DATA PRIVACY AND SECURITY

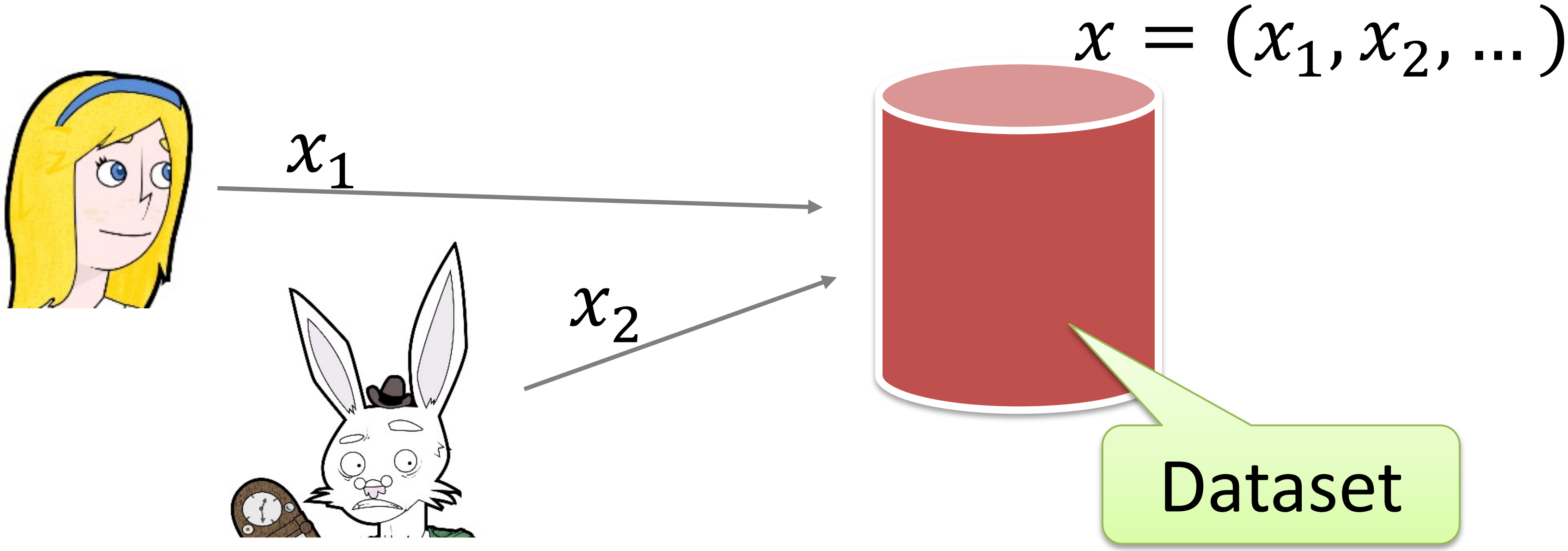
Prof. Daniele Venturi

Master's Degree in Data Science
Sapienza University of Rome

CIS SAPIENZA
RESEARCH CENTER FOR CYBER INTELLIGENCE AND INFORMATION SECURITY

CHAPTER 5: Differential Privacy

Data Exploitation

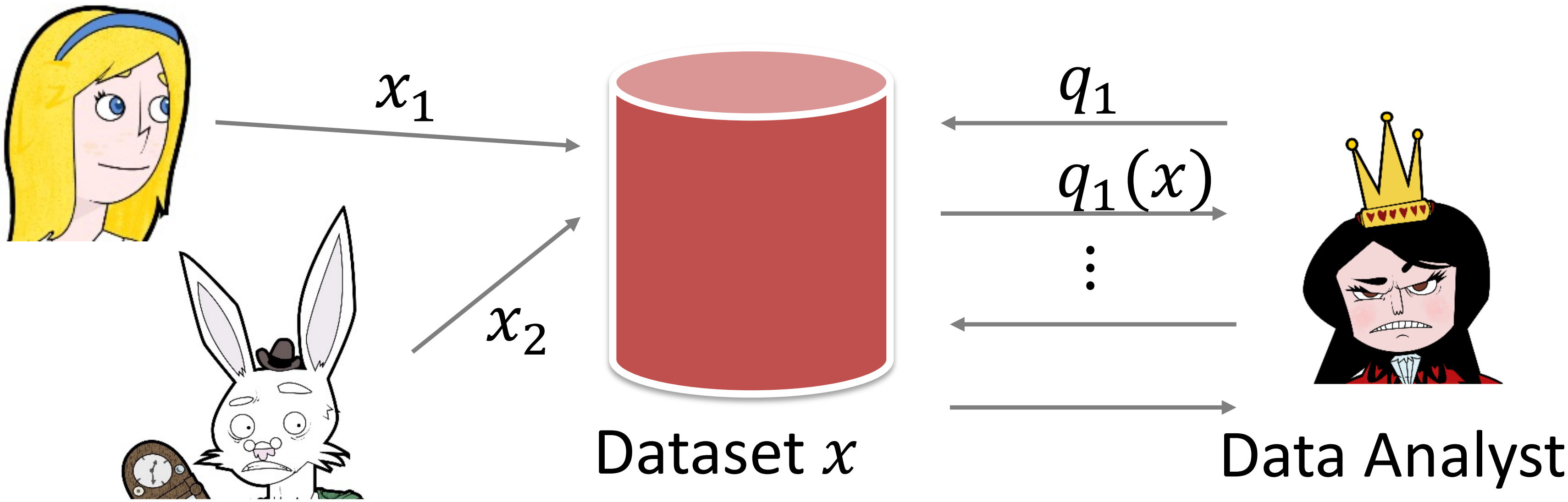


- Availability of **lots of data**
  - Social networks, financial data, medical records…
- All these data are an asset
  - We would like to **exploit them**

Applications

- Finding **statistical correlations**
  - Genotype/phenotype association
  - Correlating medical outcomes with risk factors
- Publishing **aggregate** statistics
- Noticing **events/outliers**
  - Intrusion detection
- Datamining/**learning**
  - Update **strategies** based on customers data

Data Analysis and Privacy



- How to define **privacy**?
  - Intuitively we want that published statistics do not undermine privacy of the individuals
  - After all statistics are just **aggregated data** about the overall population

The Statistics Masquerade

- Differential attack
  - How many people in the room XYZ last night?
  - How many people, **other than the speaker**, XYZ last night?
- Needle in a haystack
  - Determine presence of an individual genomic data in GWAS case group based on **aggregate stats**
- The big bang attack
  - Reconstruct sensitive attributes given statistics from **multiple overlapping datasets**

NYC Taxicab Data

- 2014: NYC Taxi & Limo Commission sharing visualization on **taxi usage statistics** on twitter
  – Chris Whong filed a FOIL request and released the dataset **publicly online**
  – 19 GB with all taxi fares and statistics in 2013
- Attempt to **anonymize** the data

  6B111958A39B24140C973B262EA9FEA5,
  D3B035A03C8A34DA17488129DA581EE7, ...

  – Someone discovered those were the MD5 hash of the driver's **medallion** and **license number**

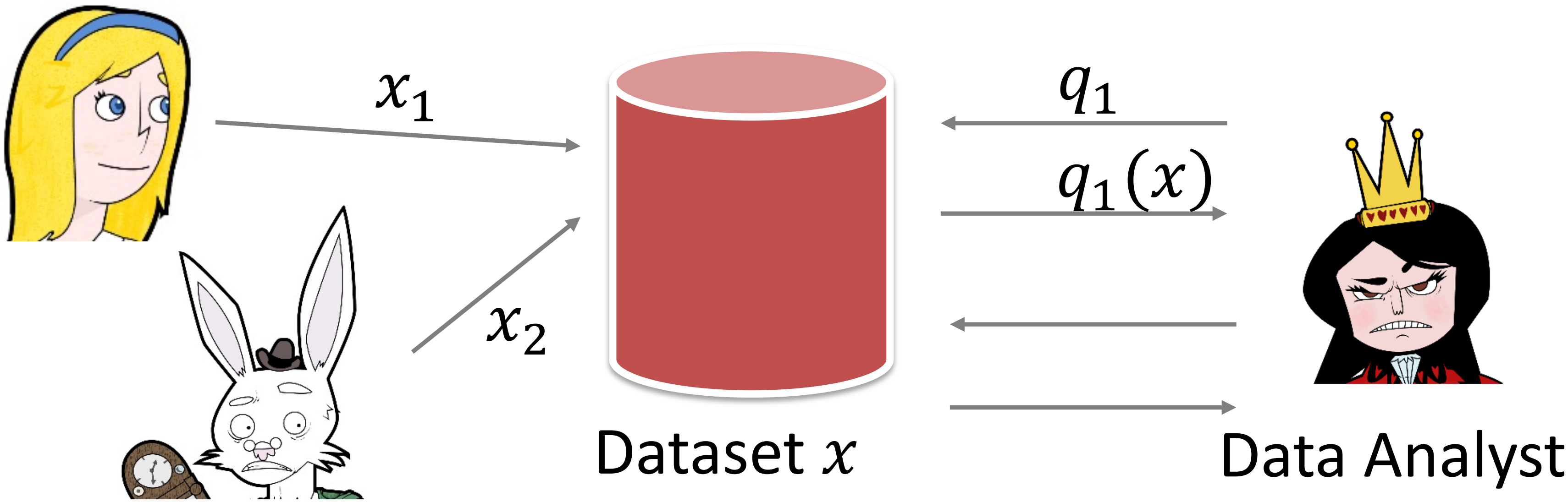
The Netflix Prize

- 2006-2009: 1M USD for improving the **recommendation engine**
- **Anonymized** dataset including movie id, user id, rating and date
- The dataset was **de-anonymized** by combining it with the **public** IMDB dataset
  – Matching users that gave **similar** preferences
  – A class action **lawsuite** was filed against Netflix

Lessons to be Learned

- Privacy is a **concern** when publishing datasets
- Wait: This does not apply to me!
  - Don't make the **entire** dataset available
  - Only publish **aggregate** statistics
- Even if only **data aggregations** are published privacy **can be broken**
- Overly accurate estimates of too many statistics is **blatantly non-private**

Privacy-Preserving Data Analysis?

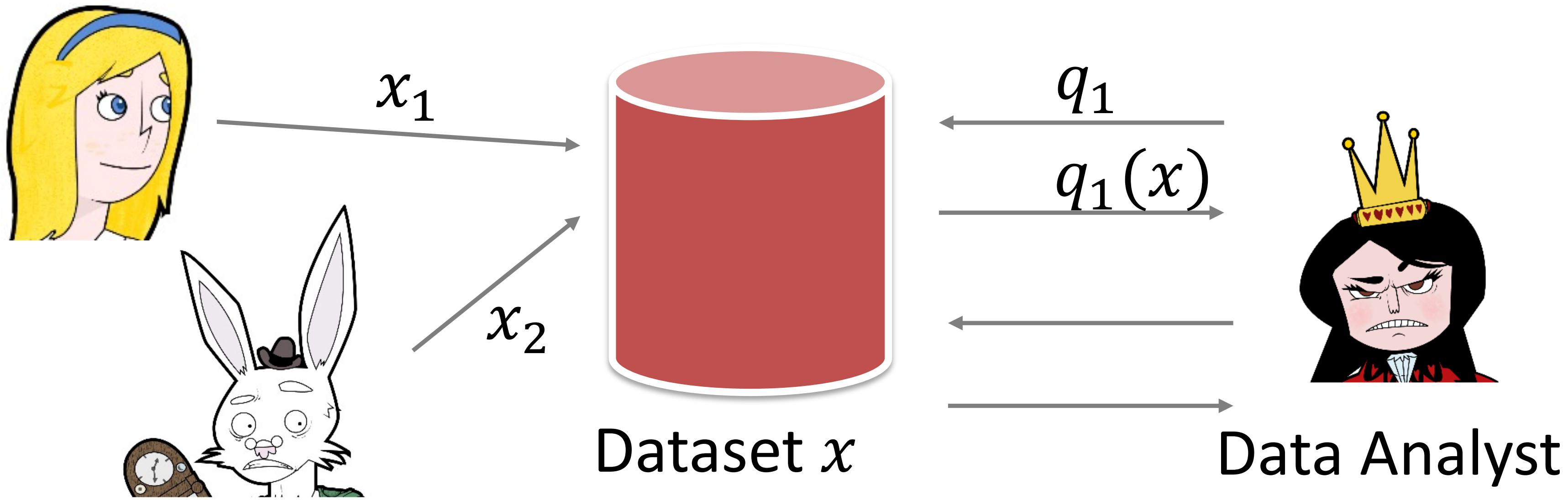


- Can't learn **anything new** about Alice?
  - Reminiscent of semantic security for encryption
- Ideally: Learn same thing if Alice is **replaced** by a **random member** of the population

