Data Confidentiality Lab 5

Isaac Kleisle-Murphy

March 1, 2020

Note that other materials were submitted in the previous lab, and project materials are in the KleisleMurphy_Xie markdown document.

```
suppressMessages(require(dplyr))
options(scipen = 999)
acs_org = read.csv("~/Documents/Swat_2020/Data_Privacy/Data-Confidentiality/datasets/ACSdata_org.csv")
acs_syn = read.csv("~/Documents/Swat_2020/Data_Privacy/Data-Confidentiality/datasets/ACSdata_syn.csv")

#org_df = acs_org; syn_df = acs_syn; ref = "id"; c_vars = knwn
```

First, we define a helper function to compute our three risk measures:

```
# @param: org_df, the true dataset, in a dataframe
# @param: syn_df, the synthesized dataset, in a dataframe
# @param: c_vars, a character vector of variables (in both syn_df and org_df) of items known
#                 to the intruder; these help identify c_i
# @param: ref, the string column name of the entry id reference/label in both org_df and syn_df,
#              used to pair a synthetic entry with a true entry.
# @return: a dataframe with the three variables/measures of interest.
# @dontrun: match_risk_helper(org_df = acs_org, syn_df = acs_syn, c_vars = knwn, ref = "id")

match_risk_helper <- function(org_df, syn_df, c_vars, ref){

  match_df = left_join(syn_df, org_df, by = c_vars, suffix =  c(".s", ".o"))
  match_df$is_match = match_df%>%pull(paste0(ref, ".s")) == match_df%>%pull(paste0(ref, ".o"))

  match_score = match_df%>%
    group_by_at(paste0(ref, ".s"))%>%
    summarise(t_i = sum(is_match), c_i = n())%>%
    ungroup()%>%
    mutate(k_i = ifelse(c_i*t_i==1, 1, 0),
           f_i = ifelse(c_i*(1-t_i)==1, 1, 0))%>%
    summarise(exp_match_risk = sum(t_i/c_i),
              true_match_rate = sum(k_i)/n(),
              false_match_rate = sum(f_i)/n())

  return(match_score)

}
```

Then, we run it on the data

```
knwn = c('SEX', 'RACE', 'MAR') #known to intruder variables
syn = c('WAOB', 'DIS', 'HICOV', 'MIG', 'SCH') #synthesized variables
```

```r
acs_org$id = 1:nrow(acs_org);  acs_syn$id = 1:nrow(acs_syn) #give each entry an id to identify matches

#compute
match_analysis = match_risk_helper(org_df = acs_org, syn_df = acs_syn, c_vars = knwn, ref = "id")

#print
knitr::kable(match_analysis)
```

exp_match_risk	true_match_rate	false_match_rate
57	0.0004	0