Identification Disclosure Risk- Hu Paper

MATH 301 Data Confidentiality

Henrik Olsson

March 1, 2020

```
acsdata_org <- read.csv("ACSdata_org.csv")
acsdata_syn <- read.csv("ACSdata_syn.csv")
```

Let c_i be the number of records with the highest match probability for the target t_i; let $T_i = 1$ if the true match is among the c_i units and $T_i = 0$ otherwise. Let $K_i = 1$ when $c_i T_i = 1$ and $K_i = 0$ otherwise, and let N denote the total number of target records. Finally, let $F_i = 1$ when $c_i(1 - T_i) = 1$ and $F_i = 0$ otherwise, and let s equal the number of records with $c_i = 1$.

(i) The expected match risk:

```
## Reference from Kevin Ros
syn_sex <- acsdata_syn$SEX
syn_race <- acsdata_syn$RACE
syn_mar <- acsdata_syn$MAR

N = nrow(acsdata_syn)
n = nrow(acsdata_org)

expected_match_risk = 0
true_match_rate = 0

for(i in 1:N){
  c_i = 0
  data = acsdata_org[acsdata_org$SEX == syn_sex[i] & acsdata_org$RACE == syn_race[i] & acsdata_org$MAR

  if(nrow(data) != 0){
    expected_match_risk = expected_match_risk + (1/nrow(data))
  }
  if(nrow(data) == 1)
    true_match_rate = true_match_rate + 1/N
}
print(expected_match_risk)
```

```
## [1] 57
```

(ii) The true match rate:

```
print(true_match_rate)
```

```
## [1] 4e-04
```

(iii) The false match rate:

```
false_match_rate = 0
for(i in 1:N){
  c_i = 1
  s=nrow(c_i)
  data = acsdata_org[acsdata_org$SEX == syn_sex[i] & acsdata_org$RACE == syn_race[i] & acsdata_org$MAR

  if(nrow(data) == 1){
    false_match_rate = sum(false_match_rate/s)
  }
}
print(false_match_rate)
```

```
## [1] 0
```