TU Graz

SCIENCE
PASSION
TECHNOLOGY

1.

Privacy-Preserving Data Science

João Machado de Freitas
KDDM2

Motivation

What is motion? A problem from the 300's B.C.

Aristotle's Motion

- *Motion is the fulfillment of that which exists potentially*
- As many types of *motion* as there are meaning of the word *is*.

Newton's Laws of Motion, XVII century

- Do little to answer many of the questions about motion which Aristotle considered.
- It is about the movement of point particles

Motivation

What is Privacy? A problem for the 2000's A.D.

$$\text{Newton's Laws of Motion} = \text{Motion} \cap \text{Point Particles}$$

$$\text{Privacy} \cap \text{Data Science} = ?$$

Motivation

Dataset anonymization

1.

Dataset anonymization substitute/remove identifiers and sensitive information.

Netflix Challenge (2008) by finding the best match ⇒
Netflix anonymized data + public ImDB data = Re-identified Netflix data

Anonymized Data Isn't

Dataset anonymization is a fundamentally broken technique and should not be used.

k-anonymity (1998) 1.

Idea Make sure there are more people with the same set of combinations of *pseudo-identifiers*.

Terminology

- **Identifiers** *name* or *ssn* – unique
- **Pseudo identifiers** *(zip, dob, gender)* – not unique, but together they identify a person
- **Sensitive attributes** *diagnostic, income, ...*

k-anonymity (1998)

Solution

- Redact information from individual records so that a set of characteristics matches at least $k-1$ individuals.
- If for any setting of pseudo-IDs, there are at least $k-1$ other subjects with the same setting of pseudo-IDs, then we have k-anonymity.

k-anonymity (1998) – a faulty solution

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Problems

- **Does not** prevents **record re-identification** if multiple datasets are released – **linkage attacks**
- The k-anonymous sets with homogeneous sensitive attribute leak information – no **plausible deniability** (ability to deny something)

Reconstruction Attack 1.

Diffix and Aircloack Challenge (2017) Reconstruct private database with unlimited number of queries, but limiting the query type

1. Get aggregated statistics by querying database
   - How many rows satisfy [CONDITION] and have *has_secret* = *True*
2. Generate constraints (e.g. 0<age<125)
3. Find feasible point using constrained optimization solver.
   - **NP-Hard** because of integer constrained values
   - In practice, easy to solve

Aggregated Statistics

1.

Genome Wide Association Studies (GWAS) release relative proportions of each allele frequency

- There are hundreds of thousands or millions of Single Nucleotide Polymorphisms (SNPs)
- Minor allele frequency, χ^2-statistics, p-values, ...

Homer et al. (2008) Simple *correlation test* is enough to test whether a particular individual was part of the GWAS group – a ***membership inference attack***

National Institutes of Health (NIH) ended up restricting free access

Memorization in Neural Networks

1.

- **Model parameters** are also vulnerable, since they are another kind of aggregated data.
- *Are training set observations predicted with higher confidence than observations in the test set?* Low perplexity ⇒ NN memorized data point.
- **Membership inference attacks** determine if a target individual is in the dataset or training set.

Memorization in Neural Networks

1.

On the left there is an image recovered using a new model inversion attack and, on the right, a training set image. The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Cryptography

1.

- **Cryptography** solves a different problem. A lot of times it deal with the security, not the privacy of the data
- *Privacy guarantees in case the encryption is compromised?* The lifetime of cryptosystems is usually short.
- Cryptographic techniques increase computation and communication cost a lot

Privacy Preserving Data Science

- In the data science life cycle we want to ask arbitrary queries, visualize, manipulate the data at will
- We want to publish the **data** or **statistics** or **models**.
- Generally we want to release some properties about data to the world – we worry about **unwanted inferences** by an adversary.

Privacy Expectations

Unreasonable

- **Privacy for free** Removal of information without accuracy loss
- **Absolute privacy** your friend and family habits are correlated with yours, they leak your information with theirs.

Reasonable

- **Quantitative** control accuracy vs. privacy and quantify accuracy loss.
- **Plausible deniability** yours presence in a database cannot be ascertained.
- **Prevent targeted attacks** limit information leaked even in the presence of side knowledge.

1. Let's set the expectation for Privacy-preserving statistics or data science ...
2. There is no removal of information without loss of accuracy in statistical privacy. Meaning there is no privacy for free.

