A Firm Foundation for Private Data Analysis

wbg231

January 2023

1 Introduction

- statistical data protection has a long history because it is important
- must focus on rigorous privacy guarantees
- accessing the database should reveal nothing about any individual in it. need to serperate the utility of a database from the increased risk of hard due to joining the database
- this can be achived with low distortion
- the key idea is to randomize responses so as to effectively hide the presence or absence of any individual over the course of the life time of the database
- what does it mean to preserve privacy and how and it be accomplished

how hard is hard

- lets think about common approaches and there short comings

large query set

- one idea is to prevent user from making queries about specific individuals
- should not be able to search on users at all.

query auditing

- each query to a database is evaluated in the context of the query history to determine if the response would discos information if so then refuse the query
- this is bad because it is computationally infeasible and the fact a query wont be discoed its gives information

sub sampling

- can only query from a randomized subset of individuals
- but this is not secure for those who appear in a subsample

input perturbation

- either the data or the queries are modefied before the repone is given
- randomized response flip a coin if it is heads always respond negativly if it is true respond truthly half the time
- this does not work well for complex data

adding radnom noise to the output

- will fail if done naievly
- if the noise is added at the output and is mean zero can repeate a query adn average to get an aproximate true value

non private database

- a system is blatantly non private if an adversary can constrct a canadite databse that agrees with teh real database D in some large proprtion of enteries by querying our system
- if noise is bounded by some uppper bound e (that is a response cal only be e away from the truth) than an adversary can reconstruct the database within 4 E positions of the true values
- bounded noise can be defeated by some number of queries
- so need to bound the number of quries

what is hard

- linkage attack released data linked with auxilarray data to capture ifomratoion about the responsests other than what is released from the database
- need to take into account auxilarray information
- anything that can be learned about an indivuadl from a statistcal database should be learnable with out acess to it

Differential Privacy

- DP ensures that the ability of an adversary to inflect harm (or good) to any set of people should be teh same indepednt of wheter a person is in the dataset
- consider two data bases D, D' which differ by only one row
- a function K is ϵ- Differential private if $\forall (D, D')$ and all $S \subseteq rank(K)$

$$P(K(D) \in S) \leq e^{\epsilon} P(K(D') \in S)$$

- this holds travially for anyone not in the dataset
- if any one particpant were to be added to the database under this regine no output would become more or less liekely

achiving DP

- for a function f the ℓ_1 sensativity of f is gvien by

$$\delta f = max_{D,D'} ||f(D) - F(D')||_1$$

- that is the sensativity is the max l1 difrent between the function with or without any one perosn
- with leplacian noise we have highest density at zero, and for any z, z' : $|z - z'| / leq 1$ the density is at most e^{ϵ}
- also symetric about zero
- can be expanded ot more dimesnions
- need to add leplacian noise to each query even if the queries are chosen adaotily wih each sucessive queiry dpedont on the dprevious awnser this will work
- our expected error is the same regardless of the size of teh dataset
- i think the rest is kinda in the weeds
-