Differential Privacy: What, Why and When

A Tutorial

Moni Naor מוני נאור

Weizmann Institute of Science

Slides credit: Guy Rothblum, Kobbi Nissim, Cynthia Dwork…

Crypto Innovation School (CIS 2018) Shenzhen Nov 29th 2018

What is Differential Privacy?

- Differential Privacy is a concept
  - Motivation
  - Rigorous mathematical definition
  - Properties
  - A measurable quantity
- Set of algorithmic techniques for achieving it
- First defined in:
  - *Dwork, McSherry, Nissim, and Smith*, **Calibrating Noise to Sensitivity in Private Data Analysis**, Third Theory of Cryptography Conference, TCC 2006.
  - Earlier roots: *Warner*, **Randomized Response,** 1965

Why Differential Privacy?

- DP: Strong, quantifiable, composable mathematical privacy **guarantee**
- **Provably resilient** to known and unknown attack modes!
- Theoretically: DP enables many computations with personal data while preserving personal privacy
  - Practicality in first stages of validation

Not a panacea

Good References

- **The Algorithmic Foundations of Differential Privacy**

  Cynthia Dwork and Aaron Roth

  http://www.cis.upenn.edu/~aaroth/privacybook.html

- **The Complexity of Differential Privacy**, Salil Vadhan

- Differential Privacy: A Primer for a Non-technical Audience

  https://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_new.pdf

Privacy-Preserving Analysis: The Problem

- Given dataset with **sensitive personal** info

Health, social n/w, location, communication,

- How to compute and release **functions of the dataset**

- While protecting **individual privacy**

Academic research, informed policy, national security

Glorious Failures of Traditional Approaches to Data Privacy

- Re-identification [Sweeney ’00, …]
- Auditors [Kenthapadi, Mishra, Nissim ’05]
- Genome-Wide association studies (**GWAS**) [Homer et al. ’08]
- Netflix Prize [Narayanan, Shmatikov ‘08]
- Social networks [Backstrom, Dwork, Kleinberg ‘11]
- Attack on statistical aggregates [Dwork, Smith, Steinke, Ullman Vadhan ‘15]

The Netflix Prize

- **Netflix Recommends Movies to its Subscribers**
  - Seek an improved recommendation system
  - Offered $1,000,000 for “10% improvement”
  - Published training data

Prize won in September 2009
“BellKor's Pragmatic Chaos team”

Very influential competition in machine learning

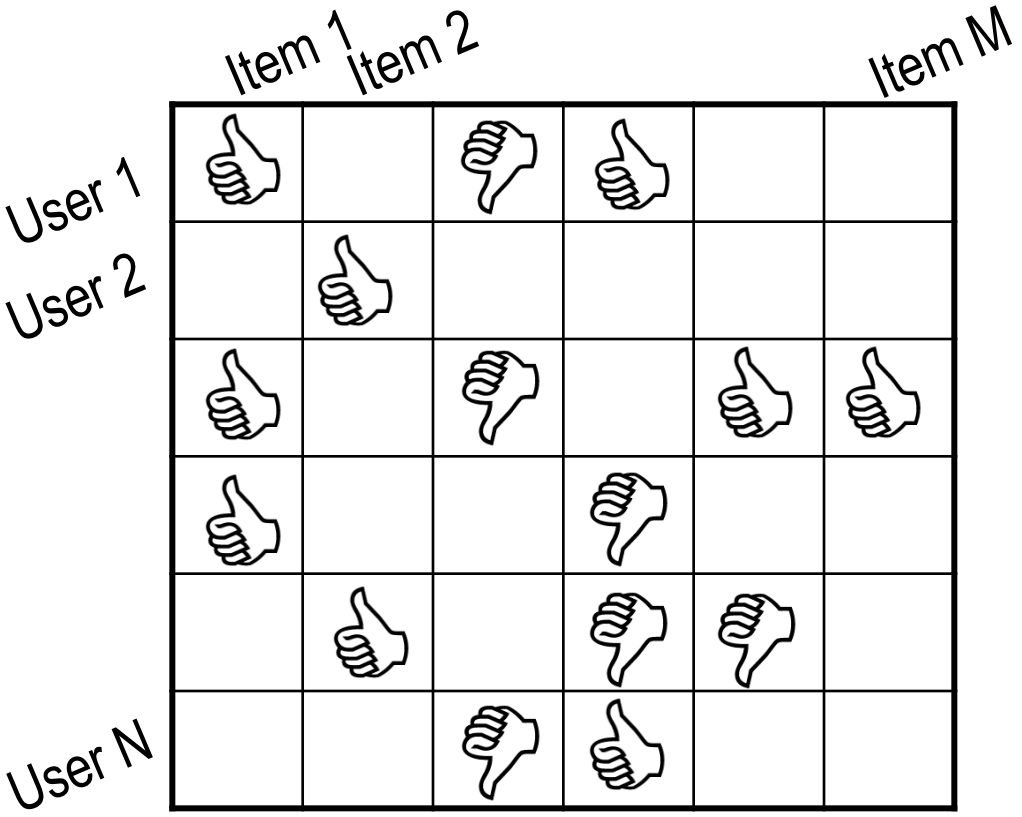
From the Netflix Prize Rules Page…

- "The training data set consists of more than 100 million ratings from over 480 thousand randomly-chosen, **anonymous** customers on nearly 18 thousand movie titles."

- "The ratings are on a scale from 1 to 5 (integral) stars. **To protect customer privacy, all *personal information identifying* individual customers has been removed and all customer ids have been replaced by randomly-assigned ids.** The **date** of each rating and the title and year of release for each movie are provided."

Netflix Data Release [Narayanan-Shmatikov 2008]

- Ratings for subset of movies and users
- Usernames replaced with random IDs
- Some additional perturbation

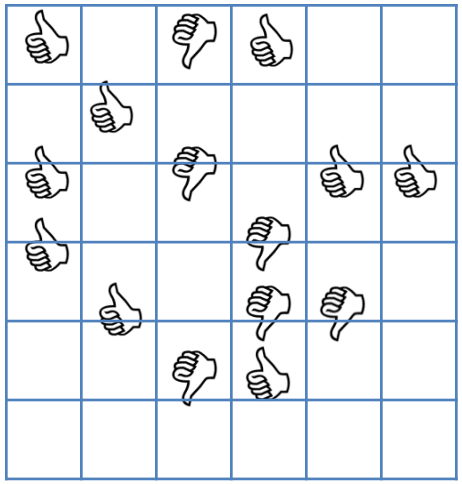


Credit: Arvind Narayanan via Adam Smith

A Source of Auxiliary Information

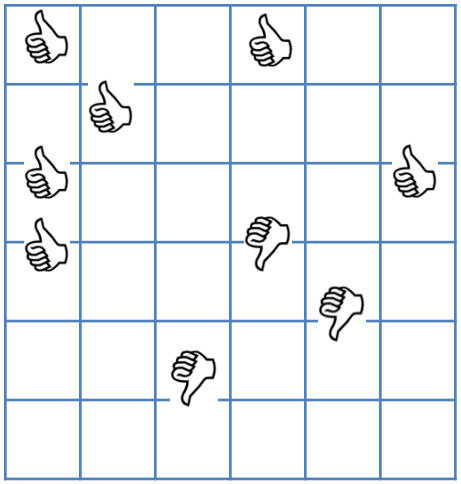
- Internet Movie Database (IMDb)
  - Individuals may register for an account and rate movies
  - **Need not be anonymous**
    - **Probably want to create some web presence**
  - Visible material includes ratings, **dates**, comments

