Differential Privacy

BY LEO LIU

Introduction to Data Privacy

Alice has a dataset that she wants to share with analyst Bob

- This dataset contains sensitive data
- Alice wants Bob to be able to learn about trends in the dataset and draw conclusions based on the data
- Alice does not want to reveal any information specific to individuals

A few initial approaches

De-identify the individuals?

Just remove the columns containing identifying information!

- E.g.: name, SSN, address, phone numbers, etc.

No formal definition of “identifying information”

- Therefore, it has no formal privacy guarantees

Subject to linking attacks

- If we know some auxiliary data, it could be combined with the de-identified data to learn sensitive information
- Hospital visit data released by Massachusetts State Group Insurance Commission was linked with public voter records to de-identify then-governor William Weld’s personal health records in 1997.
- Netflix Prize anonymized dataset was linked with records of IMDb users, and watching history is shown to be de-anonymizeable

A few initial approaches (cont.)

k-Anonymity

A dataset is *k*-anonymous if the information of each individual cannot be distinguished from at least *k*-1 other individuals within the dataset based on their quasi-identifiers, where quasi-identifiers are a selected subset of attributes in the dataset.

- If name, zip-code, gender, and birthdate make up the quasi-identifiers, then a dataset is *k*-anonymous if there are at least *k* individuals with the same name, zip-code, and birthdate for every combination within the dataset

Anonymize dataset by aggregating the data

- Suppression: Remove names
- Generalization: use first 4 digits of zip-codes, and use age ranges instead of birthdates

A few initial approaches (cont.)

k-Anonymity

Computationally to even check if a dataset satisfies *k*-anonymity

Optimal generalization is extremely difficult, especially with outliers

- Automatic generalization is NP-hard

Still subject to attacks

- Homogeneity attack: if everyone that shares a quasi-identifier have the same or similar entries for sensitive information, that sensitive information is leaked for the entire group
- Background knowledge attack: attributes outside of the quasi-identifier set that the attacker may know can let the attacker identify individuals

Differential privacy

A property of algorithms, not data.

A function F satisfies differential privacy (aka is a *mechanism*) if:

- For any neighboring datasets x and x' and all possible outputs S,

$$\frac{P(F(x) = S)}{P(F(x') = S)} \leq e^{\epsilon}$$

Informally, given the output of F, it is hard to tell if it came from the full dataset or a subset of the dataset with one individual removed

If an adversary cannot even be sure about presence of an entry, they cannot learn anything about the content of the entry

Differential privacy (cont.)

Even more informally, add enough noise to query results so that the presence and absence of any single entry doesn't really affect them

Differential privacy (cont.)

How much noise to add?

The Laplace mechanism

- For a function $f(x)$ which returns a number,

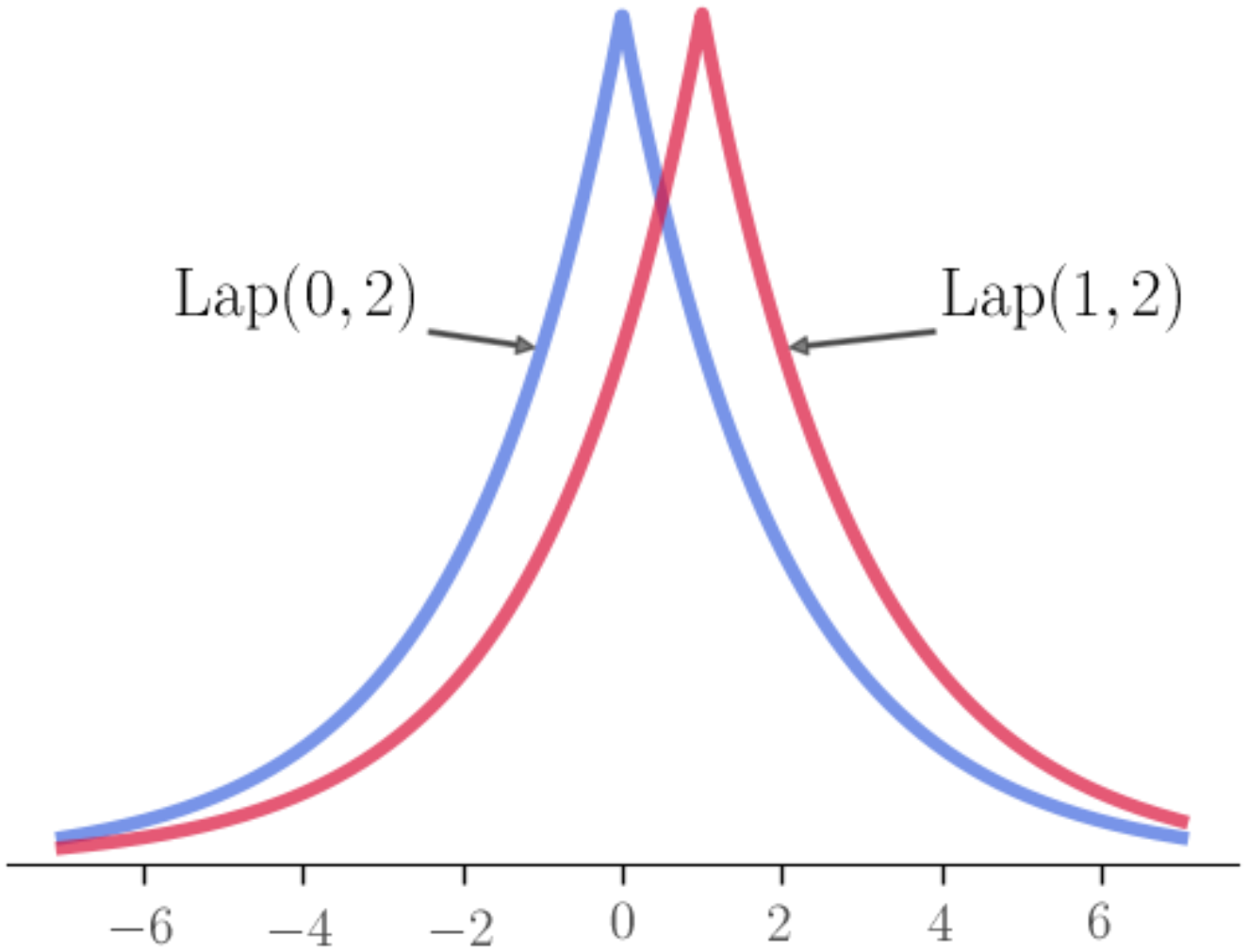
$$F(x) = f(x) + Laplace\left(\frac{S}{\epsilon}\right)$$

