Basics on Differential Privacy

Changyeol Lee (Computer Science, Yonsei University)

Motivations and Backgrounds

Fundamental Limit

- For **all** techniques for *privacy-preserving data analysis*, *overly accurate* answers to *too many* questions will destroy privacy.

- Goal: postpone this as long as possible

Problematic Approaches

- Anonymization
  - removal of personally identifiable information
- Vulnerable to *linkage attack*.
  - the medical records of the governor were identified by matching anonymized medical data with publicly available voter registration records

Problematic Approaches

- Usage of queries over large set
  - reject questions about specific individuals
- Vulnerable to *differencing attack*.
  - “How many people have disease D?” 900
  - “How many people except Mr. X have disease D?” 899
  - Auditing can be disclosive and/or computationally infeasible.

Differential Privacy, a Promise

- A **promise** made by a *data curator*:

A data subject will **not** be *affected*

by allowing his/her data to be used in any data analysis,

no matter what other information sources are available.

Differential Privacy, a Promise

- A **promise** made by a *data curator*.
- Any sequence of responses to queries is “essentially” equally likely to occur, independent of the presence or absence of any individual.

Terminologies and Definitions

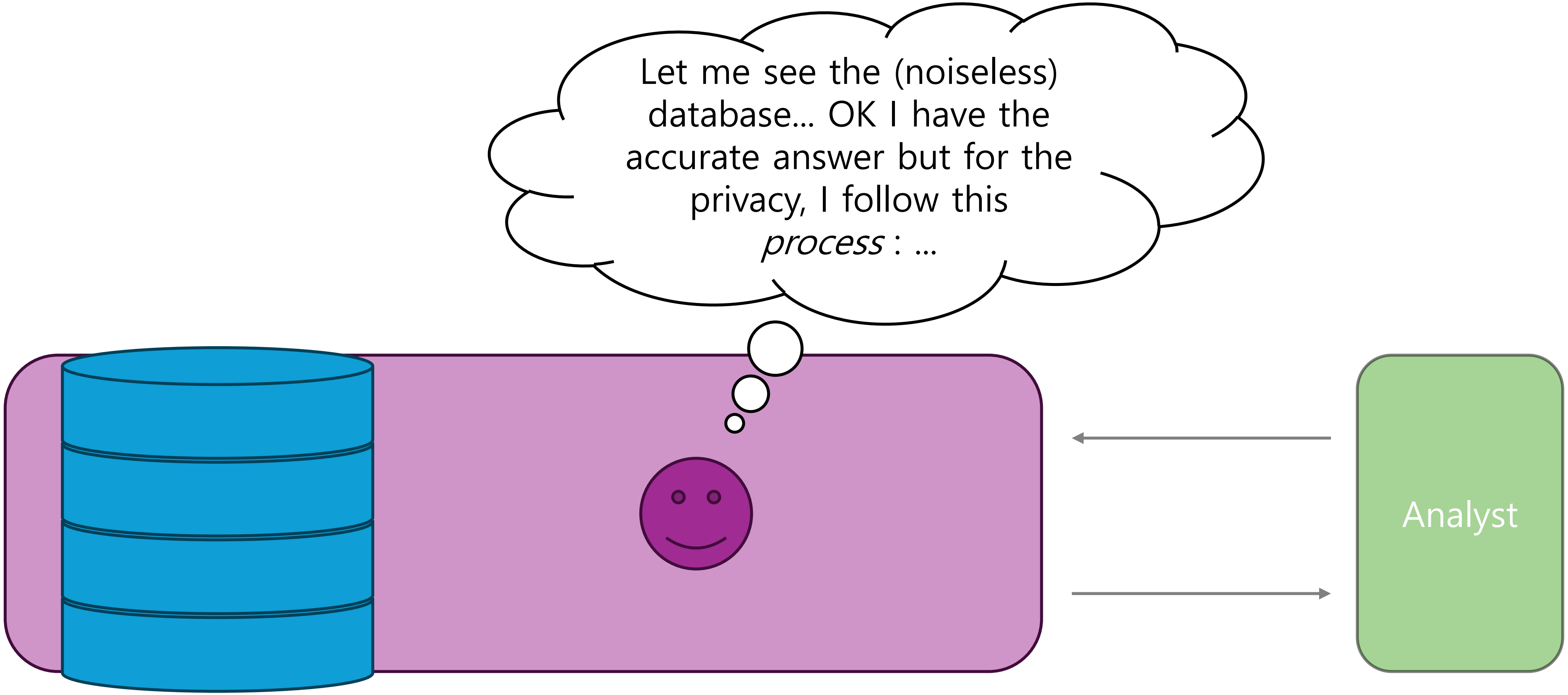
and some properties

The Model of Computation

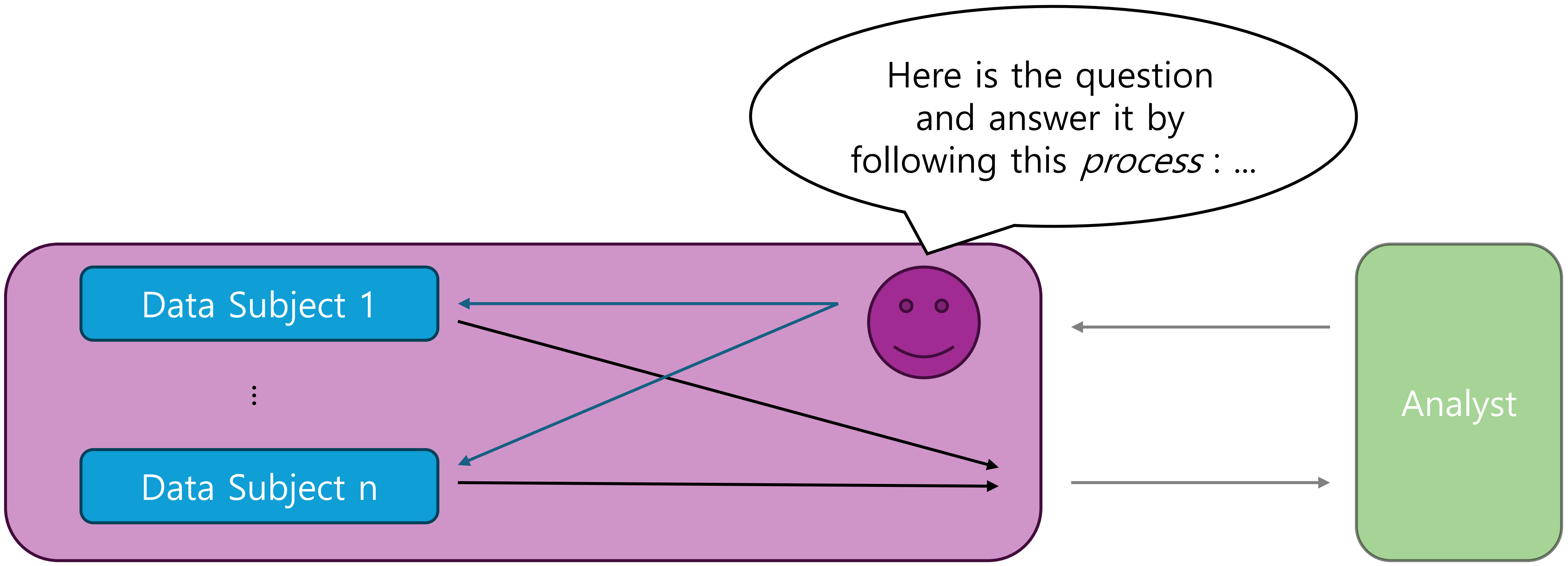
- A curator C outputs an object. (e.g., statistics, data table, histogram)
  - Offline or non-interactive model: C outputs an object once for all.
  - Online or interactive model: Allows multiple queries. (which can be adaptive)
- Privacy-preserving data analysis: An analyst A knows "no more" about any individual after the analysis is done than A knew before the analysis was begun.

