School of Computer Science and Technology
University of Science and Technology of China

Homework 1 for
Methodology, Ethics and Practice of Data privacy
2024 Autumn

Exercise 1 *K-Anonymity (20')*

	Zip Code	Salary	Nationality	Condition
1	$130**$	$15k-25k$	Chinese	Heart Disease
2	$130**$	$15k-25k$	Chinese	Heart Disease
3	$130**$	$25k-30k$	American	Viral Infection
4	$130**$	$25k-30k$	American	Viral Infection
5	$148**$	$15k-25k$	American	Cancer
6	$148**$	$15k-25k$	American	Heart Disease
7	$148**$	$15k-25k$	American	Viral Infection
8	$148**$	$15k-25k$	American	Viral Infection
9	$130**$	$15k-25k$	Chinese	Cancer
10	$130**$	$15k-25k$	Chinese	Cancer
11	$130**$	$25k-30k$	American	Cancer
12	$130**$	$25k-30k$	American	Cancer

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

(b) (5') How many QI-clusters are in the table above?

(c) (5') The table above satisfies p-sensitive k-anonymity, where p=?,k=?

(d) (5') List 5 attacks against k-anonymity.

Exercise 2 *ℓ-Diversity (20')*
The introduction of ℓ-diversity ensures the diversity of sensitive attributes to prevent attacks such as homogeneity attack.

(a) (5') Whether the attributes in Exercise 1 meet recursive (1, 2)-diversity, and provide reasons.

(b) (15') Prove the monotonicity of entropy ℓ-diversity. That is, if a table T satisfies entropy ℓ-diversity, then any generalization T^* of T also satisfies entropy ℓ-diversity.

Exercise 3 *t-Closeness (25')*
(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. The most important point of t-closeness is the distance measure of the probability distribution. Explore the properties of the following two commonly used probability distribution distance measures.

(a) (10') (Earth Mover's Distance) Let $\{v_1, v_2, \cdots, v_m\}$ be an ordered list of the values of the target attribute. Ordered Distance between two values in $\{v_1, v_2, \cdots, v_m\}$ is based on the number of values between them in the total order, i.e., ordered_list $(v_i, v_j) = \frac{|i-j|}{m-1}$. Consider $\boldsymbol{P} = \{p_1, p_2, \cdots, p_m\}$ and $Q = \{q_1, q_2, \cdots, q_m\}$ as two distributions over $\{v_1, v_2, \cdots, v_m\}$,where p_i and q_i represent the probabilities of v_i under distributions $\boldsymbol{P}$ and $\boldsymbol{Q}$ respectively. Define $r_i = p_i - q_i (i = 1, 2, \cdots, m)$, prove that the EMD between $\boldsymbol{P}$ and $\boldsymbol{Q}$ induced by the Ordered Distance can be calculate as:

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

(b) (5') (Kullback-Leibler divergence) The KL divergence of two probability distributions $\boldsymbol{P} = (p_1, p_2, \ldots, p_m)$ and $\boldsymbol{Q} = (q_1, q_2, \ldots, q_m)$ is defined as follows:

$$D_{KL}(\boldsymbol{P} || \boldsymbol{Q}) = \sum_{i=1}^{m} p_i \log \frac{p_{\mathrm{i}}}{q_i}$$

Prove that the KL divergence satisfies positivity but is not a true distance. (Hint: KL divergence does not satisfy symmetry.)

(c) (10') (Jensen-Shannon divergence) The JS divergence is defined based on KL divergence and addresses the asymmetry limitation of KL divergence.

$$D_{JS}(P || Q) = \frac{1}{2} D_{KL}\left(P \middle\| \frac{P+Q}{2}\right) + \frac{1}{2} D_{KL}\left(Q \middle\| \frac{P+Q}{2}\right)$$

Try to prove that the JS divergence satisfies the triangle inequality

$$D_{JS}(P_1 \parallel P_2) + D_{JS}(P_2 \parallel P_3) \geq D_{JS}(P_1 \parallel P_3)$$

if and only if

$$H\left(\frac{P_1 + P_2}{2}\right) + H\left(\frac{P_2 + P_3}{2}\right) \geq H\left(\frac{P_1 + P_3}{2}\right) + H(P_2)$$

for any probability distributions P_1, P_2, P_3. Here, $H(P) = -\sum_x P(x) \log P(x)\, dx$ represents the entropy.

