

PRIVACY IN FEDERATED LEARNING

It is never too much!

Diego Klabjan

Acknowledgments

- Fanfei Meng, ECE at Northwestern
 - Former student of mine
- Veena Mendiratta, Adjunct Professor in MLDS
 - Provides tutorials on b/f/p (p=privacy)

1. **Model Inversion Attack** : An adversary attempts to infer sensitive information about the data used in the training set by analyzing the trained model's outputs. This can be done by feeding the model with inputs and observing its responses, attempting to reverse-engineer sensitive data.
2. **Gradient-Based Attack** (by server) : Adversaries may try to exploit the gradients shared during the federated learning process. By analyzing the gradients, they could gain insights into the data on individual devices, potentially leading to data reconstruction attacks.
3. **Model Stealing Attack** (by local client) : A malicious participant attempts to mimic the global model by repeatedly querying it and using these queries to construct a copy of the model. This can compromise the intellectual property of the model owner and the privacy of the data sources.
4. **Data Poisoning Attack** (by local client) : Data poisoning attacks involve manipulating the data used by a participant to influence the global model. Adversaries may intentionally introduce biased or malicious data to skew the model's output, potentially compromising its accuracy.

NEED FOR DIFFERENTIAL PRIVACY



Motivation

- Even well-meaning data collection can go bad
 - Netflix competition to [develop a movie recommendation algorithm](#).
 - Released an “anonymized” viewing dataset
 - The datasets could be used to re-identify specific users — and even predict their political affiliation — if you knew a little bit of additional information about a user.
- Concern
 - Companies share data
 - Breaches happen
 - Statistics about a dataset can even leak information about the individual samples used for computation

➔ Differential Privacy is a set of tools designed to address this problem

A Simple Explanation of DP

Imagine you have two otherwise identical databases:

- one **with** your information in it, and 
- one **without** it. 

DP ensures that:

- the probability that a statistical query will produce a given result
- is (**nearly**) the same with each database

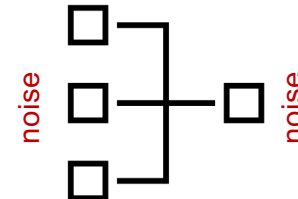
Why Nearly the Same

- Computing average
 - User opts out
 - Value of user must be the same as average
 - Highly unlikely
 - Needs to hold for every individual
 - All users must have the same value
- Opt out = remove from database
 - Every participant
 - End up with empty database



DP: Algorithms and Queries

- DP applies to **queries**, not databases.
- DP adds **noise** to protect privacy
 - Noise on input or
 - Noise on output
- More intrusive query → More noise added
- Referred to them as algorithms and not queries
 - Point is to emphasize that a query is not always a simple SQL query.
 - Can be clustering
 - Can be classification



Data Privacy: the Problem

- Given a dataset with sensitive personal information:
 - how can we **compute and release functions** of the dataset,
 - while **protecting individual privacy**?
- Common intuitive solutions:
 - **Anonymization** - trusted data curator removes **identifiers** such as SSN, name, etc. to get anonymity and, hence, privacy
 - **Aggregation (statistical analysis)** - counts, averages, statistical models, classifiers, etc., are safe

Anonymization

A hospital publishes **person-specific** patient data:
- information is practically useful
- individual identity is not disclosed

	Non-Sensitive Data			Sensitive Data	
#	Zip	Age	Nationality	Name	Condition
1	13053	28	Indian	Kumar	Heart Disease
2	13067	29	American	Bob	Heart Disease
3	13053	35	Canadian	Ivan	Viral Infection
4	13067	36	Japanese	Umeko	Cancer

Published
Data

	Non-Sensitive Data			Sensitive Data
#	Zip	Age	Nationality	Condition
1	13053	28	Indian	Heart Disease
2	13067	29	American	Heart Disease
3	13053	35	Canadian	Viral Infection
4	13067	36	Japanese	Cancer

Voter List
(public)

#	Name	Zip	Age	Nationality
1	John	13053	28	American
2	Bob	13067	29	American
3	Chris	13053	23	American

Data
Leak

Massachusetts GLC Linkage Attack



Anonymized Health Data for research

- Ethnicity, Visit Date, Diagnosis, Procedure, Medication, Charge
- **Zip, Birthdate, Sex**



