CS291A Introduction to Differential Privacy (and differentially private machine learning)

Instructor: Prof. Yu-Xiang Wang

Fall 2021

COMPUTER SCIENCE
UC SANTA BARBARA
Computing. ReInvented.

What is privacy? What are the differences between "privacy", "confidentiality" and "security"?

- Philosophically,
  - "Confidentiality" = "Don't tell"
  - "Privacy" = "Don't ask"

- Legally speaking: Privacy is about the right to be left alone (from public scrutiny); Confidentiality is about a promise from people who have privileged access.

- "Security" vs "Privacy":
  - Security is to prevent risks due to unintended system use.
  - Privacy prevents risks due to intended system use.

Importance of privacy by the United Nations

Universal declaration of human rights

Article 12. No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.

https://www.un.org/en/about-us/universal-declaration-of-human-rights

Personal data in Big Data era

- Government, company, research centers collect personal information and analyze them.
- Social networks: Facebook, LinkedIn
- YouTube & Amazon use viewing/buying records for recommendations.
- Emails in Gmail are used for targeted Ads and for completing your sentence!

Source: https://www.garyfox.co/social-media-statistics/

Recent legislations on privacy forces businesses to revise their data practice

- I can’t keep personal data for more than three weeks?
- I will have to delete all traces of a user upon request?

How about my machine learning models trained on user data?

Scientists need to have access to data, but

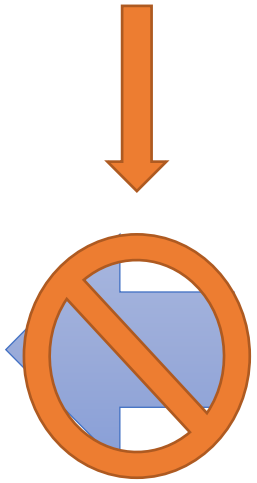
Privacy law

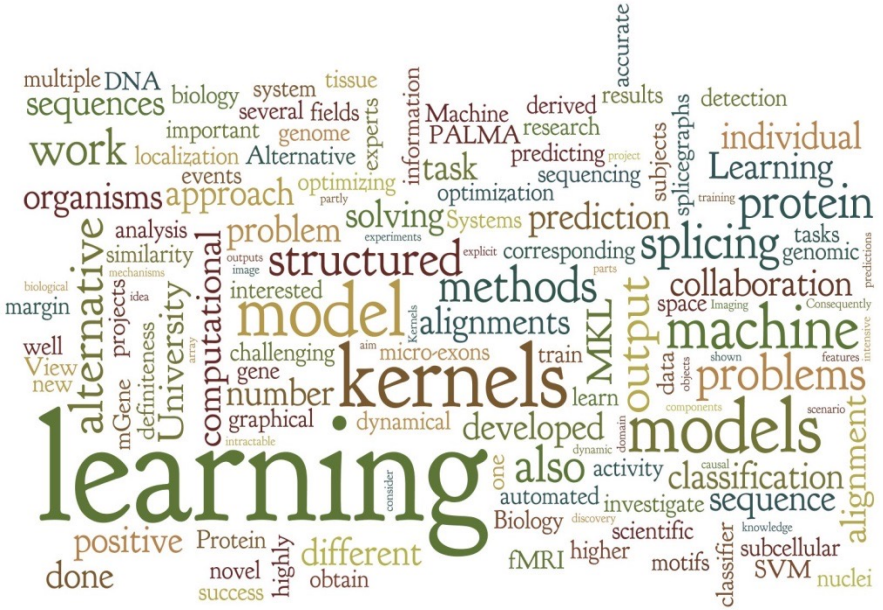
Food safety

Understand how brain works

Better treatment

Better education





Medical diagnosis

Clean energy

Epidemic forecast/prevention

Fraud detection

Increasing privacy awareness

The census bureau's dilemma

Imaged by Heritage Auctions, HA.com

- Article I, Section 2 of the United States Constitution requires the US Census to count individuals **accurately** for tasks such as "**congressional apportionment**".

