the root of the taxonomy for A_j, $Subtr(a)$ denotes the subtree under a node a, and $Leaf(a)$ denotes the set of all leaf nodes in $Subtr(a)$. For two values a and a', $a' \succeq a$ means that either $a' = a$ or a' is an ancestor of a.

A *base table* over U, denoted T, contains only leaf values. A *generalized table* over U, denoted T^*, is produced by applying zero or more generalization to T. Each *generalization* replaces *all* child values $v_1, \cdots, v_p$ with their parent v in *all* records containing $v_1, \cdots, v_p$. T^* can also be produced by applying a sequence of specializations to $T^\top$, where $T^\top$ denotes the most generalized table containing ANY_i for every attribute A_i. Each *specialization* replaces a parent value v in every record containing v with the child value that generalizes the original value in the record. For two generalized tables T_2^* and T_1^*, we say that T_2^* is more general than T_1^* if T_2^* is obtained from T_1^* by zero or more generalization.

Let $Cut(A_i, T^*)$ denote the set of values in T^* on A_i, and $Cut^+(A_i, T^*)$ denote the set of values in $Cut(A_i, T^*)$ plus all ancestors. For a set of attributes $Attset$, $Cut^+(Attset, T^*)$ denotes the cross product $\times_{A_i \in Attset} Cut^+(A_i, T^*)$. A *valueset*

in $Cut^+(Attset, T^*)$ is a tuple containing exactly one value from $Cut^+(A_i, T^*)$ for each attribute A_i in $Attset$. For two valuesets X and X' in $Cut^+(Attset, T^*)$, $X' \succeq X$ means that for every A_i in $Attset$, $X'[A_i] \succeq X[A_i]$, where $X[A_i]$ denotes the value in X on A_i. For a record r in T^*, we say that r *matches* X if $X \succeq r[Attset]$. $sup(X)$ denotes the number of records in T^* that match X.

B. Sensitivity model

Consider a value a for some attribute A_i. The sensitivity of a, denoted $s(a)$, measures the degree to which the publisher considers a sensitive. We assume that the publisher is able to specify each leaf node a as either sensitive or non-sensitive, i.e., $s(a) = 1$ or $s(a) = 0$, respectively. This can be done by specifying a set of highest possible nodes to cover all sensitive leaf nodes, called a *guarding set*. A leaf node a is sensitive if and only if a is in $Leaf(g)$ for some node g in the guarding set. Let $sLeaf(a)$ denote the set of sensitive leaf nodes in $Leaf(a)$.

Example 2.1: Consider the taxonomy for Disease in Figure 1. The guarding set $GS1 = \{AIDS, TB\}$ specifies the sensitive leaf nodes $AIDS$ and TB. $GS2 = \{AIDS, BI\}$ specifies the sensitive leaf nodes $AIDS$, $E.Coli$ and TB.

The publisher does not need to specify $s(a)$ for non-leaf nodes a because this will be specified automatically by our model. For a non-leaf node a, $s(a)$ depends on what a generalizes, i.e., the leaf nodes in $Leaf(a)$. For instance, $ANY\_Disease$ is more sensitive if most leaf nodes in $Leaf(ANY\_Disease)$ are sensitive. This motivates the following principle.

Definition 2.1 (Sensitivity Principle): The sensitivity of a, $s(a)$, conveys the probability that a originates from sensitive leaf values in $Leaf(a)$. This probability is the publisher's interpretation of the sensitivity of a.

The exact interpretation of probability depends on the instantiation for "a originates from sensitive leaf values in $Leaf(a)$". We leave this instantiation open so that a new instantiation can be "plugged in" through $s(a)$. Below, we demonstrate this flexibility by considering two instantiations.

The aggregate instantiation. In this instantiation, $s(a)$ measures the probability that a comes from *any* sensitive leaf node in $Leaf(a)$, defined as:

$$s(a) = |sLeaf(a)|/|Leaf(a)|. \quad (1)$$

That is, $s(a)$ is the fraction of sensitive leaf nodes in $Leaf(a)$. In particular, this instantiation makes no distinction among individual sensitive leaf nodes and the publisher considers it sensitive to infer that a comes from *any* sensitive leaf node $sLeaf(a)$.

Example 2.2: Continue with the guarding set $GS1 = \{AIDS, TB\}$. $AIDS$ and TB are sensitive leaf nodes and Flu and $E.Coli$ are non-sensitive leaf nodes. $s(AIDS) = s(TB) = 1$, $s(Flu) = s(E.Coli) = 0$. $s(a) = 1/2$ for $VI, BI, ANY\_Disease$. For example, $s(ANY\_Disease) = 1/2$ because 50% of leaf nodes are sensitive.