Differential Privacy

- Outcome of analysis is roughly **equally likely**
  - **Independent** of whether any individual joins, or refrains from joining, the dataset
  - Alice goes away, Bob joins, Alice replaced by Bob
  - Small perturbations **do not matter**
- Note that instead if we **completely change** the dataset we get **completely different** answers!
- Adopted in **real-world** applications by Apple, Google and Microsoft

More Formally…



Definition: Mechanism $\mathbf{M}: \mathcal{X}^n \times \mathcal{Q} \to \mathcal{Y}$ gives ε-**differential privacy** if for all pairs of **adjacent** datasets $x, x' \in \mathcal{X}^n$, and for all queries $q \in \mathcal{Q}$:

$$\forall y \in \mathcal{Y}, \mathbb{P}[\mathbf{M}(x,q) = y] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathbf{M}(x',q) = y]$$

Notes on the Definition

- All that an adversary learns about **you**, it could have learned from the **rest of the dataset**
  - Even if you don't participate
  - This **doesn't mean** nothing about you is leaked
  - Can't use DP to take actions on **specific** individuals
- **Worst-case** guarantee
  - For all datasets, against **unbounded** adversaries
- Probability over the **randomness** of the algorithm, not over the choice of the dataset

Counting Queries

- Simply a **predicate on rows** $q\colon \mathcal{X} \to \{0,1\}$
  - Can be extended to datasets $\mathcal{X}^n$ by counting the fraction of people satisfying the predicate, i.e.

$$\boxed{q(x) = \frac{1}{n}\sum_{i=1}^{n} q(x_i)}$$

- **Point functions:** $\mathcal{Q}_{\mathrm{pt}}(\mathcal{X}) = \{q_y\}_{y\in\mathcal{X}}$ s.t. $q_y(w) = 1$ iff $w = y$
  - Answering all queries in $\mathcal{Q}_{\mathrm{pt}}(\mathcal{X})$ amounts to computing the **histogram** of the dataset

Counting Queries

- **Threshold functions:** $\mathcal{Q}_{\text{thr}}(\mathcal{X}) = \{q_y\}_{y \in \mathcal{X}}$ s.t. $q_y(w) = 1$ iff $w \leq y$ (with $\mathcal{X}$ totally ordered)
  - Answering all queries in $\mathcal{Q}_{\text{thr}}(\mathcal{X})$ amounts to the **cumulative distribution function** of the dataset
- **Attribute means:** $\mathcal{Q}_{\text{means}}(d) = \{q_j\}_{j \in [d]}$ s.t. $q_j(w) = w_j$, where $w \in \mathcal{X} = \{0,1\}^d$
  - Answering all queries in $\mathcal{Q}_{\text{means}}(d)$ amounts to computing the **fraction** of the dataset possessing each of the d attributes (1-way marginal statistics)

Counting Queries

- **Conjunctions:** $\mathcal{Q}^t_{\text{conj}}(d)$ with all conjunctions of $t \in [0, d]$ literals on $\mathcal{X} = \{0,1\}^d$
  - E.g., $\mathcal{Q}^2_{\text{conj}}(5)$ contains $q(w) = w_2 \wedge \neg w_4$ (what fraction of individual in the dataset have lug cancer and are non-smokers?)
  - These are called ***t*-way marginal**
  - Answering all queries in $\mathcal{Q}^t_{\text{conj}}(d)$ amounts to computing the ***t*-way contingency table**

Postprocessing

- **Theorem:** If $\mathbf{M}: \mathcal{X}^n \times \mathcal{Q} \to \mathcal{Y}$ is ε-DP, and $\Psi: \mathcal{Y} \to \mathcal{Z}$ is any **randomized function**, then $\Psi \circ \mathbf{M}: \mathcal{X}^n \times \mathcal{Q} \to \mathcal{Z}$ is ε-DP
- Let Ψ be a **distribution** on **deterministic** $\psi: \mathcal{Y} \to \mathcal{Z}$. For any $z \in \mathcal{Z}$:

$$
\begin{aligned}
&\mathbb{P}[(\Psi \circ \mathbf{M})(x) = z] \\
&= \mathbb{E}_{\psi \leftarrow \Psi}[\mathbb{P}[\mathbf{M}(x) = \psi^{-1}(z)]] \\
&\leq \mathbb{E}_{\psi \leftarrow \Psi}[e^{\varepsilon} \cdot \mathbb{P}[\mathbf{M}(x') = \psi^{-1}(z)]] \\
&= e^{\varepsilon} \cdot \mathbb{P}[(\Psi \circ \mathbf{M})(x') = z]
\end{aligned}
$$

Group Privacy

- **Theorem:** If $\mathbf{M}$ is ε-DP, then for all pairs of datasets $x, x' \in \mathcal{X}^n$, $\mathbf{M}(x)$ and $\mathbf{M}(x')$ are $k\varepsilon$-DP for $k = d(x, x')$
  - Here, $d(x, x')$ is the number of rows that **need to be changed** to go from x to x'
  - Let x_{i+1} be obtained from x_i by changing **one row**

$$\begin{aligned}
&\mathbb{P}[\mathbf{M}(x_0) = y] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathbf{M}(x_1) = y] \\
&\leq e^{\varepsilon} \cdot (e^{\varepsilon} \cdot \mathbb{P}[\mathbf{M}(x_2) = y]) \\
&\vdots \\
&\leq e^{k\varepsilon} \cdot \mathbb{P}[\mathbf{M}(x_k) = y]
\end{aligned}$$

Basic Composition

- **Theorem:** If $\mathbf{M}_1, \dots, \mathbf{M}_k$ are ε-DP, then $\mathbf{M}$ s.t. $\mathbf{M}(x) = (\mathbf{M}_1(x), \dots, \mathbf{M}_k(x))$ is $k\varepsilon$-DP
- Fix $x \sim x'$. For $y \in \mathcal{Y}$, define

$$\boxed{\Lambda_{\mathbf{M}(x)||\mathbf{M}(x')}(y) = \ln\left(\frac{\mathbb{P}[\mathbf{M}(x) = y]}{\mathbb{P}[\mathbf{M}(x') = y]}\right)}$$

  - When $\Lambda_{\mathbf{M}(x)||\mathbf{M}(x')}(y) > 0$, the outcome y is **"evidence"** that the dataset is x rather than x'
  - Thus, ε-DP means that for all $x \sim x'$, and for all y, $\left|\Lambda_{\mathbf{M}(x)||\mathbf{M}(x')}(y)\right| \leq \varepsilon$

Basic Composition

- In our case:

$$
\begin{aligned}
&\Lambda_{\mathbf{M}(x)||\mathbf{M}(x')}(y) \\
&= \ln\left(\frac{\mathbb{P}[\mathbf{M}_1(x) = y_1 \wedge \ldots \wedge \mathbf{M}_k(x) = y_k]}{\mathbb{P}[\mathbf{M}_1(x') = y_1 \wedge \ldots \wedge \mathbf{M}_k(x') = y_k]}\right) \\
&= \ln\left(\frac{\prod_{i=1}^{k} \mathbb{P}[\mathbf{M}_i(x) = y_i]}{\prod_{i=1}^{k} \mathbb{P}[\mathbf{M}_i(x') = y_i]}\right) \\
&= \sum_{i=1}^{k} \Lambda_{\mathbf{M}_i(x)||\mathbf{M}_i(x')}(y_i) \\
&\Rightarrow \left|\Lambda_{\mathbf{M}(x)||\mathbf{M}(x')}(y)\right| \leq \sum_{i=1}^{k} \left|\Lambda_{\mathbf{M}_i(x)||\mathbf{M}_i(x')}(y_i)\right| \leq k\varepsilon
\end{aligned}
$$

Summary of Properties

- Immune to **auxiliary information**
  – Current and future side information
- Automatically yields **group privacy**
  – Privacy loss $k\varepsilon$ for groups of size k
- Composition
  – Can bound cumulative privacy loss over **multiple analysis** (the epsilons add up)
  – Can combine a few differentially private mechanisms to solve **complex analytical tasks**

Did You XYZ Last Night?

- Flip a coin
  - If heads, **flip again** and return YES if heads, and else return NO
  - If tails, **answer honestly**
- $\frac{\mathbb{P}[\text{YES}|\text{Truth}=\text{YES}]}{\mathbb{P}[\text{YES}|\text{Truth}=\text{NO}]} = \frac{1/2+1/2\cdot(1/2+0)}{0+1/2\cdot(1/2)} = 3$
- $\frac{\mathbb{P}[\text{NO}|\text{Truth}=\text{NO}]}{\mathbb{P}[\text{NO}|\text{Truth}=\text{YES}]} = 3$
- Gives ε-DP for $\varepsilon = \ln 3 \approx 1.098$
- Expected #YES: $1/4\,(1-p) + 3/4\,p$

p = fraction of people that XYZ $= 2(\#\text{YES}) - 1/4$

Randomized Response: Privacy

- Let $q: \mathcal{X} \to \{0,1\}$ be a **counting query**
- For each row x_i, let $\mathbf{M}(x_i) = q(x_i)$ w.p. $1/2 + \varepsilon$ and $\mathbf{M}(x_i) = \overline{q(x_i)}$ w.p. $1/2 - \varepsilon$
- Consider $\mathbf{M}(x) = \mathbf{M}(x_1, \ldots, x_n) = (y_1, \ldots, y_n)$
  - Assume $x \sim x'$ **differ only** in the j-th row

$$\frac{\mathbb{P}[\mathbf{M}(x) = y]}{\mathbb{P}[\mathbf{M}(x') = y]} = \frac{\prod_i \mathbb{P}[\mathbf{M}(x_i) = y_i]}{\prod_i \mathbb{P}[\mathbf{M}(x'_i) = y_i]}$$

$$= \frac{\mathbb{P}[\mathbf{M}(x_j) = y_j]}{\mathbb{P}[\mathbf{M}(x'_j) = y_j]} \leq \frac{1/2 + \varepsilon}{1/2 - \varepsilon}$$

Randomized Response: Accuracy

- The latter is $\leq e^{O(\varepsilon)}$ when, say, $\varepsilon \leq 1/4$ and thus the mechanism $\mathbf{M}$ has $O(\varepsilon)$-DP
- As for **accuracy**, note that
  - $\mathbb{E}[y_i] = (1/2 + \varepsilon) \cdot q(x_i) + (1/2 - \varepsilon) \cdot \big(1 - q(x_i)\big) = 2\varepsilon \cdot q(x_i) + 1/2 - \varepsilon$
  - Thus, $q(x_i) = 1/(2\varepsilon) \cdot \mathbb{E}[(y_i - 1/2 + \varepsilon)]$
- This suggests the following **estimator**

$$\tilde{y} = \frac{1}{n}\sum_{i=1}^{n} \left[\frac{1}{2\varepsilon} \cdot (y_i - 1/2 + \varepsilon)\right]$$

$\mathbb{E}[\tilde{y}] = q(x)$

Randomized Response: Accuracy

- Next, we analyze the **variance**
  - $\mathrm{V}[\tilde{y}] = \mathrm{V}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2\varepsilon}\cdot(y_i - 1/2 + \varepsilon)\right]\right] = \frac{1}{4\varepsilon^2 n^2}\cdot \sum_{i=1}^{n}\mathrm{V}[y_i] \leq \frac{1}{4\varepsilon^2 n^2}\cdot n \cdot \frac{1}{4} = \frac{1}{16\varepsilon^2 n}$
  - Finally, by **Chebyshev**'s inequality

$$\boxed{|\tilde{y} - y| \leq O\left(\frac{1}{\sqrt{n}\cdot\varepsilon}\right)}$$

  - As $n \to \infty$, we get an **increasingly accurate** estimate of the **result**

Differential Privacy by Adding Noise

- Let q be a **counting query**
- Answer with $\mathbf{M}(x) = q(x) + \mathbf{noise}$
  - But according to which distribution?
- Note that if $x{\sim}x'$, then $|q(x) - q(x')| \leq 1/n$
- At every y, the density of output distribution should be same for x, x' **up to a factor** e^{ε}
  - Density of $\mathbf{M}(x)$ (resp. $\mathbf{M}(x')$) at y is that of the noise at $z = y - q(x)$ (resp. $z = y - q(x')$)
  - Again, $|z - z'| \leq 1/n$

