Data Privacy Homework 1

Daoyu Wang PB21030794

November 7

Contents

1 K-anonymity

1.1 (a)

The quasi-identifier attributes are **Zip Code**, **Age**, **Salary**, **Nationality**.

1.2 (b)

Assume **H** represents **Heart Disease**, **V** represents **Viral Infection**, **C** represents **Cancer**.

Assume **R** represents **Russian**, **A** represents **American**, **C** represents **Chinese**, **I** represents **Indian**, **J** represents **Japanese**.

1.2.1 generalization hierarchies

When $k = 2$, the generalization hierarchies are as follows:

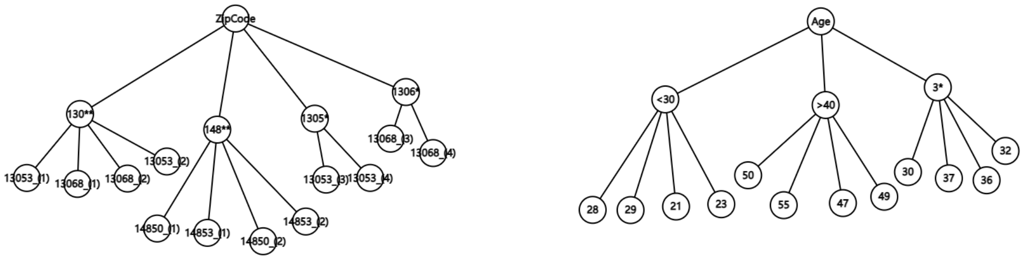

Figure 1: generalization hierarchies: Zip Code, Age

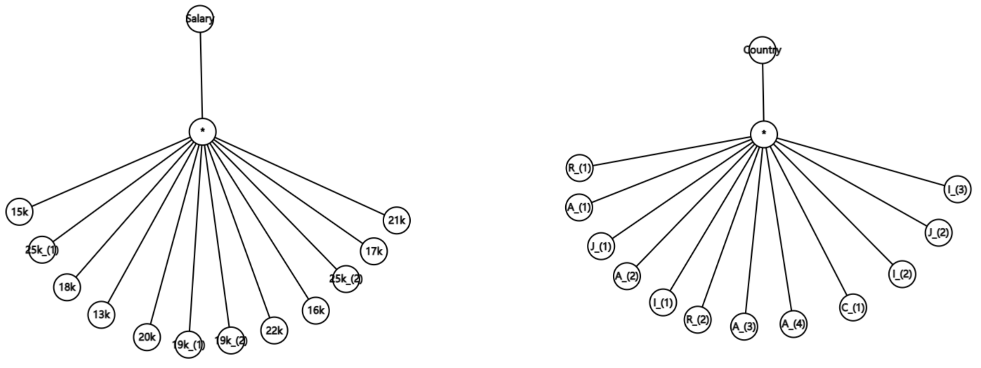

Figure 2: generalization hierarchies: Salary, Nationality

1.2.2 released table

The released table is as follows:

	Zip Code	Age	Salary	Nationality	Condition
1	130**	< 30	*	*	H
2	130**	< 30	*	*	H
3	130**	< 30	*	*	V
4	130**	< 30	*	*	V
5	1485*	> 40	*	*	C
6	1485*	> 40	*	*	H
7	1485*	> 40	*	*	V
8	1485*	> 40	*	*	V
9	1305*	3*	*	*	C
10	1305*	3*	*	*	C
11	1306*	3*	*	*	C
12	1306*	3*	*	*	C

1.2.3 loss metric

For a tuple t, suppose the value of $t[A]$ has been generalized to x. Letting $|A|$ represent the total number of leaf nodes in the tree; letting M represent the number of leaf nodes in the subtree rooted at x, then the loss for $t[A]$ is $\frac{M-1}{|A|-1}$.

The loss for attribute A is the average of the loss for all tuples t. The **LM** for the entire data set is the sum of the losses for each attribute. The loss metric is as follows:

- **Zip Code:** $|A| = 16$
  - **130**:** $M = 4$, $loss = \frac{4-1}{16-1} = \frac{1}{5}$
  - **148**:** $M = 4$, $M = 4$, $loss = \frac{4-1}{16-1} = \frac{1}{5}$
  - **1305*:** $M = 2$, $M = 2$, $loss = \frac{2-1}{16-1} = \frac{1}{15}$
  - **1306*:** $M = 2$, $M = 2$, $loss = \frac{2-1}{16-1} = \frac{1}{15}$

