The University of Newcastle
School of Electrical Engineering and Computer Science

COMP3260/COMP6360 Data Security
Week 12 Workshop – 30 & 31 May 2019
Solutions

Privacy

1. For each of the following attack models, describe the Attack Model, and name a Privacy Model that addresses that kind of attack.

a. Record Linkage
b. Attribute Linkage
c. Table Linkage

Solution:

a. For a record linkage attack the attacker manages to uniquely identify the target individual's row in the dataset. A privacy model that address record linkage is k-anonymity.

b. For an attribute linkage attack the attacker cannot uniquely identify the target individual's row in the dataset but is still able to learn some sensitive attribute of the individual. A privacy model that address record linkage is ℓ-diversity.

c. For a table linkage attack the attacker does not manage to identify the target individual's row in the dataset but is able to determine that the target individual is in the dataset. A privacy model that address record linkage is ε-differential privacy.

2. One set of techniques for privacy involves restricting access to the dataset: query set size control, query set overlap control, maximum order control, partitioning, cell suppression and auditing. In this context, what is partitioning?

Solution:

In the context of restricting access to a dataset, the technique of partitioning involves partitioning the dataset into groups of records. When a query is made of the dataset, the answer either contains all rows in the group of records, or none of the rows from that record. This is the approach of the k-anonymity model – it generalises the values of the quasi-identifiers such that for every group of quasi-identifiers, there are at least k records with the same quasi-identifiers.

3. In information theory, what is entropy, and how is it calculated? What is equivocation (conditional entropy), and how is it calculated?

Solution:

Entropy is a measure of uncertainty (or information content). Entropy can be interpreted as the minimum number of bits required to encode the possible outcomes of an event (messages). Entropy is calculated as follows:

$$H(X) = -\left(\sum_{i=1}^{n} p(X_i)\ \log_2 p(X_i)\right)$$
$$= \sum_{i=1}^{n} p(X_i) \log_2 \frac{1}{p(X_i)}$$

Equivocation (also called Conditional Entropy) is a measure of how many bits is required to encode the outcome of event Y given that event X has happened. Equivocation is calculated as follows:

$$H_Y(X) = -\left(\sum_{X,Y} p(X,Y) \log_2 p_y(X)\right)$$
$$= \sum_{Y} p(Y) \sum_{X} p_y(X) \log_2 \frac{1}{p_y(X)}$$

4. Consider the following dataset:

a. Categorise the attributes into Identifier, Quasi-identifier, Non-sensitive attribute and Sensitive attribute

b. What level of k-anonymity is achieved by the original table? (What is the smallest equivalence class?)

c. Create your own taxonomy and generalize the data values so that 4-anonymity is achieved.

d. Find ℓ so that the anonymized data set achieves ℓ-diversity.

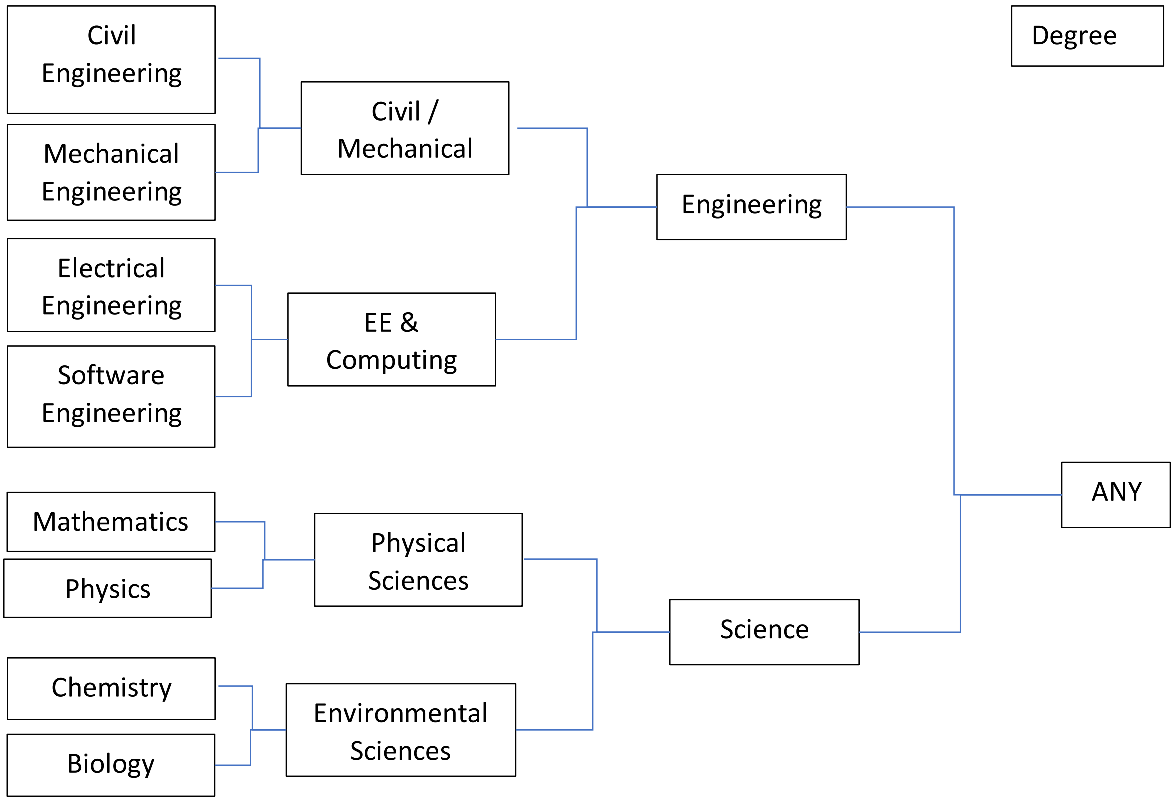
Degree	Sex	Name	Age	Average grade
Civil Engineering	Female	Anne	20	HD
Electrical Engineering	Female	Betty	23	D
Mechanical Engineering	Female	Claire	25	D
Software Engineering	Female	Donna	22	HD
Mathematics	Male	Andrew	21	C
Chemistry	Male	Bob	23	HD
Biology	Male	Charlie	25	HD
Physics	Male	Dennis	20	D