Laplace Mechanism: Privacy

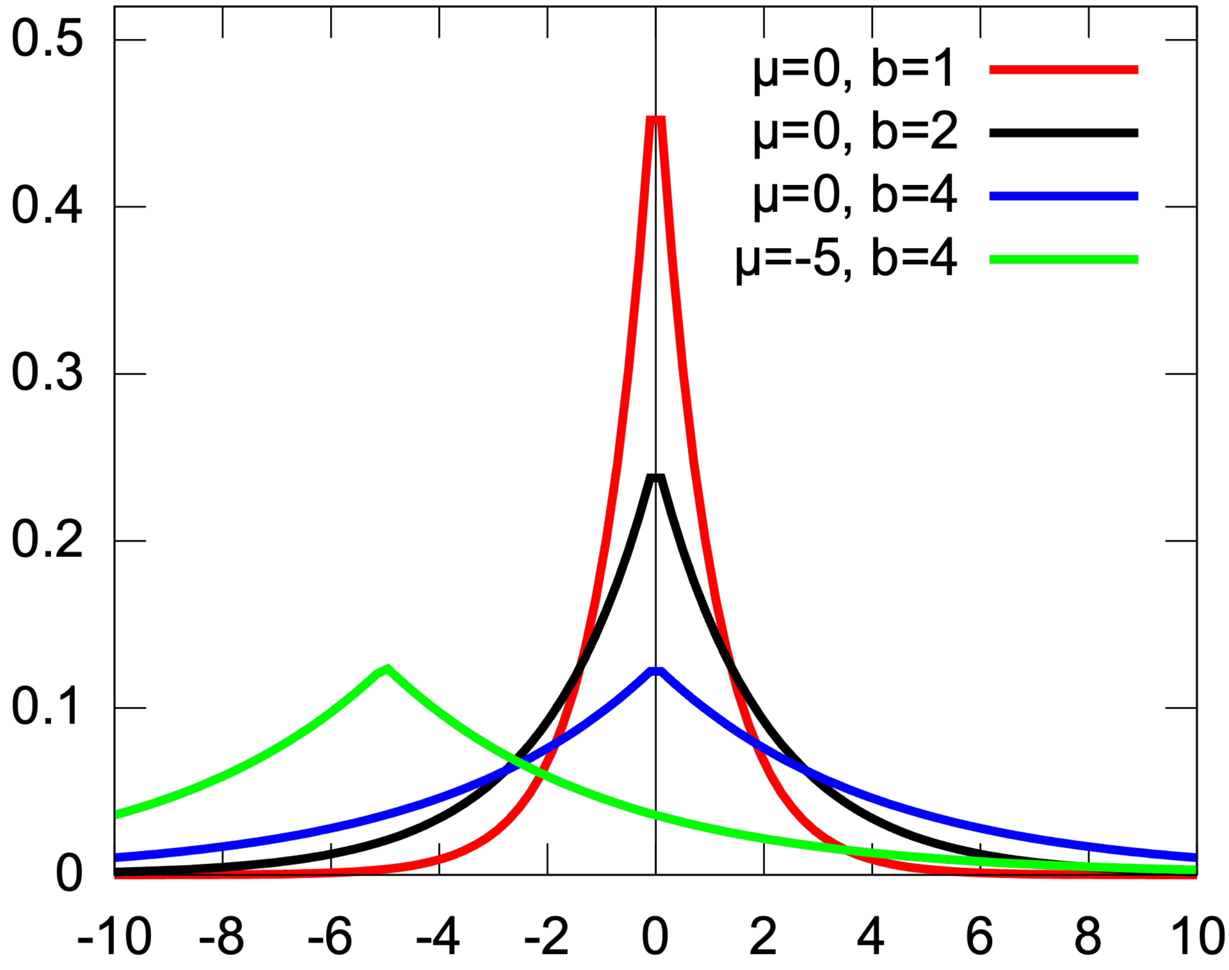
- Let $L(\mu, \sigma)$ at z be $1/(2\sigma) \cdot e^{-|z-\mu|/\sigma}$
- If we set $\mu = 0, \sigma = 1/\varepsilon n$, we have:

$$\frac{\mathbb{P}[\mathbf{M}(x) = y]}{\mathbb{P}[\mathbf{M}(x') = y]}$$

$$= e^{\frac{|y-q(x')|-|y-q(x)|}{\sigma}}$$

$$\leq e^{\frac{|q(x)-q(x')|}{\sigma}} \leq e^{\frac{1}{n\sigma}} = e^{\varepsilon}$$

Laplace Distribution



Laplace Mechanism: Accuracy

- Note that $L(0,\sigma)$ has mean 0 and standard deviation $\sqrt{2}\sigma$, and **exponentially vanishing tails**: $\mathbb{P}[|L(0,\sigma)| > \sigma t] \leq e^{-t}$
- Hence, for any $0 < \beta \leq 1$:

$$\boxed{\mathbb{P}[|q(x) - y| > \ln(1/\beta) \cdot 1/(\varepsilon n)] \leq \beta}$$

- With **high probability** we get error $O(1/(\varepsilon n))$
  - Compare this with accuracy $O(1/\varepsilon\sqrt{n})$ of randomized responses

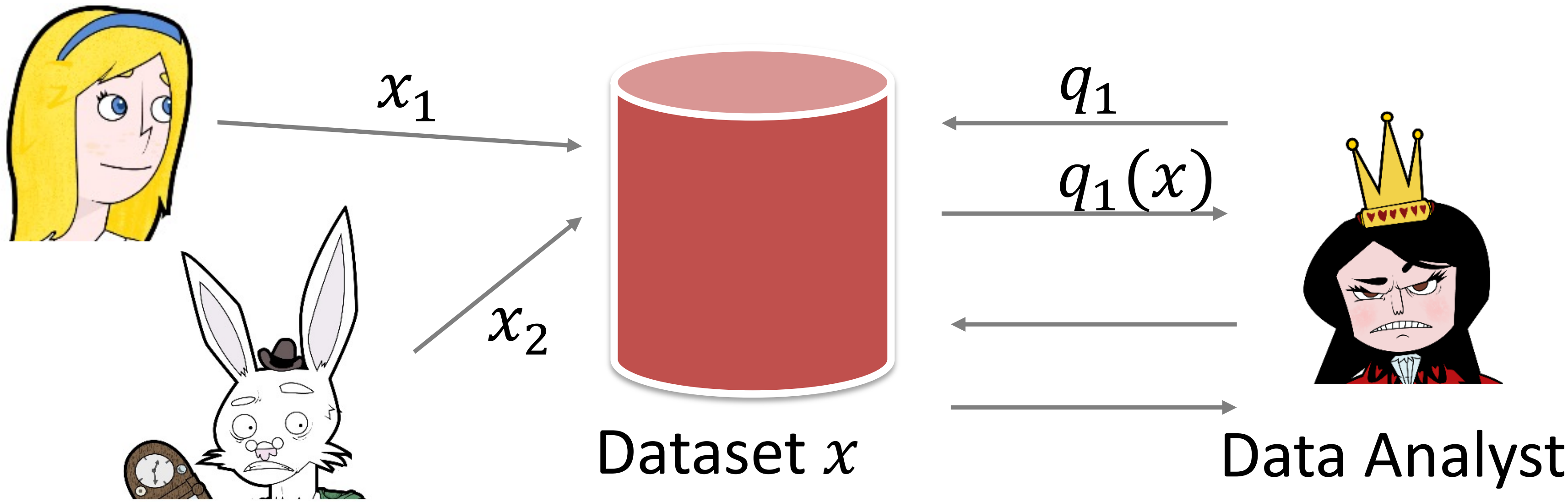
Laplace Mechanism: Multivariate Case

- **Not specific** to counting queries
  - All we used is that $|q(x) - q(x')| \leq 1/n$ for $x{\sim}x'$
- For arbitrary $q\colon \mathcal{X}^n \to \mathbb{R}^d$ scale the noise to **global ℓ_1-sensitivity**

$$\Delta_1 = \max_{x\sim x'} \|q(x) - q(x')\|_1 = \sum_{i=1}^{n} |y_i - y_i'|$$

- **Theorem:** Let $q\colon \mathcal{X}^n \to \mathbb{R}^d$. The mechanism $\mathbf{M}(x) = q(x) + (z_1, \dots, z_d)$ where each $z_i \leftarrow L(0, \Delta_1/\varepsilon)$ satisfies ε-DP

Approximate Differential Privacy



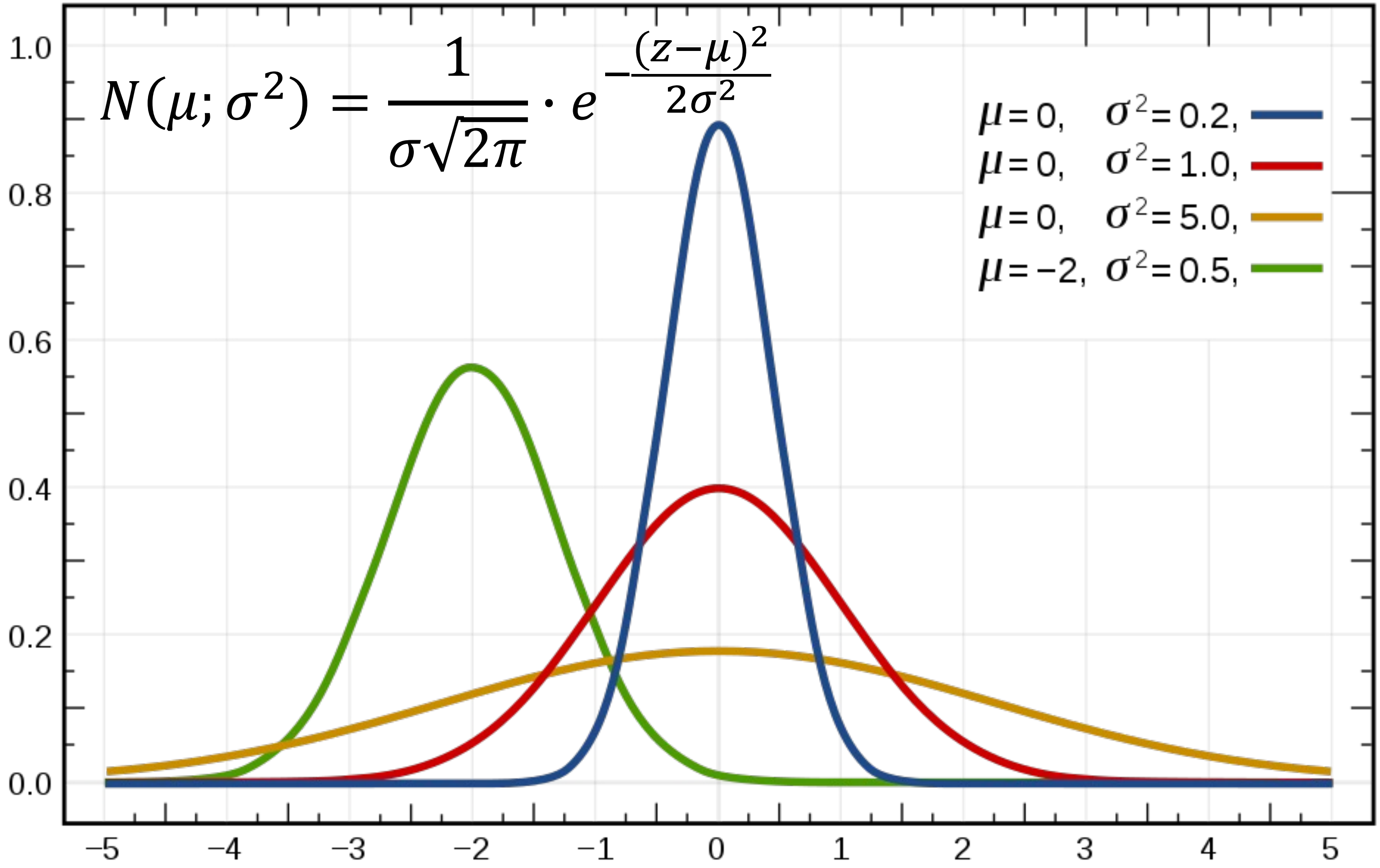
Definition: Mechanism $\mathbf{M}: \mathcal{X}^n \times \mathcal{Q} \rightarrow \mathcal{Y}$ gives (ε, δ)-**differential privacy** if for all pairs of **adjacent** datasets $x, x' \in \mathcal{X}^n$, and for all queries $q \in \mathcal{Q}$:

$$\forall y \in \mathcal{Y}, \mathbb{P}[\mathbf{M}(x,q) = y] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathbf{M}(x',q) = y] + \delta$$

Benefits of the Relaxation

- **Gaussian** noise
  - Leading to **better accuracy**
- **Advanced composition**
  - Can answer k queries with **cumulative** loss $\sqrt{k} \cdot \varepsilon$
  - Instead of $k\varepsilon$ as in **pure** differential privacy
- Can use cryptography to **simulate trusted center** (see a later lecture)

Gaussian Distribution



Gaussian Mechanism

- Let $q: \mathcal{X}^n \to \mathbb{R}^d$. The **global ℓ_2-sensitivity** is:

$$\Delta_2 = \max_{x \sim x'} \|q(x) - q(x')\|_2 = \sum_{i=1}^{n} \sqrt{(y_i - y_i')^2}$$

- **Theorem:** Let $q: \mathcal{X}^n \to \mathbb{R}^d$. The mechanism $\mathbf{M}(x) = q(x) + (z_1, \ldots, z_d)$ where **each** $z_i \sim N\left(0, \frac{2\ln(1.25/\delta) \cdot \Delta_2^2}{\varepsilon^2}\right)$ satisfies (ε, δ)-DP

