Data Privacy

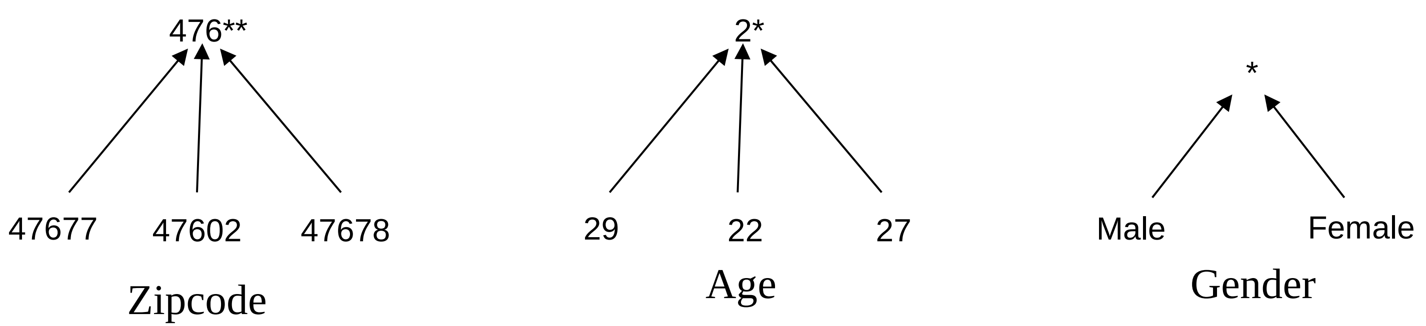
Tianhao Wang

Agenda

- Review
- Differential Privacy
- Local Differential Privacy

k-Anonymity [Sweeney, Samarati]

- Privacy is "protection from being **brought to the attention of others**."
- k-Anonymity
  - Each record is indistinguishable from ≥ k-1 other records when only "quasi-identifiers" are considered
  - These k records form an equivalence class
- To achieve k-Anonymity, uses
  - Generalization: Replace with less-specific values
  - Suppression: Remove outliers



k-Anonymity [Sweeney, Samarati]

The Microdata

QID			SA
Zipcode	Age	Gen	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

A 3-Anonymous Table

QID			SA
Zipcode	Age	Gen	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

- k-Anonymity
  - Each record is indistinguishable from $\geq$ k-1 other records when only "quasi-identifiers" are considered
  - These k records form an equivalence class

Attacks on *k*-Anonymity

- k-anonymity does not provide privacy if:
  - Sensitive values lack diversity
  - The attacker has background knowledge

Homogeneity Attack

Bob	
Zipcode	***Age***
47678	27

Background Knowledge Attack

Carl does not have heart disease

Carl	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

l –Diversity: [Machanavajjhala et al. 2006]

- Principle
  - Each equi-class contains at least l well-represented sensitive values
- Instantiation
  - Distinct l-diversity
    - Each equi-class contains l distinct sensitive values
  - Entropy l-diversity

$$H(X) = \mathrm{E}(I(X)) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

    - *entropy(equi-class)≥*$\log_2(l)$

The Skewness Attack on *l*-Diversity

- Two values for the sensitive attribute
  - HIV positive (1%) and HIV negative (99%)
- Highest diversity still has serious privacy risk
  - Consider an equi-class that contains an equal number of positive records and negative records.
- l-diversity does not differentiate:
  - Equi-class 1: 49 positive + 1 negative
  - Equi-class 2: 1 positive + 49 negative

l-diversity does not consider the overall distribution of sensitive values

The Similarity Attack on *l*-Diversity

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

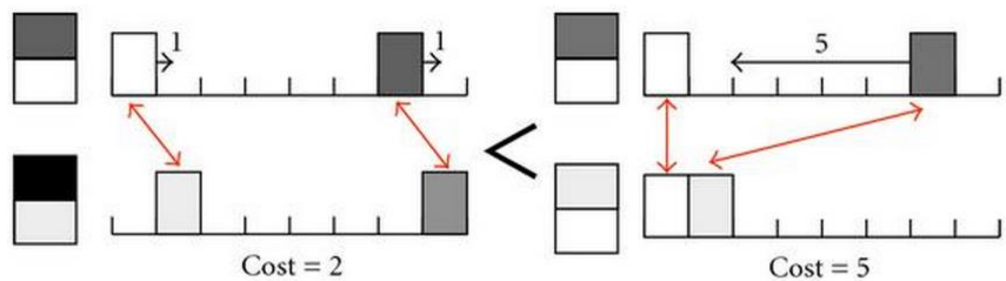
1. Bob's salary is in [20k,40k], which is relative low.
2. Bob has some stomach-related disease.