  loss of **Zip Code** is $\frac{\frac{1}{5}+\frac{1}{5}+\frac{1}{15}+\frac{1}{15}}{4} = \frac{2}{15}$.
- **Age:** $|A| = 15$
  - **<30:** $M = 4$, $loss = \frac{4-1}{15-1} = \frac{3}{14}$
  - **>40:** $M = 4$, $loss = \frac{4-1}{15-1} = \frac{3}{14}$
  - **3*:** $M = 4$, $loss = \frac{4-1}{15-1} = \frac{3}{14}$

  loss of **Age** is $\frac{\frac{3}{14}+\frac{3}{14}+\frac{3}{14}}{3} = \frac{3}{14}$.
- **Salary:** $|A| = 13$
  - ***:** $M = 12$, $loss = \frac{12-1}{13-1} = \frac{11}{12}$

  loss of **Salary** is $\frac{11}{12}$.
- **Nationality:** $|A| = 13$
  - ***:** $M = 12$, $loss = \frac{12-1}{13-1} = \frac{11}{12}$

  loss of **Nationality** is $\frac{11}{12}$.

So the **LM** for the entire data set is the sum of the losses for each attribute: $\frac{2}{15} + \frac{3}{14} + \frac{11}{12} + \frac{11}{12} \approx 2.181$.

2 L-Diversity

2.1 (a)

If the attributes meet recursive (2,2)-diversity, they should meet the following formula:

$$r_1 < 2(r_2 + r_3 + \cdots + r_m) \tag{1}$$

Firstly, we should count the frequency of each attribute value in each group:

Group1

$$\begin{aligned} n(q^*, HeartDisease) &= 1/4 \\ n(q^*, ViralInfection) &= 1/4 \\ n(q^*, Cancer) &= 2/4 \end{aligned} \tag{2}$$

Group2

$$\begin{aligned} n(q^*, HeartDisease) &= 1/4 \\ n(q^*, ViralInfection) &= 2/4 \\ n(q^*, Cancer) &= 1/4 \end{aligned} \tag{3}$$

Group3

$$\begin{aligned} n(q^*, HeartDisease) &= 1/4 \\ n(q^*, ViralInfection) &= 1/4 \\ n(q^*, Cancer) &= 2/4 \end{aligned} \tag{4}$$

We can find that:$\frac{2}{4} < 2(\frac{1}{4} + \frac{1}{4})$, and this is suitable for each group. So the attributes meet recursive (2,2)-diversity.

2.2 (b)

First we can define the entropy of l-diversity: Assume that the sequence of l-diversity in table T is $L = l_1, l_2, \cdots, l_m$:

$$H = \sum_{i=1}^{m} -l_i \log_2 l_i \tag{5}$$

After generalization, the sequence of l-diversity is $L' = l'_1, l'_2, \cdots, l'_m$ in table T^*: Assume that some items in L are merged, and the merged item in L' is l_i: Assume the items in L is $l_{j_1}, l_{j_2}, \cdots l_{j_k}$. Obviously, due to repeated sensitive information, the item l_i after merging is not as large as the sum of items before merging, as $l_i < \sum\limits_{r=1}^{k} l_{j_r}$.

For the function $f(x) = -x \log_2 x$, it can be proved that it is a convex function. Anyway, it has a property that:

$$f(\sum_{i=1}^{n} x_i) > \sum_{i=1}^{n} f(x_i) \tag{6}$$

So:

$$f(\sum_{r=1}^{k} l_{j_r}) > \sum_{r=1}^{k} f(l_{j_r}) \tag{7}$$

Then, when analysing l-diversity, the parameter l is larger than 1. Meanwhile, when $x > 1$, the function $f(x)$ is a decreasing function, so:

$$f(l_i) > f(\sum_{r=1}^{k} l_{j_r}) \tag{8}$$

Combining the above two inequalities formula (7) and (8), we can get:

$$f(l_i) > \sum_{r=1}^{k} f(l_{j_r}) \tag{9}$$

So, after merging, the entropy is larger than before merging which means T^* also satifies entropy l-diversity.

3 T-closeness

3.1 (a)

Numerical attribute values are ordered. Let the attribute domain be $v_1, v_2, \cdots, v_m$ and assume v_i is the i^{th} smallest value.

To calculate the EMD under ordered distance, we only need to consider the flow of mass transfer between adjacent elements, as any transfer between two more distant elements can be effectively decomposed into multiple transfers between adjacent elements. According to this conclusion, the minimum workflow f_{ij} can be achieved by sequentially satisfying all elements of Q. We can first consider **element 1** greedily, which has an extra amount of $p_1 - q_1$. Assume, without loss of generality, which represents that $p_1 - q_1 < 0$, an amount of $q_1 - p_1$ should be transported from other elements to **element 1** to make up for this increase. We can surely transport this from **element 2**. After this transportation, **element 1** is satisfied and **element 2** has an extra amount of $(p_1 - q_1) + (p_2 - q_2)$. Similarly, we can satisfy **element 2** by transporting an amount of $(p_1 - q_1) + (p_2 - q_2)$ between **element 2** and **element 3**. This process continues until element m is satisfied and Q is reached.

