cs208 Homework 1

Anthony Rentsch

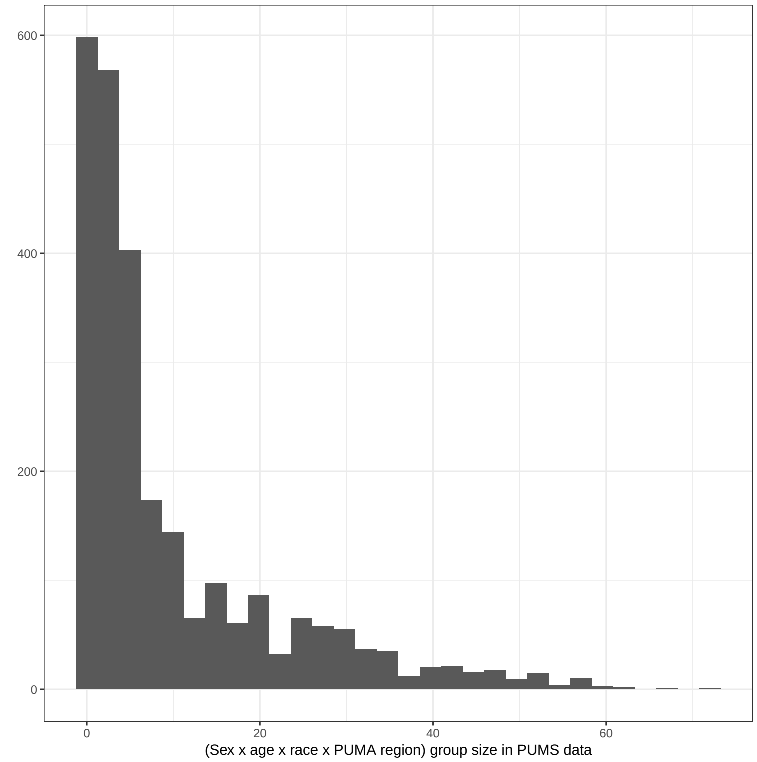
2/26/2019

For this assignment, I collaborated with Bhaven Patel, Lipika Ramaswamy, and Karina Huang. Find the code I wrote for this assignment in my Github repository.

Question 1

Similar to Latanya Sweeney's original attack, I would use the Georgia state voter file to attempt to re-identify individuals in the 2010 Census PUMS dataset for Georgia. As a result of the Voting Rights Act of 1965, several states - including Georgia - began asking people to self-report their race when filling out voter registration forms. Thus the statewide Georgia voter file, which costs just $250, could be used to conduct an attack on the PUMS dataset with race, gender, age, and PUMA region (derived from address) as available quasi-identifiers.

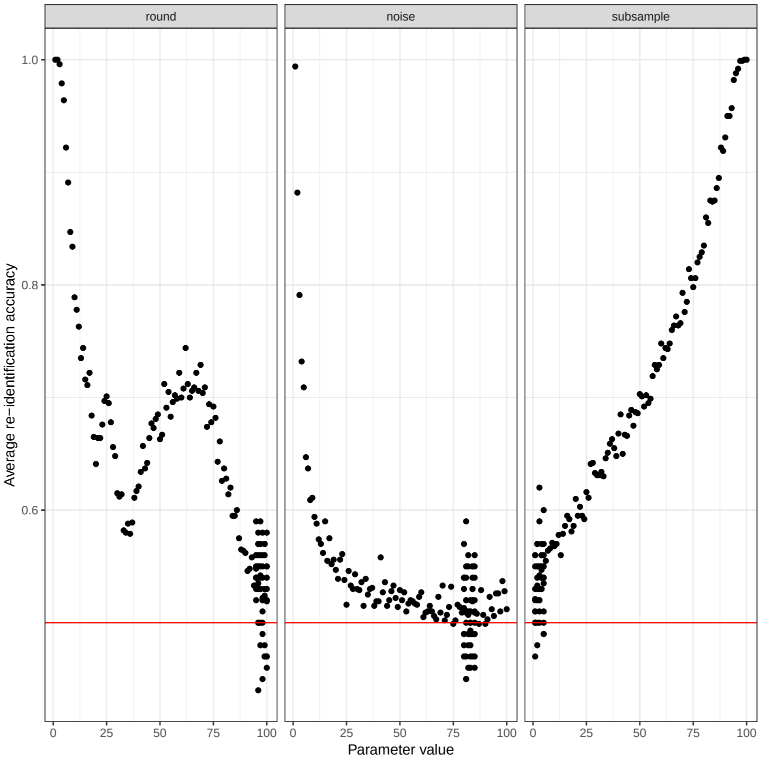
When we have these four attributes at our disposal, we are able to uniquely identify 2.3% of the individuals in the PUMS data. Furthermore, roughly 7% of individuals in the sample share their combination of sex, race, age, and PUMA region with less than 10 other people.