l-diversity does not consider semantic meanings of sensitive values

t-Closeness

- Principle: Distribution of sensitive attribute value in each equi-class should be close to that of the overall dataset (distance $\leq$ t)
  - Assuming that publishing a completely generalized table is always acceptable
  - We use Earth Mover Distance to capture semantic relationship among sensitive attribute values
- **(n,t)-closeness:** Distribution of sensitive attribute value in each equi-class should be close to that of some natural super-group consisting at least n tuples

N. Li, T. Li, S. Venkatasubramanian: t-Closeness: Privacy Beyond k-Anonymity and *l*-diversity. In ICDE 2007. Journal version in TKDE 2010.



From Syntactical Privacy Notions to Differential Privacy

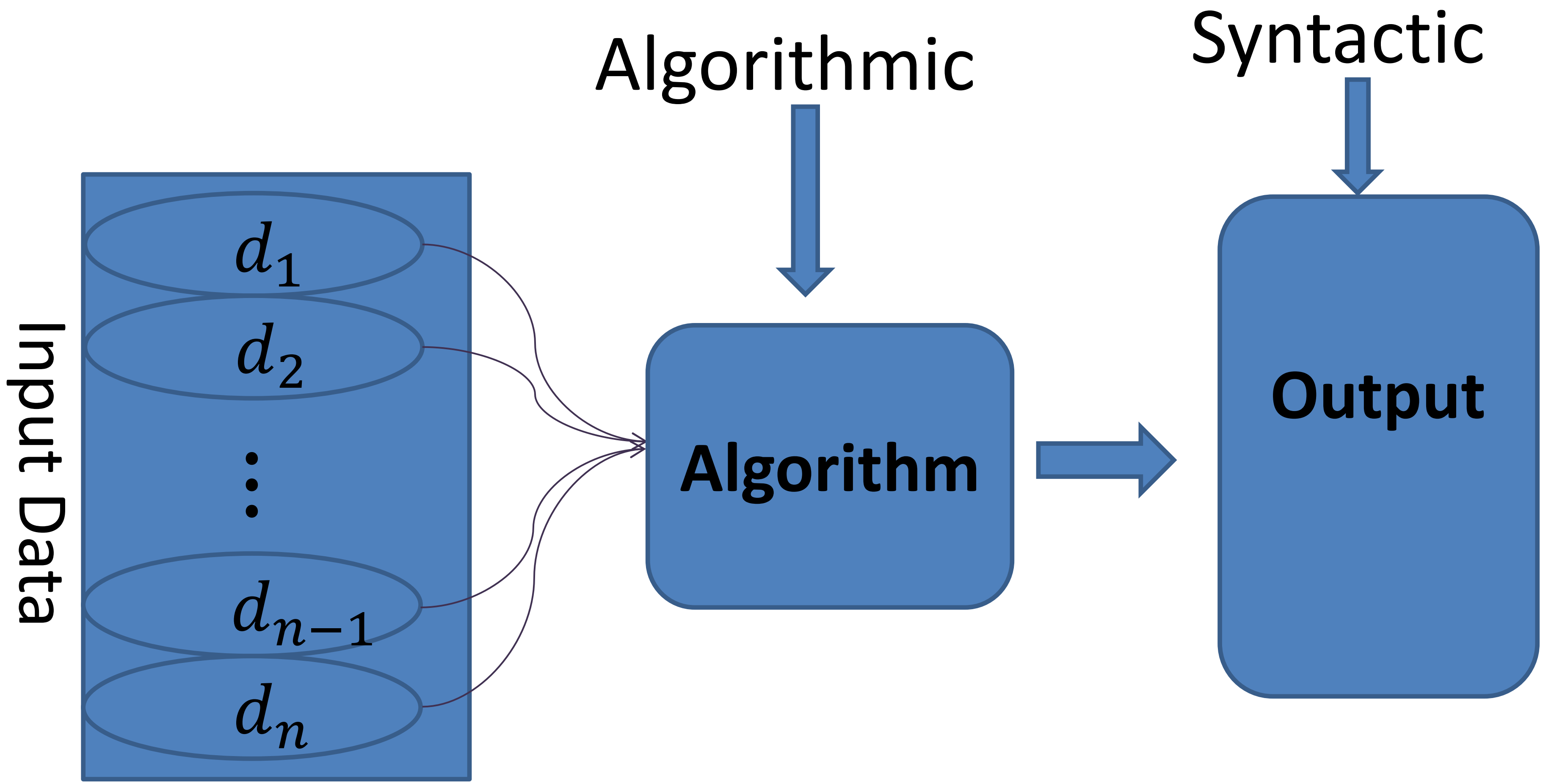
- Limitation of previous privacy notions:
  - Requires identifying which attributes are quasi-identifier or sensitive, not always possible
  - Difficult to pin down due to background knowledge
  - Syntactic in nature (property of anonymized dataset)
    - Not exhaustive in inference prevented
- Differential Privacy [Dwork et al. 2006]
  - Privacy is not violated if one's information is not included
  - Output does not overly depend on any single tuple