Gaussian versus Laplace

- Note that for **every** vector $y \in \mathbb{R}^d$, $\|y\|_2 \leq \|y\|_1 \leq \sqrt{d} \cdot \|y\|_2$
- Suppose that $x \in \{0,1\}^{n \times d}$ and take the query $q(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ for the **multivariate mean**
  - Here, $\Delta_1 \leq d/n$ and $\Delta_2 \leq \sqrt{d}/n$
  - The Laplace mechanism would add noise of magnitude $O(d/n\varepsilon)$ whereas the Gaussian mechanism needs **less noise** $O(\sqrt{d \cdot \ln(1/\delta)}/n\varepsilon)$ for roughly the **same accuracy**

Gaussian Mechanism: Privacy

- We first show that for $\mathbf{M}(x) = q(x) + z$ where $z \leftarrow N(0, \sigma^2 \cdot I\ )$ the **privacy loss** is **distributed as** $N\left(\frac{\|q(x)-q(x\prime)\|_2^2}{2\sigma^2}, \frac{\|q(x)-q(x\prime)\|_2^2}{\sigma^2}\right)$

$$\ln\left(\frac{\mathbb{P}[\mathbf{M}(x) = q(x) + z]}{\mathbb{P}[\mathbf{M}(x') = q(x) + z]}\right) = \ln\left(\frac{\exp(-\|z\|_2^2/2\sigma^2)}{\exp(-\|z+v\|_2^2/2\sigma^2)}\right)$$

$$= -\frac{1}{2\sigma^2} \cdot (\|z\|_2^2 - \|z+v\|_2^2)$$

$$= -\frac{1}{2\sigma^2} \cdot \left(\sum_{j=1}^{d} \left(z_j^2 - \left(z_j + v_j\right)^2\right)\right)$$

Gaussian Mechanism: Privacy

- **Fact:** $a \cdot N(0,1) + b \cdot N(0,1) \sim N(0, a^2 + b^2)$
- Simplifying, we get:

$$\ln\left(\frac{\mathbb{P}[\mathbf{M}(x) = q(x) + z]}{\mathbb{P}[\mathbf{M}(x') = q(x) + z]}\right) = \frac{1}{2\sigma^2} \cdot \left(\sum_{j=1}^{d} (2z_j v_j + v_j^2)\right)$$

  – The **constant** term is $\frac{\|v\|_2^2}{2\sigma^2}$ and matches the **mean**

  – The **other** term is $\frac{1}{\sigma^2} \cdot \sum_j z_j v_j = \sum_j z_j' = z'$, where $z_j' \sim N(0, \sigma^2 \cdot v_j^2)$ and $z' \sim N\left(0, \frac{\|v\|_2^2}{\sigma^2}\right)$

Gaussian Mechanism: Privacy

- To finish the proof, we need to show that the **privacy loss** is $\leq \varepsilon$ with **probability** $\geq 1-\delta$
- For $\tilde{z} \sim N(0,1)$ and $\sigma = \Delta_2 \cdot t/\varepsilon$, we can write:

$$\mathbb{P}\left[\left|\frac{\|q(x)-q(x')\|_2}{2\sigma} \cdot \tilde{z} + \frac{\|q(x)-q(x')\|_2^2}{2\sigma^2}\right| \geq \varepsilon\right]$$

$$= \mathbb{P}\left[|\tilde{z}| \geq \frac{\varepsilon\sigma}{\|q(x)-q(x')\|_2} - \frac{\|q(x)-q(x')\|_2}{2\sigma}\right]$$

$$\leq \mathbb{P}\left[|\tilde{z}| \geq t - \frac{\varepsilon}{2t}\right]$$

  – For simplicity we **drop** the **small term** $\varepsilon/(2t)$

Gaussian Mechanism: Privacy

- By standard **tail bounds** $\mathbb{P}[|\tilde{z}| \geq t] \leq e^{-t^2/2}$
- If we **set** $t = \sqrt{2\ln(2/\delta)}$ we obtain that $\mathbb{P}[|\tilde{z}| \geq t] \leq \delta$, which implies (ε, δ)-DP
- Note that the latter corresponds **roughly** to

$$\boxed{\sigma \approx \frac{\Delta_2}{\varepsilon} \cdot \sqrt{\ln(1/\delta)}}$$

Properties of Approximate DP

- **Post processing:** If $\mathbf{M}: \mathcal{X}^n \times \mathcal{Q} \to \mathcal{Y}$ is (ε, δ)-DP, and $F: \mathcal{Y} \to \mathcal{Z}$ is any **randomized function**, then $F \circ \mathbf{M}: \mathcal{X}^n \times \mathcal{Q} \to \mathcal{Z}$ is (ε, δ)-DP
- **Group privacy:** If $\mathbf{M}$ is (ε, δ)-DP, then for all pairs of datasets $x, x' \in \mathcal{X}^n$, $\mathbf{M}(x)$ and $\mathbf{M}(x')$ are $(k\varepsilon, k\delta \cdot e^{(k-1)\varepsilon})$-DP for $k = d(x, x')$
- **Basic composition:** If $\mathbf{M}_1, \dots, \mathbf{M}_k$ are (ε, δ)-DP, then $\mathbf{M}$ s.t. $\mathbf{M}(x) = (\mathbf{M}_1(x), \dots, \mathbf{M}_k(x))$ is $(k\varepsilon, k\delta)$-DP

Advanced Composition

- **Theorem:** For all $\varepsilon, \delta, \delta' > 0$, if $\mathbf{M}_1, \dots, \mathbf{M}_k$ are (ε, δ)-DP, then $\mathbf{M}(x) = (\mathbf{M}_1(x), \dots, \mathbf{M}_k(x))$ is $(\tilde{\varepsilon}, \tilde{\delta})$-DP for

$$\tilde{\varepsilon} = \varepsilon\sqrt{2k \cdot \log(1/\delta')} + k\varepsilon \cdot \frac{e^{\varepsilon} - 1}{e^{\varepsilon} + 1}$$
$$\tilde{\delta} = k\delta + \delta'$$

  - In the **high-privacy** regime, $(e^{\varepsilon}-1)/\,(e^{\varepsilon} + 1) \approx \varepsilon/2$ and thus we can **ignore** the **second term** in $\tilde{\varepsilon}$
  - The above holds even if in the **adaptive** setting

CIS SAPIENZA
Research Center for Cyber Intelligence and Information Security

Reduction to Binary(ish) Mechanisms

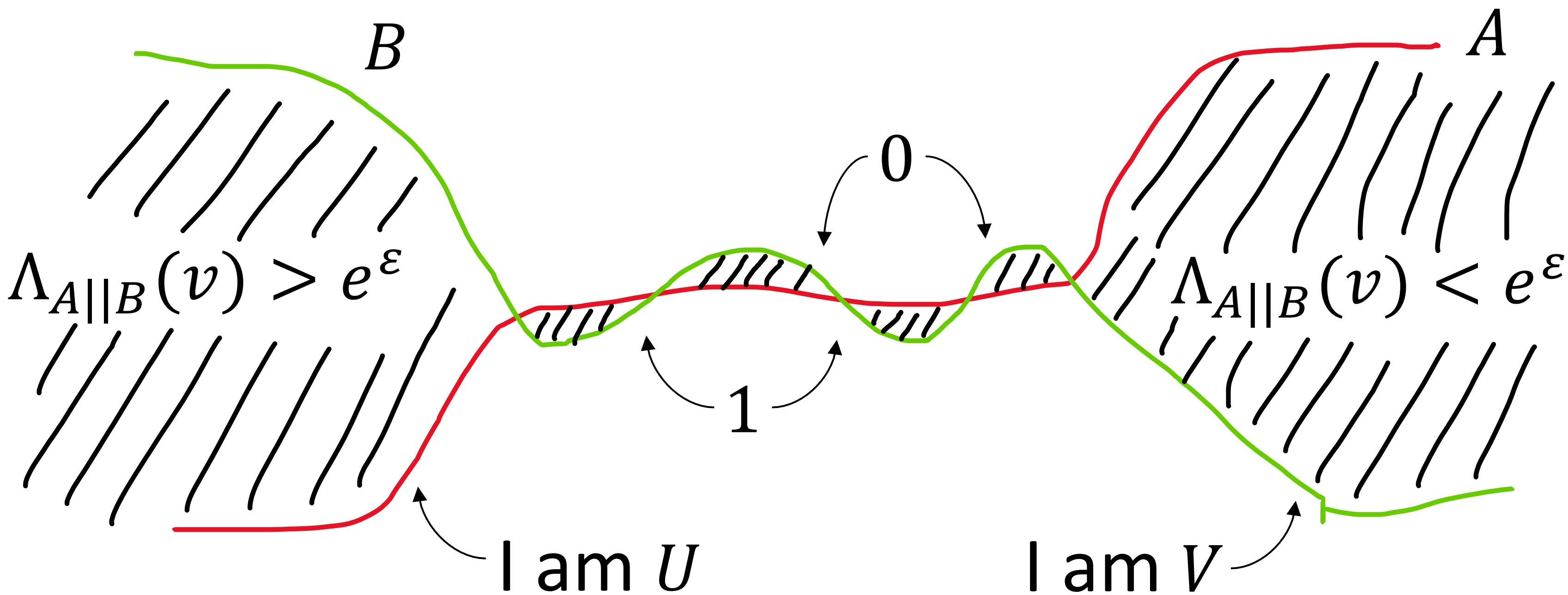
- What is the **simplest** pair of random variables (U, V) satisfying the definition of (ε, δ)-DP?
  - I.e., with probability $\geq 1 - \delta$

$$\Lambda_{U||V} = \left| \ln\left(\frac{\mathbb{P}[U = v]}{\mathbb{P}[V = v]}\right) \right| \leq \varepsilon$$

v	$\mathbb{P}[U = v]$	$\mathbb{P}[V = v]$
0	$e^{\varepsilon}(1-\delta)/(1+e^{\varepsilon})$	$(1-\delta)/(1+e^{\varepsilon})$
1	$(1-\delta)/(1+e^{\varepsilon})$	$e^{\varepsilon}(1-\delta)/(1+e^{\varepsilon})$
I am U	δ	0
I am V	0	δ

Reduction to Binary(ish) Mechanisms

- **Lemma:** Let A and B be s.t. $|\Lambda_{A||B}| \leq \varepsilon$ and $|\Lambda_{B||A}| \leq \varepsilon$ w.p. $1-\delta$; there is a **randomized mapping** Ψ s.t. $\Psi(A) \sim U$ and $\Psi(B) \sim V$



Reduction to Binary(ish) Mechanisms

- We can think $\mathbf{M}_j$ takes as input $x \in \mathcal{X}^n$ as well as the **transcript** τ_{j-1} of **previous outputs**
- **Corollary:** There is a **randomized mapping** Ψ^* s.t. $\mathbf{M}(x) = (\mathbf{M}_1(x), \dots, \mathbf{M}_k(x))$ satisfies
  - $\mathbf{M}(x) \sim \Psi^*(U_1, \dots, U_k)$, with $U_1, \dots, U_k \sim U$
  - $\mathbf{M}(x') \sim \Psi^*(V_1, \dots, V_k)$, with $V_1, \dots, V_k \sim V$
- By **post-processing**, it suffices to bound the privacy loss between $U_1, \dots, U_k$ and $V_1, \dots, V_k$

Composition of Binary(ish) Mechanisms

- Let $v_j \in \{0,1, \text{I am } U\}$ be the j-th **realization**
  - That is, $v_j \sim U$
  - When $v_j = \text{I am } U$, **privacy is violated**, but $\mathbb{P}[\exists v_j \text{ s.t. } v_j = \text{I am } U] = 1 - (1-\delta)^k \leq k\delta$

    Call this event E_1
  - Next, we condition on E_1 **not happening**

$$\ln\left(\frac{\mathbb{P}[(U_1, \dots, U_k) = v]}{\mathbb{P}[(V_1, \dots, V_k) = v]}\right) = \sum_{j=1}^{k} \ln\left(\frac{\mathbb{P}[U_j = v_j]}{\mathbb{P}[V_j = v_j]}\right)$$

$$\sum_{j=1}^{k} \frac{(1-\delta)e^{\varepsilon(1-v_j)}/(e^{\varepsilon}+1)}{(1-\delta)e^{\varepsilon v_j}/(e^{\varepsilon}+1)} = \sum_{j=1}^{k} \varepsilon(1-2v_j)$$

Composition of Binary(ish) Mechanisms

- Note that (always conditioning on $\overline{E_1}$

$$1 - 2v_j = \begin{cases} 1 \text{ w.p. } e^{\varepsilon}/(1+e^{\varepsilon}) \\ -1 \text{ w.p. } 1/(1+e^{\varepsilon}) \end{cases}$$