The Model of Computation

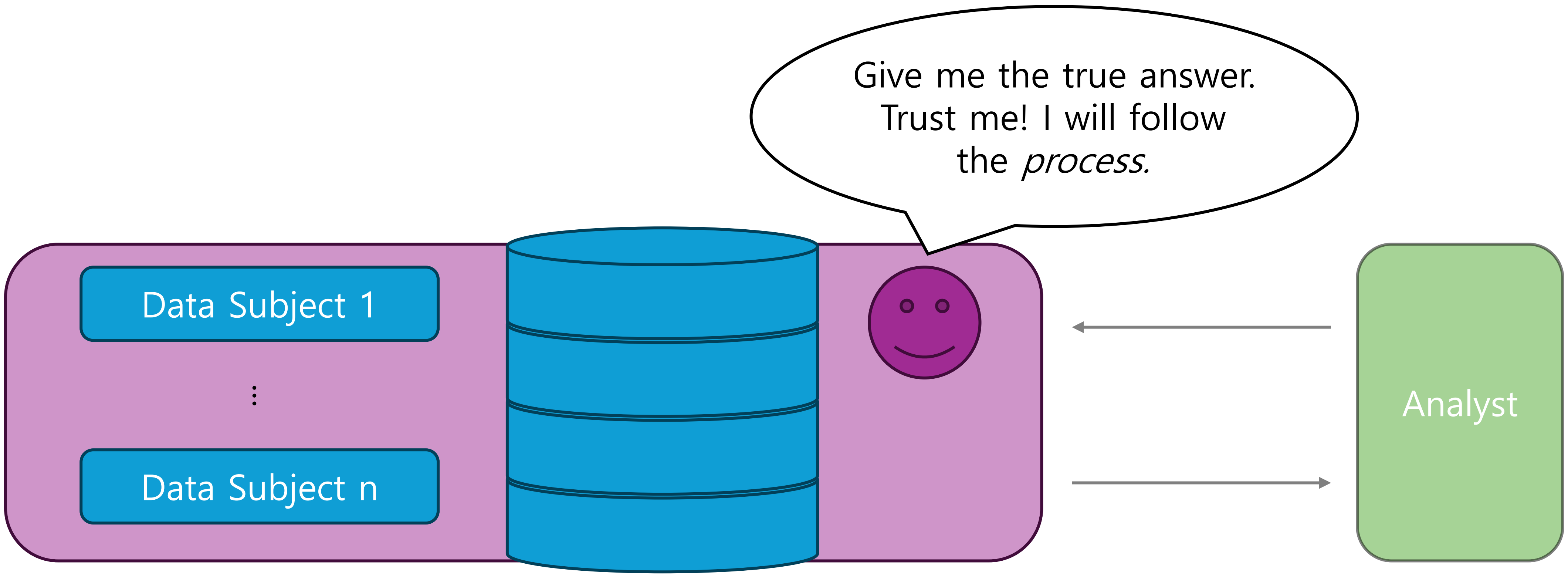
The Model of Computation



The Model of Computation



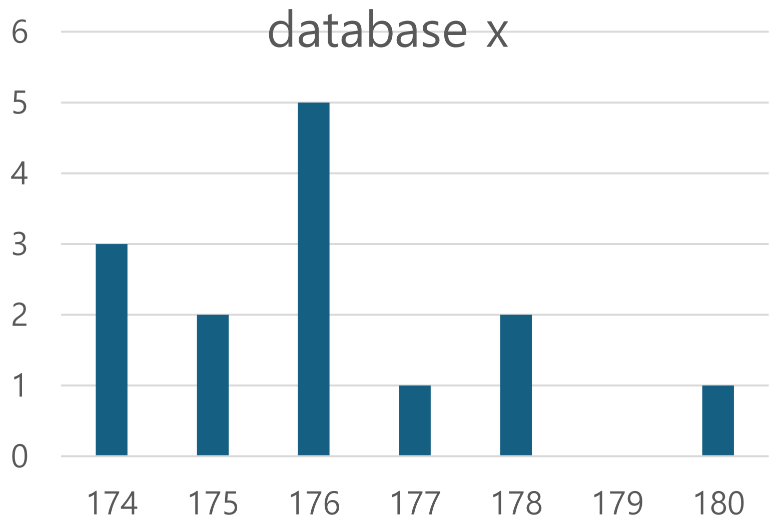
The Model of Computation



Mechanism

- A universe $\mathcal{X}$ of data types
  - Heights) $\mathcal{X} = \{\dots, 174,175,176, \dots\}$
  - Disease D) $\mathcal{X} = \{(Alice\ has\ D), (Bob\ has\ D), (Chris\ has\ D), \dots\}$
- A database x is multiset of $\mathcal{X}$
  - $x = \{\dots, 174,174,174,175,175,176,176, \dots\}$
  - $x = \{0,1,1, \dots\}$