Definition (ε-Differential Privacy)

A randomized algorithm $\mathcal{A}$ satisfies ε-differential privacy, if for any pair of neighboring datasets D and D' and for any $O \subseteq \text{Range}(\mathcal{A})$:

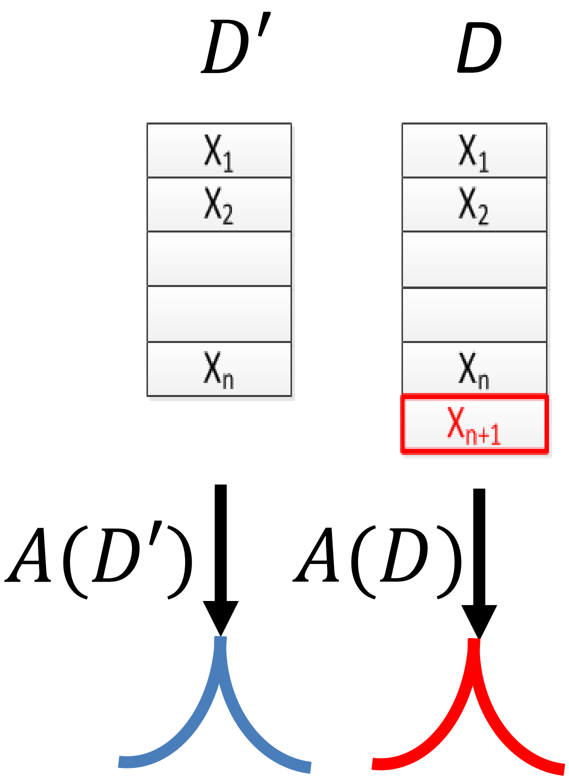
$$e^{-\varepsilon} Pr[\mathcal{A}(D') \in O] \leq Pr[\mathcal{A}(D) \in O] \leq e^{\varepsilon} Pr[\mathcal{A}(D') \in O]$$

Syntactic versus Algorithmic Privacy Notions



Differential Privacy [Dwork et al. 2006]

- Idea: Any output should be about as likely regardless of whether or not I am in the dataset