- Hence, we can compute the **expectation**

$$\mathbb{E}\left[\ln\left(\frac{\mathbb{P}[(U_1,\dots,U_k)=v]}{\mathbb{P}[(V_1,\dots,V_k)=v]}\right)\right] = k\varepsilon \cdot \frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}$$

  – Finally, we apply the **Chernoff bound** in order to prove that the **privacy loss** does **not exceed** its **expectation** with probability more than δ'

Composition of Binary(ish) Mechanisms

- **Hoeffding bound:** For $X_1, \dots, X_k$ i.i.d. and **bounded** in the range $[a, b]$, we have:

$$\mathbb{P}\left[\sum_{j=1}^{k} X_j \geq \mathbb{E}[X_j] + \gamma\right] \leq e^{-\frac{2\gamma^2}{k(b-a)^2}}$$

- Define the event that the privacy loss **goes too far** from its **mean**

Call this event E_2

$$\ln\left(\frac{\mathbb{P}[(U_1, \dots, U_k) = v]}{\mathbb{P}[(V_1, \dots, V_k) = v]}\right) > k\varepsilon \cdot \frac{e^{\varepsilon} - 1}{e^{\varepsilon} + 1} + \beta\varepsilon\sqrt{k}$$

Composition of Binary(ish) Mechanisms

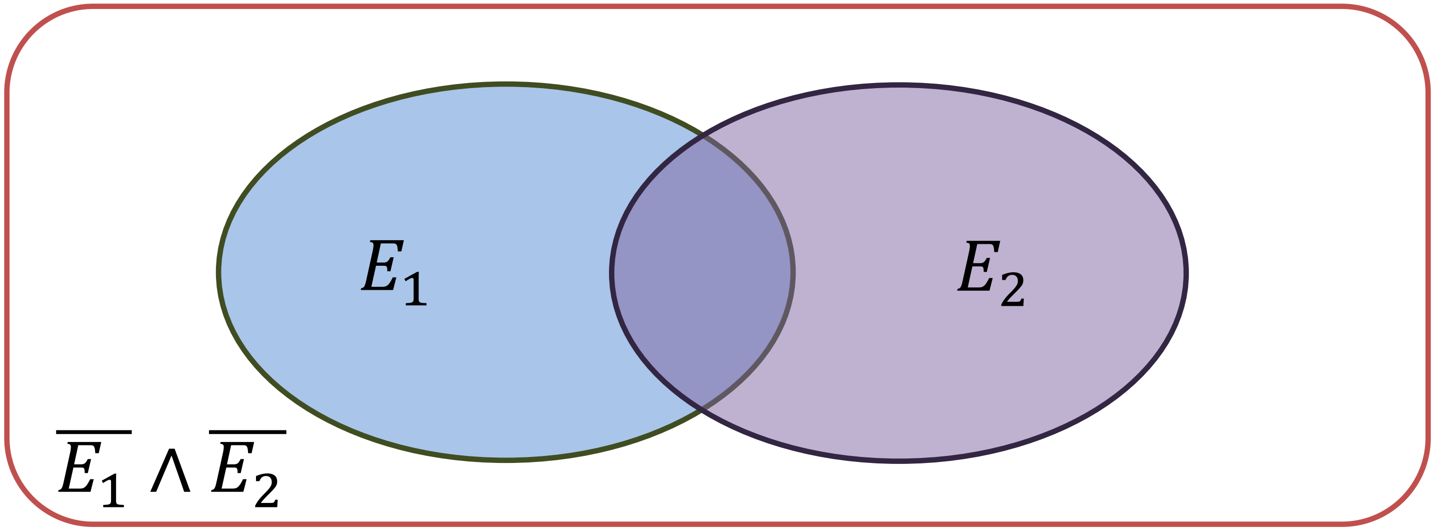
- By **setting** $[a, b] = [-\varepsilon, \varepsilon]$ and $\gamma = \beta\varepsilon\sqrt{k}$

$$\mathbb{P}[E_2 | \overline{E_1}] \leq e^{-\beta^2/2}$$

- Putting it **all together** using $\beta = \sqrt{2\ln(1/\delta')}$

$$\begin{aligned}
&\mathbb{P}[(U_1, \ldots, U_k) = v \wedge \overline{E_1} \wedge \overline{E_2}] \\
&\leq e^{\tilde{\varepsilon}} \cdot \mathbb{P}[(V_1, \ldots, V_k) = v \wedge \overline{E_1} \wedge \overline{E_2}] \\
&\leq e^{\tilde{\varepsilon}} \cdot \mathbb{P}[(V_1, \ldots, V_k) = v]
\end{aligned}$$

Composition of Binary(ish) Mechanisms



$$\mathbb{P}[U^* = v] = \mathbb{P}[U^* = v \wedge \overline{E_1} \wedge \overline{E_2}] + \mathbb{P}[U^* = v \wedge E_1]$$
$$+ \ \ \mathbb{P}[U^* = v \wedge \overline{E_1} \wedge E_2] \leq \mathbb{P}[U^* = v \wedge \overline{E_1} \wedge \overline{E_2}] + \mathbb{P}[E_1]$$
$$+ \ \ \mathbb{P}[E_2 | \overline{E_1}] \cdot \mathbb{P}[\overline{E_1}]$$
$$\leq e^{\tilde{\varepsilon}} \cdot \mathbb{P}[V^* = v] + k\delta + e^{-\beta^2/2} \cdot 1$$
$$e^{\tilde{\varepsilon}} \cdot \mathbb{P}[V^* = v] + k\delta + \delta' = e^{\tilde{\varepsilon}} \cdot \mathbb{P}[V^* = v] + \tilde{\delta}$$

Exponential Mechanism

- Until now, we focused on **numerical** queries
- In some situations, we wish to output **objects**
- Example: **Digital auction**
  - One seller having **infinite copies** of digital good
  - n buyers each with **valuation** v_i
  - What's the price p max. **the revenue** $\sum_{i:v_i \leq p} p$?
- Idea: Use **differential privacy**
  - If $v_1 = v_2 = 1$ and $v_3 = 3.01$, the **revenue drops** from 3 to 1.01 **increasing** p from 1 to 1.01

Exponential Mechanism

- More formally, the mechanism takes as input
  - A **dataset** $x \in \mathcal{X}^n$, a set of **objects** $\mathcal{H}$ and a **score function** $s: \mathcal{X}^n \times \mathcal{H} \to \mathbb{R}$
  - Only the **dataset is private**
- Define the **sensitivity** of the score function:

$$\boxed{\Delta s \leq \max_{h \in \mathcal{H}} \max_{x, x' : x \sim x'} |s(x,h) - s(x',h)|}$$

- **Definition:** The **exponential mechanism** outputs $h \in \mathcal{H}$ w.p. $\propto \exp(\varepsilon \cdot s(x,h)/(2\Delta s))$

Exponential Mechanism: Privacy

- **<u>Theorem:</u>** The **exponential mechanism** is ε-differentially private
- Fix any $x \sim x'$ and $h \in \mathcal{H}$

$$\frac{\mathbb{P}[\mathbf{M}(x,\mathcal{H},s)=y]}{\mathbb{P}[\mathbf{M}(x',\mathcal{H},s)=y]} = \frac{\frac{\exp(\varepsilon \cdot s(x,h)/(2\Delta s))}{\sum_{h'\in\mathcal{H}} \exp(\varepsilon \cdot s(x,h')/(2\Delta s))}}{\frac{\exp(\varepsilon \cdot s(x',h)/(2\Delta s))}{\sum_{h'\in\mathcal{H}} \exp(\varepsilon \cdot s(x',h')/(2\Delta s))}}$$

$$= \exp(\varepsilon \cdot (s(x,h) - s(x',h))/(2\Delta s)) \cdot \frac{\sum_{h'\in\mathcal{H}} \exp(\varepsilon \cdot s(x',h')/(2\Delta s))}{\sum_{h'\in\mathcal{H}} \exp(\varepsilon \cdot s(x,h')/(2\Delta s))}$$

$$\leq \exp(\varepsilon/2) \cdot \frac{\sum_{h'\in\mathcal{H}} \exp(\varepsilon/2) \cdot \exp(\varepsilon \cdot s(x,h')/(2\Delta s))}{\sum_{h'\in\mathcal{H}} \exp(\varepsilon \cdot s(x,h')/(2\Delta s))} = \exp(\varepsilon)$$

Exponential Mechanism: Accuracy

- **Theorem:** Let $s^*(x) = \max_{h \in \mathcal{H}} s(x,h)$ and $\mathcal{H}^*$ be the set containing all $h \in \mathcal{H}$ such that $s(x,h) = s^*(x)$. Then:

$$\mathbb{P}\left[s\big(\mathbf{M}(x,\mathcal{H},s)\big) \leq s^*(x) - \frac{2\Delta s}{\varepsilon} \cdot \left(\ln\left(\frac{|\mathcal{H}|}{|\mathcal{H}^*|}\right) + \beta\right)\right] \leq e^{-\beta}$$

- **Corollary:** Since $|\mathcal{H}^*| \geq 1$, we get

$$\mathbb{P}\left[s\big(\mathbf{M}(x,\mathcal{H},s)\big) \leq s^*(x) - \frac{2\Delta s}{\varepsilon} \cdot (\ln(|\mathcal{H}|) + \beta)\right] \leq e^{-\beta}$$

Exponential Mechanism: Accuracy

- By **definition**:

$$\begin{aligned}
&\mathbb{P}\left[s\left(\mathbf{M}(x, \mathcal{H}, s)\right) \leq \gamma\right] \\
&= \frac{\sum_{h \in \mathcal{H}: s(x,h) \leq \gamma} \exp(\varepsilon \cdot s(x,h)/(2\Delta s))}{\sum_{h' \in \mathcal{H}} \exp(\varepsilon \cdot s(x,h')/(2\Delta s))} \\
&\leq \frac{|\mathcal{H}| \cdot \exp(\varepsilon\gamma/(2\Delta s))}{|\mathcal{H}^*| \cdot \exp(\varepsilon s^*(x,h)/(2\Delta s))} \\
&= \frac{|\mathcal{H}|}{|\mathcal{H}^*|} \cdot \exp(\varepsilon(\gamma - s^*(x,h))/(2\Delta s))
\end{aligned}$$

  – Set $\gamma = s^*(x) - 2\Delta s/\varepsilon \cdot (\ln(|\mathcal{H}|/|\mathcal{H}^*|) + \beta)$

Application: Laplace Mechanism

- Let $x \in \mathcal{X}^n$ and $q: \mathcal{X}^n \to \mathbb{R}$ with **sensitivity** Δ
  - Set x to be the **dataset**, $\mathbb{R} = \mathcal{H}$ be the **objects**, and $s(x, h) = -|q(x) - h|$ be the **score**

$$\boxed{\mathbb{P}[\mathbf{M}(x, \mathcal{H}, s) = h] \propto \exp(\varepsilon \cdot -|q(x) - h|/(2\Delta s))}$$

  - The latter is identical to the **Laplace mechanism** up to a factor of 2 (resulting in **twice** the noise)
  - Actually, the factor of 2 can be **removed** by revisiting the **privacy proof**

Application: Selling Digital Goods

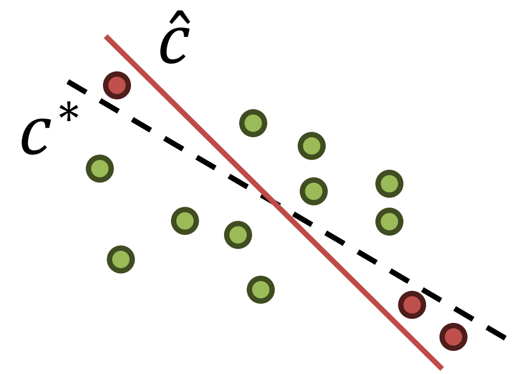
- Example: **Digital auction**
  - One seller having **infinite copies** of digital good
  - n buyers each with **valuation** $v_i \in [0,1]$
  - Price $p \in [0,1]$ maxim. **the revenue** $p \cdot |i: v_i \leq p|$
- We first **discretize** $\mathcal{H} = \{\alpha, 2\alpha, \dots, 1\}$ for some α, so that $|\mathcal{H}| = 1/\alpha$
  - Letting $p^* = \max_p p \cdot |i: v_i \leq p|$, we get $s^*(v_1, \dots, v_n) \geq p^* - \alpha n$ (**round down** p to the **closest multiple** of α, **loosing at most** αn)

Application: Selling Digital Goods

- Example: **Digital auction**
  - One seller having **infinite copies** of digital good
  - n buyers each with **valuation** $v_i \in [0,1]$
  - Price $p \in [0,1]$ maxim. **the revenue** $p \cdot |i: v_i \leq p|$
- We let $s(x, p)$ be **the revenue** $p \cdot |i: v_i \leq p|$
  - Since $p \leq 1$ and changing **an individual** only affects $|i: v_i \leq p|$ by one, $\Delta s \leq 1$
  - Thus, $s(x, p)$ is **at least** $p^* - \alpha n - \ln(1/\alpha)/\varepsilon$ resulting in $p^* - \ln(n)/\varepsilon$ when $\alpha = \ln(n)\,/(n\varepsilon)$

