Privacy Enhancing Technologies

1 Data Anonymization

Def. K-Anonymity For each individual, there are at least $k-1$ with the same value for all ≪quasi-identifiers≫ (zip, age, ...).

Attacks

- *Homogeneity*: all individuals of anonymity group have same sensitive attribute
- *Background knowledge*: combine knowledge of multiple quasi-identifiers
- *Composition*: combine multiple datasets
- *Downcoding*: learn from used algorithm

Aggregate Statistics Aggregate information across different subsets. Problem: very much redundancy that can be used to recover all information!

Counting queries for some predicate $\phi : \mathcal{Z} \to \{0,1\}$, where $z_i \in \mathcal{Z}$ are identifiers and $s_i \in \{0,1\}$ is the private bit.

$$f_\phi(D) = \frac{1}{n}\sum_{i=1}^{n} \phi(z_i)\cdot s_i = \frac{\phi(z)\cdot s}{n}$$

The curator answers with $\texttt{Ans}(\phi) = f_\phi(D) \pm e$ with $0 \le e \le E$.

Def. Blatant non-privacy Adversary can, w/ high prob. $1 - o(1)$ construct $\tilde{s} \in \{0,1\}^n$ s.t. s and $\tilde{s}$ differ in $o(n)$ coords.

Thm. 1 Dinur/Nissim If adv can ask 2^n subset queries on DB of size n, and noise per query $\le E$, then adv can output $\tilde{s}$ s.t. s and $\tilde{s}$ differ in at most $4En$ positions.

Thm. 2 Dinur/Nissim If adv can ask n count. queries on DB of size n, and noise/query bounded by $E = o(1/\sqrt{n})$, then curator's algo is blatantly non-private.

2 Differential Privacy

Each individual has data $x \in \mathcal{X}$, which form n rows of the database $D \in \mathcal{X}^n$. Adversary can make queries, for which the curator runs algo M over the DB D to produce response $y = M(D)$.

Def. Neighboring DB Two databases are neighboring $D \sim D'$, if $|D| = |D'|$ and they differ in at most 1 row.

Def. Differential Privacy Randomized algo $M : \mathcal{X}^n \to \mathcal{Y}$ is ϵ-differentially private if, for all neighboring DBs $D \sim D'$ and for all events $S \subseteq \mathcal{Y}$:

$$\Pr[M(D) \in S] \le e^\epsilon \cdot \Pr[M(D') \in S]$$

ϵ is the privacy loss, we want as small as possible, e.g. $e^\epsilon = 2$. If $\mathcal{Y}$ is discrete, the above definition is equivalent to $\Pr[M(D) = y] \le e^\epsilon \cdot \Pr[M(D') = y]$ for all $y \in \mathcal{Y}$. If $\mathcal{Y}$ continuous, use the PDF $f_D(y)$ instead of Pr[...].

Thm. Post-Processing If M is ϵ-differentially private, then any function of the output of M (i.e. $f \circ M$) is also ϵ-DP.

Thm. Basic composition Let $M = (M_1, ..., M_k)$ sequence of algos where M_i is ϵ_i-DP, where M_i can depend on $M_1, ..., M_{i-1}$. Then M is $(\sum_{i=1}^k \epsilon_i)$-DP.

Proof:

$$\frac{\Pr[M(D)=y]}{\Pr[M(D')=y]} = \prod_{i=1}^{k} \frac{\Pr[M_i(D)=y_i | M_1(D)=y_1, ...]}{\Pr[M_i(D')=y_i | M_1(D')=y_1, ...]}$$
$$\le \prod_{i=1}^{k} e^{\epsilon_i} = \exp\left(\sum_{i=1}^{k}\epsilon_i\right)$$

Thm. Group privacy For $M : \mathcal{X}^n \to \mathcal{Y}$, ϵ-DP and DBs D and D' that differ in up to $k$rows:

$$\Pr[M(D) \in S] \le e^{k\epsilon} \cdot \Pr[M(D') \in S]$$

Proof: hybrid argument, definition applied k times.