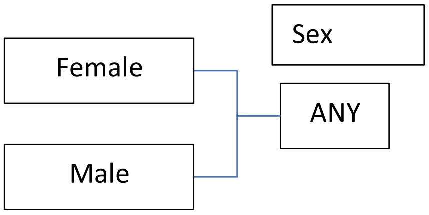
Solution: Note that there are multiple possible answers for parts (a) and (c) of question 4

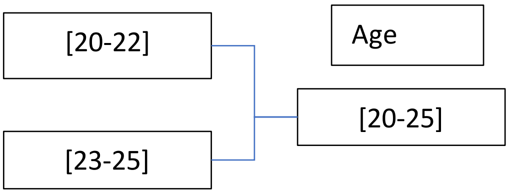
a. Identifier: Name. Quasi-identifier: Degree, Sex, Age. Sensitive attribute: Average Grade

b. The original table has a k value of 1, i.e. it achieves 1-anonymity, which is not at all.

c.







Degree	Sex	Name	Age	Average grade
Engineering	Female	-	[20-25]	HD
Engineering	Female	-	[20-25]]	D
Engineering	Female	-	[20-25]	D
Engineering	Female	-	[20-25]	HD
Science	Male	-	[20-25]	C
Science	Male	-	[20-25]	HD
Science	Male	-	[20-25]	HD
Science	Male	-	[20-25]	D

d.

We will first calculate distinct ℓ-diversity:

The equivalence class (Engineering, Female, [20-25]) contains 2 values for the sensitive attribute "Average Grade"

The equivalence class (Science, Male, [20-25]) contains 3 values for the sensitive attribute "Average Grade"

For distinct ℓ-diversity, we have $\ell = 2$

We now calculate entropy ℓ-diversity:

The equivalence class (Engineering, Female, [20-25]):

$$\begin{aligned}
-\left(\sum_{i=1}^{n} p(X_i)\ \log_2 p(X_i)\right) &\geq \log_2 \ell \\
-\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) &\geq \log_2 \ell \\
-\left(\log_2\frac{1}{2}\right) &\geq \log_2 \ell \\
1 &\geq \log_2 \ell \\
2^1 &\geq \ell \\
2 &\geq \ell
\end{aligned}$$

The equivalence class (Science, Male, [20-25]):

$$\begin{aligned}
-\left(\sum_{i=1}^{n} p(X_i)\ \log_2 p(X_i)\right) &\geq \log_2 \ell \\
-\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4}\right) &\geq \log_2 \ell \\
-\left(\frac{1}{2}\log_2\frac{1}{4} + \frac{1}{2}\log_2\frac{1}{2}\right) &\geq \log_2 \ell \\
1.5 &\geq \log_2 \ell \\
2^{1.5} &\geq \ell \\
2.8284 &\geq \ell
\end{aligned}$$

For entropy ℓ-diversity, we have $\ell \leq 2$.

5.

a. What is the basic idea behind ε-differential privacy? What problem is it addressing?

b. If we have $P(F(T_1) = S) = 0.5$ and $P(F(T_2) = S) = 0.4$, for $\varepsilon = 1$ then is the ε-differential privacy model satisfied for that particular query?

c. If we have $P(F(T_1) = S) = 0.8$ and $P(F(T_2) = S) = 0.4$, for $\varepsilon = 1$ then is the ε-differential privacy model satisfied for that particular query?

Solution:

a.

The main idea is to switch from an absolute protection from disclosure to a relative protection of disclosure. We do not attempt to claim that an adversary cannot successfully obtain any information about an individual from the published data, but we can claim that what an adversary can obtain about an individual does not significantly change if the individual participates in the database or not.

One way to think about this is that it is protecting against a table linkage attack. It becomes difficult for an adversary to know if an individual is in a dataset or not.

One motivating factor behind ε-differential privacy is to encourage participation in a database by being able to give confidence that participating in a database will not substantially increase the risk to an individual. This is the balance of privacy and utility – the goal of ε-differential privacy is to allow aggregate trends in the population can be observed without an adversary being able to extract or deduce the data of any particular user.

b.

$$\left|\ln\frac{P(F(T_1)=S)}{P(F(T_2)=S)}\right| \leq \varepsilon$$
$$\left|\ln\frac{0.5}{0.4}\right| \leq 1$$
$$0.2231 \leq 1$$

The equation holds, so it is satisfied for this particular S

c.

$$\left|\ln\frac{P(F(T_1)=S)}{P(F(T_2)=S)}\right| \leq \varepsilon$$
$$\left|\ln\frac{0.8}{0.4}\right| \leq 1$$
$$0.6931 \leq 1$$

The equation holds, so it is satisfied for this particular S