Application: Private PAC Learning

- Probably **approximately** correct **learning**
  - **Concept** class $\mathcal{C} = \{c: \{0,1\}^d \rightarrow \{0,1\}\}$
  - n elements $(x_i, y_i = c^*(x_i))$ for some (**unknown**) $c^* \in \mathcal{C}$, where $x_i \sim D$ (also **unknown**)
- **Goal:** Output $\hat{c}$ s.t. $\mathbb{P}_{x\sim D}[\hat{c}(x) \neq c^*(x)]$ is **minimized**
- Example: **Learning halfspaces**
  - Classic task in **machine learning**
  - Intractable **with noise**



Application: Private PAC Learning

- **Theorem:** If $n = \Omega(\log(|\mathcal{C}|)/2\alpha^2)$, then there exists $\hat{c}$ s.t. $\mathbb{P}_{x\sim D}[\hat{c}(x) \neq c^*(x)] \leq \alpha/2$
- Get the **training data** and see how **every function** in $\mathcal{C}$ **classifies** the dataset
  - Output any function $\hat{c} \in \mathcal{C}$ that **never errs**
- Fix $h \in \mathcal{C}, n = 2t/\alpha^2$. By a **Chernoff bound**:

$$\mathbb{P}_{x_1,\ldots,x_n\sim D}\left[\left|\mathbb{P}_{x\sim D}[h(x) = c^*(x)] - \frac{|i: h(x_i) = c^*(x_i)|}{n}\right| \geq \frac{\alpha}{2}\right] \leq e^{-t}$$

  - With $t = \Omega(\log(|\mathcal{C}|))$, the above holds $\forall h \in \mathcal{C}$

Application: Private PAC Learning

- Turning to **differential privacy**, for $x \sim x'$ we have the change of a **single** point (x_i', y_i')
  - **Worst case:** x_i' may **not** follow D and $y_i' \neq c^*(x_i')$
- **Theorem:** If $n = \Omega(2\log(|\mathcal{C}|)/\alpha^2 + \log(|\mathcal{C}|)/(4\alpha\varepsilon))$, then $\exists$ an ε-**DP** algorithm outputting $\hat{c}$ s.t. $\mathbb{P}_{x \sim D}[\hat{c}(x) \neq c^*(x)] \leq \alpha$
- Apply the **exponential mechanism**
  - Set $\mathcal{C} = \mathcal{H}$ and $s\big((x,y),h\big) = -|i : h(x_i) \neq y_i|/n$
  - So, $\Delta s = 1/n$ and $s^*\big((x,y)\big) = 0$

Application: Private PAC Learning

- The **exponential mechanism** guarantees that we output $\hat{c}$ such that **w.h.p.**:

$$s\left((x,y),\hat{c}\right) \geq -\frac{2\Delta s}{\varepsilon} \cdot \ln(|\mathcal{H}|) = -\frac{2}{\varepsilon n} \cdot \ln(|\mathcal{C}|) \geq -\alpha/2$$

  – Thus $|i: h(x_i) \neq y_i|/n \leq \alpha/2$
  – Putting the two together: $\mathbb{P}_{x\sim D}[\hat{c}(x) \neq c^*(x)] \leq \alpha/2 + \alpha/2 = \alpha$

Answering Many Queries

- Assume we are given a set Q of $k = |Q|$ queries, and we wish to answer all with ε-DP
  - First add Laplace noise to achieve ε_0-DP
  - By **basic composition** can set $\varepsilon_0 = \varepsilon/k$ so that the noise per query has scale $O(1/(\varepsilon_0 n)) = O(k/\varepsilon n)$
- The above implies that we can answer **all queries** in Q with ε-DP to within error

$$\alpha \leq O\left(\frac{k\log k}{\varepsilon n}\right)$$

Setting $\beta = 1/O(k)$

Answering Many Queries

- For (ε, δ)-DP, we can use the **Gaussian** mechanism and **advanced composition** to get:

$$\alpha \leq O\left(\frac{\sqrt{k\log k \cdot \log(1/\delta)}}{\varepsilon n}\right)$$

  – Thus, we can **accurately** answer $k = o(n)$

  – However, note that whenever $|Q|$ is **larger** than n^2, the error is **too large**

- Next, we show how to answer **much more** than n^2 **counting queries**

SmallDB Mechanism

- **Theorem:** There exists an ε-DP mechanism $\mathbf{M}$ such that for all datasets $x \in \mathcal{X}^n$ w.h.p. $\mathbf{M}(x)$ answers **all queries** in Q to within error

$$\boxed{\alpha \leq O\left(\frac{\log|\mathcal{X}| \cdot \log|Q|}{\varepsilon n}\right)^{1/3}}$$

Can handle $\gg n^2$ queries

- Moreover, $\mathbf{M}(x)$ outputs a **synthetic dataset** $y \in \mathcal{X}^m$ with $m = O(\log|Q|/\alpha^2)$ s.t. $\forall q \in Q$ w.h.p. $|q(y) - q(x)| \leq \alpha$

SmallDB Mechanism

- **For each** $y \in \mathcal{X}^m$, let

$$\mathrm{weight}_x(y) = \exp(-\varepsilon n \cdot \max_{q \in \mathcal{Q}} |q(y) - q(x)|)$$

- Output y w.p. $\propto \mathrm{weight}_x(y)$, i.e.

$$\boxed{\Pr[\mathbf{M}(x) = y] = \frac{\mathrm{weight}_x(y)}{\sum_{z \in \mathcal{X}^m} \mathrm{weight}_x(z)}}$$

  - **Exponential mechanism** with **dataset** $x \in \mathcal{X}^n$, **objects** $\mathcal{H} = \{y \in \mathcal{X}^m : m = O(\log|\mathcal{Q}|/\alpha^2)\}$, and **score function** $s(x, y) = -\max_{q \in \mathcal{Q}} |q(y) - q(x)|$

SmallDB Mechanism: Privacy & Accuracy

- **Corollary:** The SmallDB mechanism is 2ε-DP
  - The proof follows directly from the **privacy property** of the **exponential mechanism**
- The **accuracy** proof is more involved
  - First, we show there is at least one **good small dataset** $y \in \mathcal{X}^m$ s.t. $|s(x, y)| \leq \alpha$
  - Then, we show the **exponential mechanism** outputs such a **good** dataset **w.h.p.**

SmallDB Mechanism: Accuracy

- **Chernoff bound:** For $X_1, \ldots, X_m$ i.i.d. in $[0,1]$ and $X = \sum_{j=1}^{m} X_j$ with $\mu = \mathbb{E}[X]$

$$\mathbb{P}[X \geq \mu + \varepsilon] \leq e^{-2m\varepsilon^2} \text{ and } \mathbb{P}[X \leq \mu - \varepsilon] \leq e^{-2m\varepsilon^2}$$

- Let y^* be a **random sample** of m rows from x
  - Then $q(y^*) = \sum_{j=1}^{m} q(x_j)$ and $\mathbb{E}[q(y^*)] = q(x)$
- By the **union bound**, and invoking the Chernoff bound with $m = O(\log |Q|/\alpha^2)$:

$$\Pr[\exists q \in Q \text{ s.t. } |q(y^*) - q(x)| > \alpha] \leq 2|Q| \cdot 2^{-2m\alpha^2}$$

SmallDB Mechanism: Accuracy

- By **accuracy** of the **exponential mechanism** with $\Delta s = 1/n$ and $|\mathcal{H}| = |\mathcal{X}|^{\log|\mathcal{Q}|/\alpha^2}$

$$\mathbb{P}\left[s(\mathbf{M}(x,\mathcal{H},s)) \le s^*(x) - \frac{2}{\varepsilon n}\cdot\left(\frac{\log|\mathcal{X}|\cdot\log|\mathcal{Q}|}{\alpha^2} + \log(1/\beta)\right)\right] \le \beta$$

$$\Rightarrow \mathbb{P}\left[\max_{q\in\mathcal{Q}}|q(y) - q(x)| \ge \alpha + \frac{2}{\varepsilon n}\cdot\left(\frac{\log|\mathcal{X}|\cdot\log|\mathcal{Q}|}{\alpha^2} + \log(1/\beta)\right)\right] \le \beta$$

SmallDB Mechanism: Accuracy

- By replacing α with $\alpha/2$ and **setting** $\alpha/2 = 2/(\varepsilon n) \cdot \left(\frac{4\log |\mathcal{X}| \cdot \log |\mathcal{Q}|}{\alpha^2} + \log(1/\beta)\right)$

$$\mathbb{P}\left[\max_{q \in \mathcal{Q}} |q(y) - q(x)| \geq \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha\right] \leq \beta$$

- Thus, w.p. **at least** $1 - \beta$ the **accuracy** is:

$$\boxed{\alpha \leq O\left(\frac{\log |\mathcal{X}| \cdot \log |\mathcal{Q}|}{\varepsilon n}\right)^{1/3}}$$

The Downside

- The exponential mechanism can be **very expensive**
  - Need to **enumerate over all** $y \in \mathcal{Y}$
- Computation time is

$$\begin{aligned}\Omega(|\mathcal{Y}|) &= \Omega(|\mathcal{X}|^m) \\ &= \Omega\left(|\mathcal{X}|^{O(\log|\mathcal{Q}|/\alpha^2)}\right)\end{aligned}$$

  - Answering all queries in the family $\mathcal{Q}_{\text{conj}}(d)$ **with error tending to zero** requires $n = \omega(d^2/\varepsilon)$

Private Multiplicative Weights

- We now present the state of the art mechanism for **linear queries**
  - Query $q: \mathcal{X} \rightarrow [0,1]$ instead of $q: \mathcal{X} \rightarrow \{0,1\}$
  - For a **dataset** $x \in \mathcal{X}^n$, $q(x) = \frac{1}{n}\sum_{i=1}^{n} q(x_i)$
- **Theorem:** There is a mechanism that answers a **set** $\mathcal{Q}$ of **linear queries** on a dataset with (ε, δ)-**DP** and **accuracy**

Running time $\tilde{\mathrm{O}}(|\mathcal{Q}| \cdot |\mathcal{X}| \cdot n/\alpha^2)$

$$\alpha \leq O\left(\frac{\sqrt{\log |\mathcal{X}| \cdot \log 1/\delta} \cdot \log|\mathcal{Q}|}{\varepsilon n}\right)^{1/2}$$

Lower Bounds

- So far, we have seen DP mechanisms able to answer **many queries** with **good accuracy**
- Next, we look at **lower bounds** essentially telling that these algorithms are **optimal**
- We will consider both
  - Information-theoretic lower bounds
  - Computational lower bounds

Blatant Non-Privacy

- A mechanism $\mathbf{M}: \mathcal{X}^n \to \mathcal{Y}$ is **blatantly non-private** if for every $x \in \mathcal{X}^n$, one can use $\mathbf{M}(x)$ to compute $x' \in \mathcal{X}^n$ s.t. x and x' differ in at most $n/10$ coordinates w.h.p.
  - A very **weak privacy notion**, ruling out attacks that can reconstruct **almost all** of the dataset
  - **Exercise:** A mechanism that is $(1, .1)$-DP **cannot be** blatantly non-private

Reconstruction Attacks

- Let $\mathcal{X} = \{0,1\}$, so that a dataset of n people is a vector $x \in \{0,1\}^n$
- Consider normalized **inner-product queries** $q \in \{0,1\}^n$, with answer $\langle q, x \rangle / n \in [0,1]$
  - Bits of x are **attributes** of the n members, and q specifies a subset of the population according to some **demographics**
  - The value $\langle q, x \rangle / n$ measures the **correlation** between the demographics and the attributes
  - Can be **transformed** into **counting queries**

Reconstruction Attacks

- **Theorem:** If we are given **for each** $q \in \{0,1\}^n$ a value $y_q \in \mathbb{R}$ s.t. $|y_q - \langle q, x\rangle / n| \leq \alpha$, then we can use the y_q's to compute x' **differing** from x in $\leq 4\alpha$ fraction of the coordinates
- **Corollary:** If $\mathbf{M}(x)$ outputs y_q as above with $\alpha \leq 1/40$, then $\mathbf{M}$ is **blatantly non-private**
  - Thus, additive error $\Omega(1)$ is **necessary** for answering all 2^n normalized inner-product queries
  - This shows that the error in SmallDB is **tight**