Randomized response $M(x_1, ..., x_n) = (y_1, ..., y_n)$ is DP: for some set of randomized responses $\vec{b} \in \{0,1\}^n$ and neighboring D, D' differing only in x_j:

$$\frac{\Pr[M(D)=\vec{b}]}{\Pr[M(D')=\vec{b}]} = \frac{\prod_{i=1}^n \Pr[y_i = b_i]}{\prod_{i=1}^n \Pr[y'_i = b_i]} = \frac{\Pr[y_j = b_j]}{\Pr[y'_j = b_j]} \le \frac{1/2+\gamma}{1/2-\gamma}$$

RR is ϵ-DP for $\epsilon = \ln(\frac{1/2+\gamma}{1/2-\gamma})$. If $\gamma = 1/4$, $\epsilon = \ln 3 \approx 1.1$.

Def. l_1-Sensitivity of a function $f : \mathcal{X}^n \to \mathbb{R}^k$ is defined as the max difference in outputs of f in neighboring databases:

$$\Delta_1 = \max_{D\sim D'} \left\|f(D) - f(D')\right\|_1 = \max_{D\sim D'} \sum_{i=1}^{k} |f(D)_i - f(D')_i|$$

To achieve DP, we want to add noise to f from a prob. dist. whose density changes by at most e^ϵ when the input changes by Δ_1.

Def. Laplace Distribution $\texttt{Laplace}(\mu, b)$ with location μ and scale $b > 0$ has mean μ and variance $2b^2$. Density function:

$$f(x|\mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}, \quad x \in \mathbb{R}$$

Def. Laplace Mechanism Let $f : \mathcal{X}^n \to \mathbb{R}^k$ with l_1-sensitivity Δ_1 and $\epsilon > 0$. Then the following mechanism is ϵ-DP:

$$M(D) = f(D) + (\mathcal{Y}_1, ..., \mathcal{Y}_k), \quad \mathcal{Y}_i \text{ iid. Laplace}(0, \Delta_1/\epsilon)$$

To achieve a privacy level ϵ, add Laplace noise with scale parameter $b = \Delta_1/\epsilon$, which is $O(\frac{1}{\epsilon n})$ for the mean, which has sensitivity $1/n$.

Multiple queries Swapping an individual might affect the result of all queries, which results in sensitivity k/n for the mean. Hence, add noise $\mathcal{Y}_i \sim \texttt{Laplace}(0, \frac{k}{\epsilon n})$.

Lower bound Any ϵ-DP algorithm that answers k counting queries must incur error of magnitude $\Omega\left(\frac{k}{\epsilon n}\right)$.

3 Approximate Differential Privacy

Def. Approximage DP Algo $M : \mathcal{X}^n \to \mathcal{Y}$ is (ϵ, δ)-DP if, for all neighboring DBs $D \sim D'$ and for all events $S \subseteq \mathcal{Y}$:

$$\Pr[M(D) \in S] \le e^\epsilon \cdot \Pr[M(D') \in S] + \delta$$

$\delta \in o(1/n)$ to get any meaningful privacy guarantee.

Thm. Basic composition Let $M = (M_1, ..., M_k)$ sequence of algos where M_i is (ϵ_i, δ_i)-DP, where M_i can depend on $M_1, ..., M_{i-1}$. Then M is $(\sum_{i=1}^k \epsilon_i, \sum_{i=1}^k \delta_i)$-DP.

Def. Privacy loss random variable Let D, D' be two DBs. The privacy loss random variable $\mathcal{L}_{M(D)||M(D')}$ is distributed by drawing $y \leftarrow M(D)$, and outputting

$$\mathcal{L}_{M(D)||M(D')}(y) \coloneqq \ln\left(\frac{\Pr[M(D)=y]}{\Pr[M(D')=y]}\right)$$

If M is ϵ-DP, then $\mathcal{L}$ is bounded by ϵ for neighboring DBs.