Mechanism



- A universe $\mathcal{X}$ of data types
  - Heights) $\mathcal{X} = \{\ldots, 174, 175, 176, \ldots\}$
  - Disease D) $\mathcal{X} = \{(Alice\ has\ D), (Bob\ has\ D), (Chris\ has\ D), \ldots\}$
- A database $x \in \mathbb{N}^{|\mathcal{X}|}$ is a histogram of $\mathcal{X}$ $\mathbb{N}$: nonneg int
  - $x = \{\ldots, 3, 2, 5, \ldots\}$
  - $x = \{0, 1, 1, \ldots\}$

Mechanism

- A universe $\mathcal{X}$ and a database $x \in \mathbb{N}^{|\mathcal{X}|}$
- randomness (i.e., some random bits)
- a set of queries
  - "How many 177?" "Does Alice have disease D?"
- Output: a string (an object)
  - an output string can be a *synthetic database* $x' \in \mathbb{N}^{|\mathcal{X}|}$

Distance between databases

- The distance btw $x, y \in \mathbb{N}^{|\mathcal{X}|}$ is $\|x - y\|_1 = \sum_{i=1,\dots,|\mathcal{X}|} |x_i - y_i|$.
- We say x and y are neighboring (or $x \sim y$) if $\|x - y\|_1 \leq 1$.
  - For the privacy of an individual.
  - For the privacy of a group of size k, $\|x - y\|_1 \leq k$.

Randomized Algorithm

set of probability distributions

Definition 2.1 (Probability Simplex). Given a discrete set B, the *probability simplex* over B, denoted $\Delta(B)$ is defined to be:

$$\Delta(B) = \left\{ x \in \mathbb{R}^{|B|} : x_i \geq 0 \text{ for all } i \text{ and } \sum_{i=1}^{|B|} x_i = 1 \right\}$$

Definition 2.2 (Randomized Algorithm). A randomized algorithm $\mathcal{M}$ with domain A and discrete range B is associated with a mapping $M : A \to \Delta(B)$. On input $a \in A$, the algorithm $\mathcal{M}$ outputs $\mathcal{M}(a) = b$ with probability $(M(a))_b$ for each $b \in B$. The probability space is over the coin flips of the algorithm $\mathcal{M}$.

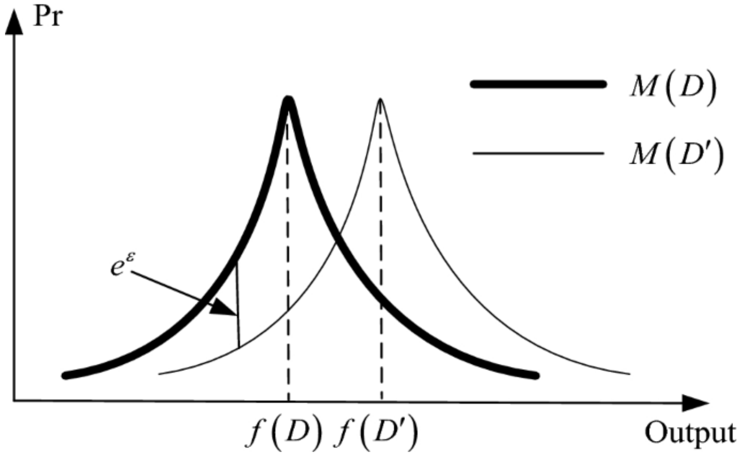
Differential Privacy

Definition 2.4 (Differential Privacy). A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ε, δ)-differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$:

*for continuous case

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

where the probability space is over the coin flips of the mechanism $\mathcal{M}$. If $\delta = 0$, we say that $\mathcal{M}$ is ε-differentially private.



*Randomness is essential!