Use Public Reviews from IMDb.com



Anonymized
NetFlix data

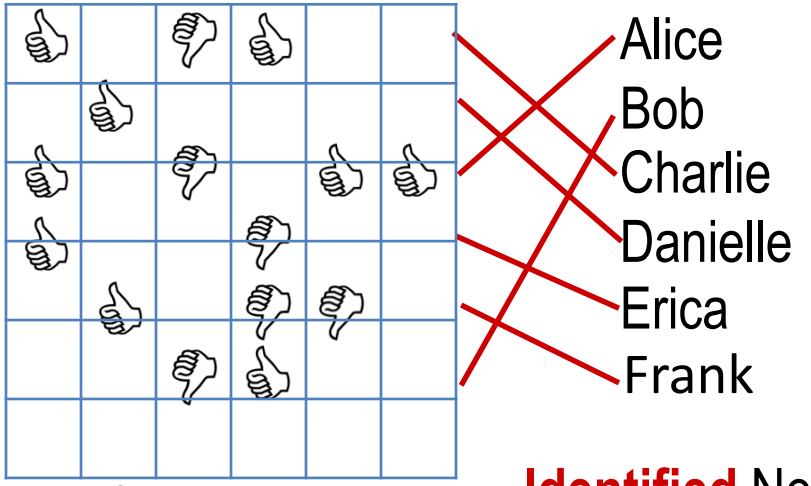
+



Alice
Bob
Charlie
Danielle
Erica
Frank

Public, incomplete
IMDB data

=



Identified Netflix Data

Credit: Arvind Narayanan via Adam Smith

De-anonymizing the Netflix Dataset

Results

of which 2 may be completely wrong

- "With 8 movie ratings and **dates** that may have a 3-day error, 96% of Netflix subscribers whose records have been released can be uniquely identified in the dataset."
- "For 89%, 2 ratings and **dates** are enough to reduce the set of plausible records to 8 out of almost 500,000, which can then be inspected by a human for further deanonymization."

Consequences?

Settled, March 2010

- Learn about movies that IMDB users didn't want to tell the world about...

  Sexual orientation, religious beliefs

- **Subject of lawsuits**

US Video Privacy Protection Act 1988

Credit: Arvind Narayanan via Adam Smith

Perfect Privacy?

Why not **"Semantic Security"**?
[a la Goldwasser Micali]

Anything that can be learned about a participant **from sanitized data**, can be **learned without it**
[Dalenius77]

Unachievable: **Auxiliary information** is a problem
[Dwork Naor]

Common theme in privacy horror stories

A “New” Approach to Privacy

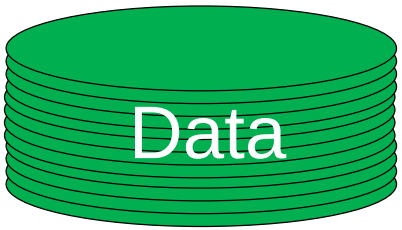
Differential Privacy [DMNS06]

Any outcome is **equally likely** when I’m
in the database or **out of the database**

Risk incurred **by participation** is low




Learning Can Hurt



q_1

a_1

q_2

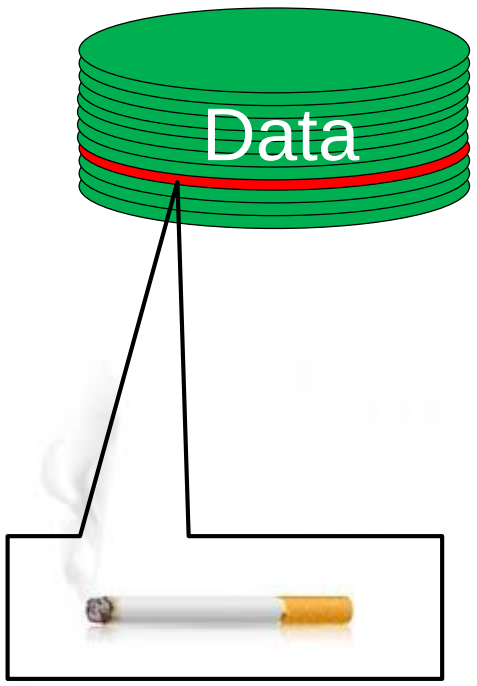
a_2

⋮

Data Analyst

SURGEON GENERAL'S WARNING: Quitting Smoking Now Greatly Reduces Serious Risks to Your Health.

Teachings vs. Participation



q_1

a_1

q_2

a_2

⋮

Data Analyst

SURGEON GENERAL'S WARNING: Quitting Smoking Now Greatly Reduces Serious Risks to Your Health.

Differential Privacy

Dwork, McSherry
Nissim & Smith
2006

Any outcome is equally likely when I'm
in the database or out of the database

Algorithm A guarantees ε-differential privacy

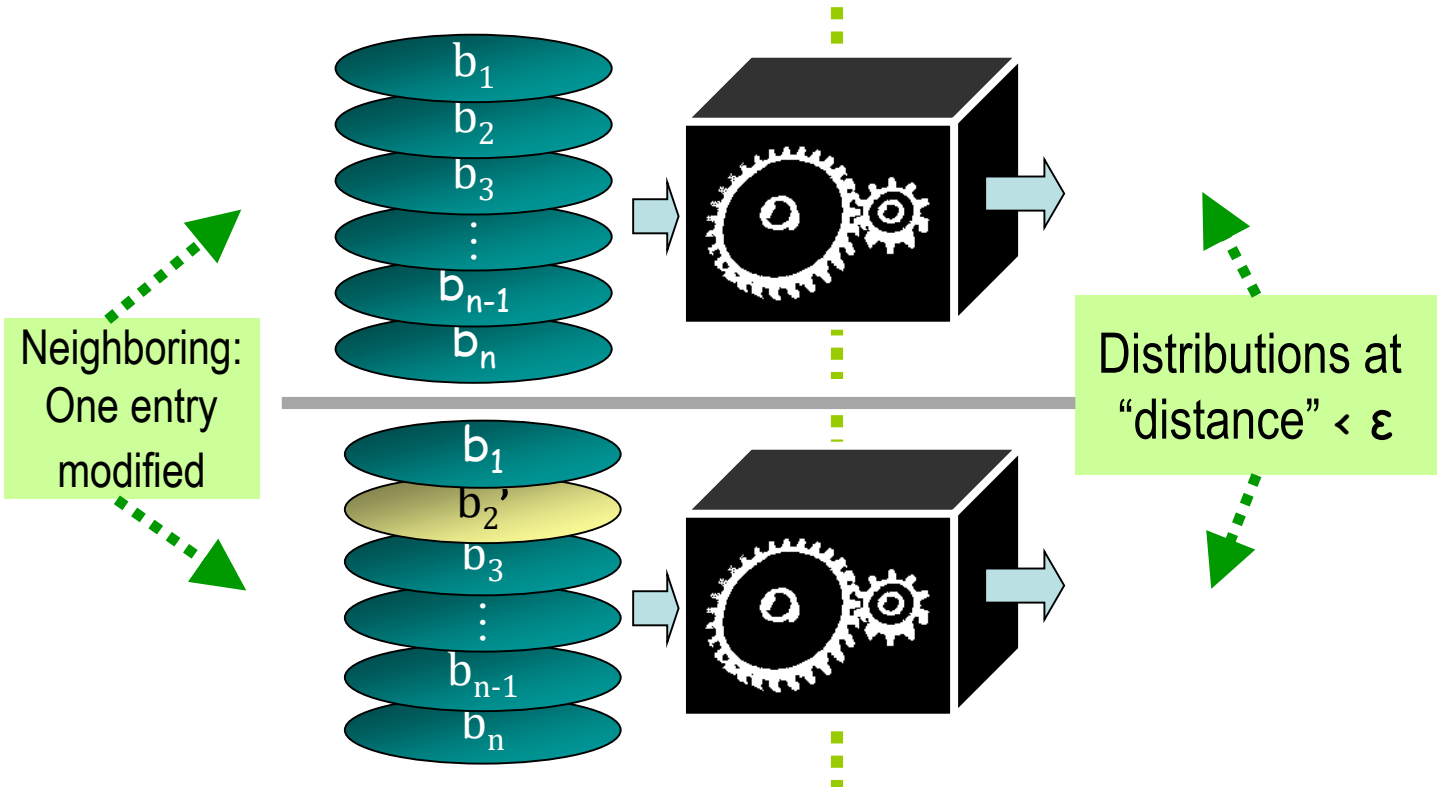
if for all DBs D and all events S:

$$Pr_A[A(D + me) \in S] \leq e^{\varepsilon} \cdot Pr_A[A(D - me) \in S]$$

$1 + \varepsilon$

Randomness introduced by A

Differential Privacy



Slide credit: Kobbi Nissim

Differential Privacy

Dwork, McSherry
Nissim & Smith
2006

Any outcome is equally likely when I'm
in the database or out of the database

Algorithm A guarantees ε-differential privacy

if for all DBs D and all events S:

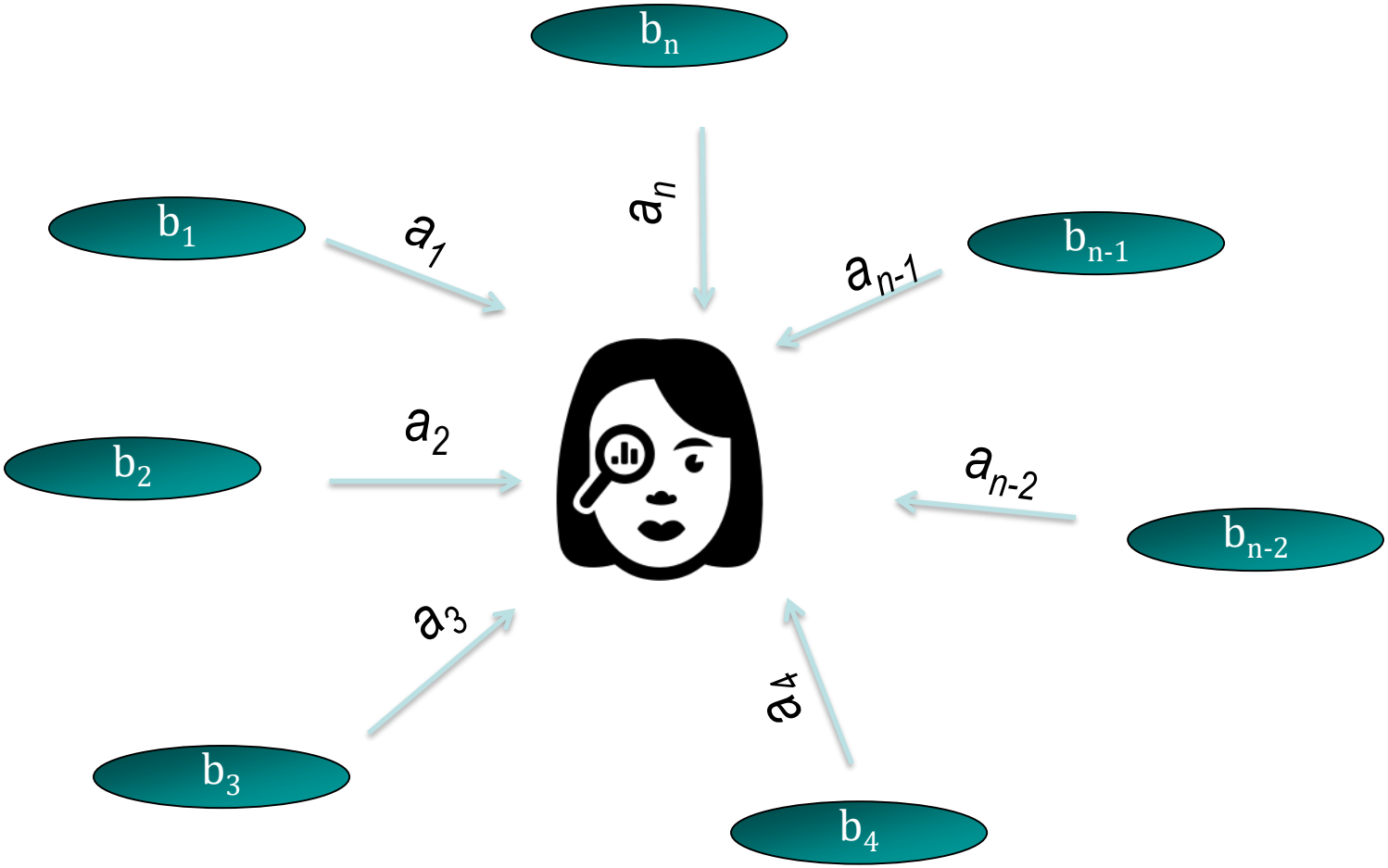
$$Pr_A[A(D + me) \in S] \leq e^{\varepsilon} \cdot Pr_A[A(D - me) \in S] + \delta$$

Randomness introduced by A

(ε, δ)-differential privacy

Dwork, Kenthapady, McSherry, Mironov and Naor, 2007

Local Model



Differential Privacy is a Success

- **Algorithms in many setting and for many tasks**

Important Properties:

Programmable!

- **Group privacy**: εk privacy for a group of size k
- **Composability**
  - Applying the sanitization several time: graceful degradation
  - proportional to number of applications
  - **even prop. to squareroot of number of applications**.
- Robustness to **side information**

Hard to quantify

  - No need to specify **exactly** what the adversary knows
  - Postprocessing

Differential Privacy: A Tutorial

- Basic composition
  Answering small numbers of queries

- Advanced composition
  Answering moderate numbers of queries

- Coordinated mechanisms
  Answering huge number of queries

- Example of Mixing MPC and DP for passwords

Composition

Privacy maintained even under **multiple analyses**

Core issue

The key to differential privacy's success!