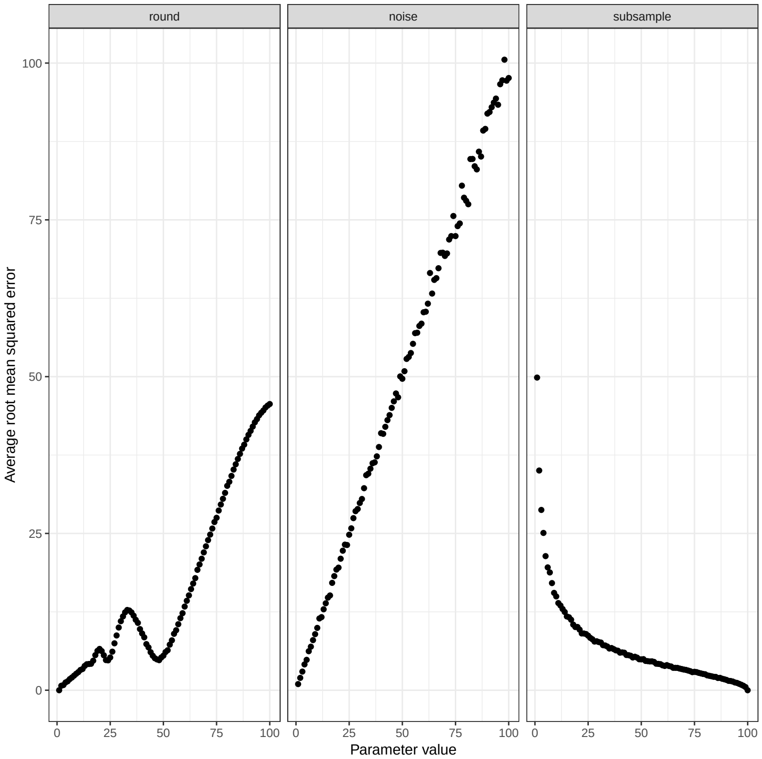
Question 2

The plot below shows the relationship between the defense parameter (R for rounding, ϵ for noise, and T for subsampling) and the average re-identification accuracy for 10 experiments run at each parameter value. In general, we observe that re-identification accuracy goes up when privacy is worse, which is the expected behavior. As we increase the standard deviation of the added noise the accuracy goes down and as we increase the number of people we subsample the re-identification accuracy goes up, as expected. For rounding there is a spike in the re-identification accuracy rate around 50, which makes sense because 96% of the 100-person sample are U.S. citizens and when we randomly subsample roughly 50 people the expected value of the query is approximately 50.

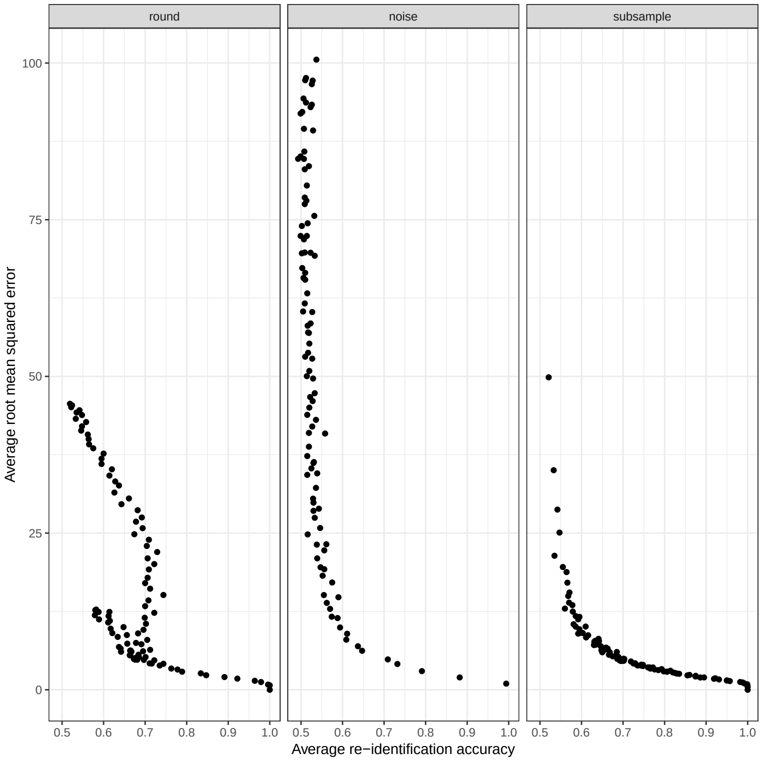
Also note that the attack would generally fail at 50% accuracy, which occurs at (roughly) $R = 100$, $T = 1$, and $\epsilon = 83$. I've added additional data points around these values, i.e., the results from all of the individual experiments within 5 parameter values of this point.