For approx. DP, we have that M is (ϵ, δ)-DP if for all neighboring databases $\Pr[|\mathcal{L}_{M(D)||M(D')}| > \epsilon] \le \delta$

Def. l_2-Sensitivity of $f : \mathcal{X}^n \to \mathbb{R}^k$:

$$\Delta_2 = \max_{D\sim D'} \left\|f(D) - f(D')\right\|_2 = \max_{D\sim D'} \sqrt{\sum_{i=1}^{k}(f(D)_i - f(D')_i)^2}$$

We have that $\Delta_2 \le \Delta_1 \le \sqrt{k}\cdot\Delta_2$

Def. Gausian Distribution with mean μ and variance σ^2:

$$f(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \quad x \in \mathbb{R}$$

Def. Gaussian Mechanism Let $f : \mathcal{X}^n \to \mathbb{R}^k$ with l_2-sensitivity Δ_2, $\epsilon, \delta > 0$. Gaussian mechanism:

$$M(D) = f(D) + (\mathcal{Y}_1, ..., \mathcal{Y}_k)$$

where $\mathcal{Y}_i$ iid. from $\mathcal{N}(0, \frac{2\log(1.25/\delta)\Delta_2^2}{\epsilon^2})$

Thm. For $\epsilon \le 1, \delta > 0$ the Gauss. mech. is (ϵ, δ)-DP

Gaussian tail bound If $Z \sim \mathcal{N}(0, \sigma^2)$, then for every $t > 0$:

$$\Pr[|Z| > t\cdot\sigma] \le \exp(-t^2/2), \quad \Pr[Z \ge t] \le \exp(-t^2/2\sigma^2)$$

Thm. Advanced composition For $\epsilon > 0$, $\delta \in [0,1]$, $\delta' \in (0,1]$, let $M = (M_1, ..., M_k)$ sequence of (ϵ, δ)-DP algos. M is $(\tilde{\epsilon}, \tilde{\delta})$-DP, where $\tilde{\epsilon} = \epsilon\sqrt{2k\log(1/\delta')} + k\epsilon\frac{e^\epsilon - 1}{e^\epsilon + 1}$, $\tilde{\delta} = k\delta + \delta'$.

If we have $\epsilon < 1/\sqrt{k}$ and we set $\delta' = \delta$,
we get $\tilde{\epsilon} = \Theta(\epsilon\sqrt{k\log(1/\delta)})$ and $\tilde{\delta} = \Theta(k\delta)$

Guarantee	Error per query	Mechanism	Lower bound
Local ε-DP	$\tilde{O}\left(\frac{1}{\varepsilon\sqrt{n}}\right)$	Randomized Response	$\Omega\left(\frac{1}{\varepsilon\sqrt{n}}\right)$
ε-DP	$\tilde{O}\left(\frac{k}{\varepsilon n}\right)$	Laplace	$\Omega\left(\frac{k}{\varepsilon n}\right)$
(ε,δ)-DP	$\tilde{O}\left(\frac{\sqrt{k}}{\varepsilon n}\right)$	Gaussian	$\Omega\left(\frac{\sqrt{k}}{\varepsilon n}\right)$
	$\tilde{O}\left(\frac{\sqrt{\log\lvert\mathcal{X}\rvert}\log k}{\varepsilon n}\right)^{1/2}$	Private Multiplicative weights	$\tilde{\Omega}\left(\frac{\sqrt{\log\lvert\mathcal{X}\rvert}\log k}{\varepsilon n}\right)^{1/2}$

Def. Selection problem is specified by:

- Set $\mathcal{Y}$ of possible outcomes
- Score function $q : \mathcal{Y} \times \mathcal{X}^n \to \mathbb{R}$, measures ≪quality≫ of output- for given input data set $D \in \mathcal{X}^n$
- Sensitivity bound Δ s.t. $q(y;\cdot)$ has sensitivity $\leq \Delta$. I.e. for each $D \sim D'$, we have $|q(y;D) - q(y;D')| \leq \Delta$

Def. Exponential Mechanism Let $(\mathcal{X}, \mathcal{Y}, q, \Delta)$ selection problem. Return sample from following distribution over $\mathcal{Y}$

$$\Pr[\mathcal{Y} = y] \propto \exp\left(\frac{\epsilon}{2\Delta} q(y;D)\right)$$

Thm. The Exponential Mechanism is ϵ-DP.