- Title 13 of the United States Code. "It is **against the law** for any Census Bureau employee to disclose or **publish any** census or survey **information that identifies an individual** or business. This is true even for inter-agency communication: the FBI and other government entities do not have the legal right to access this information. "

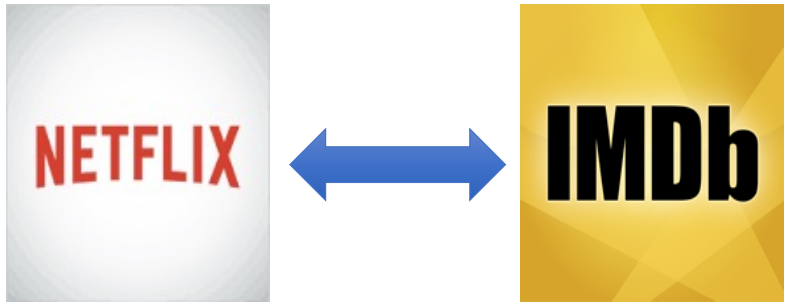
- What if the two goals contradict each other?

Do we really need math / science for “privacy”?

- Can’t we just remove personal identifiable information from the data so that it is de-identified?
- We are only seeing aggregate statistics usually?
- Secure multi-party computation (MPC) and federated learning have made it possible for companies to train ML models with my data while keeping my data on my device.

Removing/modifying personal identifiable information

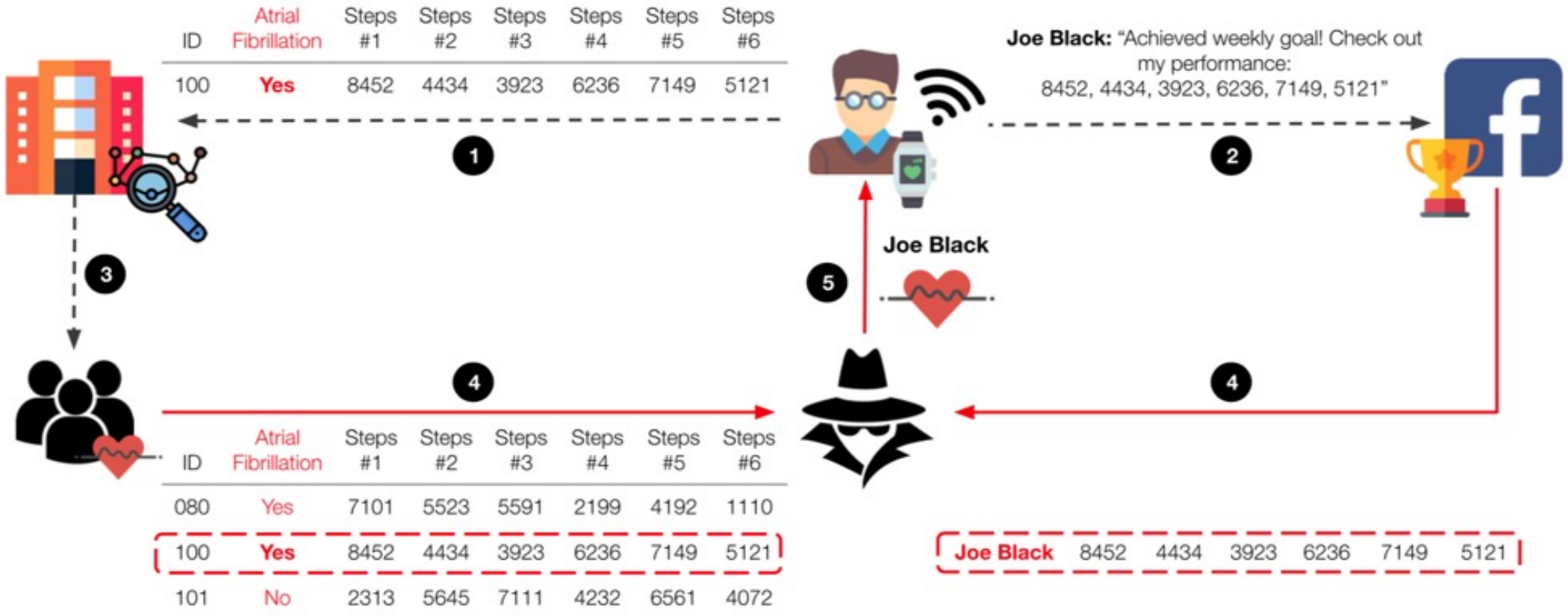
- Name? Gender? Zip code? Watched movies?