The non-aggregate instantiation. In many cases, sensitivity arises from the specificity of a property. Location privacy is such an example. This notion of sensitivity $s(a)$ can be measured by the probability that a comes from a *particular* sensitive leaf node in $Leaf(a)$. Unlike the aggregate instantiation, this instantiation distinguishes among sensitive leaf nodes. Under the uniform distribution, the probability that a comes from a particular leaf node in $Leaf(a)$ is $1/|Leaf(a)|$ and the probability that a given leaf node in $Leaf(a)$ is sensitive is $|sLeaf(a)|/|Leaf(a)|$. Therefore, under the independence assumption, the probability that a comes from a particular sensitive leaf node is

$$s(a) = |sLeaf(a)|/|Leaf(a)|^2. \quad (2)$$

There is no requirement that the same instantiation be used for all attributes. The choice of instantiation should be based on what probabilistic interpretation is desired for an attribute.

C. Observability model

For a valueset X in $Cut^+(Attset, T^*)$, the observability of X, denoted $o(X)$, measures the degree to which the publisher considers X as observable to the adversary. Our observation is that, for two valuesets X and X' with $X' \succeq X$, whenever X is observed on some individual, X' is observed on the same individual because X is an instance of X'. For example, observing the exact income 170K implies observing the income bracket $High$. This observation leads to the following principle.

Definition 2.2 (Observability Principle): The observability of a valueset X, denoted $o(X)$, conveys some notion of the degree that X may be observed on an individual with the following property: for $X' \succeq X$, $o(X') \geq o(X)$; in other words, a general valueset is as observed as a special one.

Since observing a valueset X entails observing *every* value in X, the "bottleneck" is the least observed value in X. So we define $o(X)$ by $min\{o(a) \mid a \in X\}$. For a given threshold σ_o, we say that X is *observable* if $o(X) \geq \sigma_o$. Below, we consider two instantiations for specifying $o(a)$ for a single value a on some attribute.

The sensitivity based instantiation. The first instantiation is based on the intuition that a sensitive value is hard to observe and a non-sensitive value is easy to observe. Therefore, if all the leaf nodes in $Leaf(a)$ are sensitive, it is hard to observe a, so $o(a) = 0$; if some leaf node in $Leaf(a)$ is non-sensitive, such leaf nodes can be observed, and a can be observed following Observability Principle, so $o(a) = 1$.

Example 2.3: Continue with the guarding set $GS1 = \{AIDS, TB\}$ in Example 2.1. $o(AIDS) = o(TB) = 0$ because these nodes are sensitive. $o(a) = 1$ for $Flu, E.Coli, VI, BI, ANY\_Disease$ because these nodes have some non-sensitive leaf nodes.

The belief based instantiation. Sometimes the publisher wants to specify $o(a)$, in the range [0..1], according to her own belief. Suppose that the publisher has specified $o(a)$ for all leaf nodes a. For a non-leaf node a, Observability Principle requires $o(a) \geq o(a_i)$ for all child nodes a_i of a. A valid

specification is $o(a) = min\{1, \alpha + max_i o(a_i)\}$, where α is a "boost" of observability when going from the child level to the parent. For example, $o(a_i) = 0$ and $\alpha = 1$ specifies that exact incomes a_i cannot be observed but the income bracket a can be observed. However, the specification $o(a) = avg_i o(a_i)$ violates Observability Principle because $avg_i o(a_i)$ may be smaller than $o(a_i)$.

Corollary 2.1: The sensitivity based and belief based instantiations satisfy Observability Principle. ■

Our approach does not depend on the exact instantiation for $o(X)$ and $s(a)$. The only requirement is that they follow Observability Principle and Sensitivity Principle. Therefore, a new instantiation for $o(X)$ and $s(a)$ can be easily incorporated into our approach.

III. Problem Statements

A. Privacy breaches

Consider a published table T^*. The adversary attempts to infer some value a in $Cut(A, T^*)$ for some attribute A by observing a valueset X in $Cut^+(U - A, T^*)$. In the following discussion, we assume that X is observable (otherwise, X cannot be linked to any individual) and a is non-observable (otherwise, there is no need to infer a).

Definition 3.1 (Attacks): Given a threshold σ_o on observability, a freeform attack *wrt T^* has the form $X \rightarrow a$, where a is a non-observable value in $Cut(A, T^*)$ for some attribute A and X is an observable valueset in $Cut^+(U - A, T^*)$.*

Definition 3.2 (Threat): Given T^* and the taxonomies for all attributes, the threat *of $X \rightarrow a$ wrt T^*, denoted $t(X \rightarrow a)$, is the probability with which the adversary can infer that a record in T^* matching X "originates" from sensitive leaf nodes in $Leaf(a)$.*

$t(X \rightarrow a)$ depends on two probabilities. The first is the probability that a record in T^* matching X matches a, i.e., $P(a|X) = \frac{sup(Xa)}{sup(X)}$. The second is the probability that a originates from a sensitive leaf node in $Leaf(a)$, i.e., $s(a)$.

