Assignment 3: Anonymisation Techniques

Privacy Enhancing Technologies (201500042)

(Total number of achievable points: 20)

Issue date: 29 May 2016; **Due date: 06 June 2016, 23:59 CET** *(hand in via BB)*

1 Introduction

In this assignment you are going to practice how to apply disclosure attacks on published data and how to use anonymisation techniques to prevent such attacks. You are going to work on UCI Machine Learning Repository's **Adult** dataset and implement your solutions using **sdcMicro** library of **R** programming language. Instructions about dataset and usage of **R** is provided in Section 3.

You have to hand in a document (PDF) with your answers and your source code (plain text) on Blackboard. Please mention your name and student number in submission. You can submit an archive (ZIP, GZip, etc.) containing both the report and the source code, but please do not submit a RAR file.

2 Assignment

1. **Record Linkage Attacks** *(3 points)*

   In the first step of assignment you are going to apply record linkage attack on Adult dataset. Record linkage attacks aim to identify a record owner from published database using some external knowledge. Table 1 shows a public data table which provides you external knowledge to apply record linkage.

Name	Age	Sex	Race	Education	ZIP Code	...
...	...	...	...	...	...	...
Patricia Conner	39	Female	Black	Masters	23456	...
Benjamin Hodges	78	Male	White	Some-college	34567	...
Sergio Townsend	24	Male	Asian-Pac-Islander	Bachelor	45678	...
Dianne Joseph	31	Female	Black	10th	56789	...
Rufus George	52	Male	Amer-Indian-Eskimo	7th-8th	67890	...
...	...	...	...	...	...	...

Table 1: Non de-identified public data table

   (a) *(0 points)* Using instructions in Section 3, upload Adult dataset to your workspace.

   (b) *(3 points)* Using Table 1, apply a record linkage attack on Adult dataset. Show how many people you can identify.

   Hint: To filter Adult dataset, you can use subset() function in R (See Section 3).

2. **Domain Generalization Hierarchies** *(4 points)*

   k-anonymity is a solution to prevent record linkage attacks [Swe02] as in Question 1. Generalisation is a well-known method for k-anonymity which recodes the value of

an item to a general value to prevent linkage. [Swe02] utilises Domain Generalisation Hierarchies (DGH) to decide the degree of generalisation.

In the second step of assignment, you are going to construct DGHs for given quasi-identifiers.

(a) *(1 point)* Using instructions in Section 3, create an sdcMicro object. Quasi-identifiers (QID) and sensitive attributes (SE) in the object are as follows:

- age → QID
- education → QID
- race → QID
- sex → QID
- occupation → SE
- inocme → SE

(b) *(3 points)* Draw a DGH for each QID and explicitly show the domain for each level of hierarchy in your report. Do not forget that the degree of generalisation affects the utility of anonymisation operations. Thus, try to optimise generalisation while constructing DGHs.

> **Note 1:** This part of question does not require any coding in R.
>
> **Note 2:** You can refer to lecture slides, notes and [Swe02] for details of constructing DGHs.
>
> **Note 3:** We do not expect a combined DGH for all QIDs. Draw a separate DGH for each QID.

3. **k-anonymity with Generalisation and Suppression** *(5 points)*

k-anonymity can be achieved by combining generalisation and suppression methods. In this step, you have to anonymise Adult dataset for given k values using generalisation and suppression functions in **R**.

For the following k values:

- $k = 3$
- $k = 5$
- $k = 10$
- $k = 50$
- $k = 100$

(a) *(3 points)* Apply generalisation functions on Adult dataset using the DGHs in Question 2. Each k value requires a different degree of generalisation. In your solutions, select optimal degrees for generalisation domains and show them explicitly in your report.

(b) *(2 points)* Apply suppression on generalised data to remove the items that violate given k-anonymity requirement.

> **Note 1:** For generalisation, you have to use globalRecode and groupVars functions in **R**. The usage of functions is explained in Section 3.