$F(x)$ satisfies ϵ-differential privacy where s is the sensitivity of $f(x)$ and $Laplace(S)$ denotes a random sample from the Laplace distribution with center 0 and scale S.

Differential privacy (cont.)

Laplace mechanism offering .5-differential privacy for a function with sensitivity 1.

(source: Wikipedia)



Sensitivity?

Global sensitivity is defined as:

$$GS(f) = \max|f(x) - f(x')|$$

For all x and x' that are neighbors, i.e., if their *distance* is 1.

Informally, sensitivity is the change in $f(x)$ when x changes by 1 entry

- Sensitivity is 1 for counting queries
- Sensitivity is $max - min$ for summation queries on bounded numerical attributes
- What about mean?

Queries on a mock dataset

	A	B	C	D	E	F	G
1	id	first_name	last_name	gender	department	age	salary
2	0	Angela	David	female	Mechanical and Aerospace Engineering	47	83565
3	1	Jean	Munden	female	Mechanical and Aerospace Engineering	36	133332
4	2	Tony	Dinham	male	Finance and Risk Engineering	73	97091
5	3	Cornelia	Wood	female	Applied Physics	22	154483
6	4	Micheal	Olson	male	Chemical and Biomolecular Engineering	27	68116
7	5	Betty	Young	female	Technology Management and Innovation	50	153970
8	6	Horace	Burnett	male	Technology Management and Innovation	64	66305
9	7	Frederick	Gary	male	Chemical and Biomolecular Engineering	43	119809
10	8	John	Brice	male	Finance and Risk Engineering	41	104378
11	9	Lillie	Prucha	female	Applied Physics	33	167237

Complete dataset is held by Alice

- Contains sensitive information like salary

Bob can be an analyst querying the dataset to learn about trends

- What's the average age of all men in the Applied Physics and Biomedical Engineering department who earns between 30k and 60k?

Bob can also be an adversary querying the dataset to obtain sensitive information

- What department is Betty Young in and how much does she make?

Queries on a mock dataset (cont.)

What's the average age of all men in the Applied Physics and Biomedical Engineering department who earns between 30k and 60k?

number of individuals	average age
94.91760056737708	50.82965106725618
99.28460466233581	50.723596539233455
102.64317947825745	47.946052813632505
103.13331394715405	46.005192738560545
100.96577814138537	51.515825647753694

Queries on a mock dataset (cont.)

What department is Betty Young in and how much does she make?

- Department query made by asking “how many Betty Youngs are there in the Computer Science department?”

number of Betty Youngs in CS	Salary
-0.740799704223247	-178641.42491814902
-4.410852318949979	388985.4267772836
-0.6247250774515986	-729264.4705212039
2.303202319721806	-382720.60724892234
0.8743265043087393	657062.3796867629

Limitations

Distance?

- Distance is easy to define for tabular data, like US Census data, where one response is a *row*
- Distance is hard to formalize for others, like trajectory data, social network graphs, time series, etc.

Say “I’m doing my part” and choose a high ϵ?

- Apple reportedly uses $\epsilon = 6$ for MacOS, $\epsilon = 14$ in iOS 10 (and $\epsilon = 43$ in some beta versions)
- Google uses up to $\epsilon = 9$

Works Cited

Near, J., Abuah C. Programming Differential Privacy. 2021. https://programming-dp.com/

Dwork, C., McSherry, F., Nissim, K., Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. 2006

Desfontaines, D. A Friendly, Non-technical Introduction to Differential Privacy. Sep. 9, 2021. https://desfontain.es/privacy/friendly-intro-to-differential-privacy.html

Greenberg, A. How one of Apple's key privacy safeguards falls short. Wired, Sep. 15, 2017. https://www.wired.com/story/apple-differential-privacy-shortcomings/

Questions?

PROJECT AT:
HTTPS://GITHUB.COM/LEOLIU1999/DIFFERENTIALPRIVACYDEMO