Note: Here, adding two probability distributions and dividing by 2 refers to adding their probability density functions and dividing by 2, which corresponds to the density function of a new probability distribution.

Aside: We can show that Jensen-Shannon is not a metric by constructing three simple distributions P_1, P_2, and P_3 for which the above inequality does not hold. (Distributions over a discrete space of size two suffice.) While the Jensen-Shannon divergence is not a metric, it can be shown that the square root of the Jensen-Shannon divergence is a metric. https://ieeexplore.ieee.org/document/1207388

Exercise 4 *Loss Metric (10')*
Let the valid range of age be $\{0, \cdots, 100\}$. Given the health condition as the sensitive attribute, design a cell-level generalization solution to achieve 4-Anonymity. Please give the generalization hierarchies, released table and calculate the loss metric (LM) of your solution.

Name	Age	Gender	Nationality	Condition
Alan	38	M	Chinese	Heart Disease
Bruce	18	M	Chinese	Heart Disease
Cindy	20	F	Japanese	Viral Infection
David	32	M	Korean	Viral Infection
Eric	40	M	American	Cancer
Frank	36	M	India	Heart Disease
Grace	18	F	American	Viral Infection
Helen	27	F	American	Viral Infection
Irene	48	M	Chinese	Cancer
Jack	20	M	American	Cancer
Ken	25	F	American	Cancer
Lewis	52	F	American	Cancer

Exercise 5 *Reconstruct single column aggregates (25')*
Try to prove THEOREM 1 on page 209 of the PPT. Below are some definitions that might be useful.
Definition 1 **(Retention Replacement Perturbation)** Retention replacement perturbation is a perturbation algorithm, where each element in column j is retained with probability p_j, and with probability $(1-p_j)$ replaced with an element selected from the replacing p.d.f. on D_j. That is,

$$t'_{ij} = \begin{cases} t_{ij} \text{ with probability } p_j \\ \text{element from replacing p.d.f. on } D_j \text{ with probability } (l - p_j). \end{cases}$$

Definition 2 **(Reconstructible Function)** Given a perturbation α converting table T to T', a numeric function f on T is said to be (n, ϵ, δ) reconstructible by a function f', if f' can be evaluated on the perturbed table T' so that $|f - f'| < \max(\epsilon, \epsilon f)$ with probability greater than $(1 - \delta)$ whenever the table T has more than n rows.
Consider the uniform retention replacement perturbation with retention probability p applied on a database with n rows and a single column, C, with domain $[min, max]$. Consider the predicate $P = C[low, high]$. Given the perturbed table T', we explore how to estimate an answer to the query $count(P)$ on T.

(a) Let tables T, T' each have n rows. Let $n_r = count(P)$ evaluated on table T', while $n_o = count(P)$ estimated for table T. Given n_r we estimate n_o as

$$n_o = \frac{1}{p}(n_r - n(1-p)b), \text{ where } b = \frac{high - low}{max - min}.$$

(b) The fraction f of rows originally in $[low, high]$ is therefore estimated as

$$f' = \frac{n_o}{n} = \frac{n_r}{pn} - \frac{(1-p)(high - low)}{p(max - min)}.$$

Now we can prove THEOREM 1:
Theorem 1 Let the fraction of rows in $[low, high]$ in the original table f be estimated by f', then f' is a (n, ϵ, δ) estimator for f if $n \geq 4\log(\frac{2}{\delta})(p\epsilon)^{-2}$.
Hint 1: Consider the indicator variable for the event that the i^{th} row $(1 \leq i \leq n)$ is perturbed and the perturbed value falls within $[low, high]$; and the indicator variable for the event that the i^{th} row is not perturbed and it falls within $[low, high]$. Then consider the indicator variable for the event that the i^{th} randomized row falls in $[low, high]$.
Hint 2: **Multiplicative Chernoff bound.** Suppose $X_1, \ldots, X_n$ are independent random variables taking values in $\{0, 1\}$. Let X denote their sum and let $\mu = E[X]$ denote the sum's expected value. Then for any $0 < \delta < 1$, $\Pr(|X - \mu| \geq \delta\mu) < 2e^{-\delta^2\mu/4}$.