Noise on the order of $\tilde{O}(\frac{\ln k}{\epsilon})$ is needed.

4 Differentially Private Learning

Def. Risk $\mathcal{L}(\theta)$ of a model is expected loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[l(\theta, x, y)]$$

Def. Empirical risk $\hat{\mathcal{L}}_D(\theta)$ of a model on dataset D:

$$\hat{\mathcal{L}}_D(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta, x_i, y_i)$$

Learning a model: find parameters $\theta \in \Theta$ that minimize $\hat{\mathcal{L}}_D(\theta)$.

Generalization gap, where D training set: $|\mathcal{L}(\theta^*) - \hat{\mathcal{L}}_D(\theta^*)|$

Private learning approaches

- **Output perturbation**: add noise to loss function
- **Objective perturbation**: add $b^\top\theta$ to optimization objective (i.e. objective fn), where b is sampled from Gaussian
- **Gradient perturbation**: add noise to gradient

Gradient clipping In each step of gradient descent, clip each example's gradient to some max. l_2 norm C. This gives us a bound on l_2 sensitivity.

Privacy loss after T steps of (ϵ, δ)-DP guarantee per step:

- Basic composition: $(\epsilon T, \delta T)$-DP
- Advanced composition: $(\epsilon\sqrt{T\log(1/\delta)}, \delta T)$-DP
- Amplification by subsampling: Let $\rho = m/n$ be fraction of selected samples in SGD. Then M' is (ϵ', δ')-DP for $\epsilon' \approx \rho\epsilon$, $\delta' \approx \rho\delta$ $\implies$ $O(\epsilon\rho\sqrt{T\log(1/\delta)}, \delta\rho T)$. ρ typ. small

5 Membership Inference

Membership Inference Game

Adversary		Challenger
	$D, x \longrightarrow$	$D_0 = D, D_1 = D \cup \{x\}$
	$\longleftarrow M(D_b)$	$b \leftarrow^{\$} \{0,1\}$
output $b' \in \{0,1\}$		

Adversary wins if $b' = b$

$\text{TPR} = \Pr[b' = 1 | b = 1]$ $\quad$ $\text{TNR} = \Pr[b' = 0 | b = 0]$
$\text{FPR} = \Pr[b' = 1 | b = 0]$ $\quad$ $\text{FNR} = \Pr[b' = 0 | b = 1]$

Thm. M is (ϵ, δ)-DP iff. for all adversaries in the MI game:

$e^\epsilon \cdot \text{FNR} \geq \text{TNR} - \delta$ and $e^\epsilon \cdot \text{FPR} \geq \text{TPR} - \delta$

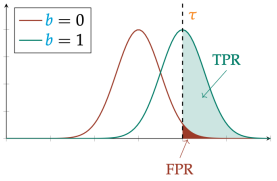
Cor. M is (ϵ, δ)-DP iff. for all adversaries in the MI game:

$$e^\epsilon \geq \max\left(\frac{\text{TNR} - \delta}{\text{FNR}}, \frac{\text{TPR} - \delta}{\text{FPR}}\right)$$

Thm. This Gaussian mechanism is **not** $(O(1), \delta)$-DP:

$$M(D) = \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i + \mathcal{N}(\vec{0}, \sigma^2 \cdot I_{k\times k}), \quad \sigma = o(\sqrt{k}/n)$$

To determine if $x = \vec{1}$ was used in a all-zero DB we can fix a threshold $\tau = k/n$. TPR = 50% by symmetry and FPR = $o(1)$ since gaussian tail bound and $\sigma = o(\sqrt{k}/n)$. Hence, we need $\sigma = \Omega(\sqrt{k}/n)$, i.e. this error per query for k counting queries.



MI Attack against Neural Net We can only reliably detect outliers (data that is not similar to the rest), both for easy and hard to fit examples. Neural nets generalize very well to inliers.

Likelihood ratio test Null hypothesis H_0 ($x \notin D$) and alt. hypo. H_1. $\texttt{View} = M(D_b)$ the output of M. Reject, if $\Lambda(\texttt{View}) \leq \tau$.

$$\Lambda(\texttt{View}) = \frac{\Pr[\texttt{View}|H_0]}{\Pr[\texttt{View}|H_1]}$$

To do this, train ≪IN≫ and ≪OUT≫ models and fit gaussians to the respective loss. Difference of gaussians is again gaussian.

