HW 2: Reconstruction and Membership Attacks

CS 208 Applied Privacy for Data Science, Spring 2022

Version 1.0: Due Fri, Feb. 11, 5:00pm.

Instructions: Submit a single PDF file to Gradescope containing your solutions, plots, and analyses. Submit any code files and notebooks separately on Gradescope. Make sure to list all collaborators and references.

1. **Reconstruction Attack**

   Among the variables in the 2000 PUMS dataset above is `uscitizen`, which asks the resident about their US Citizenship status.

   This is a sensitive piece of information, and including this question on the regular Census questionnaire has been a topic of recent controversy.[1] This PUMS dataset is public, but makes a good stand-in for a database that might be secured behind a query interface. We've provided a sample of size $n = 100$.

   In this problem, you will run experiments to evaluate the performance of the reconstruction attack on determining individuals' citizenship status. Treat the following variables in the dataset as public (so as an attacker you know them for all of the individuals in the dataset):

   $$\texttt{pub} = (\texttt{sex}, \texttt{age}, \texttt{educ}, \texttt{latino}, \texttt{black}, \texttt{asian}, \texttt{married}, \texttt{divorced}, \texttt{children}, \texttt{disability}, \texttt{militaryservice}, \texttt{employed}, \texttt{englishability}).$$

   Each query in your attack should specify a "random" boolean predicate q on the public variables (e.g. $q(\texttt{pub}_i) = [\texttt{age}_i/\texttt{educ}_i > 4 \;\&\&\; \texttt{sex}_i == 0]$), and receive as an answer an approximation to the subset sum query:

   $$\sum_{i:q(\texttt{pub}_i)=1} \texttt{uscitizen}_i, \tag{1}$$

   where i ranges over the $n = 100$ individuals in the PUMS dataset sample, `FultonPUMS5reconstruction.csv`, that we have provided.

   We have provided you with some (optional) starter code in `hw2_starter.py` [2] in the GitHub repo; the code from section on 2/7–2/8 may also be helpful.

   This includes definitions of:

   - `data`, a DataFrame containing the dataset you will be attacking.
   - `pub`, the names of the columns we are treating as public knowledge.
   - `execute_subsetsums_exact(predicates)`, a function that takes as input a list `predicates` on the `pub` variables and returns the list of (exact) answers to the corresponding subset sum queries on `data`, computed as in Equation (1).

[1] See e.g. `https://www.nytimes.com/2019/01/15/us/census-citizenship-question.html`

[2] Starter code at `https://github.com/opendp/cs208/blob/main/spring2022/homework/hw2_starter.py`

- An example of using `execute_subsetsums_exact` to count both the number of female citizens in the dataset and the number of married citizens in the dataset.
- A function `make_random_predicate()` that returns a (pseudo)random predicate q, which you can use to emulate the random subset sums that are used in the reconstruction attack as presented in class.

Carry out your attack and experiments in the following steps:

(a) Write a function `reconstruction_attack(data_pub,predicates,answers)` that takes as input `data_pub`, a DataFrame restricted to public columns, a list `predicates` of predicates on the public attributes, and a list of (possibly approximate) `answers` to the queries, and returns an attempted reconstruction of the sensitive (`uscitizen`) column as an array of boolean values of length n, where n is the number of rows in `data_pub`. Test your attack against `data` using $2n$ random queries generated by `make_random_predicate` and answered by `execute_queries_exact`. It should, with high probability, reconstruct all of the sensitive bits correctly.

(b) Implement the following defenses by modifying `execute_subsetsums_exact`:

i. `execute_subsetsums_round(R,predicates)`: round each result to the nearest multiple of R.

ii. `execute_subsetsums_noise(sigma,predicates)`: add independent Gaussian noise of mean zero and variance σ^2 to each result.

iii. `execute_subsetsums_sample(t,predicates)`: given a parameter $t \in \{1, \ldots, n\}$, randomly subsample a set T consisting of t out of the n rows and calculate all of the answers using only the rows in T (scaling up answers by a factor of n/t).

(c) Finally, run experiments on how your attack performs against the three defenses above.

i. Create functions to compute the accuracy of the answers returned by each of the `execute_subsetsums_*` functions (root-mean-squared-error between answers and exact values) and success of the attack (average fraction of values $\texttt{uscitizen}_i$ that are successfully reconstructed).

ii. Vary parameters R, σ, and t as integers from 1 to n. For each parameter setting, run 10 experiments with fresh randomness and plot the averages of the accuracy and reconstruction success fractions.

iii. Compare the trade-off between accuracy and success of the attack. Make sure to identify the regime where your attack transitions from near-perfect reconstruction (fraction close to 1) to failed reconstruction (fraction reconstructed is no higher than the proportion of the majority value).

(d) In section on 2/7–2/8, similar coding and experiments as above will be carried out for membership inference attacks against the subsampling defense, also on a dataset of size $n = 100$. Set the subsampling parameter to be $t = 50$. Run the code from section with a growing number d of queries (starting at $d = 2n$ and increasing from there) until the true positive probability of the reconstruction attack either converges or it becomes too computationally expensive to continue. Based on your experiments above and/or analytical reasoning, how do you think a reconstruction attack would perform against subsampling

with $t = 50$ and the same number d of queries? Which attack do you consider to be a more worrisome privacy violation, and why? Discuss.

2. **Research Access vs. Privacy Protection**

   Recall the reading on "Public Access to Genome-Wide Data: Five Views on Balancing Research with Privacy and Protection." Select two of the five views, and in a short paragraph (100-300 words), reflect on how the authors characterize uncertainty about the future when discussing risks or possibilities (eg. around privacy of human subjects, family disclosures, public participation in genomic studies, and scientific breakthroughs that benefit the public). Which future events are depicted as more certain or uncertain than others? What risks or possibilities are downplayed versus emphasized? What are the implicit assumptions made by the authors, and how do these assumptions shape their arguments?