Voter registration records publicly available

- Name, Address, Date Registered, Party Affiliation, Date Voted
- **Zip, Birthdate, Sex**

Massachusetts GLC Linkage Attack

The medical records of William Weld
(governor of MA) were re-identified



Per the Cambridge Voter list:

- 6 people had his same birth date
- 3 of the 6 were men
- He was the only one in his 5-digit ZIP code

A common phenomenon

dob+5 digit zip

re-identify 69% of Americans

dob+9 digit zip

re-identify 97% of Americans

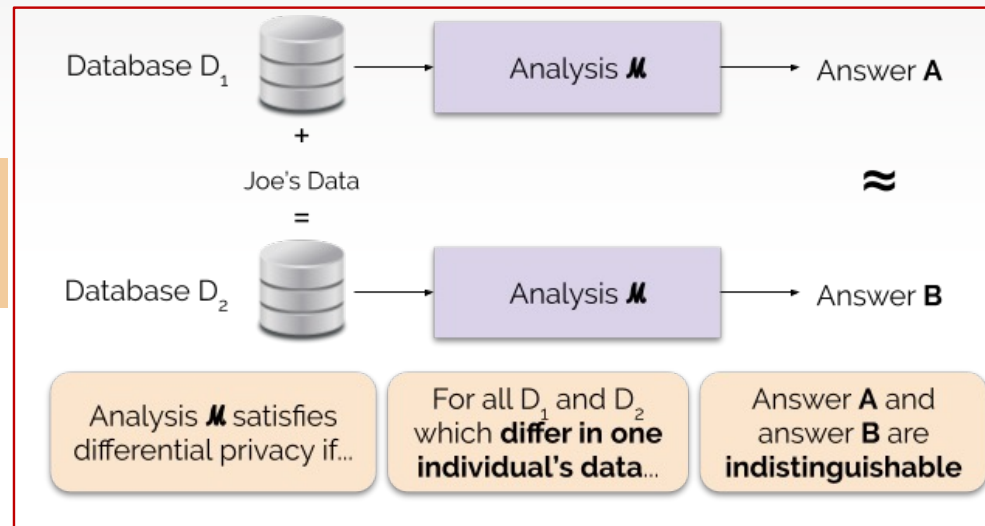
DEFINITIONS

DP Properties: informal definition

Guarantee:

for each individual, who contributes data for analysis, the output of a DP analysis will be roughly the same, whether they contribute their data or not.

\mathcal{M} is a randomized algorithm



DP Properties: Formal Definition

The strength of the privacy guarantee is controlled by tuning the **privacy parameter ϵ** , also called a **privacy loss or privacy budget**.

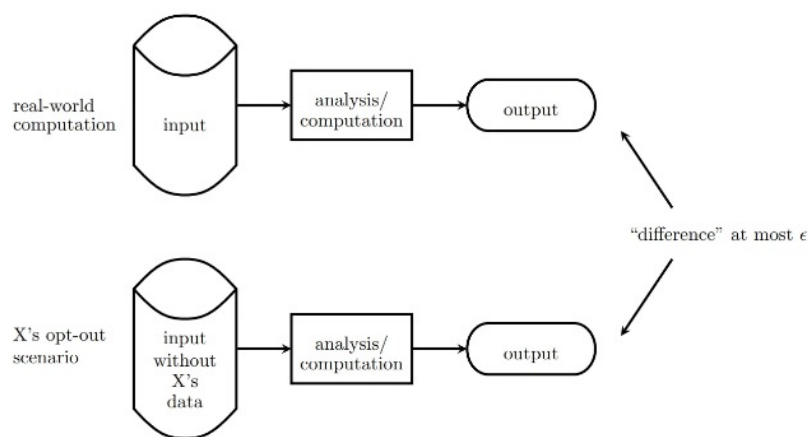
$$\frac{\text{Probability of seeing output } O \text{ on input } D_1}{\text{Probability of seeing output } O \text{ on input } D_2} = \frac{\Pr[\mathcal{M}(D_1) \in O]}{\Pr[\mathcal{M}(D_2) \in O]} \leq e^\epsilon$$

Indistinguishability:
bounded ratio of probabilities

ϵ = maximum distance between a query on database D1 and the same query on database D2

The lower the value of ϵ , the more indistinguishable the results, and, therefore, the more each individual's data is protected.

