R Notebook

Dataset for this project was obtained from https://www.kaggle.com/osmi/mental-health-in-tech-survey. R queries will be the primary mode to access data from the database. To identify individuals on data I employed inferential attack and background knowledge attack. I designed specific scenarios to better understand the consequence of the attacks.

Obtaining input

```
mental_health_data<-read.csv("/Users/meenakshinagarajan/Desktop/Privacy aware computing/survey.csv", header=TRUE, sep=",")
head(mental_health_data)
```

```
##             Timestamp Age Gender        Country state self_employed
## 1 2014-08-27 11:29:31  37 Female  United States    IL          <NA>
## 2 2014-08-27 11:29:37  44      M  United States    IN          <NA>
## 3 2014-08-27 11:29:44  32   Male         Canada  <NA>          <NA>
## 4 2014-08-27 11:29:46  31   Male United Kingdom  <NA>          <NA>
## 5 2014-08-27 11:30:22  31   Male  United States    TX          <NA>
## 6 2014-08-27 11:31:22  33   Male  United States    TN          <NA>
##   family_history treatment work_interfere   no_employees remote_work
## 1             No       Yes          Often           6-25          No
## 2             No        No         Rarely More than 1000          No
## 3             No        No         Rarely           6-25          No
## 4            Yes       Yes          Often         26-100          No
## 5             No        No          Never        100-500         Yes
## 6            Yes        No      Sometimes           6-25          No
##   tech_company   benefits care_options wellness_program  seek_help
## 1          Yes        Yes     Not sure               No        Yes
## 2           No Don't know           No       Don't know Don't know
## 3          Yes         No           No               No         No
## 4          Yes         No          Yes               No         No
## 5          Yes        Yes           No       Don't know Don't know
## 6          Yes        Yes     Not sure               No Don't know
##    anonymity              leave mental_health_consequence
## 1        Yes      Somewhat easy                         No
## 2 Don't know         Don't know                      Maybe
## 3 Don't know Somewhat difficult                         No
## 4         No Somewhat difficult                        Yes
## 5 Don't know         Don't know                         No
## 6 Don't know         Don't know                         No
##   phys_health_consequence    coworkers supervisor mental_health_interview
## 1                      No Some of them        Yes                      No
## 2                      No           No         No                      No
## 3                      No          Yes        Yes                     Yes
## 4                     Yes Some of them         No                   Maybe
## 5                      No Some of them        Yes                     Yes
```

```
## 6                      No          Yes        Yes                        No
##   phys_health_interview mental_vs_physical obs_consequence comments
## 1                 Maybe               Yes              No     <NA>
## 2                    No        Don't know              No     <NA>
## 3                   Yes                No              No     <NA>
## 4                 Maybe                No             Yes     <NA>
## 5                   Yes        Don't know              No     <NA>
## 6                 Maybe        Don't know              No     <NA>
```

From this large dataset, we are choosing only a random sample of rows and columns that are vulnerable to privacy attacks for the purpose of demonstration. We consider columns 'Age', 'Gender', 'Country', and 'treatment' from this dataset as potential elements that are vulnerable to attacks. Below are the two scenarios which describes the possibility of privacy attacks, when the data is exposed to public. The attacker could gain knowledge on user and his/her mental health treatment in below scenarios even though their anonymity is protected.

Inference attack: Finding if an individual has mental health issue

Scenario 1: Revealing Time Critical Survey Data

A survey company is planning to take a survey regarding user's mental health across regions to construct mental health awareness camps. While taking survey, they have decided to publish health related data excluding user's private information. They have decided to tour Company A for 2 days. Let the person who is answering the survey have health issues and is the attacker.

```
#Day 1 survey
myvars <- c("Age", "Gender", "Country","treatment")
newdata <- mental_health_data[myvars]
mysample <- newdata[11:20,]
head(mysample)
```

```
##    Age Gender       Country treatment
## 11  31   Male United States       Yes
## 12  29   male      Bulgaria        No
## 13  42 female United States       Yes
## 14  36   Male United States        No
## 15  27   Male        Canada        No
## 16  29 female United States       Yes
```

```
#Selecting observations from day 1 survey where mental treatment is 'Yes'
mysample_yes <- mysample[ which(mysample$treatment=='Yes'),]

#number of female who answered 'Yes'
print("Number of females who took the mental treatment in day 1 survey:")
```

```
## [1] "Number of females who took the mental treatment in day 1 survey:"
```

```
nrow(mysample_yes[ which(mysample_yes$Gender=='female'), ])

## [1] 2

#Day 2 survey
mynewsample <- newdata[31:40,]

#Selecting observations from day 2 survey where mental treatment is 'Yes'
mynewsample_yes <- mynewsample[ which(mynewsample$treatment=='Yes'),]

#Selecting observations from day 2 survey where mental treatment is 'Yes' an
Gender is 'Female'
print("Number of females who took the mental treatment in day 2 survey:")

## [1] "Number of females who took the mental treatment in day 2 survey:"

nrow(mynewsample_yes[ which(mynewsample_yes$Gender=='female'), ])

## [1] 1
```

The total number of female respondents who took the mental health treatment on day 1 and day 2 together is 3. If the attacker is one among the survey respondents and he has knowledge on statistics of day 1 survey, then on combining the data obtained from day 1 and day 2, he can easily conclude that only one female respondent attended day 2 survey and he could access her mental treatment data.

Background Knowledge attach: Finding if an individual has mental health issue

Scenario 2: Company has only 1 employee is equal to or above 50

```
mynewsample

##    Age Gender        Country treatment
## 31  32   Male United Kingdom        No
## 32  31   Male  United States        No
## 33  30   male United Kingdom       Yes
## 34  42   Male  United States       Yes
## 35  40 female  United States       Yes
## 36  27   Male  United States       Yes
## 37  29   Male         Canada        No
## 38  38   Male       Portugal        No
## 39  50      M  United States        No
## 40  35      M  United States       Yes

print("Number of persons who said 'Yes' to treatment in day 2 survey and 50
years of age:")

## [1] "Number of persons who said 'Yes' to treatment in day 2 survey and 50
years of age:"
```

```r
nrow(mynewsample[ which(mynewsample$Age=='50'), ])

## [1] 1
```

On Publishing this data people inside company, who has knowledge about their peers, can find out the private field of the individual.