Reconstruction Attacks

- Pick any x' such that $\forall q$: $\left|y_q - \langle q, x'\rangle / n\right| \leq \alpha$
  - **At least one** x' exists, namely x
- Let $q_1 = x$ and $q_0 = \bar{x}$
- The Hamming distance between x, x' is:

$$\frac{d(x,x')}{n} = \frac{|\langle q_0, x\rangle - \langle q_0, x'\rangle| + |\langle q_1, x\rangle - \langle q_1, x'\rangle|}{n}$$
$$\leq \left|\frac{\langle q_0, x\rangle}{n} - y_{q_0}\right| + \left|y_{q_0} - \frac{\langle q_0, x'\rangle}{n}\right|$$
$$+ \left|\frac{\langle q_1, x\rangle}{n} - y_{q_1}\right| + \left|y_{q_1} - \frac{\langle q_1, x'\rangle}{n}\right| \leq 4 \cdot \alpha$$

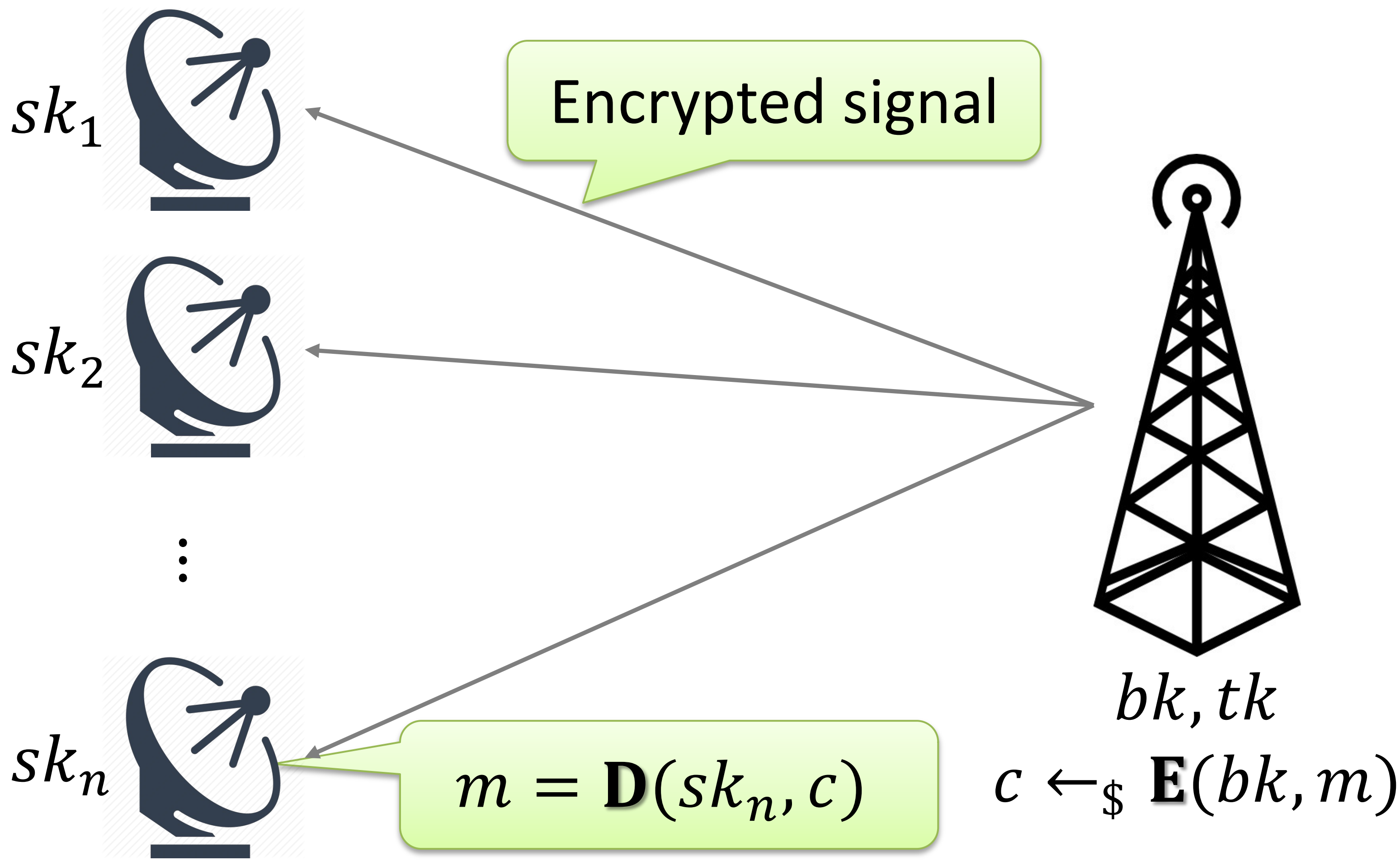
Reconstruction Attacks

- Dinur and Nissim provided a **computationally efficient** variant of the above attack
- **Theorem [DN03]:** For **every** mechanism that answers **all** normalized inner-product queries with accuracy $O(\alpha\sqrt{n})$, there is an **efficient** attacker that reconstructs the dataset in **all but** $O(\alpha^2)$ positions by asking $O(n)$ queries
  - This shows that the Laplace and Gaussian mechanisms are also **tight**

Attacks based on Traitor Tracing

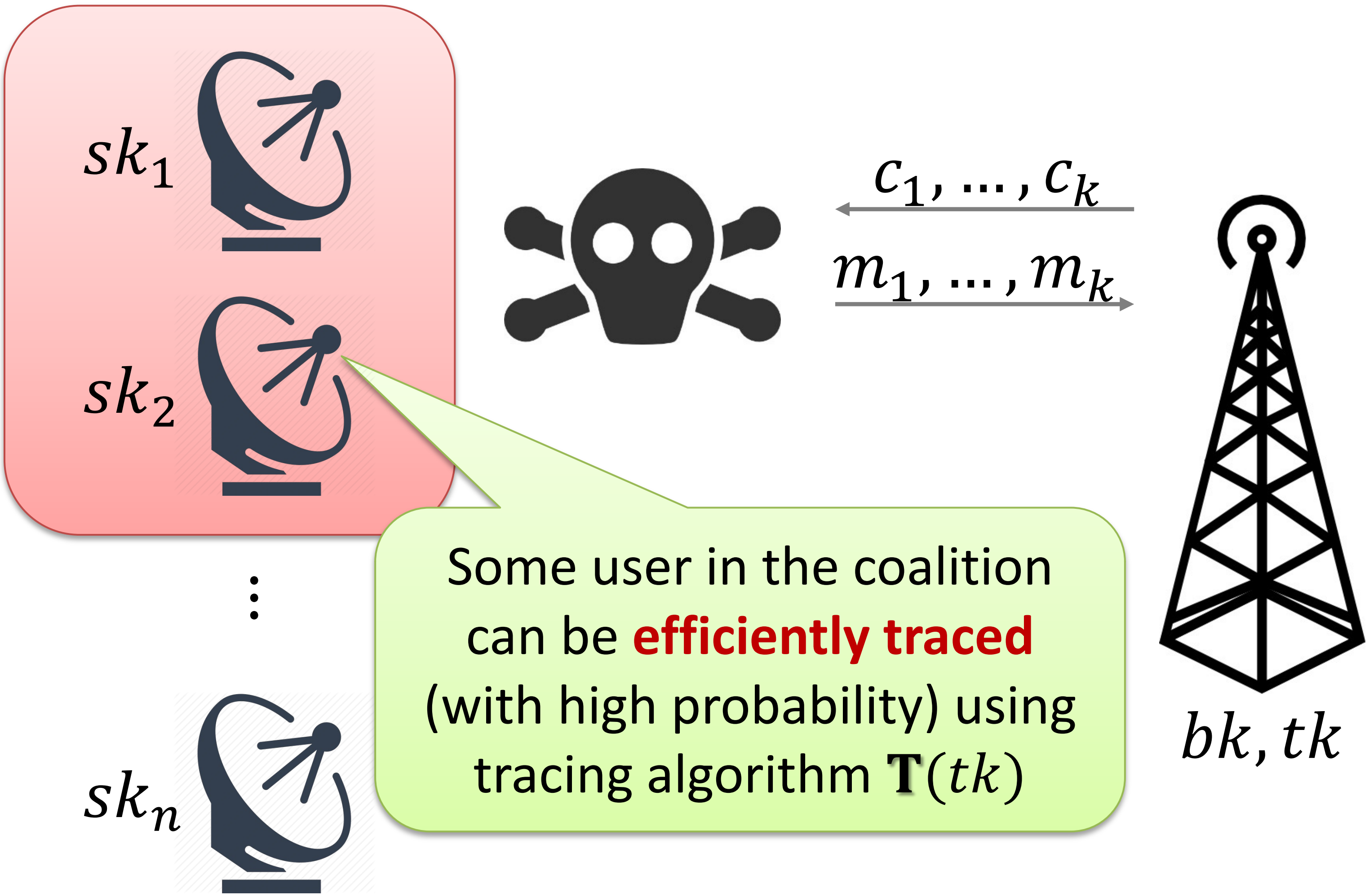
- The smallDB and private multiplicative weight mechanisms answer $\gg n^2$ queries over $\{0,1\}^d$
  - As long as n is large compared to d (e.g., $n \geq d^2$)
- But the computation time is **exponential** in d
- We now show that the above limitation is **inherent** in the **worst case**
- Proof based on **traitor tracing** schemes
  - **Cryptographic** tool for preventing piracy of digital content (using a broadcast channel)

Traitor Tracing



CIS SAPIENZA
Research Center for Cyber Intelligence and Information Security

The Tracing Algorithm



Stateless vs Stateful Pirates

- Pirate corrupts any set $S \subseteq [n]$ of decoders and produces a pirate program $\widetilde{\mathbf{P}}$
- **Stateless** pirates
  - The pirate program $\widetilde{\mathbf{P}}$ is given to the tracer
  - **Useful decryptor:** $\widetilde{\mathbf{P}}$ decrypts honest ciphertexts
- **Stateful** pirates
  - The tracing algorithm can query $\widetilde{\mathbf{P}}$ on $(c_1, \dots, c_k)$
  - **Cooperativeness:** $\widetilde{\mathbf{P}}$ decrypts honest ciphertexts, even after receiving **malformed ciphertexts**

A Computational Lower Bound

- **Theorem:** Assuming **OWFs**, there is a traitor tracing scheme secure against stateful but cooperative pirates
  - Tracing query complexity $k(n, d) = \tilde{O}(n^2)$
- **Theorem:** Every $(1, 1/10n)$-DP mechanism for answering $k = k(n + 1, d)$ counting queries within error $\alpha < 1/2$ on datasets with n individuals from $\mathcal{X} = \{0,1\}^d$ must run in time **superpolynomial** in d

Proof Sketch (1/4)

- Let $\mathbf{M}$ be as in the statement and setup the traitor tracing scheme with $n + 1$ users
- Dataset x contains the **secret keys** $sk_i \in \{0,1\}^d$ of all users but one (chosen at **random**)
- Counting queries: $q_c(sk_i) = \mathbf{D}(sk_i, c)$
  - Hence, $\mathbf{M}(x)$ yields an $\pm\alpha$ approximation a of the number of users in x whose key decrypts c to 1
  - If c is a **valid encryption** of m, then $|a - m| \leq \alpha < 1/2$ so that **rounding** a equals m

Proof Sketch (2/4)

- Define the pirate to be

$$\widetilde{\mathbf{P}}((sk_i)_{i\in S}, c_1, \ldots, c_k) = \left[\mathbf{M}(x, q_{c_1}, \ldots, q_{c_k})\right]$$

  - The accuracy of $\mathbf{M}$ implies that $\widetilde{\mathbf{P}}$ **cooperates**
  - Moreover, by **postprocessing**, $\widetilde{\mathbf{P}}$ is DP too
- Next, we show that **tracing contradicts DP**
- Thus, $\widetilde{\mathbf{P}}$ must **not be traceable** and hence must have **superpolynomial** running time

Proof Sketch (3/4)

- By **traceability** of the traitor traicing, w.p. ≈ 1, algorithm $\mathbf{T}^{\widetilde{\mathbf{P}}((sk_i)_{i\in S},\cdot)}(tk)$ outputs $i \in S$
- Thus, for large enough n, there is an i^* s.t.

$$\Pr\left[\mathbf{T}^{\widetilde{\mathbf{P}}((sk_i)_{i\in S},\cdot)}(tk) = i^*\right] \geq 1/2n$$

- Let $S' = \{1, \dots, n+1\} \backslash \{i^*\}$; by DP:

$$\Pr\left[\mathbf{T}^{\widetilde{\mathbf{P}}((sk_i)_{i\in S},\cdot)}(tk) = i^*\right]$$
$$\leq e \cdot \Pr\left[\mathbf{T}^{\widetilde{\mathbf{P}}((sk_i)_{i\in S'},\cdot)}(tk) = i^*\right] + 1/10n$$

Proof Sketch (4/4)

