Methodology, Ethics and Practice of Data Privacy
Course Exercise #1

March 13 2020

1. Given the following private table:

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

Table 1: Private table.

Please answer the following questions:

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

(b) Let the valid range of age be $\{0, \cdots, 120\}$. Given the health condition as the sensitive attribute, design a cell-level generalization solution to achieve k-Anonymity, where $k = 4$. Please give the generalization hierarchies, released table and calculation of the loss metric (LM) of your solution.

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

2. 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:

$$P[X = 0] = 0.01, P[X = k] = 0.00099, k = 1 \cdots 1000 \quad (1)$$

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

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

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

Please answer the following questions:

(a) Compute prior and posterior probabilities of two properties of x: 1) $X = 0$; 2) $x \in \{200, \cdots, 800\}$ using the above three methods respectively.

(b) Which method is better? Why?

3. [(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,

$$\exists u \in D_U, \exists v \in D_V \ s.t.\ P_f(\Phi(u)) \leq \alpha \ and \ P_f(\Phi(u)|R(u) = v) \geq \beta \quad (2)$$

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

$$\exists u \in D_U, \exists v \in D_V \ s.t.\ P_f(\Phi(u)) \geq \alpha \ and \ P_f(\Phi(u)|R(u) = v) \leq \beta \quad (3)$$

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

$$\forall v \in D_V, \forall u_1, u_2 \in D_U, \frac{P(R(u_1) = v)}{P(R(u_2) = v)} \leq \gamma \quad (4)$$

(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

$$\gamma \leq \frac{\beta}{\alpha} \frac{1 - \alpha}{1 - \beta}. \quad (5)$$