It should be noted that due to the two opposite cases ($(p_1 - q_1) + (p_2 - q_2) + \cdots + (p_i - q_i) > 0$ or < 0), the absolute value must be added when calculating each flow.

Formally, let $r_i = p_i - q_i, (i = 1, 2, ..., m)$, then the distance between P and Q can be calculated as:

$$\begin{aligned} D[\mathbf{P}, \mathbf{Q}] &= \frac{1}{m-1}(|r_1| + |r_1 + r_2| + \cdots + |r_1 + r_2 + \cdots + r_{m-1}|) \\ &= \frac{1}{m-1} \sum_{i=1}^{m} |\sum_{j=1}^{j=i} r_i| \end{aligned} \tag{10}$$

3.2 (b)

Assume that Q is the sequence of salary. Obviously, $Q = \{3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k\}$ which is ordered. The salary sequence corresponding to three groups are: $P_1 = \{4k, 5k, 6k\}$, $P_2 = \{3k, 8k, 11k\}$, $P_3 = \{7k, 9k, 10k\}$.

To calculate the EMD under ordered distance, we should search for the optimized flow between P_i and Q Tne optimal mass flow that transforms P1 to Q is to move 1/9 probability mass across the following pairs:

$$\begin{aligned} &(4k \to 3k), (4k \to 4k), (4k \to 5k) \\ &(5k \to 6k), (5k \to 7k), (5k \to 8k) \\ &(6k \to 9k), (6k \to 10k), (6k \to 11k) \end{aligned} \tag{11}$$

The cost of this flow is: $\frac{1}{9} \times (1+0+1+1+2+3+3+4+5)/(9-1) = 0.278$

Similarly, the cost of the optimal flow that transforms $P2$ to Q is: $\frac{1}{9} \times (0+1+2+2+1+0+2+1+0)/(9-1) = 0.125$. the cost of the optimal flow that transforms $P3$ to Q is: $\frac{1}{9} \times (4+3+2+3+2+1+1+0+1)/(9-1) = 0.236$.

According to three results, we can get the t-closeness of this table is 0.278. In other word, this table is 0.278-closeness.

4 Prior and posterior

4.1 (a)

Above all, to simplify the expression, we can reduce formula $P(R(x) = 0|x = k)$ to $P(R(k) = 0)$.

4.1.1 $x = 0$ (given $R_1(x) = 0$)

The prior and posterior probability of $x = 0$ (given $R_1(x) = 0$) is as follows:

By Bayesian formula and assume $P_1 = P(x = 0|R_1(x) = 0)$:

$$\begin{aligned} P_1 &= \frac{P(R_1(0) = 0) \cdot P(x = 0)}{P(R_1(0) = 0) \cdot P(x = 0) + \sum\limits_{k=1}^{100} P(R_1(k) = 0) \cdot P(x = k)} \\ &= \frac{(30\% + 70\% \times \frac{1}{101}) \times 0.01}{(30\% + 70\% \times \frac{1}{101}) \times 0.01 + \sum\limits_{k=1}^{100} 70\% \times \frac{1}{101} \times 0.099} \\ &= 30.9\% \end{aligned} \tag{12}$$

4.1.2 $x = 0$ (given $R_3(x) = 0$)

The prior and posterior probability of $x = 0$ (given $R_3(x) = 0$) is as follows:

Firstly, we can easily get the probability of $R_2(x) = 0$ when $x = k$:

$$\begin{aligned} P(R_2(k) = 0) &= P((k + \xi)mod(101)) \\ &= \begin{cases} P(k + \xi = 0) = 0 & 11 \leq k \leq 90 \\ P((k + \xi)mod(101) = 0) = \frac{1}{21} & 0 \leq k \leq 10 \\ P((k + \xi)mod(101) = 0) = \frac{1}{21} & 91 \leq k \leq 100 \end{cases} \end{aligned} \tag{13}$$

By Bayesian formula and assume $P_3 = P(x = 0|R_3(x) = 0)$:

$$\begin{aligned} P_3 &= \frac{P(R_3(0) = 0) \cdot P(x = 0)}{P(R_3(0) = 0) \cdot P(x = 0) + \sum\limits_{k=1}^{100} P(R_3(k) = 0) \cdot P(x = k)} \\ &= \frac{P(R_3(0) = 0)}{P(R_3(0) = 0) + \sum\limits_{k=1}^{100} P(R_3(k) = 0) \cdot 0.99} \end{aligned} \tag{14}$$