Note 2: Generalisation functions do not take a k value. Therefore to satisfy k-anonymity requirement for a given k, after generalisation you have to perform suppression function (see Section 3) which takes k as parameter.

4. **Utility Measures on Anonymised Data** *(5 points)*

In data publishing, the utility of data is important to perform further analysis on it. Therefore, in the fourth step of assignment you are going to perform utility measures on anonymised data.

For this assignment you are going to measure utility by using Precision metric ([Swe02]):

$$Prec(RT) = 1 - \frac{\sum_{i=1}^{N_A} \sum_{j=1}^{N} \frac{h}{|DGH_{A_i}|}}{|PT| \cdot |N_A|}$$

In the equation, RT is a generalised table and PT is original table. $|DGH_{A_i}|$ refers to the height of DGH for QID A_i where $1 \leq i \leq N_A$ and h is the current height of the generalised value in DGH. Finally, N is the size of generalised table.

(a) *(3 points)* For each k value compute Precision metric on anonymised table.

(b) *(2 points)* Plot a graph which shows the change in Precision metric with respect to k.

Note: For Precision computation you can use nrow() function to learn the size of data tables (see Section 3)

5. **Attribute Linkage Attack** *(3 points)*

Although k-anonymity is successful in preventing record linkage, it cannot prevent attribute linkage. In attribute linkage the attacker cannot directly identify data owner, but can obtain information related to his sensitive attributes. Homogeneity attack and background knowledge attack are two well known attacks for attribute linkage [Mac+07].

(a) *(3 points)* For each k-anonymous table in Question 3, using Table 1 apply homogeneity and background knowledge attacks on anonymised data for sensitive attributes Income and Occupation, if possible. Show which data owners you are able to identify.

Note: Depending on your generalisation strategy and k-value, you may not observe an attribute linkage on anonymised data. If this is the case, explain it clearly in your report.

3 Instructions for R and Dataset

3.1 Adult Dataset

You are going to implement anonymisation operations on UCI Machine Learning Repository's Adult dataset. You can download the dataset from the following link:

https://archive.ics.uci.edu/ml/machine-learning-databases/adult/

You have to download "adult.data" file from the above link. You can see the detail about names and properties of the attributes in "adult.names" file. Adult dataset consists

Name	Domain
age	continuous
workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
fnlwgt	continuous
education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
education-num	continuous
marital-status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
sex	Female, Male
capital-gain	continuous
capital-loss	continuous
hours-per-week	continuous
native-country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
income	$\leq$ 50K, >50K

Table 2: Attribute names and domains for Adult dataset

of 15 attributes. Name and domain of each attribute are illustrated in Table 2 which are necessary for construction of DGHs and generalisation functions.

3.2 R and sdcMicro Library

You can download R from the following link:

https://cran.rstudio.com/

Once you installed R in your computer, you have to install sdcMicro package to use anonymisation functions by following command:

```
install.packages("sdcMicro")
```

Some functions in sdcMicro package that you will use to operate anonymization operations are listed below:

- Load sdcMicro package to workspace:

```
library(sdcMicro)
```

- Load dataset to workspace:

```
data <- read.table("DATA_PATH", sep = ",")
```

  - **Note 1:** The dataset does not have a header which shows names of attributes. Thus, **R** assigns a name for each attribute as 'V1', 'V2', ... etc. Using str(data) function you can observe which attribute corresponds to which 'V' value.
  - **Note 2:** The dataset contains some unknown values (shown as '?'). Before applying anonymisation operations, you have to remove those unknown values. A naive way for removal is illustrated below:

```
ndata <- subset(data, V1 != "?" & V2 != "?" & V3 != "?"
& V4 != "?" & V5 != "?" & V6 != "?" & V7 != "?"
& V8 != "?" & V9 != "?" & V10 != "?" & V11 != "?"
& V12 != "?" & V13 != "?" & V14 != "?" & V15 != "?")
```

