Data Privacy Course Exercise 1

zlt0116@mail.ustc.edu.cn

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".

Answer:

If the adversary possesses $l-1$ statements, the rest one's sensitive data can be directly inferred.

There are at least l distinct sensitive values. The most frequent value does not appear too frequently, while less common values are ensured to not appear too infrequently.

Therefore, the attacker can not distinguish which sensitive data belongs to the target person because the number of the rest tuples is greater than 1. The (c, l)-diversity guards adversaries with $l-2$ statements.

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

(a) PT

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

(b) Suppression for table PT

表 1: Table for question 2

2. (15')

Consider domains R_0 (Race) and Z_0 (ZIP code) whose generalization hierarchies are

$$R_0: \ R_1 = \{\text{person}\} \leftarrow R_0 = \{\text{asian, black, white}\}$$
$$z_0: \ z_2 = \{941**\} \leftarrow z_1 = \{9413*, 9414*\} \leftarrow z_0 = \{94138, 94139, 94141, 94142\}$$

Assume $Q_I = \{\text{Race, ZIP}\}$ to be a quasi-identifier. Consider private table PT illustrated in table 1a, 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 1b).

Answer:

Applying full domain generalization and suppression ($MaxSup \leq 1$), the answers are shown in table 2a, 2, 2c, 2d, 2e, 2f, and 2g.

Race:R_0	**ZIP:**Z_0
person	9413*
person	9413*
person	9414*
person	9414*
person	9413*
person	9414*
person	9414*
person	9413*

(a) only gen

Race:R_0	**ZIP:**Z_0
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**

(b) only gen

Race:R_0	**ZIP:**Z_0
person	9413*
person	9413*
person	9414*
person	9414*
person	9413*
person	9414*
person	9414*

(c) gen+sup

Race:R_0	**ZIP:**Z_0
person	9413*
person	9413*
person	9414*
person	9414*
person	9413*
person	9414*
person	9413*

(d) gen+sup

Race:R_0	**ZIP:**Z_0
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**

(e) gen+sup

Race:R_0	**ZIP:**Z_0
person	94138
person	94138
person	94142
person	94142
person	94138
person	94142
person	94138

(f) gen+sup

Race:R_0	**ZIP:**Z_0
asian	941**
asian	941**
asian	941**
asian	941**
black	941**
black	941**
black	941**

(g) gen+sup

表 2: The answers of question 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.

(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

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

表 3: The anonymized table

use **Earth Mover's distance (EMD)** to calculate the distance between two distributions.

Answer:

The overall distribution of the Income attribute in the whole table is $Q = \{3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k\}$. Let $Q = q_i$, where $1 \leq i \leq 9$, then $\forall i$, we have $q_i = \frac{1}{9}$.

- For equivalence class $\{4767*, \leq 40\}$, $P_1 = \{3k, 5k, 9k\}$. $p_{11} = p_{13} = p_{17} = \frac{1}{3}$, and $p_{1i} = 0$ for $i = 2, 4, 5, 6, 8, 9$. Therefore, the EMD between P_1 and Q is equal to

$$\begin{aligned} D[P_1, Q] &= \frac{1}{9-1}(|p_{11} - q_1| + |p_{11} - q_1 + p_{12} - q_2| + ... + |p_{11} - q_1 + p_{12} - q_2 + ... + p_{19} - q_9|) \\ &= \frac{1}{8}(|\frac{2}{9}| + |\frac{1}{9}| + |\frac{1}{3}| + |\frac{2}{9}| + |\frac{1}{9}| + |0| + |\frac{2}{9}| + |\frac{1}{9}| + |0|) \\ &= \frac{1}{6} \end{aligned} \tag{1}$$

- For equivalence class $\{4790*, \geq 40\}$, $P_2 = \{6k, 11k, 8k\}$. $p_{24} = p_{26} = p_{29} = \frac{1}{3}$, and $p_{2i} = 0$ for $i = 1, 2, 3, 5, 7, 8$. Therefore, the EMD between P_2 and Q is equal to

$$\begin{aligned} D[P_2, Q] &= \frac{1}{9-1}(|p_{11} - q_1| + |p_{11} - q_1 + p_{12} - q_2| + ... + |p_{11} - q_1 + p_{12} - q_2 + ... + p_{19} - q_9|) \\ &= \frac{1}{8}(|-\frac{1}{9}| + |-\frac{2}{9}| + |-\frac{1}{3}| + |-\frac{1}{9}| + |-\frac{2}{9}| + |0| + |-\frac{1}{9}| + |-\frac{2}{9}| + |0|) \\ &= \frac{1}{6} \end{aligned} \tag{2}$$

- For equivalence class $\{4760*, \leq 40\}$, $P_3 = \{4k, 7k, 10k\}$. $p_{32} = p_{35} = p_{38} = \frac{1}{3}$, and $p_{3i} = 0$ for