Algo A satisfies ϵ-differential privacy if for any possible output t,

$$e^{-\epsilon} \leq \frac{\Pr[A(D)=t]}{\Pr[A(D')=t]} \leq e^{\epsilon}$$

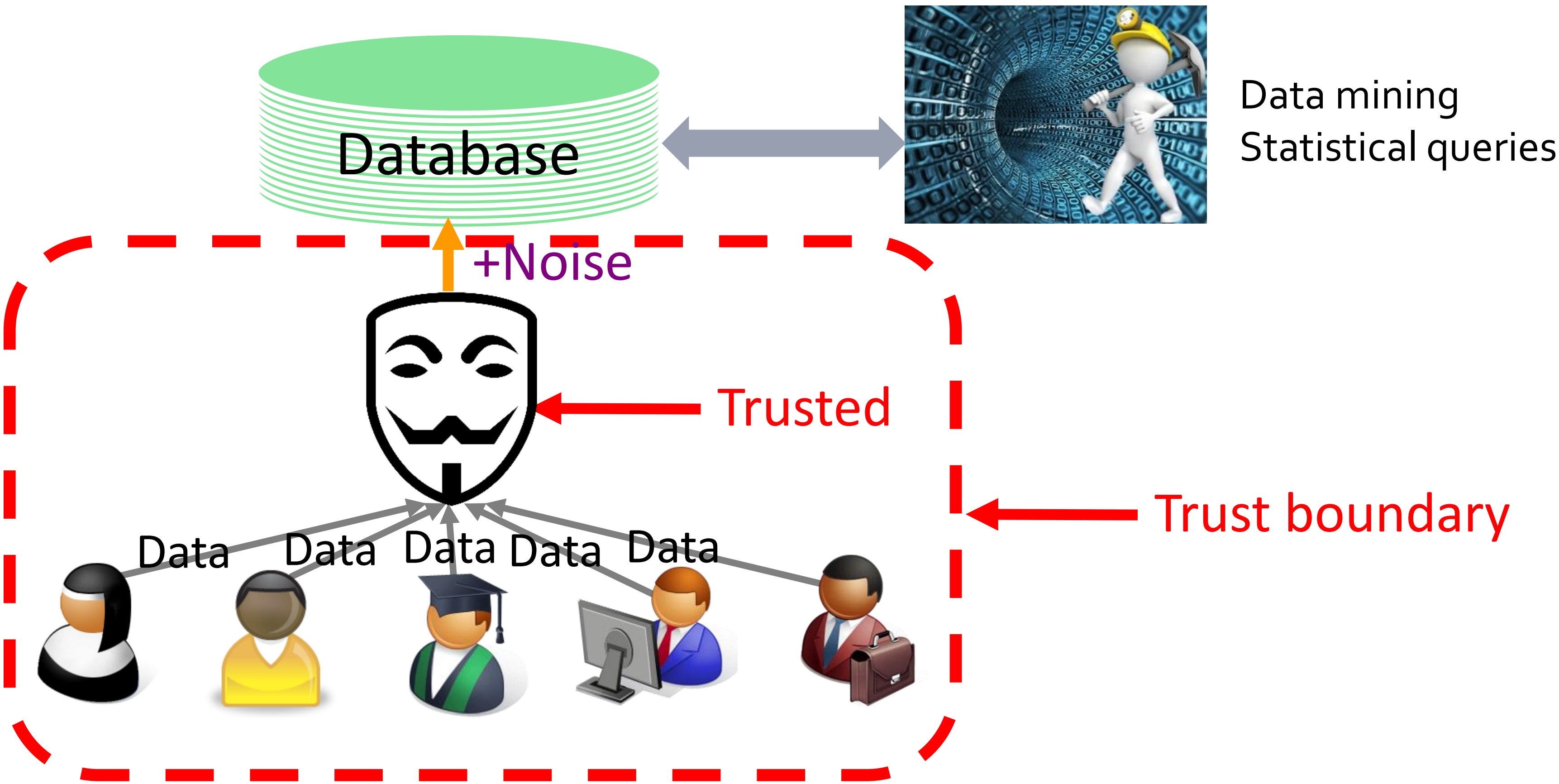
Parameter ϵ: strength of privacy protection, known as privacy budget.

Algorithm A must be randomized.

Key Assumption Behind DP: The Personal Data Principle

- After removing one individual's data, that individual's privacy is protected perfectly.
  - Even if correlation can still reveal individual info, that is not considered to be privacy violation
- In other words, for each individual, the world after removing the individual's data is an ideal world of privacy for that individual. Goal is to simulate all these ideal worlds.