ϵ-DP vs (ϵ, δ)-DP

- Consider (ϵ, δ)-DP $\mathcal{M}$ where $\delta > 0$.
- For some x, there might (rarely) exists an outcome s s.t. $\exists y \sim x$ where $\Pr[\mathcal{M}(x) = s] \approx 0.01 \cdot \delta$ and $\Pr[\mathcal{M}(y) = s] \approx \delta$.
    - The probability of observing s is *significantly* much higher on y.
    - The privacy loss is large. $\text{Privacy loss} = \ln\left(\frac{\Pr[\mathcal{M}(x) = s]}{\Pr[\mathcal{M}(y) = s]}\right)$
- In ϵ-DP, this cannot happen.

Immune to post-processing

Proposition 2.1 (Post-Processing)**.** Let $\mathcal{M} : \mathbb{N}^{|\mathcal{X}|} \to R$ be a randomized algorithm that is (ε, δ)-differentially private. Let $f : R \to R'$ be an arbitrary randomized mapping. Then $f \circ \mathcal{M} : \mathbb{N}^{|\mathcal{X}|} \to R'$ is (ε, δ)-differentially private.

Another promise, “same” utility

- Sps a (future) event is determined based on the output of $\mathcal{M}$.
- Let $\mathcal{E}$ be a set of all events and $f: \mathrm{Range}(\mathcal{M}) \to \mathcal{E}$ be a decider.
- Sps each individual i has an arbitrary utility over $\mathcal{E}$.

  Let $u_i: \mathcal{E} \to \mathbb{R}_{\geq 0}$ denote the utility function.

$$\mathbb{E}_{E \sim f(\mathcal{M}(x))}[u_i(E)] =$$

expected utility of i
when i is in the dataset

Another promise, “same” utility

- Sps a (future) event is determined based on the output of $\mathcal{M}$.
- Let $\mathcal{E}$ be a set of all events and $f: \mathrm{Range}(\mathcal{M}) \to \mathcal{E}$ be a decider.
- Sps each individual i has an arbitrary utility over $\mathcal{E}$.

  Let $u_i: \mathcal{E} \to \mathbb{R}_{\geq 0}$ denote the utility function.

$$\mathbb{E}_{E \sim f(\mathcal{M}(x))}[u_i(E)] = \sum_{E \in \mathcal{E}} u_i(E) \cdot \Pr_{f(\mathcal{M}(x))}[E]$$

expected utility of i
when i is in the dataset

Another promise, “same” utility

- Sps a (future) event is determined based on the output of $\mathcal{M}$.
- Let $\mathcal{E}$ be a set of all events and $f\colon \mathrm{Range}(\mathcal{M}) \to \mathcal{E}$ be a decider.
- Sps each individual i has an arbitrary utility over $\mathcal{E}$.

  Let $u_i\colon \mathcal{E} \to \mathbb{R}_{\geq 0}$ denote the utility function.

$$\mathbb{E}_{E\sim f(\mathcal{M}(x))}[u_i(E)] = \sum_{E\in\mathcal{E}} u_i(E) \cdot \Pr_{f(\mathcal{M}(x))}[E] \leq \sum_{E\in\mathcal{E}} u_i(E) \cdot e^{\epsilon} \Pr_{f(\mathcal{M}(y))}[E]$$

expected utility of i when i is in the dataset

(if $\mathcal{M}$ is ϵ-DP) Immune to post-processing.

Another promise, "same" utility

- Sps a (future) event is determined based on the output of $\mathcal{M}$.
- Let $\mathcal{E}$ be a set of all events and $f: \mathrm{Range}(\mathcal{M}) \to \mathcal{E}$ be a decider.
- Sps each individual i has an arbitrary utility over $\mathcal{E}$.

  Let $u_i: \mathcal{E} \to \mathbb{R}_{\geq 0}$ denote the utility function.

$$\mathbb{E}_{E\sim f(\mathcal{M}(x))}[u_i(E)] = \sum_{E\in\mathcal{E}} u_i(E) \cdot \Pr_{f(\mathcal{M}(x))}[E] \leq \sum_{E\in\mathcal{E}} u_i(E) \cdot e^{\epsilon} \Pr_{f(\mathcal{M}(y))}[E] = e^{\epsilon} \cdot \mathbb{E}_{E\sim f(\mathcal{M}(y))}[u_i(E)]$$

expected utility of i
when i is in the dataset

(if $\mathcal{M}$ is ϵ-DP)
Immune to post-processing.

expected utility of i
when i is not in the dataset

ϵ-DP for group

Theorem 2.2. Any $(\varepsilon, 0)$-differentially private mechanism $\mathcal{M}$ is $(k\varepsilon, 0)$-differentially private for groups of size k. That is, for all $\|x - y\|_1 \leq k$ and all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(k\varepsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}],$$

where the probability space is over the coin flips of the mechanism $\mathcal{M}$.

