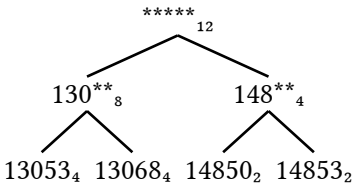
Data Privacy Homework 1

PB20000051

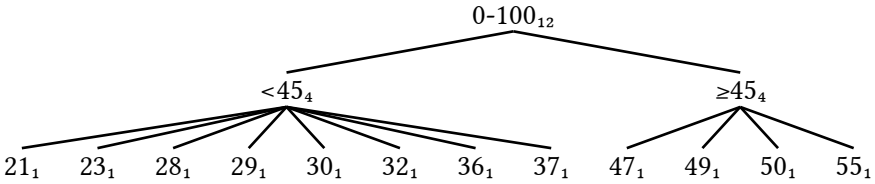
1.K-anonymity

1. The quasi-identifier attributes are **Zip Code**, **Age**, **Salary** and **Nationality**.
2. The **generalization hierarchies** are shown below. Note that the subscripted number is the item count.

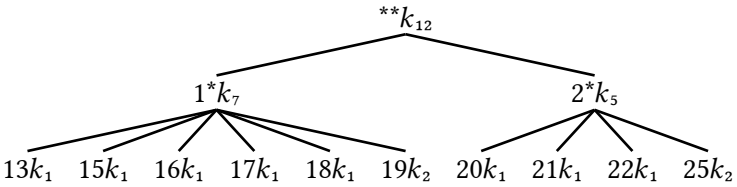
Zip Code



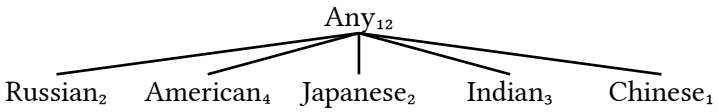
Age



Salary



Nationality



Based on the generalization hierarchies above, a cell-level generalization solution to achieve 2-anonymity can be designed. The **released table** is shown below.

	Non-Sensitive				Sensitive
	Zip Code	Age	Salary	Nationality	Condition
1	13***	<45	1*k	Any	Heart Disease
3	13***	<45	1*k	Any	Viral Infection
4	13***	<45	1*k	Any	Viral Infection
9	13***	<45	1*k	Any	Cancer
11	13***	<45	1*k	Any	Cancer
10	13***	<45	2*k	Any	Cancer
12	13***	<45	2*k	Any	Cancer
2	13***	<45	2*k	Any	Heart Disease

6	14***	≥45	1*k	Any	Heart Disease
7	14***	≥45	1*k	Any	Viral Infection
5	14***	≥45	2*k	Any	Cancer
8	14***	≥45	2*k	Any	Viral Infection

To calculate the **loss metric** (LM) of this solution, first calculate the losses of each attribute:

$$\begin{aligned}
\mathrm{LM}_{\text{Zip Code}} &= \frac{1}{12}\left(\frac{2-1}{4-1}\times 8+\frac{2-1}{4-1}\times 4\right)=\frac{1}{3}\\
\mathrm{LM}_{\text{Age}} &= \frac{1}{12}\left(\frac{44-0+1}{100-0+1}\times 8+\frac{100-45+1}{100-0+1}\times 4\right)=\frac{146}{303}\\
\mathrm{LM}_{\text{Salary}} &= \frac{1}{12}\left(\frac{6-1}{10-1}\times 7+\frac{4-1}{10-1}\times 5\right)=\frac{25}{54}\\
\mathrm{LM}_{\text{Nationality}} &= 1
\end{aligned}$$

The LM for the entrie data set is defined as the sum of the losses for each attribute, which is:

$$\mathrm{LM}=\mathrm{LM}_{\text{Zip Code}}+\mathrm{LM}_{\text{Age}}+\mathrm{LM}_{\text{Salary}}+\mathrm{LM}_{\text{Nationality}}=\frac{12425}{5454}$$

2.L-Diversity

1. For each q^*-block, the sorted sensitive attribute count sequences are all (2, 1, 1), which satisfies $r_1 = 2 < 2(r_2 + r_3) = 4$. Thus, the attributes in the figure meet recursive (2, 2)-diversity.
2. Say there is a q^*-block q^{**} merged from q_1^* and q_2^* in table T.