Among this formula:

$$\begin{aligned} P(R_3(0) = 0) &= 50\% \cdot P(R_2(0) = 0) + 50\% \times \frac{1}{101} \\ &= 50\% \times \frac{1}{21} + 50\% \times \frac{1}{101} \end{aligned} \tag{15}$$

$$\begin{aligned} P(R_3(k) = 0) &= 50\% \cdot P(R_2(k) = 0) + 50\% \times \frac{1}{101} \\ &= \begin{cases} 50\% \times \frac{1}{21} + 50\% \times \frac{1}{101} & 0 \leq k \leq 10 \\ 50\% \times \frac{1}{21} + 50\% \times \frac{1}{101} & 91 \leq k \leq 100 \\ 50\% \times \frac{1}{101} & 11 \leq k \leq 90 \end{cases} \end{aligned} \tag{16}$$

Substitute into formula (14):

$$P_3 = 2.9\% \tag{17}$$

4.1.3 $x \in [20, 80]$ (given $R_2(x) = 0$)

The prior and posterior probability of $x = i \in [20, 80]$ (given $R_2(x) = 0$) is as follows:

By Bayesian formula and assume $P_2 = P(x = i|R_2(x) = 0)$:

$$\begin{aligned} P_2 &= \frac{P(R_2(x) = 0|x = i) \cdot P(x = i)}{P(R_2(0) = 0) \cdot P(x = 0) + \sum\limits_{k=1}^{100} P(R_2(k) = 0) \cdot P(x = k)} \\ &= 0 \end{aligned} \tag{18}$$

4.2 (b)

R_3 is more suitable, the probability distribution is not as biased as R_1 (as $P(x = 0|R_1(x) = 0)$ is 70%, so large that is easy to guess), nor is the data range as small as R_2.

5 K-anonymity in graphs

5.1 (a)

Connect an edge between **Bob** and **Lucy**, like this: the degree sequence of the graph is $\{2, 3, 4, 4, 3, 2\}$ which is 2-anonymous.

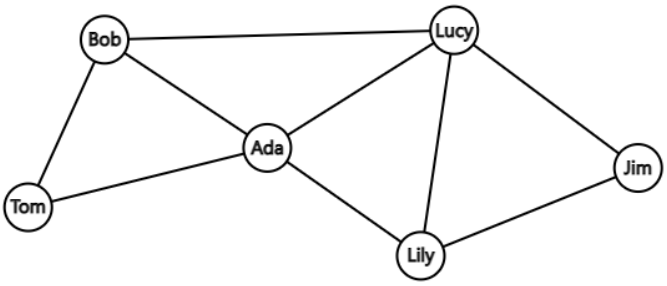

Figure 3: 2-Anonymous Graph

5.2 (b)

Because the total degree will increase by 2 because of an additional edge, and the total degree is now 16. So assuming that the final degree sequence consists of three a and three b, then $3(a+b)$ is an even number, that is, $a+b$ is an even number, taking a as 4 and b as 6, the result is as shown in the figure below:

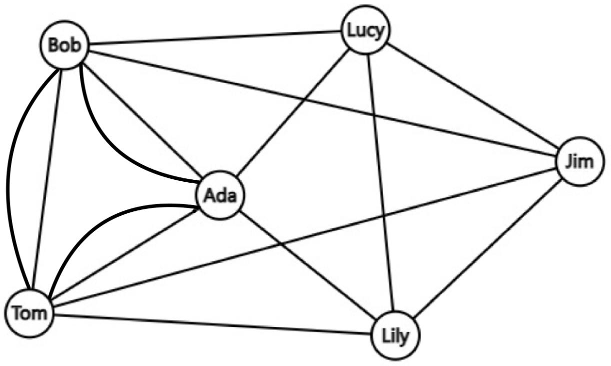

Figure 4: 3-Anonymous Graph

the degree sequence of the graph is $\{6, 6, 6, 4, 4, 4\}$ which is 3-anonymous.

5.3 (c)

In section (a), the information loss is:

$$\begin{aligned} L(G, G') &= 1 - \frac{|\Delta(E, E')|}{\max(|E|, |E'|)} \\ &= 1 - \frac{8}{9} = \frac{1}{9} \end{aligned} \tag{19}$$

In section (b), the information loss is:

$$\begin{aligned} L(G, G') &= 1 - \frac{|\Delta(E, E')|}{\max(|E|, |E'|)} \\ &= 1 - \frac{8}{15} = \frac{7}{15} \end{aligned} \tag{20}$$