$i = 1, 3, 4, 6, 7, 9$. Therefore, the EMD between P_3 and Q is equal to

$$\begin{aligned} D[P_3, Q] &= \frac{1}{9-1}(|p_{11} - q_1| + |p_{11} - q_1 + p_{12} - q_2| + ... + |p_{11} - q_1 + p_{12} - q_2 + ... + p_{19} - q_9|) \\ &= \frac{1}{8}(|-\frac{1}{9}| + |\frac{1}{9}| + |0| + |-\frac{1}{9}| + |\frac{1}{9}| + |0| + |-\frac{1}{9}| + |-\frac{1}{9}| + |0|) \\ &= \frac{1}{12} \end{aligned} \tag{3}$$

Therefore, **the value of** t is $\max\{D[P_1, Q], D[P_2, Q], D[P_3, Q]\} = \frac{1}{6}$.

4. (25')

Given the following private table (table 4):

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

表 4: Private table

Please answer the following questions:

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

Answer:

The quasi-identifiers are **Age, Gender, Nationality, and Salary**.

(b) (15') Let the valid range of age be $\{0, ..., 120\}$. 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.

Answer:

The generalization hierarchies of the 4 attributes are shown in the figure 1, and the released table is shown in table 5.

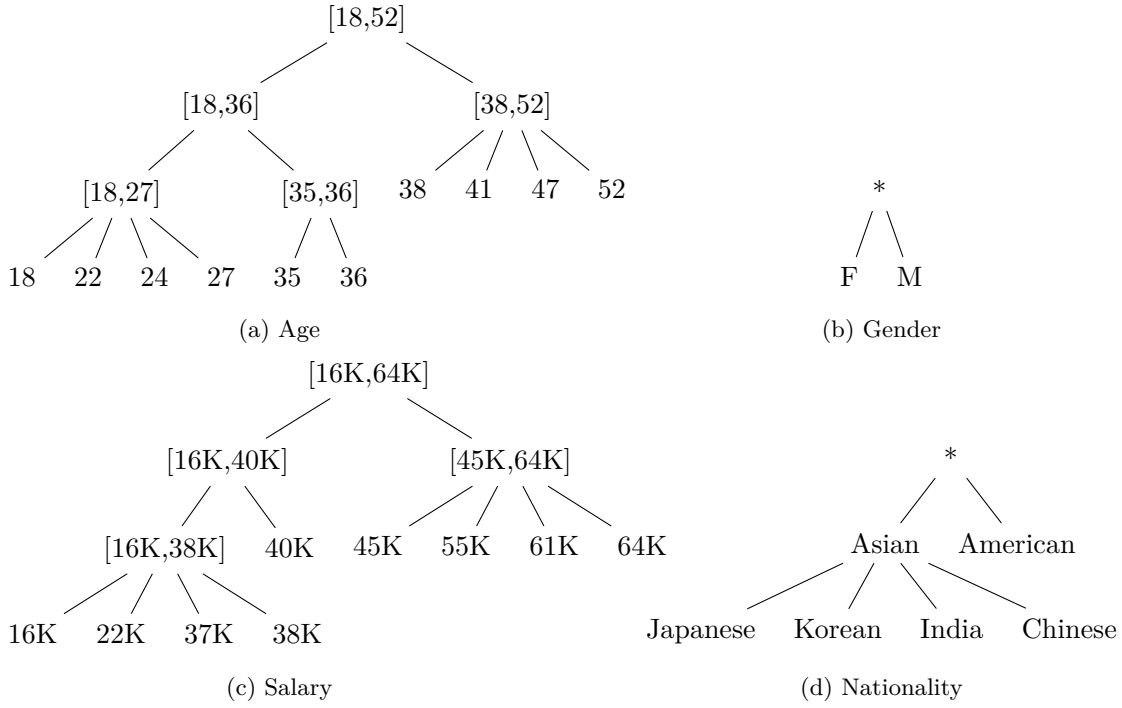

图 1: The generalization hierarchies of question 2

Age	Gender	Nationality	Salary
[38,52]	*	*	[45K,64K]
[38,52]	*	*	[45K,64K]
[38,52]	*	*	[45K,64K]
[38,52]	*	*	[45K,64K]
[18,36]	*	Asian	[16K,40K]
[18,36]	*	Asian	[16K,40K]
[18,36]	*	Asian	[16K,40K]
[18,36]	*	Asian	[16K,40K]
[18,36]	*	Asian	[16K,40K]
[18,27]	*	American	[16K,38K]
[18,27]	*	American	[16K,38K]
[18,27]	*	American	[16K,38K]
[18,27]	*	American	[16K,38K]

表 5: Released table

The loss for attribute A is defined as the average of the loss t[A] for all tuples t:

- $LM_{\text{Age}} = \frac{1}{13} * (\frac{52-38+1-1}{52-18+1-1} * 4 + \frac{36-18+1-1}{52-18+1-1} * 5 + \frac{27-18-1+1}{52-38+1-1} * 4) = \frac{90}{221}$
- $LM_{\text{Gender}} = \frac{1}{13} * \frac{2-1}{2-1} * 13 = 1$
- $LM_{\text{Nationality}} = \frac{1}{13} * (\frac{5-1}{5-1} * 4 + \frac{4-1}{5-1} * 5 + \frac{1-1}{5-1} * 4) = \frac{31}{52}$
- $LM_{\text{Salary}} = \frac{1}{13} * (\frac{64-45+1-1}{64-16+1-1} * 4 + \frac{40-16+1-1}{64-16+1-1} * 5 + \frac{38-16+1-1}{64-16+1-1} * 4) = \frac{71}{156}$

The LM for the entire data set is defined as the sum of the losses for each attribute:

$$LM = LM_{\text{Age}} + LM_{\text{Gender}} + LM_{\text{Nationality}} + LM_{\text{Salary}} = \frac{1630}{663}$$

(c) (5') Please design a k-anonymization algorithm to optimize the loss metric.

Answer:

Optimize by exhaustion (consider all possible generalization hierarchies that satisfy 2-anonymity, and select the algorithm with the optimal 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:

$$P[X=0]=0.01,\ P[X=k]=0.00099,\ k=1...1000 \tag{4}$$

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 $\{0,...,1000\}$) with 80% probability.

(b) Given x, let $R_2(x)$ be $(x+\delta)$ *mod* 1001, where δ is chosen uniformly at random in $\{-100...100\}$

(c) Given x, let $R_3(x)$ be $R_2(x)$, with 50% probability, and a uniformly random number in $\{0...1000\}$ otherwise.

Please answer the following questions:

(a) (15') Compute prior and posterior probabilities of two properties of X: 1) $X=0$; 2) $X\in\{200,...,800\}$ using the above three methods respectively. The posterior probabilities only need to be computed when $R_i(X)=0$, $i=1,2,3$, respectively.

Answer:

The prior probabilities for the three methods are:

$$P(X=0)=0.01$$
$$P(X\in\{200,...,800\})=(800-200+1)\times 0.00099=0.59499$$

Posterior probabilities:

According to the question, we only need to consider the cases that $R_i(x)=0$.

$$P(X=y|R_1(X)=0)=\frac{P(R_1(X)=0|X=y)P(X=y)}{P(R_1(X)=0)} \qquad \text{(Bayes' theorem)}$$

For method (a):

$$\begin{aligned}P(R_1(X)=0)&=\sum_{0\le z\le 1000}P(R_1(X)=0|X=z)P(X=z)\\&=(0.2+0.8\times\frac{1}{1001})\times 0.01+\frac{0.8}{1001}\times(1-0.01)\\&=0.002799\end{aligned}$$
$$P(X=0|R_1(X)=0)=\frac{(0.2+0.8\times\frac{1}{1001})\times 0.01}{P(R_1(X)=0)}=0.7173$$
$$P(X\in\{200,...,800\}|R_1(X)=0)=\frac{\frac{0.8}{1001}\times 0.59499}{P(R_1(X)=0)}=0.1699$$

For method (b):

$$\begin{aligned}
P(R_2(X)=0) &= \sum_{0\le z\le 1000} P(R_2(X)=0|X=z)P(X=z) \\
&= 0.01\times\frac{1}{201}+0.00099\times\frac{1}{201}\times 200 \\
&= 0.001035 \\
P(X=0|R_2(X)=0) &= \frac{\frac{1}{201}\times 0.01}{P(R_2(X)=0)} = 0.04808 \\
P(X\in\{200,...,800\}|R_2(X)=0) &= 0
\end{aligned}$$

For method (c):

$$\begin{aligned}
P(R_3(X)=0) &= \sum_{0\le z\le 1000} P(R_3(X)=0|X=z)P(X=z) \\
&= 0.01\times(0.5\times\frac{1}{201}+0.5\times\frac{1}{1001})+0.00099\times(0.5\times\frac{200}{201}+0.5\times\frac{1000}{1001}) \\
&= 0.001017 \\
P(X=0|R_3(X)=0) &= \frac{(0.5\times\frac{1}{201}+0.5\times\frac{1}{1001})\times 0.01}{P(R_3(X)=0)} = 0.02937 \\
P(X\in\{200,...,800\}|R_3(X)=0) &= \frac{0.5\times\frac{601}{1001}\times 0.00099}{P(R_3(X)=0)} \\
&= 0.2923
\end{aligned}$$

(b) (5') Which method is better? Why?