   Since T satisfies entropy ℓ-diversity, the entropy of q_1^* and q_2^* is greater than $\log(\ell)$. Thus,

$$\begin{aligned}
\text{entropy}(q_1^*) &= -\sum_{s\in S} p(q_1^*, s)\log(p(q_1^*, s)) \geq \log(\ell)\\
\text{entropy}(q_2^*) &= -\sum_{s\in S} p(q_2^*, s)\log(p(q_2^*, s)) \geq \log(\ell)
\end{aligned}$$

   where $p(q_n^*, s) = \dfrac{n(q_n^*, s)}{\sum_{s'\in S} n\left(q_n^*, s'\right)}$.

   Let $\boldsymbol{P}(q_n^*, \{s_1, \cdots, s_m\}) = \left(p(q_n^*, s_1) \ \cdots \ p(q_n^*, s_m)\right)^{\mathrm{T}}$, $f(\boldsymbol{X}) = -\sum_{x\in\boldsymbol{X}} x\log(x)$. Then the entropy of q_n^* equals $f(\boldsymbol{P}(q_n^*, S))$. Since $f(\boldsymbol{X})$ is a concave function, which means

$$\forall\alpha\in[0,1], f((1-\alpha)\boldsymbol{X}+\alpha\boldsymbol{Y}) \geq (1-\alpha)f(\boldsymbol{X})+\alpha f(\boldsymbol{Y})$$

   And the distribution of the merged block q^{**} satisfies

$$\boldsymbol{P}(q^{**}, S) = \frac{n(q_1^*)}{n(q_1^*)+n(q_2^*)}\boldsymbol{P}(q_1^*, S)+\frac{n(q_2^*)}{n(q_1^*)+n(q_2^*)}\boldsymbol{P}(q_2^*, S)$$

   Thus,

$$\begin{aligned}
&f(\boldsymbol{P}(q^{**}, S)) \geq \frac{n(q_1^*)}{n(q_1^*)+n(q_2^*)}f(\boldsymbol{P}(q_1^*, S))+\frac{n(q_2^*)}{n(q_1^*)+n(q_2^*)}f(\boldsymbol{P}(q_2^*, S))\\
\Leftrightarrow\ &\text{entropy}(q^{**}) \geq \frac{n(q_1^*)}{n(q_1^*)+n(q_2^*)}\text{entropy}(q_1^*)+\frac{n(q_2^*)}{n(q_1^*)+n(q_2^*)}\text{entropy}(q_2^*)\\
\Rightarrow\ &\text{entropy}(q^{**}) \geq \min(\text{entropy}(q_1^*), \text{entropy}(q_2^*)) \geq \log(\ell)
\end{aligned}$$

In other words, the merged block q^{**} satisfies entropy ℓ-diversity.

For any table T^* generalized from T, it is always obtained from table T through a finite number of q^*-block merges. Thus, the minimal entropy of T^* would never be less than the entropy of T, which implies T^* satisfies entropy ℓ-diversity.

3.T-Clossness

1. The EMD between $\boldsymbol{P}$ and $\boldsymbol{Q}$ is $D[\boldsymbol{P},\boldsymbol{Q}] = \min_{\boldsymbol{F}} \text{WORK}(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{F}) = \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{|i-j|}{m-1}f_{ij}$, where flow $\boldsymbol{F}$ satisfies

$$f_{ij} \geq 0$$
$$r_i = p_i - q_i = \sum_{j=1}^{m}(f_{ij} - f_{ji})$$
$$\sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij} = 1$$

For convience, we only need to consider flows that transport distribution between adjacent elements, since any transportation between further elements can be equivalently decomposed into several transportations between adjacent elements. So WORK(·) can be simplified as $\text{WORK}(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{F}) = \frac{1}{m-1}\sum_{i=1}^{m}(f_{i,i-1} + f_{i,i+1})$, where the flow $\boldsymbol{F}$ satisfies

$$f_{ij} \geq 0, f_{0,*} = f_{*,0} = f_{m+1,*} = f_{*,m+1} = 0$$
$$r_i = p_i - q_i = f_{i,i-1} + f_{i,i+1} - f_{i-1,i} - f_{i+1,i}$$
$$\sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij} = 1$$