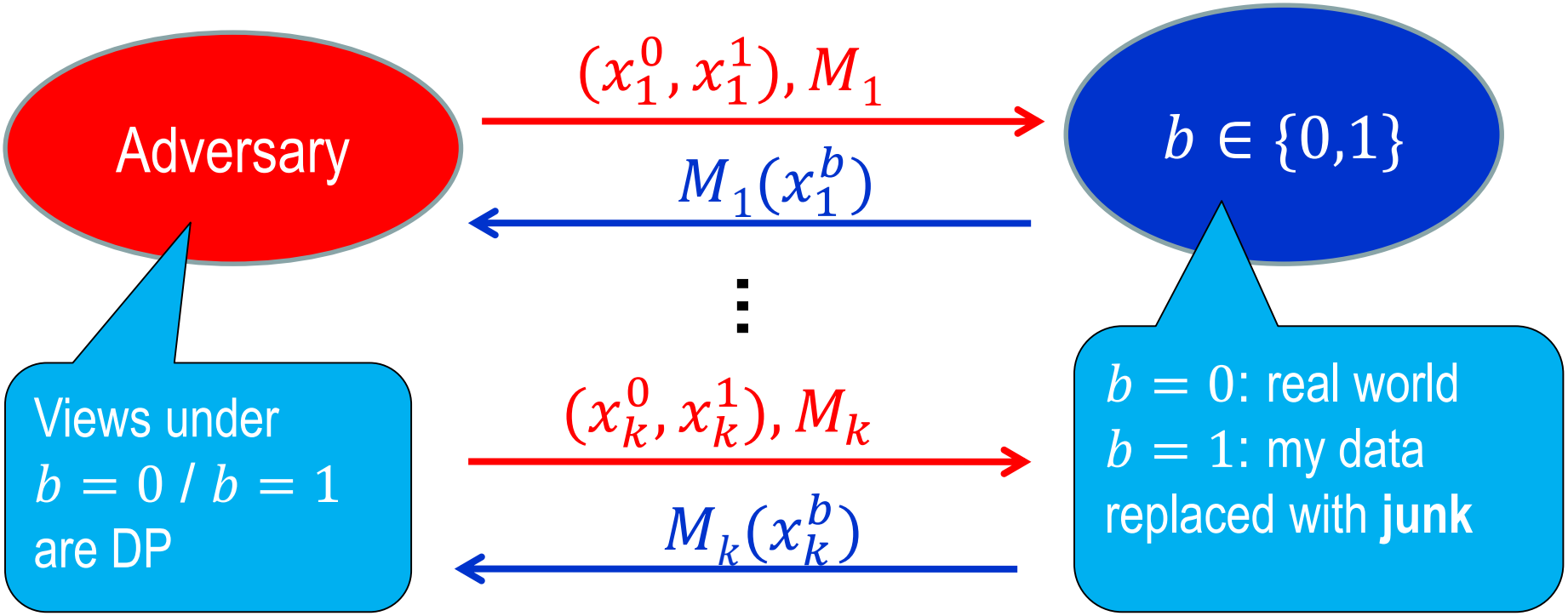
- Unavoidable
  - In reality, there are multiple analyses
- Makes DP "programmable"
  - Private subroutines make for private algorithms

Composition

Privacy maintained even under **multiple analyses**

How do we define it? [DworkRothblumVadhan10]

- Adaptive, adversarial DBs and algorithms



Basic Composition

- k (adaptively chosen) algorithms, each ε_0-DP: taken together still $k \cdot \varepsilon_0$-DP

Application: answering multiple queries

Basic Composition Proof

Define: $M_{1,2}(x) = \left(M_1(x), M_2(x)\right)$

$$\frac{\Pr[M_{1,2}(x)=(z_1,z_2)]}{\Pr[M_{1,2}(y)=(z_1,z_2)]} = \frac{\Pr[M_1(x)=z_1]\Pr[M_2(x)=z_2]}{\Pr[M_1(y)=z_1]\Pr[M_2(y)=z_2]} \leq e^{\varepsilon_1} e^{\varepsilon_2}$$

Property of the definition

- **Independent of the implementation**
- What about the adaptive case?

Statistical queries

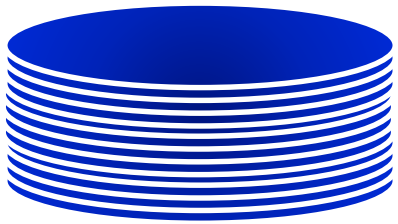
$q(D)$ = *"how many in D satisfy predicate P?"*

P is a Boolean predicate on universe U

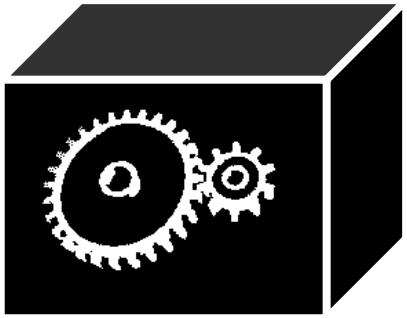
statistical queries allow powerful data analyses

- Perceptron, ID3 decision trees, PCA/SVM, k-means [BlumDworkMcSherryNissim05]
- any SQ-learning algorithm [Kearns98]
  - includes "most" known PAC-learning algorithms

Data Analysis Model



Database D
= multi-set
over universe U



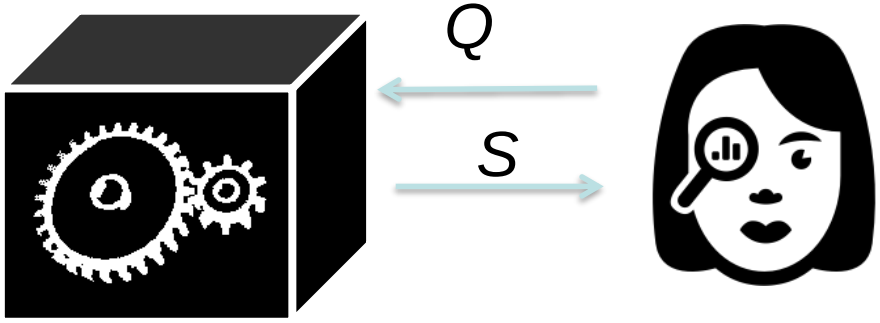
Trusted Curator

query set Q

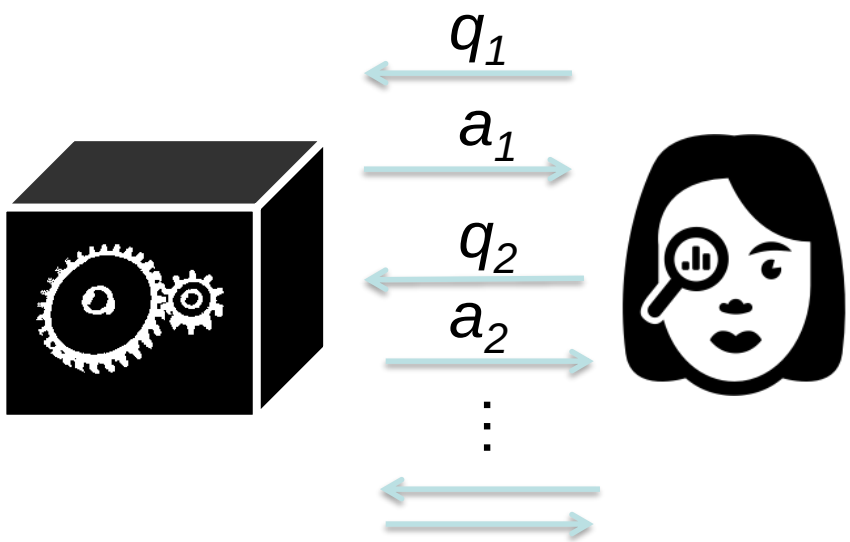
privacy-preserving
synopsis S
accurate on Q

Untrusted Analyst

Offline: non-interactive



Online: interactive



Answering a single counting query

U is set of tuples: $(name, tag \in \{0,1\})$

Counting query: # of participants with $tag = 1$

A: output **# of 1's + noise**

Differentially private! For proper noise

Choose noise from **Laplace distribution**

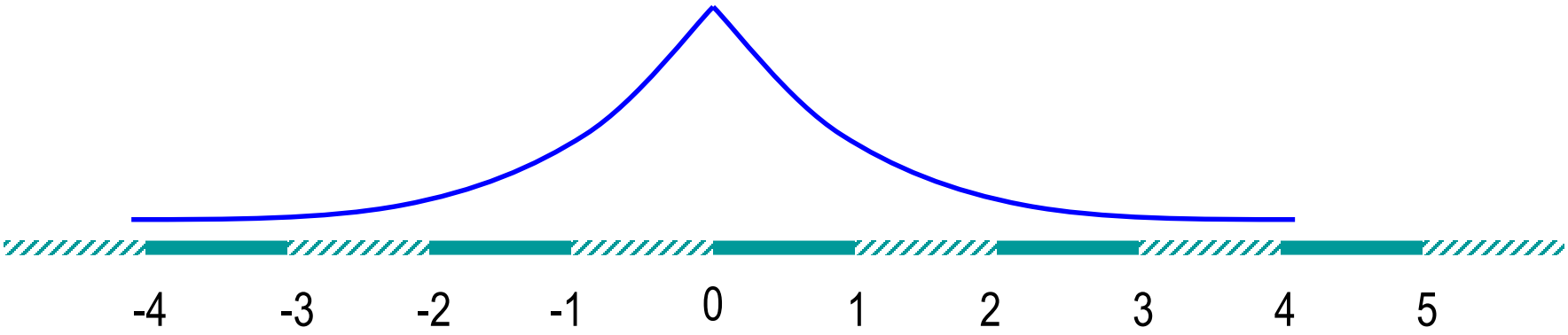
Laplacian Noise

Laplace distribution $Y = Lap(b)$ density function

$$\Pr[Y = y] = \frac{1}{2b} e^{-|y|/b}$$

Standard deviation: $O(b)$

Set $b = 1/\varepsilon$, get that $\Pr[Y = y] \propto e^{-\varepsilon \cdot |y|}$



