COSC130 Fundamentals of Cybersecurity and Privacy

Tutorial Week 5 Solutions

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 2 and knows that every person with a record in Table 1 has a record in Table 2.
    a. Would this lead to record or attribute linkage of hospital patients?
    b. Which patients would have their privacy compromised?
    c. What is the probability that Betty has HIV?

Job	Sex	Age	Disease
Engineering	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	Engineering	Female	31
Betty	Scientist	Female	33
Claire	Scientist	Female	35
Donna	Lawyer	Female	32
Anna	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

Solution:

a. 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.

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

c. 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% (assuming that all comorbidities are listed in attribute "Disease" in Table 1).

2. Consider Table 1 and Table 2.
    a. Identify the unique identifier ID and quasi identifier QID for both Table 1 and Table 2.
    b. Generate 3-anonymous tables from Table 1 and Table 2.
    c. What is the highest k you can achieve for each table?

Solution:

a. Table 1 does not contain a unique identifier as it has already been removed. The quasi identifier for Table 1 is QID_1={Job, Sex, Age}, as these attributes also exist in table 2 and can be used to link the two tables.

   In Table 2, the unique identifier as ID_2={Name} and the quasi identifier is QID_1={Job, Sex, Age}, as these attributes also exist in Table 1 and can be used to link the two tables.

b. Using generalization, we obtain the following 3-anonymus Table 1 and Table 2.

Job	**Sex**	**Age**	**Disease**
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

Name	**Job**	**Sex**	**Age**
-	Professional	Female	31-35
-	Professional	Female	31-35
-	Professional	Female	31-35
-	Professional	Female	31-35
-	Professional	Female	31-35
-	Sportsman	Male	20-25
-	Sportsman	Male	20-25
-	Sportsman	Male	20-25
-	Professional	Male	24-33
-	Professional	Male	24-33
-	Professional	Male	24-33

c. For Table 1, the highest k is equal to 8, the number of record in the data set. The highest k that can be obtained for Table 2 is 11, 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 33 and has a record in the Table 1. 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, then they her record is one of the first 5 records in Table 1 and Table 2. Out of 5 (not necessarily distinct) values, there is only one HIV so the adversary can learn that Betty has HIV with 20% chance. Note that 20% provides better protection than 0% that we obtained in question 1, as in that case the adversary knew as a fact that Betty did not have HIV and now they cannot be sure.

4. Perform independent research to answer the following questions.
   a. What are data brokers?
   b. What is the difference between first-party data brokers and third-party data brokers?
   c. What are the main products and services offered by third-party data brokers in Australia?
   d. Identify at least 5 third-party data brokers operating in Australia.
   e. Identify at least 5 sources that third-party data brokers use to collect information.
   f. Identify at least 5 ways in which the collections and use of personal information by third-party data brokers can harm individuals.

Solution:

a. Data brokers are businesses who collect personal information about individuals and sell it to others, or share it with others.

b. The main difference between first-party data brokers and third-party data brokers is related to individual whose personal data is collected. First-party data brokers collect information about their own customers, while third-party data brokers collect information about other customers from a range of sources. Both first-party and third-party data brokers sell or share customer information with others.

c. The main products and services offered by third-party data brokers in Australia include the following:
   1) Customer profiling
   2) Customer tracking
   3) Customer purchasing data
   4) Marketing optimisation and evaluation
   5) Housing and construction report
   6) Real-estate sale reports
   7) Data validation
   8) Risk and fraud management data – insurance applications, tenancy applications
   9) Identity verification

d. Some prominent data brokers operating in Australia:
   1) CoreLogic – real estate data; operates in Australia, NZ, US and UK
   2) Equifax – operates in 24 countries worldwide

3) Nielson – market research, operates in 55+ countries worldwide
4) Oracle – operates Oracle Data Marketplace containing data on 300million third=party customers
5) Experian – operates in 30 countries worldwide

e. Third-party data brokers use the following sources (among others) to collect information:
    1) Social networks
    2) Web pages (web scraping) and cookies
    3) Business (e.g., banks), other data brokers, government sources (e.g., census)
    4) Customer loyalty schemes
    5) Open data projects

f. Some ways in which the collections and use of personal information by third-party data brokers can harm individuals:
    1) Secondary use of data can lead to discrimination (e.g., data collected for another purpose can be used to influence customer access to rental properties)
    2) Personal data that is collected and sold may be incorrect, resulting in customer not being to obtain, for example, a loan
    3) Information such as address may lead to robbery, stalking or physical harm
    4) Breach of privacy by targeted advertising or customer profiling
    5) Identity theft