To minimize WORK(·), it is obivious that one of f_{ij} and f_{ji} is zero. Thus, expanding the sum in WORK and pairing each (f_{ij}, f_{ji}) gives

$$\begin{aligned}
\min_{\boldsymbol{F}} \text{WORK}(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{F}) &= \min_{\boldsymbol{F}} \frac{1}{m-1}(f_{12} + f_{21} + f_{23} + \cdots + f_{m-1,m} + f_{m,m-1}) \\
&= \min_{\boldsymbol{F}} \frac{1}{m-1}\sum_{i=1}^{m}(f_{i+1,i} + f_{i,i+1}) \\
&= \min_{\boldsymbol{F}} \frac{1}{m-1}\sum_{i=1}^{m}|f_{i+1,i} - f_{i,i+1}| \\
&= \min_{\boldsymbol{F}} \frac{1}{m-1}\sum_{i=1}^{m}\left|\sum_{j=1}^{i}(f_{j,j-1} + f_{j,j+1} - f_{j-1,j} - f_{j+1,j})\right| \\
&= \frac{1}{m-1}\sum_{i=1}^{m}\left|\sum_{j=1}^{i} r_j\right| = \frac{1}{m-1}(|r_1| + |r_1 + r_2| + \cdots + |r_1 + \cdots + r_{m-1}|)
\end{aligned}$$

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

2. The overall distribution of Salary is $\boldsymbol{Q} = \frac{1}{9}(1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1)$, each element represents $\{3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k\}$ respectively. For each QI group, calculate the EMD as below.
   - In the first QI group (Zip Code 4767* and etc.), the distribution is $\boldsymbol{P}_1 = \frac{1}{3}(0\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0)$. So the EMD is $D[\boldsymbol{P}_1,\boldsymbol{Q}] = \frac{1}{8}\left(\frac{1}{9} + \frac{1}{9} + \frac{3}{9} + \frac{5}{9} + \frac{4}{9} + \frac{3}{9} + \frac{2}{9} + \frac{1}{9}\right) = \frac{5}{18}$.

- In the second QI group (Zip Code 4790* and etc.), the distribution is $\boldsymbol{P}_1 = \frac{1}{3}(1\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1)$. So the EMD is $D[\boldsymbol{P}_1, \boldsymbol{Q}] = \frac{1}{8}\left(\frac{2}{9} + \frac{1}{9} + 0 + \frac{1}{9} + \frac{2}{9} + 0 + \frac{1}{9} + \frac{2}{9}\right) = \frac{1}{8}$.
- In the third QI group (Zip Code 4760* and etc.), the distribution is $\boldsymbol{P}_1 = \frac{1}{3}(0\ 0\ 0\ 0\ 1\ 0\ 1\ 1\ 0)$. So the EMD is $D[\boldsymbol{P}_1, \boldsymbol{Q}] = \frac{1}{8}\left(\frac{1}{9} + \frac{2}{9} + \frac{3}{9} + \frac{4}{9} + \frac{2}{9} + \frac{3}{9} + \frac{1}{9} + \frac{1}{9}\right) = \frac{17}{72}$.

Therefore, the value of t should be greater than $\frac{5}{18}$, i.e., $t \geq \frac{5}{18}$.

4.Prior and Posterior

1. **Prior and posterior probabilities of $x = 0$ given $R_1(x) = 0$**

   The **prior** probability of $x = 0$ is $\mathbf{P}[x = 0] = 0.01$. Given $R_1(x) = 0$, the **posterior** probability of $x = 0$ is

$$\begin{aligned}
\mathbf{P}[x = 0 \mid R_1(x) = 0] &= \frac{\mathbf{P}[R_1(x) = 0 \mid x = 0]\mathbf{P}[x = 0]}{\mathbf{P}[R_1(x) = 0]} \\
&= \frac{\mathbf{P}[R_1(x) = 0 \mid x = 0]\mathbf{P}[x = 0]}{\sum_{i=0}^{100} \mathbf{P}[R_1(x) = 0 \mid x = i]\mathbf{P}[x = i]} \\
&= \frac{0.3 \times 0.01}{0.3 \times 0.01 + 100 \times 0.007 \times 0.0099} \\
&= \frac{100}{331} \approx 0.302
\end{aligned}$$

   Prior and posterior probabilities of $x \in [20, 80]$ given $R_2(x) = 0$

   The **prior** probability of $x \in [20, 80]$ is $\mathbf{P}[x \in [20, 80]] = 0.0099 \times 61 = 0.6039$. Given $R_2(x) = 0$, the **posterior** probability of $x \in [20, 80]$ given $R_2(x) = 0$ is

$$\begin{aligned}
\mathbf{P}[x \in [20, 80] \mid R_2(x) = 0] &= \frac{\mathbf{P}[R_2(x) = 0 \mid x \in [20, 80]]\mathbf{P}[x \in [20, 80]]}{\mathbf{P}[R_2(x) = 0]} \\
&= \frac{0}{\mathbf{P}[R_2(x) = 0]} \\
&= 0
\end{aligned}$$