Recap

Releasing "too many" and "too accurate" aggregated statistics makes one vulnerable to:

- **Database Reconstruction**
- **Linkage attacks**
- **Membership inference attacks**

Aggregated statistics are not safe

1.

Intuition

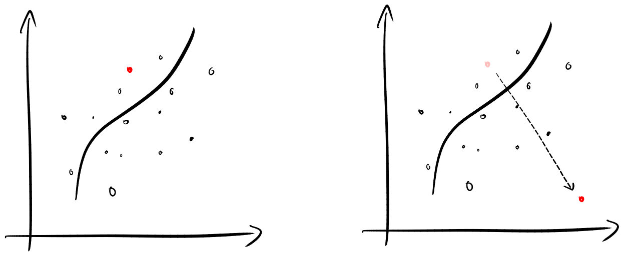
Why Differential Privacy? A quantitative theory for "too many" and "too accurate".

1. An individual data point will have almost no impact on the output of a differential private algorithm – **DP is an algorithm's property.**
2. A privacy notion centered on hiding participation in a dataset
3. Provides **plausible deniability** and doesn't ban any particular use of data
4. Protection against **linkage attacks** from multiple data releases (even future ones)

1.

Statistical Learning Theory Perspective

Related with the **generalization/stability** properties of learning algorithms



1. Your fitted model does not change much even if you change/remov
and individual point

Definition

Given the input space $\mathscr{X}$ of databases, a privacy-preserving mechanism $M : \mathscr{X} \to \mathscr{Y}$ provides ε-differential privacy ($\varepsilon \geq 0$) if for all events $\mathscr{E} \subseteq \mathscr{Y}$ and for all datasets $x, x' \in \mathscr{X}$, such that $x \approx x'$, we have:

$$\frac{P[M(x) \in \mathscr{E}]}{P[M(x') \in \mathscr{E}]} \leq \exp(\varepsilon)$$

The neighbouring relation $\approx$ is symmmetric and captures what is protected. E.g. replace/remove one entry; in location privacy it means to move by d much

1. DP is a quantitative definition of privacy. DP is also a property of algorithms.
2. Given and input space x of databases, a privacy-preserving mechanism M provides epsilon DP if for all events and for all neighboring databases, we have the following bound.
3. **Mechanism** Stochastic mapping, randomized algorithm, a random variable.
4. For location privacy, the neighboring relation mean to move by d much

Comments

- **Worst-case definition** for every pair of datasets x and x' and possible outputs: Hard to verify algorithmically!
- Quantitative definition parameterized by ε: should be small $0.1 \leq \varepsilon \leq 5$
- Any DP algorithm must be randomized: $M(x)$ needs to be a random variable.

1. We can say that DP bounds the multiplicative increase in the probability of M's output satisfying any event when you change one data point.
2. Epsilon should be small. However, there are cases in the literature where ε is even 100.
3. Also, any DP algorithm must be randomized. We see that in the definition where M is a random variable.
4. DP is an information theoretical definition of privacy, because it do not depend on computational assumptions of the adversary
5. We can also say that DP is a privacy definition against statistical inference.

Isn't

What would be known even with the individual's data removed.

- E.g. If you smoke your insurances rates go up, even if you didn't participate in any study that connect smoking with increased risk of lung cancer

What others tell about you – *family genome, social network friends, etc.*

- **Facebook likes** allow to discover political affiliation, religion, use of drugs, cigarettes, and alcohol, if parents divorces before user turned 21, etc. (Cambridge, 2013)
- **Strava case** aggregated data from fitness tracking devices revealed location of US bases (state secrets) while protecting individual jogging routes.

1. What isn't Differential Privaccy
2. DP does not protect you from study results. Whether you participat or not.
3. It also doesn't protect against what other tell about you. DP is not appropriate for social networks data, or in the case of scarce data in location data analysis.

History

1. Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith (2006) **Calibrating Noise to Sensitivity in Private Data Analysis**
2. **Test of Time Award 2016** Dwork et al.
3. **Gödel Prize 2017** Dwork et al.
4. **Knuth Prize** and **IEEE Richard W. Hamming Medal 2020** Cynthia Dwork

Oldest DP Algorithm is **Randomized Response** (Warnen, 1965)

1.

Randomized Response

Goal: find the proportion p of students that cheated in the final exam.