Differential Privacy



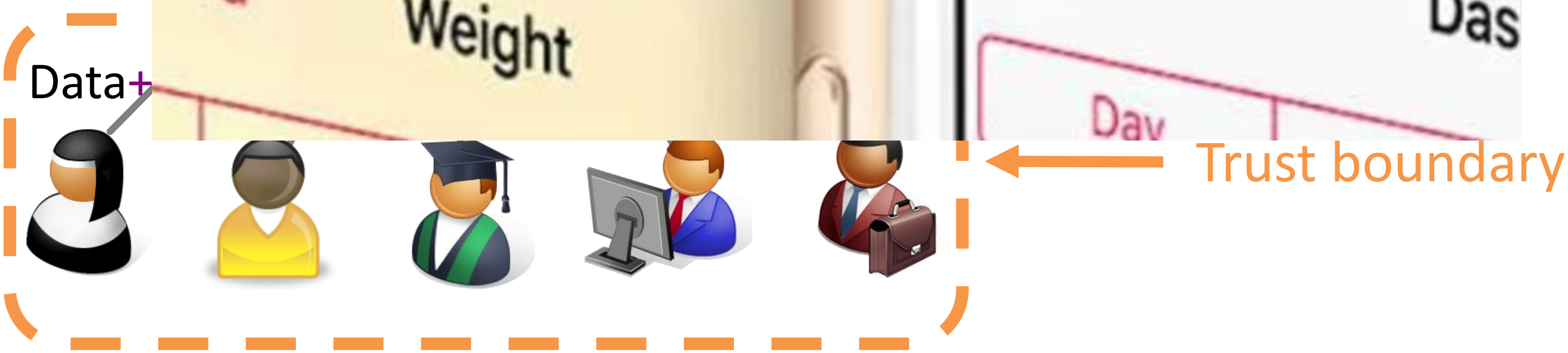
Local Differential Privacy

As Apple starts analyzing web browsing & health data, how comfortable are you with differential privacy?

RAPP

ng

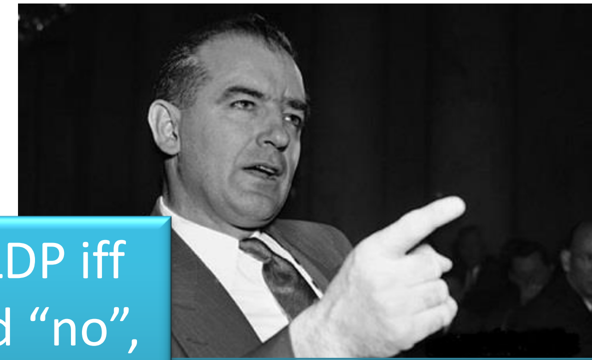
Ben Lovejoy - Jul. 7th 2017 6:59 am PT @benlovejoy



Mechanisms and Properties

- Random Response
  - Most used in the local setting
- Laplace
- Exponential
- Composition Theorem
  - Sequential composition
  - Parallel composition
  - Postprocessing
  - Advanced composition

The Warner Model (1965)



- Survey technique for private questions
- Survey people:
  - "Are you communist party?"
- Each person:
  - Flip a secre
  - Answer tru
  - Answer ra
  - E.g., a communist will answer "yes" w/p 75%, and "no" w/p 25%
- To get unbias
  - If n_v out o ee

    $E[\,I_v] =$
  - $c(n_v) = \frac{I_v - 0.25n}{0.5}$ is the unbiased estimation of number of communists
  - Since $E[c(n_v)] = \frac{E[I_v] - 0.25n}{0.5} = n_v$

tain about the secret.

We say a protocol satisfies ε -LDP iff for any v and v' from "yes" and "no",

$$\frac{\Pr[P(v) = v]}{\Pr[P(v') = v]} \le e^{\varepsilon}$$

This only handles binary attribute. We want to handle the more general setting.