Accuracy

- One (informal) definition:

  Let $x \in \mathbb{N}^{|\mathcal{X}|}$ be a database, $f: \mathbb{N}^{|\mathcal{X}|} \rightarrow R$ be a query. Let $output \in R$ be the output of the mechanism.

$$\Pr[\text{diff}(f(x), output) \text{ being large}]$$ is small.

  for some difference measure function diff.

  - Note $f(x)$ is the true answer.

Accuracy

- Another (informal) definition:

  Let $x \in \mathbb{N}^{|\mathcal{X}|}$ be a database, $f_1, \ldots, f_k : \mathbb{N}^{|\mathcal{X}|} \to R$ be a set of queries. Let $output_i \in R$ be the output for each f_i.

$$\max_i \mathrm{diff}(f_i(x), output_i) \text{ is small.}$$

  for some difference measure function diff.

Simple Mechanism for Boolean question

Randomized Response

Randomized Response

- Sps the query "Does i have disease D?" is given.

  Consider the following mechanism $\mathcal{M}$ with any database x:

  - with probability $1/2$, output x_i;

  - with probability $1/2$, output uniform random bit.

- $\Pr[\mathcal{M}(x) = x_i] = \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}$.
- Consider y s.t. $x \sim y$ and $y_i \neq x_i$. $\Pr[\mathcal{M}(y) = x_i] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.

Randomized Response

- $\mathcal{M}$ is $(\ln 3, 0)$-DP.

- You could say $\mathcal{M}$ is $(0, 1/2)$-DP but if possible, we want to analyze it as ϵ-DP.

- In general, we want $\delta = O\left(\frac{1}{\mathrm{superpoly}(\|x\|_1)}\right)$

What we will cover

Other type of queries

- It’s hard to answer numeric queries such as
  - “how many 177?”
  - “how many people in [170,175), [175,180), respectively?

Numeric queries

- Let $x \in \mathbb{N}^{|\mathcal{X}|}$ be a database, $f: \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$ be a numeric query.
- Instead returning $f(x)$, "Perturb"!
- Consider returning $f(x) + Y$ where Y is a random vector in $\mathbb{R}^k$.
  - Scale of noise? Depends on Δf, the *sensitivity* of f.

$$\Delta f = \max_{x,y \in \mathbb{N}^{|\mathcal{X}|} : x \sim y} \|f(x) - f(y)\|$$

  - captures the magnitude by which a single individual's data can change the function f in the worst case

Numeric queries

- Let $x \in \mathbb{N}^{|\mathcal{X}|}$ be a database, $f: \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$ be a numeric query.
- Instead returning $f(x)$, "Perturb"!
- Consider returning $f(x) + Y$ where Y is a random vector in $\mathbb{R}^k$.
- Laplacian mech $Y_i \sim Lap(\Delta_1 f / \epsilon)$ is ϵ-DP
- Gauss. mech $Y_i \sim N(0, \sigma)$ w/ $\sigma \geq O\left(\ln \frac{1}{\delta}\right) \cdot \Delta_2(f)/\epsilon$ is (ϵ, δ)-DP.

Nonnumeric Queries with utility

- Random noise might be problematic in some cases.

  Example 3.5 (Pumpkins.). Suppose we have an abundant supply of pumpkins and four bidders: A, F, I, K, where A, F, I each bid \$1.00 and K bids \$3.01. What is the optimal price? At \$3.01 the revenue is \$3.01, at \$3.00 and at \$1.00 the revenue is \$3.00, but at \$3.02 the revenue is zero!

- Output an object with probability based on its utility

- Exponential distribution is ϵ-DP.