- Creating sdcMicroObj object: To operate generalisation and suppression operations on the dataset, you need an sdcMicroObj object. While creating the object, you should specify the quasi identifiers (keyVars) and sensitive attributes (sensibleVar) in the dataset.

```
sdc <- createSdcObj(ndata, keyVars = c('QID_1',..., 'QID_n'),
sensibleVar = c('SE_1',..., 'SE_m'))
```

- Applying k-anonymity:
  - Generalisation:
    - Generalisation on categorical attribute:

```
groupVars(obj, var, before, after)
```

Example: Generalise "sex" attribute:

```
sdc <- groupVars(sdc, var = "sex",
        before = c("Female","Male"),
        after=c("Person", "Person"))
```

Note 1 : **R** partitions the dataset according to position of "," and it does not take care of space characters. Therefore, you need to add a space character before the value while applying generalisation operation for each QID. For example, for sex attribute, you have to write " Female" instead of "Female".

Note 2 : To perform generalisation on categorical attributes you have to list all categories of the attribute in before() function. For example, to generalise sex attribute you should list Male and Female as before(" Male", " Female"). Further, in after() function you have to list all generalised values which correspond to each category of before() function, as after("Person", "Person").

* Generalisation on continuous attribute:

```
globalRecode(obj, qid, break, labels (optional))
```

Example: Generalise "age" attribute:

```
sdc <- globalRecode(sdc, column = "age",
breaks = c(1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100))
```

- Suppression:

```
sdc <- localSuppression(sdc, k)
```

- Display functions
  - Display dataset:

    data

  - Display properties of dataset (attribute name, levels, type, etc.)

    ```
    str(data)
    ```

  - Display properties of sdcMicroObj (list of qids, observations that violating k-anonymity, number of suppressions, etc.)

    ```
    sdc
    print(sdc)
    print(sdc, type = "ls")          #suppression
    print(sdc, type = "recode")      #generalization
    ```

  - Read modified data from sdcMicroObj:

    ```
    newdata <- extractManipData(sdc)
    ```

- Using subset function to filter dataset

  ```
  newdata2 <- subset(newdata, age == '(30,40]' )
  ```

- Using nrow function to get size of data table (for Precision metric)

  nrow(data)

- For further help on sdcMicro library functions:

  help(package = sdcMicro) **or**
  ?function_name

Note: Please do not copy paste **R** commands from assignment document. **R** and Latex interprets some symbols (e.g. <, -, ', ") differently. Therefore you may get errors in your code.

3.3 Example for usage of R functions

To clarify usage of **R** functions, in this section we provide you a short sample with one QID (sex) and one SE (income).

```
#load library
library(sdcMicro)

#load data and remove unknowns
data <- read.table("xxxx/xxxx/xxxx/adult.data", sep = ",")

ndata <- subset(data, V1 != "?" & V2 != "?" & V3 != "?"
 & V4 != "?" & V5 != "?" & V6 != "?" & V7 != "?" & V8 != "?"
 & V9 != "?" & V10 != "?" & V11 != "?" & V12 != "?"
 & V13 != "?" & V14 != "?" & V15 != "?")

 #filter dataset
 ndata2 <- subset(ndata, V10 == ' Female' )

 #create sdcObject for anonymisation functions
 sdc <- createSdcObj(ndata, keyVars = c('V10'),
                          sensibleVar = c('V15'))

 #apply k-anonymity for k = 5
 #first step generalisation:

 sdc <- groupVars(sdc, var = "sex",
         before = c(" Female"," Male"),
         after=c("Person", "Person"))

 #second step suppression:
 sdc <- localSuppression(sdc, 5)

 #observe the change in data after suppression
 print(sdc, type = "ls")

 #extract data to compute precision metric
newdata <- extractManipData(sdc)

#get size of original data table for precision metric
nrow(data)

#get size of anonymised data table for precision metric
nrow(newdata)
```

References

[Mac+07] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. "l-diversity: Privacy beyond k-anonymity." In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007), p. 3.

[Swe02] Latanya Sweeney. "Achieving k-anonymity privacy protection using generalization and suppression." In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 571–588.