- Fragile under appropriate **side information**
- Easier to get in Big Data era

Narayanan, Arvind, and Vitaly Shmatikov. "How to break anonymity of the netflix prize dataset." (2006).

“Just six days of step counts are enough to uniquely identify you among 100 million other people.”



https://www.mobihealthnews.com/news/contributed-when-fitness-data-becomes-research-data-your-privacy-may-be-risk

Differencing attack and side information identifies individuals from aggregate statistics

- “Who likes Justin Bieber?”

- Questionaire: “Year, Program, Gender, Like Bieber or not?”
  - Results as of Monday: “How many like Bieber” 16
  - Results as of Tuesday: “How many like Bieber” 17
  - Side information (available to the instructor): You enrolled late on Tuesday.

ML models memorize training datasets, even though they are generalizing well!

Membership Inference Attacks Against Machine Learning Models

Reza Shokri, Cornell Tech
Marco Stronati*, INRIA
Congzheng Song, Cornell
Vitaly Shmatikov, Cornell Tech

***Abstract*—We quantitatively investigate how machine learning models leak information about the individual data records on which they were trained. We focus on the basic membership inference attack: given a data record and black-box access to a model, determine if the record was in the model's training dataset. To perform membership inference against a target model, we make adversarial use of machine learning and train our own inference model to recognize differences in the target model's predictions on the inputs that it trained on versus the inputs that it did not train on.**

We empirically evaluate our inference techniques on classification models trained by commercial "machine learning as a service" providers such as Google and Amazon. Using realistic datasets and classification tasks, including a hospital discharge dataset whose membership is sensitive from the privacy perspective, we show that these models can be vulnerable to membership inference attacks. We then investigate the factors that influence this leakage and evaluate mitigation strategies.

Security and Privacy, 2017

The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets

Nicholas Carlini, *University of California, Berkeley*
Chang Liu, *University of California, Berkeley*
Jernej Kos, *National University of Singapore*
Úlfar Erlingsson, *Google Brain*
Dawn Song, *University of California, Berkeley*

This paper presents *exposure*, a simple-to-compute metric that can be applied to any deep learning model for measuring the memorization of secrets. Using this metric, we show how to extract those secrets efficiently using black-box API access. Further, we show that unintended memorization occurs early, is not due to overfitting, and is a persistent issue across different types of models, hyperparameters, and training strategies. We experiment with both real-world models (e.g., a state-of-the-art translation model) and datasets (e.g., the Enron email dataset, which contains users' credit card numbers) to demonstrate both the utility of measuring exposure and the ability to extract secrets.

Finally, we consider many defenses, finding some ineffective (like regularization), and others to lack guarantees. However, by instantiating our own differentially-private recurrent model, we validate that by appropriately investing in the use of state-of-the-art techniques, the problem can be resolved, with high utility.

USENIX Security 19

How do these attacks work?

- Membership inference attack:
  - Train a ML model to predict whether individuals are used for training.
  - Often obvious from the confidence of the ML-predictions alone.

- Unintended memorization attack:
  - Prompt a language model: Alice's SSN is **????-??**-7452
  - Ask the language model to fill-in the question marks.

Remark: Modern DP learning models memorizes the entire dataset using their billions of parameters. They can be thought of as an implicit transformation of the data into an efficient data-structure. In fact, memorization might be the very reason why deep models work well. See (Feldman, 2019) https://arxiv.org/abs/1906.05271

Conclusions so far: Privacy is challenging!