- Thus,

$$\Pr\left[\mathbf{T}^{\widetilde{\mathbf{P}}((sk_i)_{i\in S'},\cdot)}(tk) = i^*\right] \geq 1/2en - 1/10en \geq \Omega(1/n)$$

- **Corollary:** Assuming **OWFs**, for every $n = \text{poly}(d)$ there is no **poly-time** $(1, 1/10n)$-DP mechanism for answering more than $\tilde{O}(n^2)$ queries over $\mathcal{X} = \{0,1\}^d$ within $\alpha < 1/2$
  - This is **tight**, as we can **accurately** answer $k = \widetilde{\Omega}(n^2)$ counting queries in **polynomial time**

Simple Traitor Tracing

- Let $(\mathbf{E}, \mathbf{D})$ be any **symmetric encryption**
- The broadcast key $bk = (sk_1, \dots, sk_n)$ consists of n **independent secret keys**, and $tk = bk$
- To encrypt $b \in \{0,1\}$, output

$$\boxed{c = (\mathbf{E}(sk_1, b), \dots, \mathbf{E}(sk_n, b))}$$

- To decrypt $c = (c^{(1)}, \dots, c^{(n)})$ use sk_i
  - Suffices to know **which portion** corresponds to the i-th user

How to Trace (1/3)

- Tracing exploits ciphertexts that **different users** would **decrypt differently**

$$\mathbf{TrE}(sk, i) = (\mathbf{E}(sk_1, 1), \dots, \mathbf{E}(sk_i, 1), \mathbf{E}(sk_{i+1}, 0), \dots, \mathbf{E}(sk_n, 0))$$

  - Note that users $j \leq i$ would output 1, but users $j > i$ would output 0

How to Trace (2/3)

- Consider the matrix below

$$\begin{pmatrix} 00\ldots00 & 11\ldots11 & 11\ldots11 & 11\ldots11 & 11\ldots11 \\ & 00\ldots00 & 11\ldots11 & 11\ldots11 & \\ & & 00\ldots00 & 11\ldots11 & \\ & & & \ldots & \\ 00\ldots00 & & & 00\ldots00 & 11\ldots11 \end{pmatrix}$$

- Encrypt each column and **randomly permute**
  - I.e., generate **random** $i_1, \ldots, i_k \in [0, n]$ for $k = (n+1) \cdot s$ s.t. each of $[0, n]$ appears s times
  - Then $C = (c_j)_{j \in [k]}$ with $c_j \leftarrow_\$ \mathbf{TrE}(sk, i_j)$

How to Trace (3/3)

- Next, query $\widetilde{\mathbf{P}}((sk_i)_{i\in S},\cdot)$ with $(c_1, \ldots, c_k)$ obtaining $(b_1, \ldots, b_k)$, and compute

$$\forall i \in [0,n]: p_i = \frac{1}{s} \cdot \sum_{j:i_j=i} b_j$$

- Output any i^* such that $p_{i^*} - p_{i^*-1} \geq 1/n$
  - Note that if $c \leftarrow_\$ \mathbf{TrE}(sk, 0)$, then **every user** would return $b = 0$ (similarly for $\mathbf{TrE}(sk, n)$)
  - Thus, $p_0 = 0$ and $p_n = 1$ and so **there exists** i^* such that $p_{i^*} - p_{i^*-1} \geq 1/n$

Analysis (1/2)

- It remains to show that w.h.p. $i^* \in S$
- Note that $\mathbf{TrE}(sk, i^*)$ and $\mathbf{TrE}(sk, i^* - 1)$ differ for the message encrypted under sk_{i^*}
  - And if $i^* \notin S$ this key is **unknown** to the pirate
- By security of encryption, we can replace k repetitions of $\mathbf{E}(sk_{i^*}, 1)$ with $\mathbf{E}(sk_{i^*}, 0)$ **without effecting** the success of the pirate
  - After this change $\mathbf{TrE}(sk, i^*)$ and $\mathbf{TrE}(sk, i^* - 1)$ are **identical**

Analysis (2/2)

- Since $i_1, \ldots, i_k$ are **random**, the pirate **does not know** which i_j is i^* and which is $i^* - 1$
  - Thus, if it wants to make p_{i^*} larger than p_{i^*-1}, for $i^* \notin S$ it can't do better than **guessing**
- Taking $s = \tilde{O}(n^2)$ and applying Chernoff yields that $\forall i \notin S$ w.h.p. $p_i - p_{i-1} = o(1/n)$
- Note that the query complexity is $k = \tilde{O}(n^3)$
  - The above can be improved to $k = \tilde{O}(n^2)$ by using **fingerprinting codes**

Hardness of Synthetic Data (1/4)

- Some mechanisms work by producing a **compact representation** of **all answers**
  - This is the case, e.g., of SmallDB
- In the traitor tracing world this corresponds to **stateless pirates**
  - If the ciphertext length is $\ell(n,d)$, the previous proof rules out **efficient** mechanisms for answering families Q of counting queries of description length $\ell(n+1,d)$ and size $2^{\ell(n+1,d)}$
  - Interesting only if $\ell \ll n$

Hardness of Synthetic Data (2/4)

- The above applies only to **"unnatural"** $\mathcal{Q}$
- Towards overcoming this limitation, consider the following database using signature $(\mathbf{S}, \mathbf{V})$
  - Choose **single** sign/verify key pair (vk^*, sk^*)
  - Database made by n rows: $(m_i, \mathbf{S}(sk^*, m_i), vk^*)$ for **random messages** m_i
  - One query for each vk: What fraction of rows are valid signatures w.r.t. vk (i.e., $q_{vk}(\cdot) = \mathbf{V}(vk,\cdot)$)?

Hardness of Synthetic Data (3/4)

- **Efficient curator** cannot generate **synthetic dataset** which is accurate w.r.t. vk^*
  - Let $\mathbf{M}$ output $\hat{x} \in (\{0,1\}^d)^{\hat{n}}$
  - By accuracy, $\hat{x}$ contains $\hat{x}_j = (\widehat{m}_j, \hat{\sigma}_j)$ such that $\mathbf{V}\left(vk^*, (\widehat{m}_j, \hat{\sigma}_j)\right) = 1$
- If $\widehat{m}_j \notin x$, then $\mathbf{M}$ violates unforgeability
- If $\widehat{m}_j \in x$, then $\mathbf{M}$ violates differential privacy
  - For each $i \in [n]$, if $\mathbf{M}$ has (ε, δ)-DP it outputs m_i w.p. $\leq e^{\varepsilon}/2^d + \delta$

Hardness of Synthetic Data (4/4)

- Finally, it is possible to express the query $\mathbf{V}(vk^*,\cdot)$ with **2-way conjunctions**
  - By means of the PCP theorem
- **Theorem:** Assuming **OWFs**, there exists $\alpha > 0$ such that there is no $n = \mathrm{poly}(d)$ and **poly-time** $(1, 1/10n)$-DP mechanism that given dataset $(\{0,1\}^d)^n$ outputs a **synthetic dataset** approximating all the queries in $Q^2_{\mathrm{conj}}(d)$ to within error at most α

Incentives

- Until now the goal was designing differentially private mechanisms, but the data is assumed to be **already there**
- But why should someone **participate** in the computation?
- Why would they give their **true data**?
- Do we need **compensation**? How much?

Game Theory and Mechanism Design

- Idea: Solve **optimization problem**
- Catch: **No access** to inputs
  - Inputs held by **self-interested** agents
- Design **incentives** and choice of solution (mechanism) that incentivizes **truth-telling**
  - No need for participants to strategize
  - Simple to predict what will happen
  - Often a non-truth-telling mechanism can be replaced by one where the coordinator does the lying on behalf of the participants

Good News

- **Composition:** Approximate truthfulness still satisfied under composition!
- **Collusion resistance:** $O(k\varepsilon)$-approximate dominant strategy, even for coalitions of k agents
- Both properties not immediate in game-theoretic mechanism design
- All done **without money**!

Bad News

- But not only truthful reporting gives an approximate dominant strategy
  - **Any** report does so, even **malicious** ones
- How do we actually properly get people to **truthfully participate**?
  - Perhaps need **compensation**
  - Much harder to achieve

Differential Privacy as a Tool

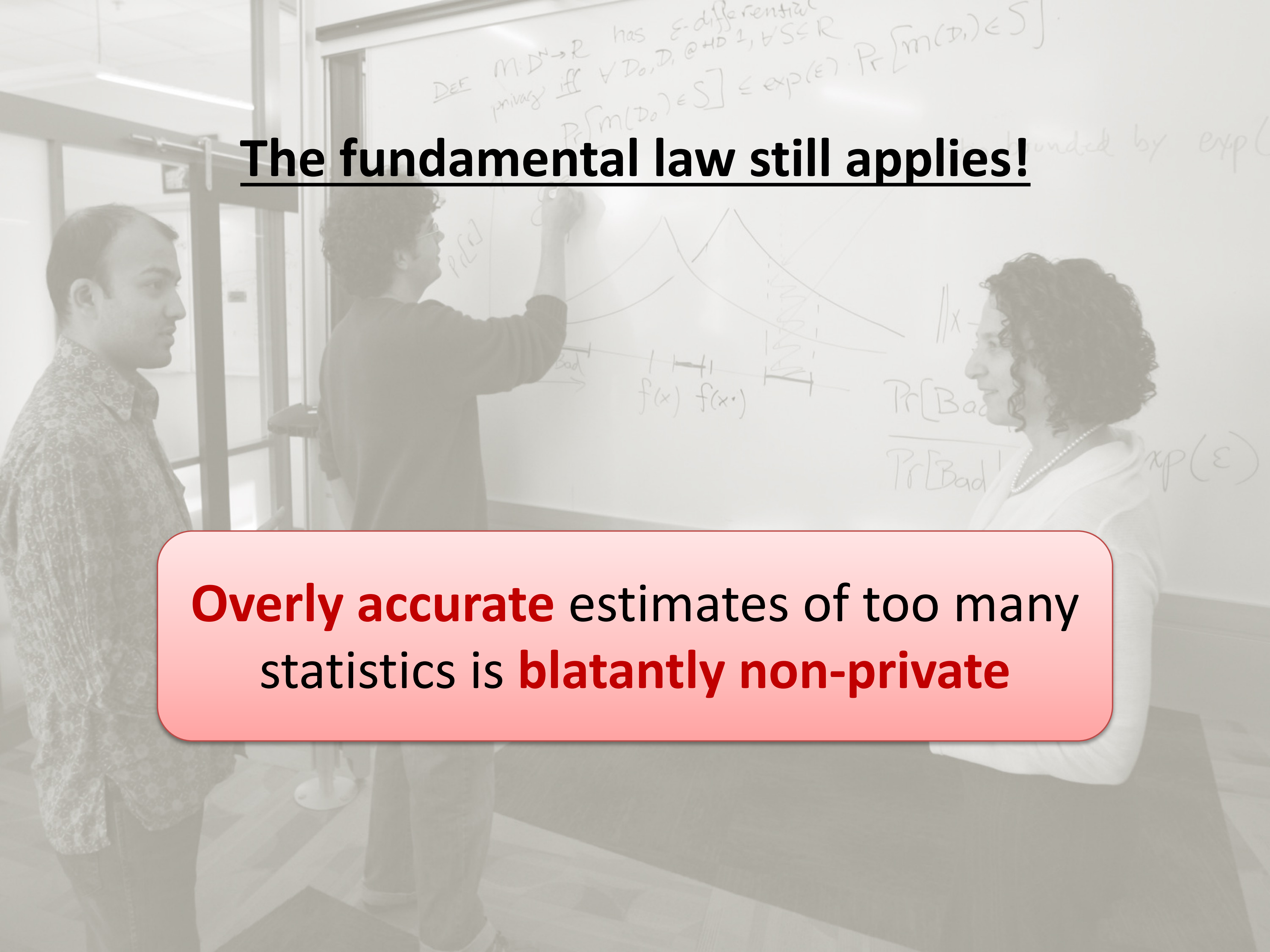
- **Nash Equilibrium:** An assignment of players to strategies so that no player would benefit by changing strategy, given how everyone else is playing
- **Correlated Equilibrium:** Players have access to correlated signals (e.g., traffic light)
  - Every Nash equilibrium is a correlated equilibrium, but not viceversa
- Differential privacy has applications to mechanism design with **correlated equilibria**

The Issue of Verification

- Challenging to strictly **incentivize truth-telling** in differentially private mechanisms design
- Exceptions:
  - Responses are **verifiable**
  - Agents **care** about outcome
- Challenge: No observed outcome
  - What is the prevalence of drug use?
  - Are students cheating in class?

Privacy and Game Theory

- **Asymptotic truthfulness**, new mechanisms design and equilibrium selection results
- Interesting challenge of modeling **costs for privacy**
- In order to design privacy for humans do we need to understand
  - How people currently value or should value it?
  - What are the right promises to give?



The fundamental law still applies!

Overly accurate estimates of too many statistics is **blatantly non-private**