

Methodology, Ethics and Practice of Data Privacy Course Exercise #1

April 13 2021

1. (10') Try to explain why recursive (c, l) -diversity guards against all adversaries who possess at most $l - 2$ statements of the form “Bob does not have heart disease”.
2. (15') Consider domains R_0 (Race) and Z_0 (ZIP code) whose generalization hierarchies are illustrated in Fig. 1a and Fig. 1b independently. Assume $QI = \{\text{Race, ZIP}\}$ to be a quasi-identifier. Consider private table PT illustrated in table 1, please give all possible 2-anonymity using **full domain generalization** and **suppression** under the condition that the maximum number of suppressed records ($MaxSup$) is less than or equal to 1. (If it is not generalized, 4 records need to be suppressed, which does not meet the requirement of $MaxSup \leq 1$, illustrated in table 2).
3. (15') [The t -closeness Principle] An equivalence class is said to have t -closeness if the distance between the distribution of a sensitive attribute in this class and distribution of the attribute in the whole table is no more than a threshold t . A table is said to have t -closeness if all equivalence classes have t -closeness.

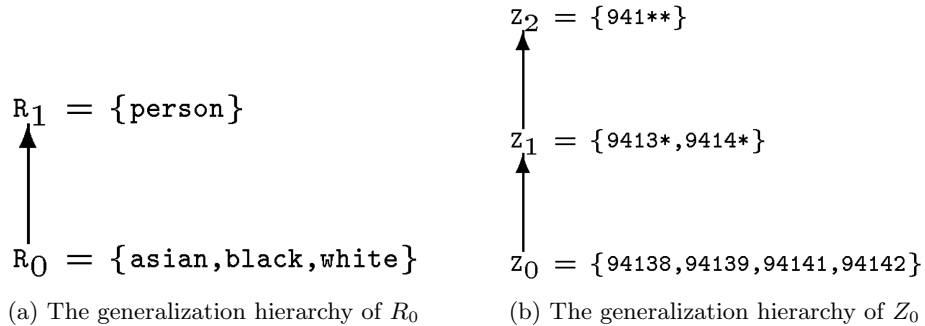


Figure 1: Generalization hierarchies

Race: R_0	ZIP: Z_0
asian	94138
asian	94138
asian	94142
asian	94142
black	94138
black	94141
black	94142
white	94138

Table 1: PT

Race: R_0	ZIP: Z_0
asian	94138
asian	94138
asian	94142
asian	94142

Table 2: Suppression for table PT

- (a) Given the anonymized table (table 3), where the quasi-identifier attributes are *ZIP Code* and *Age* and the sensitive attribute is *Salary*. Please give the value of t so that table 3 satisfies t -closeness. Please use **Earth Mover's distance (EMD)** to calculate the distance between two distributions.

Hint. The overall distribution of the Income attribute is $\mathbf{Q} = \{3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k\}$ (We use the notation $\{v_1, v_2, \dots, v_m\}$ to denote the uniform distribution where each value in $\{v_1, v_2, \dots, v_m\}$ is equally likely.) The first equivalence class in table 3 has distribution $\mathbf{P}_1 = \{3k, 5k, 9k\}$.

[Earth Mover's distance (EMD)]. The *Salary* is the numerical attribute. Numerical attribute values are ordered. Let the attribute domain be $\{v_1, v_2, \dots, v_m\}$, where v_i is the i^{th} smallest value. Let $\mathbf{P} = \{p_1, p_2, \dots, p_m\}$ and $\mathbf{Q} = \{q_1, q_2, \dots, q_m\}$ be distributions. we use *Ordered Distance* to calculate the distance between two values. Let $r_i = p_i - q_i (i = 1, 2, \dots, m)$, then EMD between \mathbf{P} and \mathbf{Q} can be calculate as:

$$\begin{aligned}
 D[\mathbf{P}, \mathbf{Q}] &= \frac{1}{m-1} (|r_1| + |r_1 + r_2| + \dots + |r_1 + r_2 + \dots + r_{m-1}|) \\
 &= \frac{1}{m-1} \sum_{i=1}^m \left| \sum_{j=1}^i r_j \right|
 \end{aligned} \tag{1}$$

[Ordered Distance] *Ordered Distance* between two values is based on the number of values between them in the total order, i.e., $ordered_list(v_i, v_j) = \frac{|i-j|}{m-1}$.

