Whale Entanglement sdcMicro Exercise

Hailey Veirs, Guillermo Romero, Alessandra Vidal Meza

2023-05-26

Your team acquired a dataset* from researchers working with whale entanglement data on the West Coast. The dataset contains both direct and indirect identifiers. Your task is to assess the risk of re-identification of the fisheries associated with the cases before considering public release. Then, you should test one technique and apply k-anonymization to help lower the disclosure risk as well as compute the information loss.

Please complete this exercise in pairs or groups of three. Each group should download the dataset and complete the rmd file, including the code and answering the questions. Remember to include your names in the YAML.

Set Up Environment

```
library(here)
library(tidyverse)
library(sdcMicro)
```

```
whale_dat <- read_csv('whale-sdc.csv')
```

Inspect the Dataset

```
## spc_tbl_ [348 x 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ case_id        : chr [1:348] "20000201Er" "20000316Er" "20000327Er" "20000330Er" ...
##  $ year           : num [1:348] 2000 2000 2000 2000 2000 ...
##  $ month          : num [1:348] 2 3 3 3 4 6 7 8 11 9 ...
##  $ type           : chr [1:348] "Gray Whale" "Gray Whale" "Gray Whale" "Gray Whale" ...
##  $ county         : chr [1:348] "San Diego" "Orange" "Los Angeles" "Los Angeles" ...
##  $ state          : chr [1:348] "CA" "CA" "CA" "CA" ...
##  $ lat            : num [1:348] 32.7 33.4 34 33.7 33.7 ...
##  $ long           : num [1:348] -117 -118 -119 -118 -118 ...
##  $ inj_level      : num [1:348] 8 8 7 10 10 5 3 3 10 3 ...
##  $ condition      : chr [1:348] "alive" "alive" "alive" "dead" ...
##  $ origin         : chr [1:348] "commercial" "commercial" "commercial" "commercial" ...
##  $ gear           : chr [1:348] "gillnet" "gillnet" "gillnet" "gillnet" ...
##  $ fishery_license: num [1:348] 4.65e+09 7.92e+09 6.62e+09 3.70e+09 7.08e+09 ...
##  $ fine           : num [1:348] 1 1 0 1 1 0 0 0 1 0 ...
##  $ infraction_type: num [1:348] 1 1 0 1 1 0 0 0 1 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   case_id = col_character(),
##   ..   year = col_double(),
##   ..   month = col_double(),
##   ..   type = col_character(),
##   ..   county = col_character(),
##   ..   state = col_character(),
```

```
##   ..   lat = col_double(),
##   ..   long = col_double(),
##   ..   inj_level = col_double(),
##   ..   condition = col_character(),
##   ..   origin = col_character(),
##   ..   gear = col_character(),
##   ..   fishery_license = col_double(),
##   ..   fine = col_double(),
##   ..   infraction_type = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Question 1

How many direct identifiers are present in this dataset? What are they? The direct identifiers present in this dataset are the *case_id* attribute for the whale entaglement event and *fishery_license* attribute for the fishing operation.

Question 2

What attributes would you consider quasi-identifiers? Why? The quasi-identifiers present in this dataset for the fishing operation are the *county*, *state*, *lat*, *long*, and *origin* attributes; these attributes can be coupled and used together to identify a fishery. The quasi-identifiers present in this dataset for the whale entaglement event are the *year*, *month*, *type*, *inj_level*, *condition*, *gear*, *fine*, and *infraction_type* attributes; these attributes can also be coupled and used together to identify a whale entaglement event, especially if the event recvied wide media coverage or foreign data repositories to cross-reference exist.

Question 3

What types of variables are they? Define them as numeric, integer, factor or string. All attributes are factor variables as they are categorical variables.

```
# Define file name
fname <- 'whale-sdc.csv'

# Read data frame with file name
file <- read_csv(fname)

# Convert to factor
file <- varToFactor(obj = file,
                    var = c('year', 'month', 'type',
                            'county', 'state', 'lat', 'long',
                            'inj_level', 'condition', 'origin',
                            'gear', 'fine', 'infraction_type'))
```

Considering your answers to questions 1, 2 and 3, and let's set up an SDC problem.

```
sdcInitial <- createSdcObj(dat = file,
                           keyVars = c('year', 'month', 'type',
                                       'county', 'state', 'lat', 'long',
                                       'inj_level', 'condition', 'origin',
                                       'gear', 'fine', 'infraction_type'),
                           weightVar = NULL,
                           hhId = NULL,
                           strataVar = NULL,
                           pramVars = NULL,
```

```
                              excludeVars = c('fishery_license', 'case_id'),
                              seed = 0,
                              randomizeRecord = FALSE,
                              alpha = c(1))
```

Question 4.1

What is the risk of re-identification for this dataset? The risk of re-identification for the entire dataset is 100%.