Layman Definition



Kobbi Nissim, et al. [Differential Privacy: A Primer for a Non-technical Audience](#). February 14, 2018

- Machine unlearning
 - Opt-out requirements
 - Spawned by GDPR
- Probability accommodates randomness in algorithms
- Must hold for any databases that differ by a single row
 - Can be generalized

Epsilon (ϵ)

- **Small ϵ :**
 - better privacy but less accurate response.
- **Small ϵ :**
 - required to provide **very similar outputs** when given similar inputs.
- **Large ϵ :**
 - allow **less similarity in the outputs**, less privacy.

$$\exp(\epsilon) \approx (1+\epsilon)$$

APPROACHES

Definitions

- δ = odds that something goes wrong
- (ϵ, δ) -DP means the algorithm is ϵ - differentially private with probability $1-\delta$

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\epsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta,$$

- Can be shown
 - Expression means with probability $1-\delta$
 - Non-trivial proof
- Meaning
 - Most of the time privacy is preserved
 - Sometimes is not (low probability)

Sensitivity

- Sensitivity

$$\Delta f = \max_{\substack{x, y \in \mathcal{X} \\ \|x - y\|_1 = 1}} \|f(x) - f(y)\|_1$$

- Similar to Lipschitz
- Interpretation
 - One record difference in databases x, y
- Can be shown
 - k records removed
 - Algorithm $(\epsilon, 0)$ becomes $(k\epsilon, 0)$
 - Examples?

Composition

- Chain several algorithms
 - Average and then cluster
- Most algorithms additive in ϵ
- Gives you privacy end-to-end
 - Need to only specify total ϵ
- Applies also to several queries for the same quantity
 - Several queries for average
 - Due to randomness each time a different outcome

Laplace Mechanism

- $(\epsilon, 0)$ -Differential Privacy
- Computes $f(x)$ and adds noise drawn from the Laplace Distribution.

$$\text{Lap}(x|b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$$

- The scale of the noise b is calibrated to the (L1) sensitivity of $b = \nabla f / \epsilon$
 - In practice, trial and error
- Works best with numeric queries with low sensitivity.

Exponential Mechanism

- $(\epsilon, 0)$ -Differential Privacy
- Applicable to numeric and categorical functional query outputs.
 - Used mostly for categorical cases
- Allows selecting the "best" element from a set while preserving its DP.
- Releases only the identity of the element with the max noisy score and not the score itself.

Exponential Mechanism

- (x, o) = (sample, category observation)
- $u(x, o)$ = utility function
 - In classification, probability of the model
 - In clustering, distance from centroid
 - In regression, residual value
- Draw outcome based on distribution

$$Pr[o] = e^{\frac{\epsilon * u(x, o)}{2\Delta u}}$$

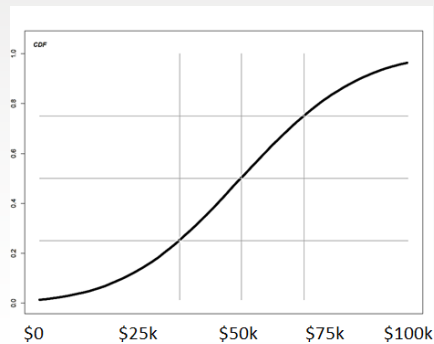
- Always member of the choice set (different from Laplacian and Gaussian)

