Data Privacy Homework 1

严禁抄袭，可以使用Latex、word或手写拍照等方式，**最终请提交pdf文件**至bb系统。

Plagiarism is strictly prohibited, and methods such as Latex, Word, or handwritten photography can be used. **Finally, please submit a PDF file to BlackBoard.**

1. (20')K-anonymity

	Zip Code	Age	Salary	Nationality	Condition
1	13053	28	15k	Russian	Heart Disease
2	13068	29	25k	American	Heart Disease
3	13068	21	18k	Japanese	Viral Infection
4	13053	23	13k	American	Viral Infection
5	14853	50	20k	Indian	Cancer
6	14853	55	19k	Russian	Heart Disease
7	14850	47	19k	American	Viral Infection
8	14850	49	22k	American	Viral Infection
9	13053	30	16k	Chinese	Cancer
10	13053	37	25k	Indian	Cancer
11	13068	36	17k	Japanese	Cancer
12	13068	32	21k	Indian	Cancer

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

(b) (15') Let the valid range of age be {0, · · · , 100}. Given the health condition as the sensitive attribute, use the quasi-identifier attributes you answered in (a), 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.

2. (20')L-Diversity

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

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

(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.

3.(20') T-closeness

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) (10') Let $\{v_1, v_2, \cdots, v_m\}$ be an odered 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 $\boldsymbol{Q} = \{q_1, q_2, \cdots, q_m\}$ 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 P and 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) (10') Given the following 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.

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

4. (20')Prior and posterior

Suppose that private information x is a number between 0 and 100 . This number is chosen as a random variable x :

$$\begin{aligned} &\mathbf{P}[X=0]=0.01 \\ &\mathbf{P}[X=k]=0.0099, \quad k \in [1,100] \end{aligned}$$

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 ways to do it:

1. Given x, let $R_1(x)$ be x with 30% probability, and some other number (chosen uniformly in [0,100] at random) with 70% probability.
2. Given x, let $R_2(x)$ be $x+\xi(\mod 101)$, where ξ is chosen uniformly at random in $\{-10,\ldots,0,\ldots 10\}$.
3. Given x, let $R_3(x)$ be $R_2(x)$ with 50% probability, and chosen uniformly in [0,100] at random otherwise.

Please answer the following questions:

(a)(15') Compute prior and posterior probabilities of $x=0$ (given $R_1(x)=0$) , $x=0$ (given $R_3(x)=0$) and prior and posterior probabilities of $x \in [20,80] (given\ R_2(x)=0)$ using the above three methods.

(b) (5') This is a prior and posterior probability table corresponding to the value of x for other R functions. Please indicate which R is more suitable and why?

Given:	$X=0$	$X \notin \{200,\ldots,800\}$
nothing	1%	$\approx 40.5\%$
$R_1(X)=0$	$\approx 71.6\%$	$\approx 83.0\%$
$R_2(X)=0$	$\approx 4.8\%$	100%
$R_3(X)=0$	$\approx 2.9\%$	$\approx 70.8\%$

5. (20') k-anonymity in graphs

We call a graph $G(V, E)$ k-degree anonymous if the degree sequence of G, d_G, is k-anonymous. For example, the degree sequence of the graph in following fig is $\{2, 2, 4, 3, 3, 2\}$. The graph is 1-degree anonymous.

(a) (5') Make the graph 2-anonymous by adding one edge.

(b) (5') Make the graph 3-anonymous by adding edges (multiple edges are permitted).

(c) (10') Let $G'(V', E')$ be a anonymized graph for G and $\Delta(E, E')$ be the identical edges in E and E'. Compute the information losses of the previous anonymized graphs as:

$$L(G, G') = 1 - \frac{|\Delta(E, E')|}{max\{|E|, |E'|\}}$$