Privacy on union of outputs

- Suppose we have a query $f: \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}$.
- Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be any ϵ-DP mechanism. (Could be $\mathcal{M}_1 = \mathcal{M}_2$)
- Let a_1, a_2 be the answer for f of each mechanism, resp.
- Knowing only a_1 and a_2 preserves privacy of an individual.
- How about knowing both a_1 and a_2?
- It still preserves privacy but less privately than before.

Privacy on union of outputs

- Suppose we have a query $f: \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}$.
- Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be any ϵ-DP mechanism. (Could be $\mathcal{M}_1 = \mathcal{M}_2$)
- Let $\boldsymbol{\mathcal{M}}$ be the mechanism s.t. when given $\boldsymbol{f} \coloneqq (f, f)$,

  it outputs $\big(\mathcal{M}_1(x, f), \mathcal{M}_2(x, f)\big)$.
- $\boldsymbol{\mathcal{M}}$ is 2ϵ-DP.

Composition

- Combination of k number of ϵ-DP mechanisms is $k\epsilon$-DP.
- Can we do better?

Composition

- Combination of k number of ϵ-DP mechanisms is $k\epsilon$-DP.
- “Strong (or Advanced) composition”
    - better analysis gives better bound

Composition

Definition 3.7. We say that the family $\mathcal{F}$ of database access mechanisms satisfies *ε-differential privacy under k-fold adaptive composition* if for every adversary A, we have $D_\infty(V^0 \| V^1) \leq \varepsilon$ where V^b denotes the view of A in k-fold Composition Experiment b above.

(ε, δ)-differential privacy under k-fold adaptive composition instead requires that $D_\infty^\delta(V^0 \| V^1) \leq \varepsilon$.

Theorem 3.20 (Advanced Composition). For all $\varepsilon, \delta, \delta' \geq 0$, the class of (ε, δ)-differentially private mechanisms satisfies $(\varepsilon', k\delta + \delta')$-differential privacy under k-fold adaptive composition for:

$$\varepsilon' = \sqrt{2k \ln(1/\delta')}\varepsilon + k\varepsilon(e^\varepsilon - 1).$$

Composition

Example 3.7. Suppose, over the course of his lifetime, Bob is a member of $k = 10,000$ $(\varepsilon_0, 0)$-differentially private databases. Assuming no coordination among these databases — the administrator of any given database may not even be aware of the existence of the other databases — what should be the value of ε_0 so that, over the course of his lifetime, Bob's cumulative privacy loss is bounded by $\varepsilon = 1$ with probability at least $1 - e^{-32}$? Theorem 3.20 says that, taking $\delta' = e^{-32}$ it suffices to have $\varepsilon_0 \leq 1/801$. This turns out to be essentially optimal against an arbitrary adversary, assuming no coordination among distinct differentially private databases.

What if too many queries?

- Need more than independent noise to preserve privacy + ensuring accuracy
  - Sps we are given a query with sensitivity 1.
  - Answering a single query as $f(x) + Lap(1/\epsilon)$ gives ϵ-DP.
  - But $\{f_i(x) + Lap(1/\epsilon)\}_{i\in[k]}$ is not "private" anymore when k is large…
    - The average converges to the true answer.
  - In this case, the magnitude of the noise need to scale with k. (Not good)

What if too many queries?

- Need more than (independent noise + strong composition)

- If we only care the (numeric) queries that lie above a certain fixed threshold, we can use the sparse vector technique.
  - Discard the numeric answer (where a random noise is added) that lie significantly below the given threshold.

What if too many queries?

- Need more than (independent noise + strong composition)
- If not…?

What if too many queries?

- Need more than (independent noise + strong composition)
- Instead of adding independent noise, add correlated noise.
- “Handle a set of query as a whole”
  - SmallDB (offline algorithm): direct application of exponential mechanism + sampling bounds (learning theory)

What if too many queries?

- Need more than (independent noise + strong composition)
- Instead of adding independent noise, add correlated noise.
- "Handle a set of query as a whole"
  - MWU, Multiplicative weight update (online algorithm): direct application of the sparse vector technique

Other

- Possibly, more generalization
  - generalization of SmallDB/MWU: net mechanism/online learning alg
- Lower bounds and trade-offs results.
  - E.g., How inaccurate must responses be in order not to completely destroy any reasonable notion of privacy?
- Application to other fields
  - ML, mechanism design, combinatorial optimization and so on…