Gaussian Mechanism

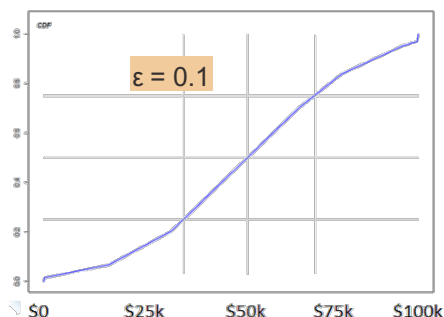
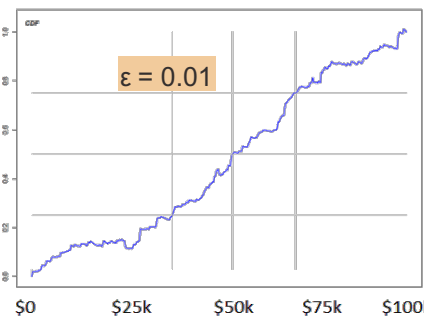
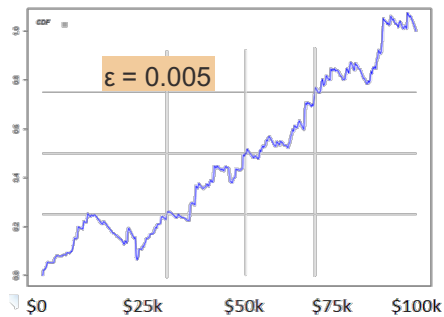
- (ϵ, δ) -Differential Privacy
- Very similar to Laplace Mechanism
 - Accounts for δ
- Both L1 or L2 sensitivity can work.
- L2 sensitivity is lower than L1 sensitivity in value
 - Allowing for less noise addition.
- $F(x) = f(x) + \mathcal{N}(\sigma^2)$
 - $\sigma^2 = \frac{2\Delta f^2 \log\left(\frac{1.25}{\delta}\right)}{\epsilon^2}$
 - Here sensitivity is L2
- Works for f being a vector
 - When L2 sensitivity is much lower than L1, use Gaussian

Differentially Private Computations

CDF of Income Distribution



Algorithms maintain DP by introduction of noise in the computation.



Kobbi Nissim, et al. [Differential Privacy: A Primer for a Non-technical Audience](#). February 14, 2018

Connection to FL

- Each client applies Laplacian or Gaussian for numeric features at input
 - Better to apply for embeddings
 - Follows theory
- For categorical features use Exponential
 - Embeddings are continuous
- Sensitivity hard to estimate
 - Use as hyperparameter
- Further imposes privacy
 - At what cost?

FL and DP

- P = differential privacy
- Listed (μ, β)
- Not big decrease due to DP

Scheme & Privacy		Datasets		MNIST		FASHION		KDD	
				Acc.	Clu.	Acc.	Clu.	Acc.	Clu.
Centralized		N/A		0.993	60k	0.895	60k	0.925	300k
FATE		Homo.		0.941	60k	0.795	60k	0.923	300k
FedEmb		Ran.		0.961	15k	0.828	15k	0.923	15k
		KM.		0.950	15k	0.812	15k	0.923	15k
		Ran.		0.954	1.2k	0.825	1.2k	0.924	1.2k
		KM.		0.951	1.2k	0.812	1.2k	0.924	1.2k
		Ran.		0.947	0.6k	0.820	0.6k	0.924	0.6k
		KM.		0.950	0.6k	0.813	0.6k	0.923	0.6k
FedEmb (G.)		Ran. + 0.21%		0.955	15k	0.798	15k	0.923	15k
		Ran. + 0.83%		0.951	15k	0.799	15k	0.923	15k
		Ran. + 3.33%		0.952	15k	0.784	15k	0.922	15k
		Ran. + 0.21%		0.952	1.2k	0.813	1.2k	0.924	1.2k
		Ran. + 0.83%		0.954	1.2k	0.793	1.2k	0.924	1.2k
		Ran. + 3.33%		0.955	1.2k	0.804	1.2k	0.923	1.2k
FedEmb (P.)		Ran. + (0, 0.25)		0.952	15k	0.804	15k	0.923	15k
		Ran. + (0, 0.50)		0.888	15k	0.803	15k	0.924	15k
		Ran. + (0, 1.00)		0.952	15k	0.815	15k	0.922	15k

DP in Practice



Recommended

Epsilon in the range [0.1 and 1].



The number of queries possible with $\epsilon < 1$ is small (maybe 10s)

Composition



Some mechanisms make assumptions about the lack of correlation between attributes and for the same attribute over time.

Allows making unlimited queries
Noise is high (1000s count errors)



Many DP implementations use $\epsilon > 10$

Resulting in user privacy loss.
Mechanism with a small epsilon can remove most of the information value.