- Revealing dataset (even if with PII removed) is a bad idea
  - Data-linkage attack, netflix prize
- Revealing aggregate statistics of the dataset have privacy risks
  - Differencing attack: With side information, even if reporting just one, may reveal information about individuals
  - An even stronger attack later: even without side-information, even with noise in the statistics.
- Machine learning models encodes information of individuals in a dataset and will spit them out when given an appropriate prompt
  - Membership inference attack
  - Unintended memorization

A bit of history of privacy protection techniques: various attempts for privacy protection

- *Since 1970s:* Statistical disclosure control (Duncan et al.; Hundepool et al)
  - e.g., Data swapping (Dalenius, Reiss, 1982) was implemented in the Census

- *2002 – 2007:* K-anonymity, l-divergence, t-closeness (Sweeney et. al., Machanavajjhala et. al., Li et. al., 2002 - 2007)
  - These attempts have been shown to be fragile against side-information and composition. See a recent revisit of this problem (and the references therein): https://aloni.net/wp-content/uploads/2021/05/Quasi-IDs-are-the-Problem-working-paper.pdf

- *2006+:* Differential privacy [Dwork, McSherry, Nissim, Smith, 2006++]

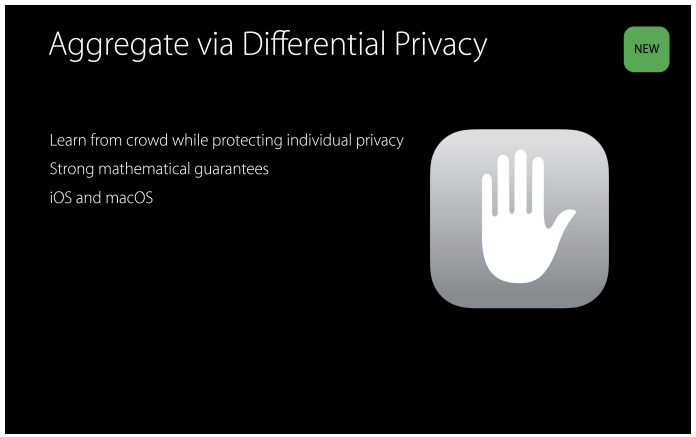
This course is about differential privacy

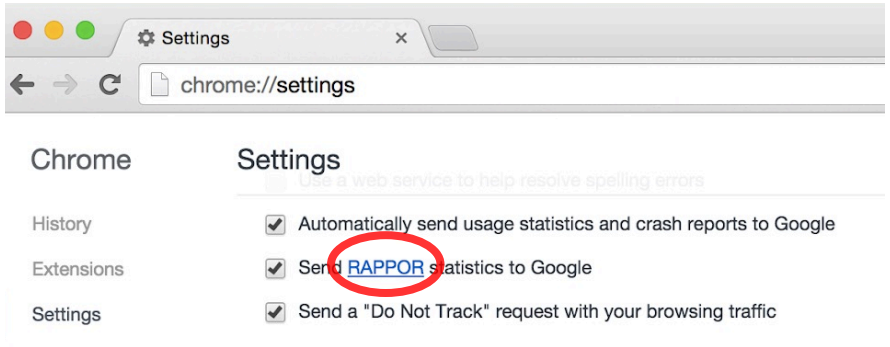
- A formal mathematical definition of privacy that provides rigorous guarantees and provably effective protections against privacy risks.
- Makes no assumption on the adversary
  - Arbitrary side info, arbitrary computational power.
- Interpretable, quantifiable, composable formalism
- The de-facto standard in privacy --- the only one still being actively researched on.

Meet the 2017 Gödel Prize winners: Dwork, McSherry, Nissim & Smith

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006, March). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference* (pp. 265-284). Springer, Berlin, Heidelberg.

Differential privacy is transitioning from a theoretical construct into a practical technology!





Topics that I will cover in this class

1. Privacy risks, privacy attacks and motivations of differential privacy.
2. The definition of differential privacy and its various interpretations.
3. Understanding the promise of DP: what does it protects and what not?
4. Fundamental building blocks of DP algorithms.
5. Modern methods in DP: Renyi DP, Privacy profiles, tradeoff functions
6. How to use autodp for privacy accounting and privacy calibration.
7. Differentially private (linear) machine learning
8. Differentially private deep learning under various models
9. Data-adaptive differentially private algorithms.

By the end of the course, you will be able to

- determine when DP is applicable and whether it will allow for sufficient utility for an application.
- design and implement DP algorithms correctly (prove DP guarantees).
- gain familiarity / proficiency in the various elements of theoretical research in CS / ML.