Theorem 3.1: For an attack $X \rightarrow a$ wrt T^*, under the independence assumption of $P(a|X)$ and $s(a)$,

$$t(X \rightarrow a) = P(a|X) * s(a). \blacksquare \quad (3)$$

With the aggregate instantiation of $s(a)$, $t(X \rightarrow a)$ is the probability that a record in T^* matching X is associated with *any* sensitive leaf node in $Leaf(a)$; with the non-aggregate instantiation of $s(a)$, $t(X \rightarrow a)$ is the probability that a record in T^* matching X is associated with a *particular* sensitive leaf node in $Leaf(a)$. We consider $X \rightarrow a$ to be a privacy breach if $t(X \rightarrow a)$ is large enough.

Definition 3.3 (Breaches): Given thresholds σ_o and σ_t on observability and threat, any attack $X \rightarrow a$ wrt T^* is said to be a breach *if $t(X \rightarrow a) > \sigma_t$. $BR(T^*)$ denotes the set of all breaches wrt T^*.*

We are interested in finding a minimally generalized T^* with $BR(T^*) = \emptyset$. The difficulty is that $BR(T^*)$ is not monotone wrt generalization: $BR(T_2^*)$ is not necessarily a subset of $BR(T_1^*)$, where T_2^* is more general than T_1^*, because a generalization may increase the threat of an attack. Consider generalizing leaf nodes a_1 and a_2 into the parent a, where $s(a_1) = 1$ and $s(a_2) = 0$. Suppose that, before the generalization, all matching records of X have a_2. So $t(X \rightarrow a_1) = t(X \rightarrow a_2) = 0$. After the generalization, $t(X \rightarrow a) = 1/2$ because $s(a) = 1/2$ (assuming the aggregate instantiation) and $P(a|X) = P(a_2|X) = 1$. To address this issue, we extend $BR(T^*)$ to $X \rightarrow a$, where a can be more general than the values in T^*. This is defined below.

Definition 3.4 (Extended Breaches): Any attack $X \rightarrow a$ wrt T^*, with a being to a non-observable value in $Cut^+(A, T^*)$ for some attribute A, is said to be a breach *if $t(X \rightarrow a) > \sigma_t$. $BR(T^*)$ denotes the set of all breaches wrt T^*.*

Note the difference between Definition 3.3 and 3.4: the latter includes those $X \rightarrow a$ with a being more general than the values in T^*. With this extended $BR(T^*)$, it follows from the definition that, if T_2^* is more general than T_1^*, $BR(T_2^*)$ is a subset of $BR(T_1^*)$.

Corollary 3.1: For $BR(T^*)$ in Definition 3.4, if T_2^* is more general than T_1^*, $BR(T_2^*) \subseteq BR(T_1^*)$.

Several remarks are in place. First, by considering the extended $BR(T^*)$, Corollary 3.1 allows a top-down search for a maximally specialized T^* with $BR(T^*) = \emptyset$: starting with the most generalized table $T^* = T^\top$, we iteratively specialize it on some attribute as long as $BR(T^*) = \emptyset$. Second, the extended $BR(T^*)$ is a superset of the breaches defined in Definition 3.3; therefore, $BR(T^*) = \emptyset$ is a sufficient condition for eliminating all the breaches in Definition 3.3. In the rest of the paper, $BR(T^*)$ refers to the extended $BR(T^*)$ in Definition 3.4. Finally, Corollary 3.1 does not depend on the instantiation of $s(a)$, therefore, the top-down approach remains applicable if a new instantiation of $s(a)$ is adapted.

B. Anonymization problems

Let $IL(T^*, T)$ denote a chosen metric for information loss of generalizing T to T^*.

Definition 3.5: We say that T^* is an FF-anonymization *of T if $BR(T^*) = \emptyset$.* Problem I*: find an* optimal FF-anonymization *T^*, that is, for every FF-anonymization T_2^*, $IL(T^*, T) \leq IL(T_2^*, T)$.* Problem II*: find a* minimal FF-anonymization *T^*, that is, no specialization on T^* leads to an FF-anonymization.*

We can show that Problem I is NP-hard by a reduction from the optimal k-anonymity problem to Problem I. With Corollary 3.1, we can find a solution to Problem II, i.e., a minimal FF-anonymization, following a top-down specialization process starting with the most generalized table $T^* = T^\top$. Corollary 3.1 implies that further specialization leads to no solution. The detailed algorithm and evaluation will be presented in the full version of this paper.

References

[1] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, 2002.

[2] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *ICDE*, Atlanta, GA, 2006.

[3] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *Proc. of the 32nd International Conference on Very Large Data Bases (VLDB)*, Seoul, Korea, September 2006.