The University of Newcastle
School of Electrical Engineering and Computer Science

COMP3260/COMP6360 Data Security
Week 12 Additional Exercises

Job	Sex	Age	Diagnosis
Engineer	Female	31	Fracture
Scientist	Female	33	Flu
Scientist	Female	35	HIV
Lawyer	Female	32	Flu
Doctor	Female	31	Flu
Cricketer	Male	23	Fracture
Cricketer	Male	25	Fracture
Golfer	Male	20	HIV

Table 1

Name	Job	Sex	Age
Anne	Engineer	Female	31
Betty	Scientist	Female	33
Claire	Scientist	Female	35
Donna	Lawyer	Female	32
Andrew	Doctor	Female	31
Bob	Cricketer	Male	23
Charlie	Cricketer	Male	25
Dennis	Golfer	Male	20
Peter	Doctor	Male	33
David	Lawyer	Male	32
Mark	Engineer	Male	24

Table 2

1. Suppose that a hospital has removed patients' names from the hospital records and intends to make these 'anonymised' patients' records in Table 1 available to a researcher. Suppose that the researcher has access to the external table Table 2 and knows that every person with a record in Table 1 has a record in Table 2. Would this lead to record or attribute linkage of hospital patients? Which patients would have their privacy compromised? With what probability can an adversary infer that Betty has HIV?

Solution:

Yes, this would lead to both attribute and record linkage.

Consider, for example, the first record in table 1. There is only one record in table 2 (the first one) that matches record 1 in table 1 in all attributes of the quasi-identifier {Job, Sex, Age}. Therefore, the first record in table 1 belongs to Anne (record linkage) and she has been treated for fracture (attribute

linkage). Therefore, her privacy has been compromised as the researcher can learn her confidential value (diagnosis) with certainty.

Since the above is the case for every record in Table 1, all patients in table 1 have their privacy compromised.

By linking Table 1 and Table 2, an adversary can learn that Betty has been treated for flu, therefore an adversary can learn that Betty has HIV with 0% (with the assumption that all comorbidities are listed in attribute "Diagnosis" in Table 1.

2. Generate k-anonymous tables from Table 1 and Table 2. What is the highest k you can achieve for each table? Where do you think the best trade-off lies?

Solution:

Using generalization, we can obtain the following 3- anonymous Table 1. It is also possible to obtain 2-anonymous, 4-anonymous and 8-anonymous tables, although in each of them the information loss may be higher than for 3-anonymity and/or attribute linkage may be possible. Therefore, the highest k is equal to 8, the number of records in the data set.

Job	Sex	Age	Diagnosis
Professional	Female	31-35	Fracture
Professional	Female	31-35	Flu
Professional	Female	31-35	HIV
Professional	Female	31-35	Flu
Professional	Female	31-35	Flu
Sportsperson	Male	20-25	Fracture
Sportsperson	Male	20-25	Fracture
Sportsperson	Male	20-25	HIV

Table 3 - Table 1 generalised to be 3-anonymous

Applying generalisation to table 2, we can make a 3-anonymous version:

Name	Job	Sex	Age
-	Professional	Female	31-35
-	Professional	Female	31-35
-	Professional	Female	31-35
-	Professional	Female	31-35
-	Professional	Female	31-35
-	Sportsperson	Male	20-25
-	Sportsperson	Male	20-25
-	Sportsperson	Male	20-25
-	Professional	Male	31-35
-	Professional	Male	31-35
-	Professional	Male	20-25

Table 4-Table 2 generalised to be 3-anonymous

The highest k that can be obtained for table 2 is 11, which is the number of records in the table.

3. Consider the anonymous data generated in Problem 2. Suppose the adversary knows that the target victim Betty is a scientist of age 30 and has a record in Table 1 and Table 2. With what probability can an adversary infer that Betty has HIV? Compare this with the case when data was not k-anonymised.

Solution:

If the adversary knows that Betty is a 33 years old scientist in Table 1 and Table 2, then they know her record is one of the first 5 records in Table 1 and Table 2. Out of the 5 records, there is only one with the value HIV, so the adversary can infer that Betty has HIV with 20% chance. Note that 20% is better than the 0% that we obtained in question 1, as previously the adversary knew for sure she did not have HIV and now they are not sure. The precise measure of how much the adversary can learn can be measured using entropy (see next question.)

4. Consider the Anonymous data generated in question 2. Calculate ℓ for Table 1 using distinct ℓ-diversity and entropy ℓ-diversity.

Solution:

Distinct ℓ-diversity:

There should be at least ℓ distinct values in each equivalence class.

Using 3-anonymous Table 1:

- In the first equivalence class {Professional, Female, 31-35} there are 3 distinct values
- In the second equivalence class {Sportsperson, Male, 20-25} there are 2 distinct values

We take the minimum of the two, so we have 2-diversity.

Entropy ℓ-diversity:

We compute adversary's uncertainty about the sensitive value for each equivalence class and take the minimum of these values.

Using 3-anonymous Table 1:

For the first equivalence class {Professional, Female, 31-35}, there are 3 distinct values with the following probabilities: $p(\text{fracture}) = \frac{1}{5}$, $p(\text{flu}) = \frac{3}{5}$ and $p(\text{HIV}) = \frac{1}{5}$; therefore, the entropy is

$$H(X) = \frac{1}{5}\log_2 5 + \frac{3}{5}\log_2 \frac{5}{3} + \frac{1}{5}\log_2 5$$

We calculate ℓ such that $\log_2\ \ell \leq H(X)$:

$$\begin{aligned}
&\log_2\ \ell \leq H(X) \\
&\log_2\ \ell \leq \frac{1}{5}\log_2 5 + \frac{3}{5}\log_2 \frac{5}{3} + \frac{1}{5}\log_2 5 \\
&\log_2\ \ell \leq 1.37 \\
&\ell \leq 2^{1.37} \\
&\ell \leq 2.58
\end{aligned}$$

For the second equivalence class we have $p(\text{fracture}) = \frac{2}{3}$, $p(\text{HIV}) = \frac{1}{3}$. So calculating ℓ we get:

$$\begin{aligned} \log_2 \ \ell &\leq H(X) \\ \log_2 \ \ell &\leq \frac{2}{3}\log_2 \frac{3}{2} + \frac{1}{3}\log_2 3 \\ \log_2 \ \ell &\leq 0.92 \\ \ell &\leq 2^{0.92} \\ \ell &\leq 1.89 \end{aligned}$$

For the whole of the 3-anonymous version of Table 1, we have $\ell \leq 1.89$ (since that satisfies both inequalities).