Prerequisites

- The course might be tough for you if you do not have working knowledge in:
  - Probability theory and statistics
  - Linear algebra, basic calculus
  - Basic data structures and algorithms
  - writing mathematical proofs
  - writing simple code (in Python / Numpy)

- Advanced knowledge that will help (but not required):
  - Linear / Convex optimization
  - Concentration inequalities
  - Statistical machine learning
  - Experience doing theoretical research
  - Experience working with a real dataset

Evaluation

- Lecture attendance: 5%
- Scribing: 10%
- Homework: 45%.
- Project: 40% (Up to 3 students per team)
  - A typical project: Differential Privacy + X
  - Example:
    - Reading an existing paper on a particular aspect of DP, reproduce the proof.
    - Empirical evaluation of differentially private methods:
      - k-means clustering, boosting
      - reinforcement learning, active learning
      - Statistical inference
    - Applying methods we learned to a real dataset
      - Privately training a model for detection of Flu from wearable device data
      - Differentially private release of COVID time series data
      - Conducting research studies with electronic patient record data

Project milestones

1. Proposal: 5%
2. Midterm report: 10%
3. Final report: 10%
4. Project presentation: 15%

I will share a list of project ideas on Piazza soon.
Also, special projects available (e.g., applying DP to a high-impact application).
Please reach out to me for about these opportunities.

Tentative schedule of the course

Part I: Differential Privacy Basics

Part II: Machine learning with DP

Part III: Beyond the worst case

Part IV: Advanced topics / Project Consultation

	Date	Lectures	Readings	Assignments
1	27-Sep	Course overview and Privacy challenges [Slides]	DR Ch 1-2, Vadhan 5.1, DR Ch 8.1	
2	29-Sep	DP Basics I: Definition, interpretation + Laplace Mechanism	DR Ch 2, Ch 3.1, Ch 3.2, Ch 3.3 (up to Page 34)	HW1 out
3	4-Oct	DP Basics II: Sparse Vector + Private Query Release	DR Ch 3.6, DR Ch 4.2	
4	6-Oct	DP Basics III: Report-Noisy-Max and Exponential Mechanism	DR Ch 3.3 (Page 34+), DR Ch 3.4	
5	11-Oct	DP Basics IV: Privacy loss RV, Advanced Composition	DR Ch 3.5	
6	13-Oct	DP Basics V: Gaussian mechanism, revisited	Bun and Steinke (2018), Balle and W. (2018), Dong et al. (2019)	
7	18-Oct	DP Basics VI: Modern tools for Optimal Privacy Accounting	W., Balle, Kasiviswanathan (2018), Zhu, Dong and W. (2021)	Project proposal due
8	20-Oct	DPML I: Introduction, Posterior sampling	Minami (2017)	HW1 due / HW2 out
9	25-Oct	DPML II: Objective Perturbation	Chaudhuri et al. , Kifer et al.	
10	27-Oct	DPML III: NoisyGD and NoisySGD	TBA	
11	1-Nov	DPML IV: PATE and PrivateKNN	TBA	Midterm report due
12	3-Nov	DPML V: Differentially Private Federated Learning	TBA	
13	8-Nov	Adaptive DP I: Smoothed Sensitivity and Median	Vadhan Ch 3.1., Nissim, Raskhodnikova, Smith (2011)	
14	10-Nov	Adaptive DP II: Propose-Test-Release	Vadhan Ch 3.2	
15	15-Nov	Adaptive DP III: pDP to DP conversion	TBA	HW2 due /HW3 out
16	17-Nov	Adaptive DP IV: Data-Adaptive Differentially private ML	TBA	
17	22-Nov	Project consultation I or Guest lecture		
18	24-Nov	Project consultation II or Guest lecture		
19	29-Nov	Project consultation III or Guest lecture		
20	1-Dec	Mini-Symposium on Practical Differential Privacy		Final project report due / HW3 due

Remainder of today's lecture

- A concrete data-reconstruction attack
- Randomized response

A simple mathematical model to consider

- We have a dataset of n individuals
- One secret bit of information per person
- (Normalized) linear query / statistical query:

We say an algorithm is blatantly non-private if one can reconstruct 90% of the dataset (secret bit vector) using its output.