1. Answer truthfully with probability $1/2+\gamma$
2. Lie with probability $1/2-\gamma$

Provide **plausible deniability** for each individual answer to illicit a honest answer.

1. Randomized Response was created for sensitive surveys. For examp we might want to find out how many students cheated. Normally they will not answer truthfully, so RR devises a strategy to illicit tha
2. It survey student with a binary question: "did you cheat?"
3. But this is not asked directly. So... This provides...
4. With $\gamma = 1/2$ we have maximum utility, but zero privacy. Students with always answer the truthfully - doesn't solve our reponse bias problem.
5. With $\gamma = 0$ we have zero utility but maximum privacy wit han uniformly random response. Answer truthfully and lies with probability one half.

Randomized Response – Analysis

$$X_1, \dots, X_n \sim Ber(p) \qquad p = E[X_i]$$

$$Y_i = RR_\gamma(X_i) = \begin{cases} X_i & \text{w.p. } 1/2 + \gamma \\ 1 - X_i & \text{w.p. } 1/2 - \gamma \end{cases} \qquad \gamma \in (0, 1/2)$$

1. Let's say we have n students, and their true response X follows a Bernoulli with parameter p
2. p is the true proportion of cheating students
3. X_i are unobserved, while Y_i, the output of the RR algorithm, are observed

Randomized Response – Analysis

Goal

$$\mathbb{E}[\tilde{p}(Y_1, \dots, Y_n)] = p$$

We know that:

$$\begin{aligned} \mathbb{E}[Y_i] &= (1+\gamma)X_i + (1-\gamma)(1-X_i) \\ &= 2\gamma X_i + 1/2 - \gamma \end{aligned}$$

1. The Goal is to find an good estimator $\tilde{p}$ of p. Remember that an estimator is a random variable, while p is the true value and a deterministic number
2. We know that Y_i and X_i are related by the expectation of Y_i

Randomized Response – Analysis

Thus, we can find the **unbiased estimator** for X_i,

$$\mathbb{E}\left[\frac{1}{2\alpha}\left(Y_i - \frac{1}{2} + \gamma\right)\right] = X_i$$

We can get a cadidate estimator $\tilde{p}$ of p if we average the the X_i estimator

$$\tilde{p} = \frac{1}{n}\sum_i^n \frac{1}{2\alpha}\left(Y_i - \frac{1}{2} + \gamma\right)$$

This estimator is also unbiased: $\mathbb{E}[\tilde{p}] = p$

Thus we can find the estimator with some simple arithmetic and properties expectation
We can also check that this estimator is unbiased

Randomized Response – Analysis

From the properties of variance and given that Y_i are independent,

$$\mathbf{Var}[\tilde{p}] = \frac{1}{4\gamma^2 n^2}\sum_i^n \mathbf{Var}[Y_i] \leq \frac{1}{16\gamma^2 n}$$

From Chebyshev's Inequality ($k > 0$)

$$\mathbb{P}\left[|\tilde{p} - p| < \frac{k}{4}\frac{1}{\gamma\sqrt{n}}\right] \geq 1 - \frac{1}{k^2}$$

1. Y_i is also Bernoulli distributed, and the variance of a Bernoulli r.v. i at most $1/4$
2. Chebyshev inequality allows to check that the absolute difference between the true parameter and the estimator decreases with squar root of n.

Randomized Response – Analysis

How ε-DP is RR_γ?

$$X = (X_1, \ldots, X_n) \qquad X' = (X_1, \ldots, X'_n)$$

For any particular binary string $b \in \{0,1\}^n$

$$\mathbb{P}[RR_\gamma(X) = b] = \prod_i^n \mathbb{P}[Y_i = b_i]$$

1. Finally, how epsilon-DP is gamma-Randomized Response?
2. Consider two datasets X and X' (X prime) that differ only in one ent the last one.

Randomized Response – Analysis

$$\frac{\prod_i^n \mathbb{P}[Y_i = b_i]}{\prod_i^n \mathbb{P}[Y'_i = b_i]} = \frac{\mathbb{P}[Y_n = b_n]}{\mathbb{P}[Y'_n = b_n]} \leq \frac{1/2 + \gamma}{1/2 - \gamma} = \exp(\varepsilon)$$

$$\text{RR}_\gamma \text{ is } \left(\log\frac{1/2+\gamma}{1/2-\gamma}\right)\text{-DP}$$

1. For all events, there is only input that changes X_n. To most factor ar eliminated as we have the following bound.

Differential Privacy
Randomized Response – Comments

Privacy of our estimate $\tilde{p}$ will follow by the **post-processing property** of DP – *essentially saying that a function of a differetially private object is also private.*

Stronger notion than **global-DP**. RR provides **local-DP**, there is not need for a **trusted curator** – *central aggregator*.

