Assignment: Introduction to information privacy

The assignment consists of discussing the application of data masking procedures to a datafile, and computing some information loss and disclosure risk measures.

For this purpose, you can use R, which is downloadable from:

http://www.r-project.org/.

R includes a package for data privacy. It is called `sdcMicro`. This package can be found e.g. in:

http://cran.r-project.org/web/packages/sdcMicro/index.html

Some of the functions of this package and some examples can be found in [1]. There is a related package with a GUI (see [2]).

In order to use this package, it has to be first installed (downloaded) and then loaded. In order to install it, use

```
Package -> Install packages
```

from the menu bar of RGui. Select a server for downloading and then select the `sdcMicro` package in the window listing all available packages. Then, in order to load this library use the following function in the R prompt.

```
library(sdcMicro)
```

I list below a few examples in R for those not familiar with it:

- **Example 1.** Read the file `census.csv`[1] and assign it to the variable `ta`.

  Read the file:

  ```
  read.table("census.csv", header = TRUE, sep = ";", row.names = 1)
  ```

  Assign the file to the variable `ta`:

  ```
  ta <- read.table("census.csv", header = TRUE, sep = ";", row.names
  = 1)
  ```

- **Example 2.** Apply rank swapping and microaggregation (with individual ranking):

  ```
  ta1 <- swappNum(ta,p=1)}
  ta2 <- microaggregation(ta, method="onedims", aggr=3)
  ```

  `ta1` and `ta2` contain the original files in `ta1$x` and `ta2$x` and the protected files in `ta1$xm` and `ta2$mx`. **Important.** Note that the protected file in rankswapping is in a variable different to the variable for microaggregation.

[1] For your information, this data can be downloaded from the following URL, where three other datasets are available:
http://neon.vb.cbs.nl/casc/..%5Ccasc%5CCASCtestsets.htm (Micro data test set/Data).

- **Example 3.** Compute disclosure risk and data utility using some of the predefined functions in the `sdcMicro` package.

```
> dRisk(obj=ta, xm=ta2$mx)
[1] 0.975
> dUtility(obj=ta, xm=ta2$mx)
[1] 0.01088046
> dRisk(obj=ta, xm=ta1$xm)
[1] 0.6055556
> dUtility(obj=ta, xm=ta1$xm)
[1] 0.04911969
```

Exercise. Select a data file, protect it using at least two data protection methods and different parameterizations, and plot a R-U map for at least 10 pairs (method, parameter) using a particular information loss and disclosure risk measure. Discuss the results.

Example. Using the file census.csv already read in the variable `ta`, apply rank swapping and microaggregation using `sdcMicro` implementations, plot the R-U maps (in both screen and ėps file) using functions `dRisk` and `dUtility`. Code for the example is listed below, and the R-U maps are displayed in Table 1.

```
resRS <- sapply(1:10,
function(param)(swappNum(ta, p=param)))
drr <- sapply(1:10, function(i) (dRisk(resRS[,i]$x, xm=resRS[,i]$xm)))
dur <- sapply(1:10, function(i) (dUtility(resRS[,i]$x, xm=resRS[,i]$xm)))
plot(drr,dur,xlim=c(0,1),ylim=c(0,1),main="R-U map")
postscript("rankswapping.eps")
plot(drr,dur,xlim=c(0,1),ylim=c(0,1),main="R-U map")
dev.off()

pMicro <- c(3,5,9,10,12,15,20,27)
resIR <- sapply(pMicro,
                function(param)(microaggregation(ta, method="onedims", aggr=param)))
resPCA <- sapply(pMicro,
                function(param) (microaggregation(ta, method="pca", aggr=param)))
drp <- sapply(1:length(pMicro), function(i) (dRisk(resPCA[,i]$x, xm=resPCA[,i]$mx)))
dup <- sapply(1: length(pMicro), function(i) (dUtility(resPCA[,i]$x, xm=resPCA[,i]$mx)))
dri <- sapply(1: length(pMicro), function(i) (dRisk(resIR[,i]$x, xm=resIR[,i]$mx)))
dui <- sapply(1: length(pMicro), function(i) (dUtility(resIR[,i]$x, xm=resIR[,i]$mx)))
plot(cbind(dri,drp),cbind(dui,dup),xlim=c(0,1),ylim=c(0,1),main="R-U map")
postscript("microaggregation.eps")
plot(cbind(dri,drp),cbind(dui,dup),xlim=c(0,1),ylim=c(0,1),main="R-U map")
dev.off()
```

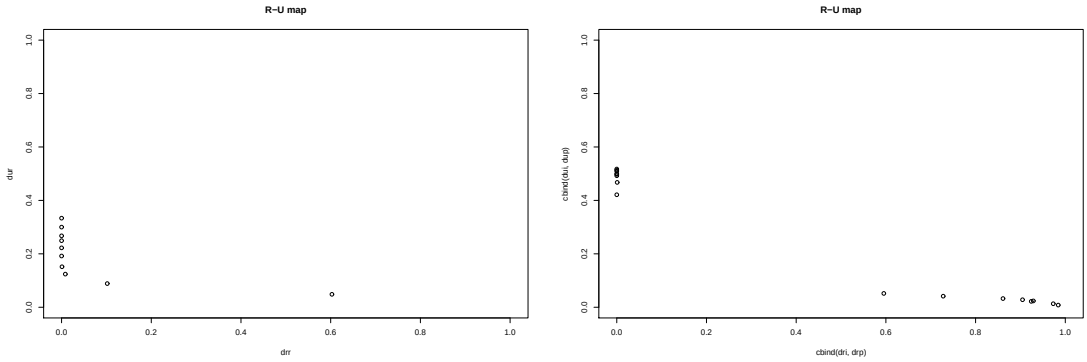

Figure 1: R-U maps for rank swapping and microaggregation using the parameters in the example.

References

[1] Templ, M. (2008) Statistical Disclosure Control for Microdata Using the R-Package sdcMicro, Transactions on Data Privacy 1:2 67-85.
http://www.tdp.cat/issues/abs.a004a08.php

[2] Templ, M., Petelin, T. (2009) A Graphical User Interface for Microdata Protection Which Provides Reproducibility and Interactions: the sdcMicro GUI, Transactions on Data Privacy 2:3 207 - 224.
http://www.tdp.cat/issues/abs.a030a09.php