Definition 5.1 (blatant non-privacy, due to Dinur and Nissim [30]). A mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ is called *blatantly non-private* if for every $x \in \mathcal{X}^n$, one can use $\mathcal{M}(x)$ to compute an $x' \in \mathcal{X}^n$, such that x' and x differ in at most $n/10$ coordinates (with high probability over the randomness of $\mathcal{M}$).

- In particular we will consider algorithms that answer a collection of linear queries about the dataset approximately.

Reconstruction Attack --- find a dataset that is consistent with the observations!

- We have k linear queries
- An algorithm returns answers that are α-accurate
- The reconstruction attack:

Any algorithm that answers all 2^n linear queries with constant error implies blatant non-privacy!

Theorem 5.2 (reconstruction from many queries with large error [30]). *Let $x \in \{0,1\}^n$. If we are given, for each $q \in \{0,1\}^n$, a value $y_q \in \mathbb{R}$ such that*

$$\left| y_q - \frac{\langle q, x \rangle}{n} \right| \leq \alpha.$$

Then one can use the y_q's to compute $x' \in \{0,1\}^n$ such that x and x' differ in at most 4α fraction of coordinates.

- Proof:

Any algorithm that answers O(n) linear queries with $O\left(\frac{1}{\sqrt{n}}\right)$ error implies blatant non-privacy.

Theorem 5.6 (reconstruction from few queries with small error [30]). *There exists $c > 0$ and $q_1, \ldots, q_n \in \{0,1\}^n$ such that any mechanism that answers the normalized inner-product queries specified by $q_1, \ldots, q_n$ to within error at most $c/\sqrt{n}$ is blatantly non-private.*

- Recall the attack is

$$x' = \arg \min_{\tilde{x} \in \{0,1\}^n} \max_{i \in [n]} \left| y_i - \frac{1}{n} q_i^T \tilde{x} \right|$$

- Observations

Proof idea / sketch

Proof idea / sketch (continuing)

Quick checkpoint

- "Any algorithms that answers too many questions too accurately will result in a blatant reconstruction of the dataset."
  - No side information needed.
  - No restriction on the dataset.
  - Fixed "design", highly generic design (iid samples)

- The attacks are not computationally efficient, but …
  - efficient attacks exist, via a Linear Programming relaxation

$$x' = \arg \min_{\tilde{x} \in [0,1]^n} \max_{i \in [k]} \left| y_i - \frac{1}{n} q_i^T \tilde{x} \right|$$

Read the original paper: Dinur-Nissim (2003) for more details.

What can we still do?

Target accuracy	k = O(2^n) linear queries	k = O(n) linear queries	k << n linear queries
α = O(1) (any non-trivial error)	Blatantly non-private	?	?
α = O(1/sqrt(n)) (statistical error)	Blatantly non-private	Blatantly non-private	?
α = o(1/sqrt(n)) (<< statistical error)	Blatantly non-private	Blatantly non-private	?

Two things to think about:

- Avoiding “Blatant-non-private” is a relatively weak privacy guarantee.
- Can we achieve the lower bound while satisfying a much stronger notion of privacy guarantee?

Randomized Response (Warner, 1965)

- Who likes Justin Bieber?
  - Space of the answer: {0,1}

1. Each individual tosses an independent coin with probability $p > 0.5$
2. If "head", keep your answer.
3. Otherwise, flip your answer.

- Intuitively each individual has a degree of plausible deniability.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*(309), 63-69.

Observations

- 1. We returns a synthetic dataset with an expected L1-error n(1-p)
  - Not blatantly non-private if p << 0.9?
  - Is it possible to post-process it to reduce the error?

- 2. We can unbiasedly estimate any fixed set of (normalized) linear queries with error: O(log(|Q|)/(2p-1)sqrt(n))
  - When |Q| = 2^n, the bound is trivial
  - When |Q| = O(n), the bound matches what we expect up to log n.
  - When p = o(1) and p >> 1/(sqrt(n) log |Q|) ➔ strong privacy + non-trivial error

Next Lecture

- Differential Privacy
- Properties of Differential Privacy
- Basic mechanisms of DP