- **Local-DP** $|\hat{\theta} - \theta| \leq O(1/\varepsilon\sqrt{n})$
- **Global-DP** $|\hat{\theta} - \theta| \leq O(1/\varepsilon n)$

1. Also RR is more than DP. Is local-differentially private, since the individual can protect it's own privacy and there is no need for a trusted curator, also known as, trusted aggregator

Differential Privacy
Laplace Mechanism

Global-DP with Laplace mechanism for computing a mean

1. Curator holds an observation x_i for each of the n observations
2. Computes sample mean $\mu = 1/n \sum_i^n x_i$
3. Sample noise $Z \sim Lap(1/\varepsilon n)$
4. Reveals the noisy mean $\tilde{\mu} = \mu + Z$

$$|\tilde{\mu} - \mu| \leq O(1/\varepsilon n)$$

1. The Laplace mechanism for computing means is the following ...
2. Part of a larger family of mechanism that perturb the output.
3. For instance, if we have Gaussian noise, we will have the Gaussian mechanism. However, the Gaussian mechanism is not ϵ-DP (next)

Differential Privacy
Approximated DP

A randomized algorithm is (ε, δ)-DP if

$$P[M(x) \in \mathscr{E}] \leq \exp(\varepsilon)\, P[M(x') \in \mathscr{E}] + \delta$$

The slack delta $\delta \in [0, 1]$ accounts for "bad events" that might result in high privacy losses. It should be very small $\delta \ll 1/n$

Laplace mechanism is ε-DP   **Gaussian mechanism** is (ε, δ)-DP

1. The Gaussian mechanism is approximated DP, which is needed also for some other mechanisms.
2. Account for the probability of releasing the true statistic of the data

Differential Privacy
Properties

1. **Robustness to post-processing** If M is (ε, δ)-DP, then $F \circ M$ is (ε, δ)-DP
2. **Composition** $(\sum_i \varepsilon_i, \sum_i \delta_i)$-DP
3. **Group privacy** If M is (ε, δ)-DP, then M is $(k\varepsilon, k\delta)$-DP for k changes.
4. **Protect against side-knowledge** if attacker has a prior and computes the posterior after observing the output of a (ε, δ)-DP mechanism M, then *distance*(prior, posterior) is bounded by ε

1. DP itself, is a property of algorithms. But what other properties DP algorithms have

Differential Privacy

Applications

Google, Apple, Microsoft, LinkedIn, US Census

E.g. Collect telemetry data from browser, operating system, etc.

There are already some commercial application using differential privacy. Apple and Microsoft use it to collect telemetry data from their operating syster
And Google uses in Chrome browser and to learn from your Android's keyboa
The US Census uses it to publish aggregated statistics.

Differential Privacy

Beyond DP

> Differential Privacy captures much what we could (reasonably) want in a privacy definition
>
> *The Ethical Algorithm*

It seems to exist a consensus that differential privacy captures much what could (reasonably) want in a privacy definition. But there are some limitatio to DP that need to be considered (next)

Differential Privacy

Limitations

- The choice of the privacy budget ε is difficult: in the literature we can find values varying from 0.01 to 100.
- Unrealistic assumption that adversary has unlimited computational and knowledge penalizes model utility too much.
- Guarantees decrease exponentially with the size of the group – it is vulnerable to correlated data.
- ε-DP algorithm does not provide any guarantee against **information leakage**.
- Applications have **large sample complexity** and provides very **limited utility** for small data.

1. It places unrealistic assumptions that adversary has unlimited computational power and knowledge and this penalizes model utili too much. Utility is not considered explicitly.
2. **Information leakage** It does not guarantee that an adversary cann learn something about a specific feature of the dataset. Often there are things we don't want the model to learn - invariances, nuisance biases, and so on.

Differential Privacy

Limitations

DP **does not protect** against what *correlations* in an whole dataset tell about you or a fact.

- How to remove confidential information from datasets?
- How to remove correlations from datasets or models before releasing or sharing them?
- How to prevent unintended inferences about a secret from the entire datasets or model?