6 Utils & Exercises

Diameter of range Largest possible value of $\|f(D_1) - f(D_2)\|_1$ for any two datasets D_1 and D_2.

Union bound For events $\{A_i\}_{i=1}^{\infty}$: $\Pr[\cup_{i=1}^{\infty} A_i] \leq \sum_{i=1}^{\infty} \Pr[A_i]$

Markov $X \geq 0, t > 0 \implies \Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t^2}$

Show that if the additive privacy parameter $\delta \geq \frac{1}{2n}$, a $(0,\delta)$-DP mechanism is not actually private and can leak individual data. **Idea:** Construct a mechanism that outputs a specific raw database entry x_i with probability exactly equal to the allowable failure probability δ (specifically sampling i uniformly and outputting x_i with probability 1/2, yielding overall probability $\frac{1}{2n} \leq \delta$).

Show that if $\delta < \frac{1}{2n}$, a $(0,\delta)$-DP mechanism cannot be useful (specifically, it incurs constant error $\Omega(1)$ when computing the mean). **Idea:** Use a hybrid argument (chaining n adjacent databases from "all zeros" to "all ones") to show that the total statistical distance between the outputs of the two extreme databases is bounded by $n\delta < 1/2$. Since the distributions are indistinguishable but the true means are far apart (0 vs 1), the algorithm must fail on at least one.

Show that for the Randomized Response mechanism (flipping bits with probability 1/4), there exists a predicate query where the adversary incurs large error $\Omega(1)$. **Idea:** Construct a biased query that selects only the subset of indices where the true value is 0. This prevents noise cancellation and forces the noise (bit flips) to accumulate in a single direction, creating a large observable bias.

Show that for the Randomized Response mechanism (flipping bits with probability 1/4), there exists a predicate query where the adversary incurs large error $\Omega(1)$. **Idea:** Construct a biased query that selects only the subset of indices where the true value is 0. This prevents noise cancellation and forces the noise (bit flips) to accumulate in a single direction, creating a large observable bias.

Is the result in the previous question consistent with the Dinur-Nissim reconstruction theorem (which allows reconstruction if error is small)? **Idea:** Yes, because the Dinur-Nissim reconstruction bound depends on the magnitude of the error. Since the error here is large (linear in n), the reconstruction bound becomes trivial (greater than n), rendering the theorem consistent but vacuous.

Construct a mechanism to estimate the mean of a property over a dataset D of size n that: 1) prevents full database reconstruction (is not blatantly non-private), 2) maintains high utility with estimation error $\tilde{O}(1/\sqrt{n})$, but 3) fails to satisfy (ϵ,δ)-differential privacy for any $\delta < 1/2$. **Idea:** The mechanism simply outputs the mean of a random subsample of size $n/2$. This prevents reconstruction of the unsampled half. However, it fails formal DP because of the "impossible event" problem: if D' contains a single unique element (e.g., a 1 among 0s) that D does not, there is a 1/2 probability that the subsample includes this element, producing an output value strictly impossible under D. This infinite likelihood ratio requires $\delta \geq 1/2$.

Show that the "Report Noisy Max" mechanism, which adds independent noise $Z_i \sim Exp(\frac{\epsilon}{2\Delta})$ to scores $q(y,D)$ and returns the index of the maximum, satisfies ϵ-differential privacy. **Idea:** Instead of analyzing the complex joint distribution of the maximum directly, the proof conditions on fixing $d-1$ noise variables. It then leverages the specific tail property of the Exponential distribution, $\Pr[X > a - b] = \exp(\lambda b)\Pr[X > a]$, to bound the probability ratio between neighboring databases. **What is the main practical advantage of using Report Noisy Max over the standard Exponential Mechanism? Idea:** Computational efficiency. Unlike the Exponential Mechanism, Report Noisy Max does not require computing the full distribution or its normalizing constant (partition function). It only requires sampling indep. vars and finding a max.