The next plot shows the relationship between the defense parameter and the utility of the query, as measured by the root mean squared error between the actual and query mechanism result. We see that larger values of R for the rounding defense lead to larger RMSE values (except for the earlier explained bump at 50), larger values of ϵ for the noise defense lead to larger RMSE values, and larger values of T for the subsampling defense lead to lower RMSE values.



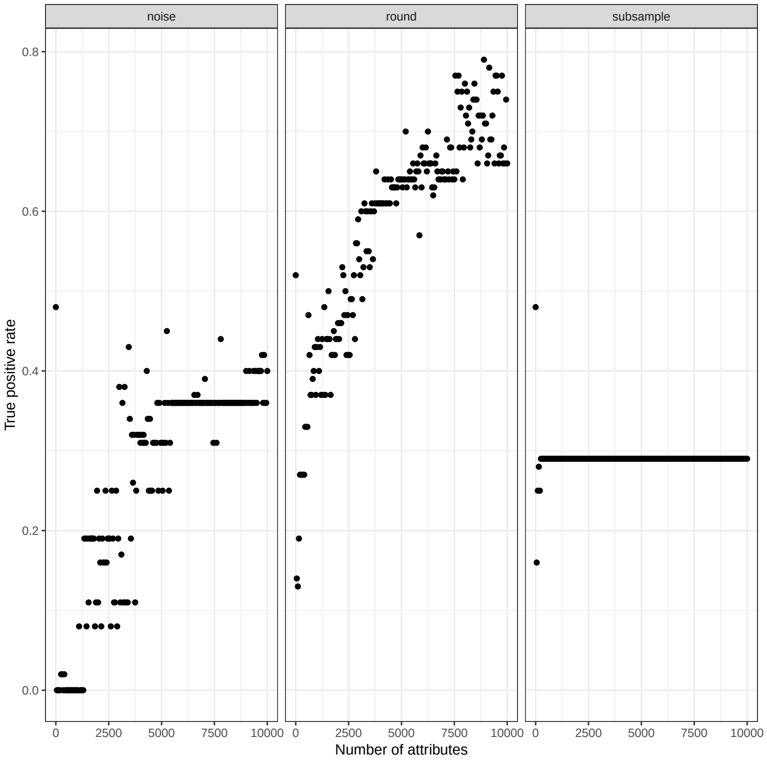
Comparing the privacy-utility tradeoff directly, we see that there is an approximately exponential decay: as we shed privacy we see huge improvements in accuracy, but those the size of those improvements tends to decrease as we approach near perfect re-identification accuracy. Again, the curve for rounding appears a bit strange, but this is a reflection of the idiosyncratic nature of the query's behavior around 50, i.e., rounding to the nearest multiple of 50 a sum with an expected value of 50 leads to high accuracy.



Question 3

We would expect membership attacks to be successful even when reconstruction attacks are unsuccessful. Membership attacks should approach true positive rates of 100% as the number of attributes released approaches n^2. In the plots below, I show the relationship between the membership attack true positive rate and the number of attributes released for the three different defense mechanisms we consider. As a reminder, I found the optimal values of the parameters to be: $R = 100$, $\epsilon = 83$, and $T = 1$ (here I consider $T = 9$ because my plot for $T = 1$ was uninteresting).

-what do i see



Question 4

For the final project, I'd like to work on an application and evaluation of a differentially private algorithm on an actual data set in the life or social sciences.

-more focus on defenses rather than attacks -experiments on privacy utility tradeoff

-take this form Google doc

Appendix

I put code snippets for all of my analyses here.

Question 1

```
grouped_pums_full <- pums_full %>% group_by(puma, sex, asian, black, latino, age) %>%
  summarise(n = n())

q1_plot <- ggplot(grouped_pums_full) + geom_histogram(aes(n)) +
  labs(x="(Sex x age x race x PUMA region) group size in PUMS data", y = "") +
  theme_bw()

sum(grouped_pums_full$n == 1)/sum(grouped_pums_full$n)
sum(grouped_pums_full$n < 10)/sum(grouped_pums_full$n)
```

Question 2