Answer:

We can estimate the utility of these 3 methods according to the posterior probabilities in question (a). The smaller the posterior probability is, the less private information is leaked, the better privacy is preserved. Hence, if $x=0$ is more sensitive, then method (c) is better since $P(X=0|R_3(X)=0)$ has the smallest value.

It is also feasible to compute the distance between the distribution of X and $R(X)$ using KL-divergence, mutual information, etc. For example, the mutual information is $I(X;R(X))=H(R(X))-H(R(X)|X)$. $I(X;R_3(X))$ is the smallest, so method (c) is better.

6. (15')

[(α,β)-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,

$$\exists u\in D_U, \exists v\in D_V \ s.t.\ P_f(\Phi(u))\le\alpha \text{ and } P_f(\Phi(u)|R(u)=v)\ge\beta \tag{5}$$

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

$$\exists u \in D_U, \exists v \in D_V \ s.t. \ P_f(\Phi(u)) \geq \beta \text{ and } P_f(\Phi(u)|R(u) = v) \leq \alpha \tag{6}$$

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

$$\forall v \in D_V \ \forall u_1, u_2 \in D_U, \frac{P(R(u_1) = v)}{P(R(u_2) = v)} \leq \gamma \tag{7}$$

(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

$$\gamma \leq \frac{\beta}{\alpha} \frac{1-\alpha}{1-\beta} \tag{8}$$

Proof:

1° For upward (α, β)-privacy, we have $P_f(\Phi(u)) \leq \alpha$, then

$$\begin{aligned} P_f(\Phi(u)|R(u) = v) &= \sum_{u \in \Phi^{-1}} P_f(U = u|R(U) = v) \\ &= \sum_{u \in \Phi^{-1}} \frac{P(R(u) = v|U = u) P_f(U = u)}{P_f(R(U) = v)} && \text{(Bayes' theorem)} \\ &= \frac{\sum_{u \in \Phi^{-1}} P(R(u) = v) P_f(U = u)}{P_f(R(u) = v)} \end{aligned}$$

According to the equation above, we take $u_1 = \arg\max_{u \in \Phi^{-1}} P(R(u) = v)$, then

$$\begin{aligned} P_f(\Phi(u)|R(u) = v) &\leq \frac{P(R(u_1) = v) \sum_{u \in \Phi^{-1}} P_f(U = u)}{P_f(R(u) = v)} \\ &= \frac{P(R(u_1) = v)}{P_f(R(u) = v)} P_f(\Phi(u)) \end{aligned}$$

Take $u_2 = \arg\min_{u \in \neg\Phi^{-1}} P(R(u) = v)$, then

$$\begin{aligned} 1 - P_f(\Phi(u)|R(u) = v) &\geq \frac{P(R(u_2) = v) \sum_{u \in \neg\Phi^{-1}} P_f(U = u)}{P_f(R(u) = v)} \\ &= \frac{P(R(u_2) = v)}{P_f(R(u) = v)} (1 - P_f(\Phi(u))) \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{P_f(\Phi(u)|R(u) = v)}{1 - P_f(\Phi(u)|R(u) = v)} &= \frac{P_f(\Phi(u)|R(u) = v)}{1 - P_f(\Phi(u)|R(u) = v)} \\ &\leq \frac{P(R(u_1) = v) P_f(\Phi(u))}{P(R(u_2) = v)(1 - P_f(\Phi(u)))} \\ &\leq \gamma \frac{\alpha}{1-\alpha} && (\gamma\text{-amplifying}) \end{aligned}$$

That is,

$$\begin{aligned} P_f(\Phi(u)|R(u)=v)(1-\alpha) &\leq \gamma\alpha(1-P_f(\Phi(u)|R(u)=v)) \\ P_f(\Phi(u)|R(u)=v) &\leq \frac{\gamma\alpha}{\gamma\alpha+1-\alpha} \\ &= \frac{\alpha}{\alpha+\frac{1-\alpha}{\gamma}} \\ &\leq \beta \qquad \text{(Property of } R\text{)} \end{aligned}$$

Obviously, for given algorithm R, there is no upward (α, β)-privacy breach.

2° For downward (α, β)-privacy, we have $P_f(\Phi(u)) \geq \beta$. Similarly,

$$\frac{P_f(\Phi(u)|R(u)=v)}{1-P_f(\Phi(u)|R(u)=v)} \geq \frac{1}{\gamma}\frac{\beta}{1-\beta} \qquad (\gamma\text{-amplifying})$$

That is,

$$\begin{aligned} P_f(\Phi(u)|R(u)=v) &\geq \frac{\beta}{(1-\beta)\gamma+\beta} \\ &\geq \alpha \qquad \text{(Property of } R\text{)} \end{aligned}$$

Obviously, for given algorithm R, there is no downward (α, β)-privacy breach.

3° Together, we have proven that R does not permit an (α, β)-privacy breach for any adversarial prior distribution with the given condition.