   Prior and posterior probabilities of $x = 0$ given $R_3(x) = 0$

   The **prior** probability of $x = 0$ is $\mathbf{P}[x = 0] = 0.01$. Given $R_3(x) = 0$, the **posterior** probability of $x = 0$ is

$$\begin{aligned}
\mathbf{P}[x = 0 \mid R_3(x) = 0] &= \frac{\mathbf{P}[R_3(x) = 0 \mid x = 0]\mathbf{P}[x = 0]}{\mathbf{P}[R_3(x) = 0]} \\
&= \frac{\mathbf{P}[R_3(x) = 0 \mid x = 0]\mathbf{P}[x = 0]}{\frac{1}{2}\sum_{i \in [0,10] \cup [91,100]} \mathbf{P}[R_2(x) = 0 \mid x = i]\mathbf{P}[x = i] + \frac{1}{202}} \\
&= \frac{\frac{1}{2} \times \frac{1}{21} \times 0.01 + \frac{1}{202}}{\frac{1}{2} \times (20 \times 0.0099 + 0.01) \times \frac{1}{21} + \frac{1}{202}} \\
&= \frac{11005}{21004} \approx 0.524
\end{aligned}$$

2. If we want to preserve better privacy, the method with less information loss is more suitable, because it makes less difference between prior and posterior probability. We can compute the distance (e.g., KL-divergence and **Hellinger Distance**) between prior and (each) posterior probability distribution as

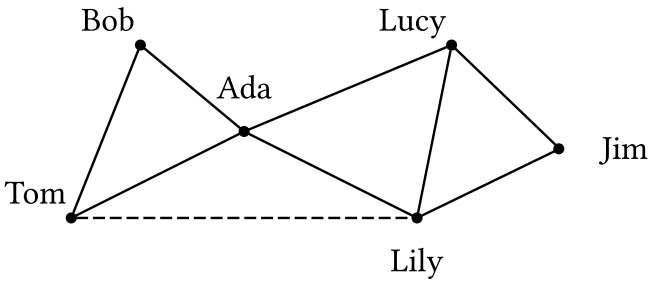
$$D_i\left[\boldsymbol{P}_{\text{prior}}, \boldsymbol{P}_{\text{post}}\right] = \sqrt{\sum_{X \in \{\{0\},[200,800],[1,199]\cup[801,1000]\}} \left(\sqrt{\mathbf{P}[x \in X]} - \sqrt{\mathbf{P}[x \in X \mid R_i(x) = 0]}\right)^2 / 2}$$

$$D_1 \approx 0.62, \quad D_2 \approx 0.60, \quad D_3 \approx 0.22$$

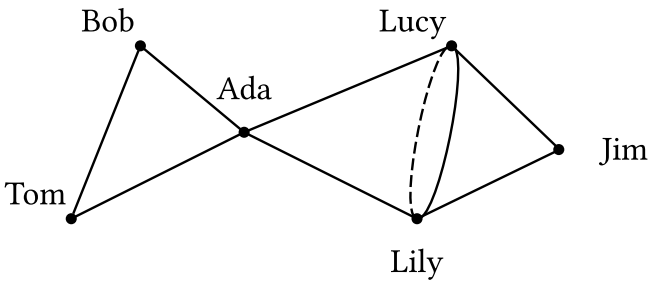
So R_3 is more suitable.

5.K-Anonymity in Graphs

1. After adding edge Tom-Lily, the degree sequence of the graph become $\{2, 3, 4, 3, 4, 2\}$, making the graph 2-anonymous.



2. After adding edge Lucy-Lily, the the degree sequence of the graph become $\{2, 2, 4, 4, 4, 2\}$, making the graph 3-anonymous.



3. For the anonymized graph in section (a), the information loss is

$$L\left(G, G_a^{'}\right) = 1 - \frac{8}{9} = \frac{1}{9}$$

For the anonymized graph in section (b), the information loss is

$$L\left(G, G_b^{'}\right) = 1 - \frac{8}{9} = \frac{1}{9}$$