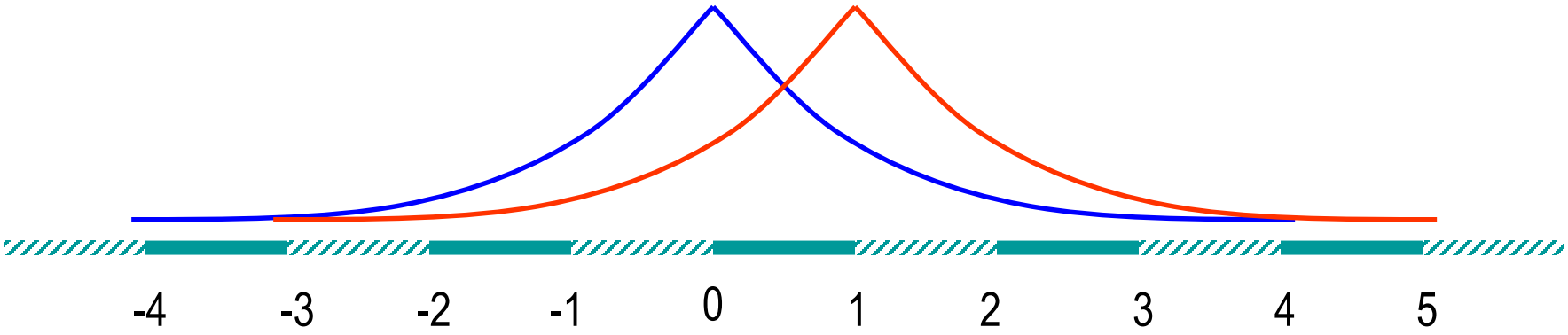
Laplacian Noise: ε-Privacy

Take $b = 1/\varepsilon$, get that $\Pr[Y = y] \propto e^{-\varepsilon \cdot |y|}$

Release: $q(D) + Lap(1/\varepsilon)$

For adjacent $\boldsymbol{D}$, $\boldsymbol{D'}$: $|\boldsymbol{q(D)} - \boldsymbol{q(D')}| \leq 1$

For any $\boldsymbol{z}$: $e^{-\boldsymbol{\varepsilon}} \leq \boldsymbol{Pr_{by\ D}[z]/Pr_{by\ D'}[z]} \leq e^{\varepsilon}$

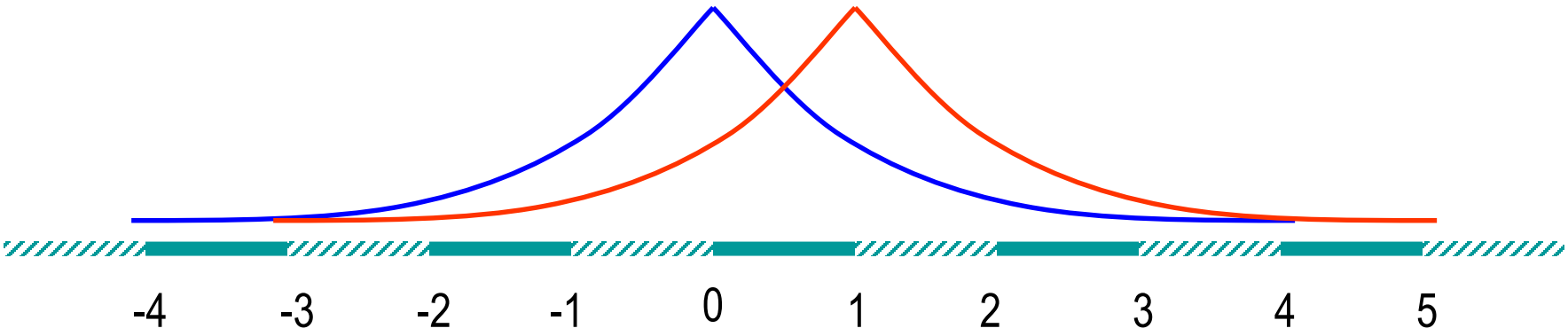


Laplacian Noise: $\tilde{O}(1/\varepsilon)$-Error

Take $b = 1/\varepsilon$, get that $\Pr[Y = y] \propto e^{-\varepsilon \cdot |y|}$

$$Pr_{y \sim Y}[|y| > k \cdot 1/\varepsilon] = O(e^{-k})$$

Expected error is $1/\varepsilon$, w.h.p error is $\tilde{O}(1/\varepsilon)$



Scaling Noise to Sensitivity [DMNS06]

Global **sensitivity** of query $q: Un \rightarrow [0, n]$

$GS_q = max_{D, D'} |q(D) - q(D')|$

For a counting query q: $GS_q = 1$

Previous argument generalizes:

For query q, release $q(D) + Lap(GS_q/\varepsilon)$

- ε-private
- error $\tilde{O}(GS_q/\varepsilon)$

Answering k Queries: Basic Composition

Answer k queries, each with sensitivity 1

- Use Laplace with $\varepsilon_0 = \varepsilon/k$ **privacy per query**

Better privacy, more noise per query ($\sim \mathrm{Lap}(k/\varepsilon)$)

- Composition: ε-privacy for all k answers

Error (roughly) **linear in number of queries**

- E.g.: can answer $\sqrt{n}$ queries with $\tilde{O}(\sqrt{n})$ error

Differential Privacy: A Tutorial

- Basic composition
  Answering small numbers of queries

- Advanced composition
  Answering moderate numbers of queries

- Coordinated mechanisms
  Answering huge number of queries

- Example of Mixing MPC and DP for passwords

Advanced Composition [DRV10]

Composing k algorithms, each ε_0-DP:

$$\varepsilon_g = O\left(\sqrt{k \cdot \ln\frac{1}{\delta_g}} \cdot \varepsilon_0 + k \cdot \varepsilon_0^2\right)$$

with all but δ_g probability.