From protecting each entry individually to protecting some secret about an ent dataset

Fair and Censored Representations

Private and Fair Presentations

Let X be a dataset, Y is the true label, $\hat{Y}$ is our prediction (the representation $Z = \hat{Y}$) and $S \in \{0, 1\}$ is the sensitive (binary) attribute we want to protect.

Statistical parity	$\hat{Y} \perp S$	$P(\hat{Y} \mid S = 0) = P(\hat{Y} \mid S = 1)$
Error parity	$\hat{Y} \perp S \mid Y$	$P(\hat{Y} \mid Y = y, S = 0) = P(\hat{Y} \mid Y = y, S = 1)$
Sufficiency	$Y \perp S \mid \hat{Y}$	$P(Y \mid \hat{Y} = \hat{y}, S = 0) = P(Y \mid \hat{Y} = \hat{y}, S = 1)$

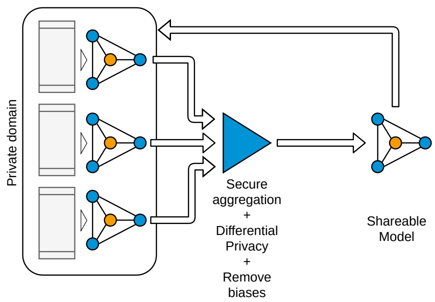
Statistical parity For any value that the sensitive attribute takes we will have the same amount predictions for each class.

Error parity For any value that the sensitive attribute takes we will have the same error rates.

1. It turns out, that the problem of removing confidential information from a dataset or model is equivalent to ensuring statistical parity i algorithmic fairness.
2. In statistical parity we want the same distribution for different valu of a sensitive attribute. This is equivalent to a independence constraint between the algorithm output and the sensitive attribute

Federated Learning

Federated Learning



1. Federated Learning is about collaborative learning, or learning from decentralized data. The main id the following,
2. In Federated learning, we have a shareable model that is sent to multiple devices, like our cellphone PCs.
3. For example, Android's keyboard suggests new word while you type. When you accept or reject the suggestion you are labelling the data. Then, during the night your cellphone computes the gradients the language model that makes the suggestions, and sends this data, encrypted, to a central server.
4. There, the gradients can be aggregated while encrypted. At this point it can also be used differential privacy to ensure that average gradients will not leak individual information. Then, they are de-encrypted and this differentially private gradients estimate is used to train the model further.
5. The gradients or parameters are being estimated with data from a single user. So we no longer have IID assumption that normally we consider in statistics and machine learning. This introduces biases while training.
6. Also some users produce much more data than others and we should ensure that the model doesn't overfit to them. (next)

Federated Learning

Federated Learning - Limitations

- Each dataset may have some bias w.r.t. the general population. *E.g. different size.*
- Local datasets vary with time - **temporal heterogeneity**.
- Nonexistence of global training data: **data is non-IID**.
- Attacker might try to poison the global model by feeding it **fake data**.

1. Also, not all devices are available at the same time.
2. The model are vunerable to data poisoning attacks where the user bot labels data incorrectly to make the model learn stuff incorrectly

Applications

Other Applications

- **Text representations** Learn privacy-preserving language models.
- **Genomic Privacy** hide sensitive genotypes e.g. that allow identification
- **User feedback** software monitors collect user statistics without compromising privacy. Enable data/model sharing.
- **Smart meters**[1] allows third-party to establish a profile of the activities being undertaken. E.g. Protect type and number of appliances.

[1] Electricity, water, heating and gas readings

Other Applications

- **Self-driving cars** Federated learning can represent a solution for limiting volume of data transfer and accelerating learning processes.
- **Personal assistant systems** Protect interactions to avoid unintended uses, like voice identification and voice cloning for speech synthesis.
- **Public Health** Public health monitors for Influenza. Privacy for contract tracing and flow modelling.
- **Census tools** to disclose data to the public.
- **Location privacy** Bike sharing, car sharing etc.
- **Vehicular networks privacy** Improving safety coordination and services in traffic management and real-time information sharing.

References

1. Michael Kearns and Aaron Roth, **The Ethical Algorithm**, 2019
2. Gautam Kamath. **CS 860 – Algorithms for Private Data Analysis**, Fall 2020
3. Borja Balle, **A short tutorial on differential privacy**, January 2018
4. federated.withgoogle.com

Thank you!