4. (25') Given the following private table (table 4):

Please answer the following questions:

- (a) (5') Given the health condition as the sensitive attribute, please name the quasi-identifier attributes.

ZIP Code	Age	Salary
4767*	≤ 40	3K
4767*	≤ 40	5K
4767*	≤ 40	9K
4790*	≥ 40	6K
4790*	≥ 40	11K
4790*	≥ 40	8K
4760*	≤ 40	4K
4760*	≤ 40	7K
4760*	≤ 40	10K

Table 3: The anonymized table.

Name	Age	Gender	Nationality	Salary	Condition
Ann	35	F	Japanese	40K	Viral Infection
Bluce	27	M	American	38K	Flu
Cary	41	F	India	45K	Heart Disease
Dick	32	M	Korean	38K	Flu
Eshwar	52	M	Japanese	61K	Heart Disease
Fox	22	M	American	22K	Flu
Gary	36	M	India	34K	Flu
Helen	26	F	Chinese	26K	Cancer
Irene	18	F	American	16K	Viral Infection
Jean	25	F	Korean	38K	Cancer
Ken	38	M	American	55K	Viral Infection
Lewis	47	M	American	64K	Heart Disease
Martin	24	M	American	37K	Viral Infection

Table 4: Private table.

- (b) (15') Let the valid range of age be $\{0, \dots, 120\}$. Given the health condition as the sensitive attribute, design a cell-level generalization solution to achieve k -Anonymity, where $k = 2$. Please give the generalization hierarchies, released table and calculation of the loss metric (LM) of your solution.
- (c) (5') Please design a k -anonymization algorithm to optimize the loss metric.
5. (20') Suppose that private information x is a number between 0 and 1000. This number is chosen as a random variable X such that 0 is 1%-likely whereas any non-zero is only about 0.1%-likely:

$$P[X = 0] = 0.01, P[X = k] = 0.00099, k = 1 \dots 1000 \quad (2)$$

Suppose we want to randomize such a number by replacing it with a new random number $y = R(x)$ that retains some information about the original

number x . Here are three possible methods to do it:

- (a) Given x , let $R_1(x)$ be x with 20% probability, and some other number (chosen uniformly at random in $\{0, \dots, 1000\}$) with 80% probability.
- (b) Given x , let $R_2(x)$ be $(x + \delta) \bmod 1001$, where δ is chosen uniformly at random in $\{-100 \dots 100\}$.
- (c) Given x , let $R_3(x)$ be $R_2(x)$ with 50% probability, and a uniformly random number in $\{0, \dots, 1000\}$ otherwise.

Please answer the following questions:

- (a) (15') Compute prior and posterior probabilities of two properties of x : 1) $X = 0$; 2) $x \in \{200, \dots, 800\}$ using the above three methods respectively.
 - (b) (5') Which method is better? Why?
6. (15') $[(\alpha, \beta)$ -Privacy] Let R be an algorithm that takes as input $u \in D_U$ and outputs $v \in D_V$. R is said to allow an upward (α, β) -privacy breach with respect to a predicate ϕ if for some probability distribution f ,

$$\exists u \in D_U, \exists v \in D_V \text{ s.t. } P_f(\Phi(u)) \leq \alpha \text{ and } P_f(\Phi(u) | R(u) = v) \geq \beta \quad (3)$$

Similarly, R is said to allow a downward (α, β) -privacy breach with respect to a predicate Φ if for some probability distribution f ,

$$\exists u \in D_U, \exists v \in D_V \text{ s.t. } P_f(\Phi(u)) \geq \alpha \text{ and } P_f(\Phi(u) | R(u) = v) \leq \beta \quad (4)$$

R is said to satisfy (α, β) -privacy if it does not allow any (α, β) -privacy breach for any predicate Φ . The necessary and sufficient conditions for R to satisfy (α, β) -privacy for any prior distribution and any property ϕ : γ -amplifying

$$\forall v \in D_V, \forall u_1, u_2 \in D_U, \frac{P(R(u_1) = v)}{P(R(u_2) = v)} \leq \gamma \quad (5)$$

- (a) Let R be an algorithm that is γ -amplifying. Please proof that R does not permit an (α, β) -privacy breach for any adversarial prior distribution if

$$\gamma \leq \frac{\beta}{\alpha} \frac{1 - \alpha}{1 - \beta}. \quad (6)$$