Simultaneously

Compare with: $\varepsilon_g = k \cdot \varepsilon_0$ (basic composition)

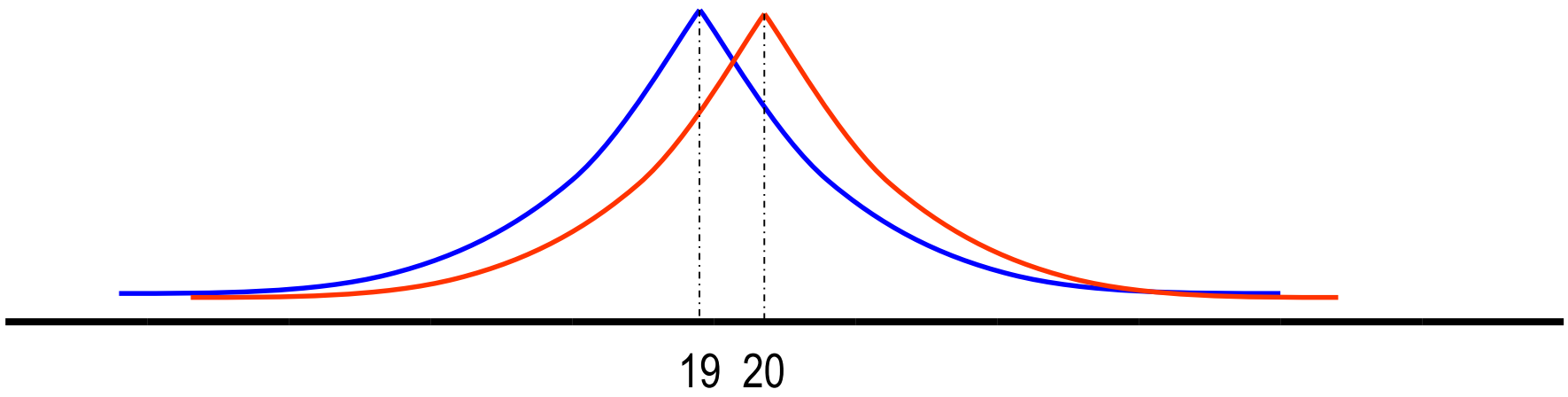
(think of $k < 1/\varepsilon_0^2$)

Privacy Loss

Fix adjacent D, D', draw $y \leftarrow M(D)$

$$PrivacyLoss(y) = \ln\left[\frac{\Pr[M(D) = y]}{\Pr[M(D') = y]}\right]$$

Can be positive, negative (or infinite)



Privacy Loss

Fix adjacent D, D', draw $y \leftarrow M(D)$

$$PrivacyLoss(y) = \ln\left[\frac{\Pr[M(D) = y]}{\Pr[M(D') = y]}\right]$$

- random variable, has a mean
- $(\varepsilon, 0) - DP$: w.p. 1 over y, $|PrivacyLoss(C)| \leq \varepsilon$
- $(\varepsilon, \delta) - DP^*$: w.p. $1 - \delta$ over y, $|PrivacyLoss(C)| \leq \varepsilon$

Advanced Composition [DRV10]

Composing k algorithms, each ε_0-DP:

$\varepsilon_g =$

with all but

Fundamental law of information recovery [DN03]:
Must have error $\Omega(\sqrt{n})$

For all δ_g simultaneously

- Better comp
- Answer n queries, error $\tilde{O}(\sqrt{n \cdot \ln(1/\delta_g)})$
  - independent Laplace noise
- Will see: Answer k queries, error $\tilde{O}(\sqrt{\log k \cdot n \cdot \ln(1/\delta_g)})$
  - coordinated noise - Private Multiplicative Weights [HR10]

Advanced Composition Proof

If M is DP, then privacy loss RV has:

- $E[PrivacyLoss(C)] = O(\varepsilon^2)$ (down to $\varepsilon^2/2$ [DR15])
- $|PrivacyLoss(C)| \leq \varepsilon$

Model **cumulative loss** from $M_1 \ldots M_k$ as **Martingale**

$$\Pr\left[\left(\sum_{i=1}^{k} Loss(C_i)\right) > k\varepsilon^2 + \sqrt{k}\varepsilon \cdot t\right] \leq \exp(-t^2/2)$$

Advanced Composition of (ε, δ)

Composing k algorithms, each $(\varepsilon_0, \delta_0)$-DP:

$$\varepsilon_g = O\left(\sqrt{k \cdot \ln \frac{1}{\delta_{err}}} \cdot \varepsilon_0 + k \cdot \varepsilon_0^2\right)$$

with all but $\delta_g = \delta_{err} + k \cdot \delta_0$ probability.

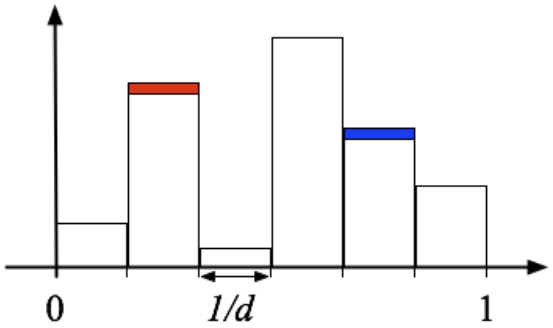
Generally: δ's add up

Do Better for Some Query Sets?

Use sensitivity of answer **vector** [DMNS06]

Example:

- Histograms, divide U into d **disjoint** bins $S_1 \ldots S_d$
  d queries: q_i counts #users in set S_i
- For adjacent D, D', only two answers can change, and each can change by 1
- Global sensitivity of answer vector is 2
- Add only $Lap(2/\varepsilon)$ noise to each query, still get ε-privacy



Further Work: Not Today

Some queries have high global sensitivity, but (usually) low local sensitivity

- Example: Median

Want noise ~ local sensitivity

Problem: local sensitivity is itself sensitive!

Smooth sensitivity [Nissim Raskhodnikova Smith 08]
Compute "smoothed" diffP upper-bound on **local sensitivity**

Propose Test Release [Dwork-Lei 09]
diffP test of local sensitivity, fail if too high

Differential Privacy: A Tutorial

- Basic composition
  Answering small numbers of queries

- Advanced composition
  Answering moderate numbers of queries

- Coordinated mechanisms
  Answering huge number of queries

k-fold **composition** of ε_0-differential privacy?

Answer 1
[DMNS06]

$\varepsilon_0 k$-differential privacy

Answer 2
[DRV10]

$\varepsilon_0 \sqrt{k}$ -differential privacy

Note: for small enough ε_0

can (privately) answer n queries with error $\sqrt{n}$

Do better for general queries?

Negative Results

Cannot answer n **counting queries** with error $o(\sqrt{n})$
[DiNi03, DwMcTa07,DwYe08]

In all these cases: **strong privacy violation**

What **can** we do?

almost entire DB compromised

Many Queries Question

Can we achieve error $\sqrt{n}$ for $k >> n$ queries

- with **any** reasonable notion of privacy

Yes!

Can answer **huge** numbers of queries with small error and differential privacy

[BLR08,DNRRV09,DRV10,RR10,HR10]

The Exponential Mechanism

Sometimes adding noise makes no sense

e.g. **output is not a number:**

- **minimum cut in a graph**
- decision tree classifier

[McSherry-Talwar 2007] Motivation: auction design

- DP implies approximate truthfulness

Subsequently applied broadly and successfully

- Can phrase any DP Mechanism as an instance of EM

The Exponential Mechanism

Input x output y artbitrary

Define for any possible input x and output y some measure of **utility** $u(y,x)$ - a real number

- The larger $u(y,x)$ the better the result
- Adjacent databases should have similar scores
  - $\Delta = \max_{x,x'y} |u(x,y) - u(x,y')|$ small

The mechanism: on input x, output y w.p $\propto e^{\varepsilon u(y,x)/\Delta}$

Simple Example: Private Lunch Preferences

DB of n individuals, lunch options $\{1,2, \ldots, k\}$,

– each individual likes/dislikes each option (1 or 0)

Goal: output a lunch option that many like

Mechanism:

Output option j with probability $\propto e^{\varepsilon \cdot \ell(j)}$

Where $\ell(j) = \#$ who like j

Actual probability: $\dfrac{e^{\varepsilon \cdot \ell(j)}}{\Sigma_i e^{\varepsilon \cdot \ell(i)}}$

Normalizer

Private Lunch: 2ε-Privacy

For each option $j \in [k]$, $\ell(j) =$ # who like j

Mechanism:

Output j with probability $\dfrac{e^{\varepsilon \cdot \ell(j)}}{\Sigma_i e^{\varepsilon \cdot \ell(i)}}$

For adjacent DBs, $\forall j$: $\ell(j)$ can differ by 1

- $e^{\varepsilon \cdot \ell(j)}$ changes by $\leq e^{\varepsilon}$ factor
- $(\Sigma_i e^{\varepsilon \cdot \ell(i)})$ changes by $\leq e^{\varepsilon}$ factor

For every j, $\Pr[\text{output} = \text{j}]$ changes by $\leq e^{2\varepsilon}$

Private Lunch: $\tilde{O}(\log k/\varepsilon)$ –Utility

For each option $j \in [k]$, $\ell(j) =$ # who like j

Mechanism:

Output j with probability $\dfrac{e^{\varepsilon \cdot \ell(j)}}{\Sigma_i e^{\varepsilon \cdot \ell(i)}}$

- If $\ell(j_1) < \ell(j_2) - d$, prob. of j_1 is $e^{\varepsilon d}$ **times smaller**
- If $\ell(j) < \max_i \ell(i) - (\ln k + b)/\varepsilon$,

$$\Pr[output\ is\ j] \leq 1/(k \cdot exp(b))$$

- **Union bound** over all $j \in [k]$:

$$\Pr[answer < \max - (\ln k + b)/\varepsilon] \leq \exp(-b)$$

Recap

- Notion of **ε-differential privacy** and **(ε, δ)-differential privacy**
- **Composition**: basic and advanced
  - For k-fold **composition** of ε_0-dp: $\varepsilon_0 k$ and $\varepsilon_0 \sqrt{k}$ respectively
  - Laplace mechanism and answering low sensitivity queries
    - The size of database as bound on number of queries
- The exponential mechanism

The Exponential Mechanism

Input x output y artbitrary

Define for any possible input x and output y some measure of **utility** $u(y,x)$ - a real number

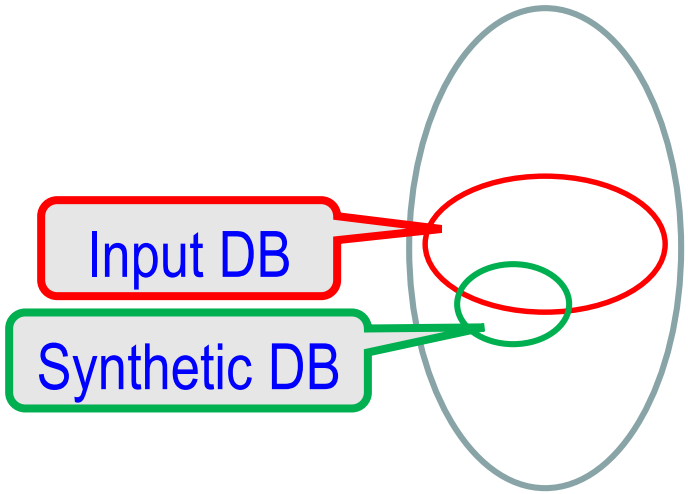
- The larger $u(y,x)$ the better the result
- **Adjacent** databases should have similar scores
  - $\Delta = \max_{x,x'y} |u(x,y) - u(x,y')|$ small

The mechanism: on input x, output y w.p $\propto e^{\varepsilon u(y,x)/\Delta}$

Answering Many Queries [BlLiRo08]

Answer **any set** Q of counting queries with diffP

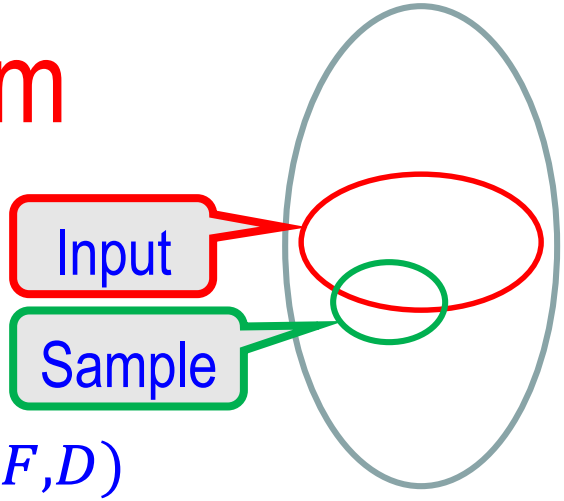
- Error is $\tilde{O}(n^{2/3} log^{1/3} |Q|)$
  - uses exponential mechanism
- algorithm outputs **synthetic DB**
  - Output is a (small) DB itself



Hope for rich private analysis of small DBs!

$$\#queries >> DB\ size$$

The BLR Algorithm



Input DB D of size n, query set Q:

Sample DB of size m: ($m < n$)

DB F gets picked w.p. $e^{-\varepsilon \cdot dist(F,D)}$

Sample

Input

Utility

For DBs F and D

$$dist(F,D) = \max_{q \in Q} |q(F) - q(D)|$$

Intuition: far DBs get smaller probability

Two samples that are **approximately the same** on Q get the **same weight**

The BLR Algorithm: 2ε-Privacy

Input DB D of size n, query set Q:

Sample DB of size m: ($m < n$)

DB F gets picked w.p. $e^{-\varepsilon \cdot dist(F,D)}$

For adjacent D, D' for every F

$$|dist(F,D) - dist(F,D')| \leq 1$$

- Probability of F by D:
  $e^{-\varepsilon \cdot dist(F,D)} / \Sigma_{G\ of\ size\ m} e^{-\varepsilon \cdot dist(G,D)}$
- Probability of F by D':
  - **numerator** and **denominator** can change by e^{ε}

2ε-differential privacy

BLR Algorithm: Error

Input DB D of size n, query set Q:

Sample DB of size m: ($m < n$)

DB F gets picked w.p. $e^{-\varepsilon \cdot dist(F,D)}$

Fix desired error α, $m = \tilde{O}((n/\alpha)^2 \cdot \log|Q|)$

- $\exists F^*$ of size m with $dist(F^*, D) \leq \alpha$

  $Pr[F^*] \propto e^{-\varepsilon\alpha}$

A random sample is good whp!

- $\forall F\ bad$ with dist 2α, $\Pr[F_{bad}] \propto e^{-2\varepsilon\alpha}$
- $\Sigma_{F bad} Pr[F_{bad}] \propto |U|^m \cdot e^{-2\varepsilon\alpha}$

Take $\alpha = \tilde{O}(n^{2/3} \cdot \log^{1/3}|Q|)$,

$$Pr[F_{good}] >> \sum Pr[F_{bad}]$$

BLR Algorithm: Running Time

Input DB D of size n, query set Q:

Sample DB of size m: ($m < n$)

DB F gets picked w.p. $e^{-\varepsilon \cdot dist(F,D)}$

Brute-force sampling:

Need to enumerate **every** size-m database,

where $m = \tilde{O}\left((n\backslash\alpha)^2 \cdot \log|Q|\right)$

Running time $\approx |U|^{\tilde{O}((n\backslash\alpha)^2 \cdot log\,|Q|)}$

BLR algorithm: Conclusion

Offline algorithm

- Error: $\tilde{O}(n^{2/3} \cdot \log^{1/3} |Q|/\varepsilon)$
- Running time: $|U|^{\tilde{O}(n^{2/3} \cdot \log^{1/3} |Q|\ /\varepsilon)}$

Private Multiplicative Weights Algorithm

Online algorithm

- Error $\tilde{O}(\sqrt{n \cdot \log |Q|}\ /\varepsilon)$
- Running time: $poly(|U|, |Q|, n)$