```
sdcInitial@risk$global$risk
```

```
## [1] 1
```

Question 4.2

Let's determine which observations have a higher risk to be re-identified:

```
head(sdcInitial@risk$individual)
```

```
##      risk fk Fk
## [1,]    1  1  1
## [2,]    1  1  1
## [3,]    1  1  1
## [4,]    1  1  1
## [5,]    1  1  1
## [6,]    1  1  1
```

And let's take a look at the frequency of the particular combination of key variables (quasi-identifiers) for each record in the sample:

```
freq(sdcInitial, type = 'fk')
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [112] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [149] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [186] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [223] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [260] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [297] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [334] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

To what extent does this dataset violate k-anonymity? All observations of the dataset violate k-anonymity.

Now, consider techniques that could reduce the risk of re-identification.

Question 5.1

Apply one non-perturbative method to a variable of your choice. How effective was it in lowering the disclosure risk?

Let's apply top and bottom recoding to de-identify and anonymize the dataset:

```
table(sdcInitial@manipKeyVars$year)
```

Recoding for *year* attribute

```
##
## 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
##    9    1    3   10    9   10    8   11    5    8   15    9   14   11   23   51
## 2016 2017 2018 2019
##   53   30   45   23
```

```
# Top recoding for variable *year*
sdcInitial <- groupAndRename(obj = sdcInitial,
                             var = c('year'),
                             before = c('2000', '2001', '2002',
                                        '2003', '2004', '2005',
                                        '2006', '2007', '2008', '2009'),
                             after = c('2000-2009'))

# Bottom recoding for variable *year*
sdcInitial <- groupAndRename(obj = sdcInitial,
                             var = c('year'),
                             before = c('2010', '2011', '2012',
                                        '2013', '2014', '2015',
                                        '2016', '2017', '2018', '2019'),
                             after = c('2010-2019'))
```

```
table(sdcInitial@manipKeyVars$inj_level)
```

Recoding for *inj_level* attribute

```
##
##  0  1  2  3  4  5  6  7  8  9 10
##  1  2 28 59 54 54 39 37 28 14 32
```

```
# Top recoding for variable *inj_level*
sdcInitial <- groupAndRename(obj = sdcInitial,
                             var = c('inj_level'),
                             before = c('0', '1', '2',
                                        '3', '4', '5'),
                             after = c('0-5'))

# Bottom recoding for variable *inj_level*
sdcInitial <- groupAndRename(obj = sdcInitial,
                             var = c('inj_level'),
                             before = c('6', '7', '8',
                                        '9', '10'),
                             after = c('6-10'))
```

```
sdcInitial@risk$global$risk
```

```
## [1] 1
```

```
print(sdcInitial, 'kAnon')
```

```
## Infos on 2/3-Anonymity:
##
## Number of observations violating
##   - 2-anonymity: 348 (100.000%)
```

```
##   - 3-anonymity: 348 (100.000%)
##   - 5-anonymity: 348 (100.000%)
##
## ----------------------------------------------------------------------
```

Top and bottom recoding of the two quasi-identifiers was not effective at lowering the risk of re-identification. The risk of re-identification remains at 100%.

Question 5.2

Apply k-3 anonymization to this dataset. After we set the parameters to aim for 3 observations sharing the same attributes in the dataset, the risk of re-identification is now 22.5%.

```r
sdcInitial <- kAnon(sdcInitial, k = c(3))
sdcInitial@risk$global$risk
```

```
## [1] 0.2254431
```

Question 6

Compute the information loss for the de-identified version of the dataset.

```r
# Extract total number of supressions for each categorical key variable
print(sdcInitial, 'ls')
```

```
##          KeyVar | Suppressions (#) | Suppressions (%)
##            year |                6 |            1.724
##           month |              125 |           35.920
##            type |               31 |            8.908
##          county |              149 |           42.816
##           state |               27 |            7.759
##             lat |              225 |           64.655
##            long |              244 |           70.115
##       inj_level |               11 |            3.161
##       condition |                6 |            1.724
##          origin |                5 |            1.437
##            gear |               60 |           17.241
##            fine |                0 |            0.000
##  infraction_type |               17 |            4.885
```

Let's compare the number of missing values (NAs) before and after anonymization.

```r
# Extract names of all categorical key variables into a vector
namesKeyVars <- names(sdcInitial@manipKeyVars)

# Create matrix to store the number of missing values
NAcount <- matrix(NA, nrow = 2, ncol = length(namesKeyVars))
# Add column names to matrix
colnames(NAcount) <- c(paste0('NA', namesKeyVars))
# Add row names to matrix
rownames(NAcount) <- c('initial', 'anonym')

# Count missing values in all key variables
for(i in 1:length(namesKeyVars)) {
  NAcount[1, i] <- sum(is.na(sdcInitial@origData[,namesKeyVars[i]]))
  NAcount[2, i] <- sum(is.na(sdcInitial@manipKeyVars[,i]))
  }
```

```r
NAcount
```

```
##         NAyear NAmonth NAtype NAcounty NAstate NAlat NAlong NAinj_level
## initial      0       0      0        0       0     0      0           0
## anonym       6     125     31      149      27   225    244          11
##         NAcondition NAorigin NAgear NAfine NAinfraction_type
## initial           0        0      0      0                 0
## anonym            6        5     60      0                17
```