Key Insight to increasing number of queries: Use **Coordinated** Noise

- If noise is added in with careful coordination,

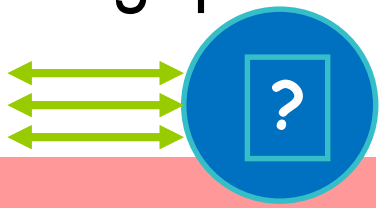
  rather than independently

  can answer **hugely** many queries

#queries >> DB size

Wave of results showing:

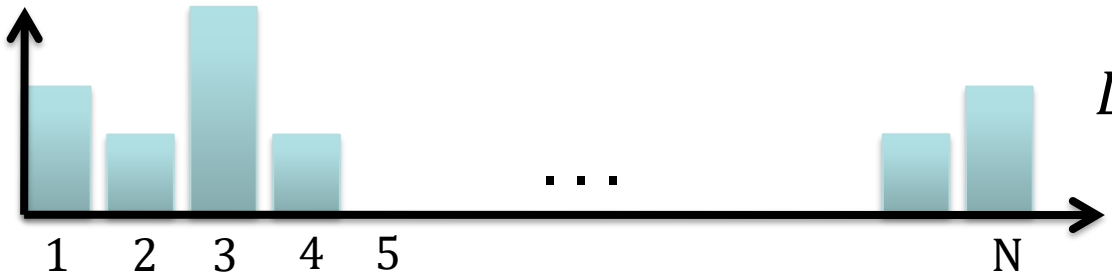
- Differential Privacy for **every** set **Q** of counting queries
- Error is $\tilde{O}(n^{1/2} \log|Q|)$



  - **Even in the *interactive* case –**
  - **Private Multiplicative Weights Algorithm**

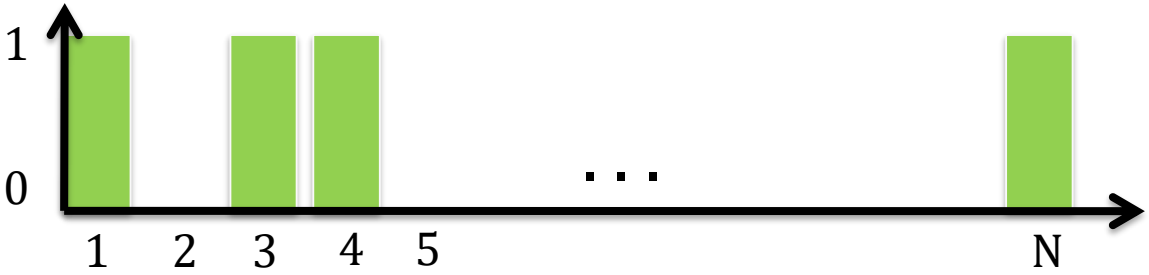
Represent database as a linear function

Data set D is *distribution* over universe $U, |U| = N$



$$D[i] = \frac{\#type\ i\ items\ in\ D}{n}$$

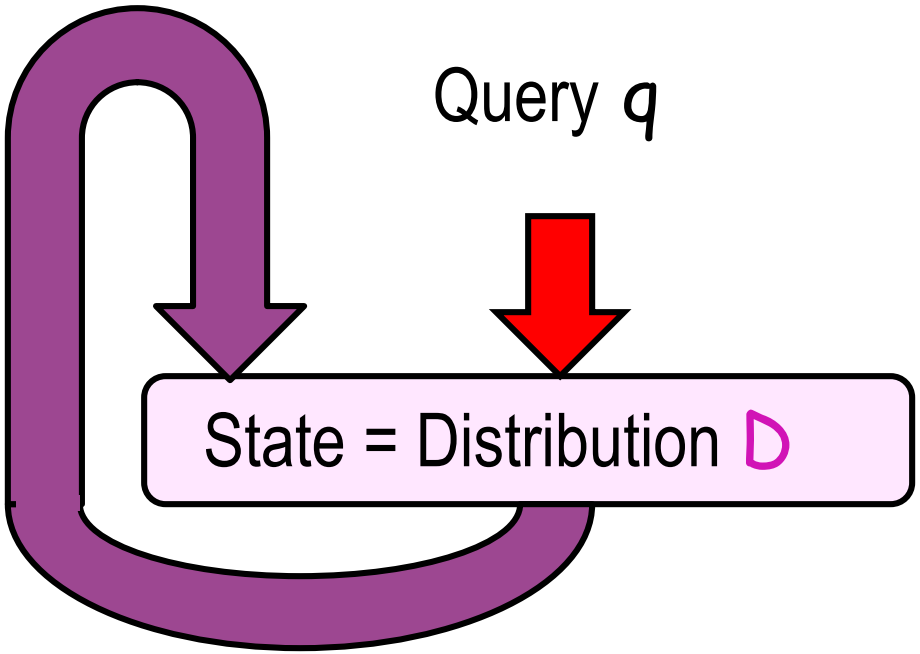
Statistical query q is vector in $\{0,1\}^N$



$$q(D) = < q, D > \cdot\, n$$

$$q(D) \in [0,1]$$

Maintaining State



Hardt & Rothblum 2010

The PMW Algorithm

Maintai on universe U

Initializ on U

Repeat

- Set μ)

This is the state.
Is completely public!

Algorithm fails if more than k updates

Multiplicative Weights

- Powerful tool in algorithms design
- Learn a Probability Distribution iteratively
- In each round:
  - **Test**: If $|q(D) - \hat{a}| \leq \hat{T}$ **output** $q(D)$.
  - **Else** (**update**):
    - **Output** $\hat{a}$
    - Update distribution
    - Update D[i] / D[i] $e^{\pm T/4q[i]}$ and re-weight.

Repeat until update occurs:

- Receive query $q \in Q$
- Let $\hat{a} = \times(q) + \mathrm{Lap}(\mu)$

The true value

the plus or minus are according to the sign of the error

either current distribution is good

or get a lot of information on distribution

Private Multiplicative Weights

Accuracy: nearly optimal (in worst case)

Privacy: differential privacy

Runtime: **linear dependence** on $|U|$

- $|U|$ exponential in # attributes of data

When can we get $poly(n)$?

Lower bounds based **on tracing traitors**

Differential Privacy: A Tutorial

- Basic composition
  Answering small numbers of queries
- Advanced composition
  Answering moderate numbers of queries
- Coordinated mechanisms
  Answering huge number of queries
- Example of Mixing MPC and DP for passwords

Applications/Implementations of Differential Privacy

- US Census Bureau **OnTheMap**: gives researchers access to agency data.
- Google's RAPPOR:
  - **R**andomized **A**ggregatable **P**rivacy-**P**reserving **O**rdinal **R**esponse
  - Open source
  - Local model
- Apple: big news coverage, commitment to privacy
- Applications to **multiple hypothesis testing**

Enabled collection of data **Chrome** avoided before

Global vs. local

How to hack Kaggle competitions

Public policy

- California Public Utilities Commission
  - Smart meters
- Interpreting and implementing FERPA by Differential Privacy

Family Educational Rights and Privacy Act

Nissim and Wood

Understand how DP **fits with existing regulatory framework**.
Problem: regulatory framework is not mathematically precise, Idea of de-identification is hard wired in it.

Challenges

- Small Datasets
- Massive Composition – global epsilon: event level vs user level
- Work in conjunction with Secure Function Evaluation
- Winning the hearts and minds of policy makers…

Winning the hearts and minds of policy makers…

- Widen scope of implementation and use.
- Identify what are the next good use cases for DP.
  - Construct DP tools matching best the practices and education of users
  - Explain shortcomings of other methods and benefits of DP
  - Need to figure how DP works as one of the layers in a suit of privacy protections.
  - Less straightforward and intuitive than anonymity/de-identification and its variants