Frequency Estimation Protocols

- Randomised response: a survey technique for eliminating evasive answer bias
  - S.L. Warner, Journal of Ame. Stat. Ass. 1965
  - Direct Encoding (Generalized Random Response)
- RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response.
  - Ú. Erlingsson, V. Pihur, A. Korolova, CCS 2014
  - Unary Encoding, Encode into a bit-vector
- Local, Private, Efficient Protocols for Succinct Histograms
  - R. Bassily, A. Smith. STOC 2015.
  - Binary Local Hash: Encode by hashing and then perturb
- Locally Differentially Private Protocols for Frequency Estimation
  - T. Wang, J. Blocki, N. Li, S. Jha: USENIX Security 2017

Direct Encoding (Random Response)

- User:
  - Encode $x = v$ (suppose v from $D = \{1,2,\ldots,d\}$)
  - Toss a coin with bias p
  - If it is head, report the true value $y = x$
  - Otherwise, report any other value with probability $q = \frac{1-p}{d-1}$ (uniformly at random)

Intuitively, the higher p, the more accurate

  - $p = \frac{\ldots}{e^{\varepsilon}+d-1}, q = \frac{\ldots}{e^{\varepsilon}+d-1} \to \frac{\ldots}{\Pr[P(v\prime)=v]} = \frac{\ldots}{q} = e^{\ldots}$

- Aggregator:
  - Suppose [illegible] reports on v.

However, when d is large, p becomes small

  - $E[I_v] = n_v \cdot p + (n - n_v) \cdot q$
  - Unbiased Estimation: $c(v) = \frac{I_v - n \cdot q}{p - q}$

Unary Encoding (Basic RAPPOR)

- Encode the value v into a bit string $\boldsymbol{x} \coloneqq \vec{0}, \boldsymbol{x}[v] \coloneqq 1$
  - e.g., $D = \{1,2,3,4\}, v = 3,$ then $\boldsymbol{x} = [0,0,1,0]$
- Perturb each bit, preserving it with probability p
  - $p_{1\to1} = p_{0\to0} = p = \frac{e^{\varepsilon/2}}{e^{\varepsilon/2}+1}$ $\qquad p_{1\to0} = p_{0\to1} = q = \frac{1}{e^{\varepsilon/2}+1}$
  - $\Rightarrow \frac{\Pr[P(E(v))=\boldsymbol{x}]}{\Pr[P(E(v'))=\boldsymbol{x}]} \leq \frac{p_{1\to1}}{p_{0\to1}} \times \frac{p_{0\to0}}{p_{1\to0}} = e^{\varepsilon}$
    - Since $\boldsymbol{x}$ is unary encoding of v, $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in two locations
- Intuition:
  - By unary encoding, each location can only be 0 or 1, effectively reducing d in each location to 2.
  - When d is large, UE is better than DE.
- To estimate frequency of each value, do it for each bit.

Binary Local Hash

- The protocol description in [Bassily-Smith ’15] is complicated
- This is an equivalent description
- Each user uses a random hash function from D to $\{0,1\}$
- The user then perturbs the bit with probabilities
  - $p = \frac{e^{\varepsilon}}{e^{\varepsilon}+g-1} = \frac{e^{\varepsilon}}{e^{\varepsilon}+1}, q = \frac{1}{e^{\varepsilon}+g-1} = \frac{1}{e^{\varepsilon}+1}$

$$\Rightarrow \frac{\Pr[P(E(\boldsymbol{v})) = b]}{\Pr[P(E(\boldsymbol{v}')) = b]} = \frac{p}{q} = e^{\varepsilon}$$

- The user then reports the bit and the hash function
- The aggregator increments the reported group
- $E[I_v] = n_v \cdot p + (n - n_v) \cdot (\frac{1}{2} q + \frac{1}{2} p)$
- Unbiased Estimation: $c(v) = \frac{I_v - n \cdot \frac{1}{2}}{p - \frac{1}{2}}$

Our Work

- We measure utility of a mechanism by its variance
  - E.g., in Random Response,
    - $Var[c(v)] = Var\left[\frac{I_v - n\cdot q}{p-q}\right] = \frac{Var[I_v]}{(p-q)^2} \approx \frac{n\cdot q\cdot(1-q)}{(p-q)^2}$
- We ... and cast exis... rk.
  - E ...
    - ... ach value with a corresponding 1
    - E.g., In BLH, Support(y) = $\{v \mid H(v) = y\}$
  - A pure protocol is specified by p' and q'
    - Each input is perturbed into a value "supporting it" with p', and into a value not supporting it with q'

$min_{q'} Var[c(v)]$

or $min_{q'} \frac{n\cdot q'\cdot(1-q')}{(p'-q)'^2}$

where p', q' satisfy ε-LDP

Optimized Unary Encoding (UE)

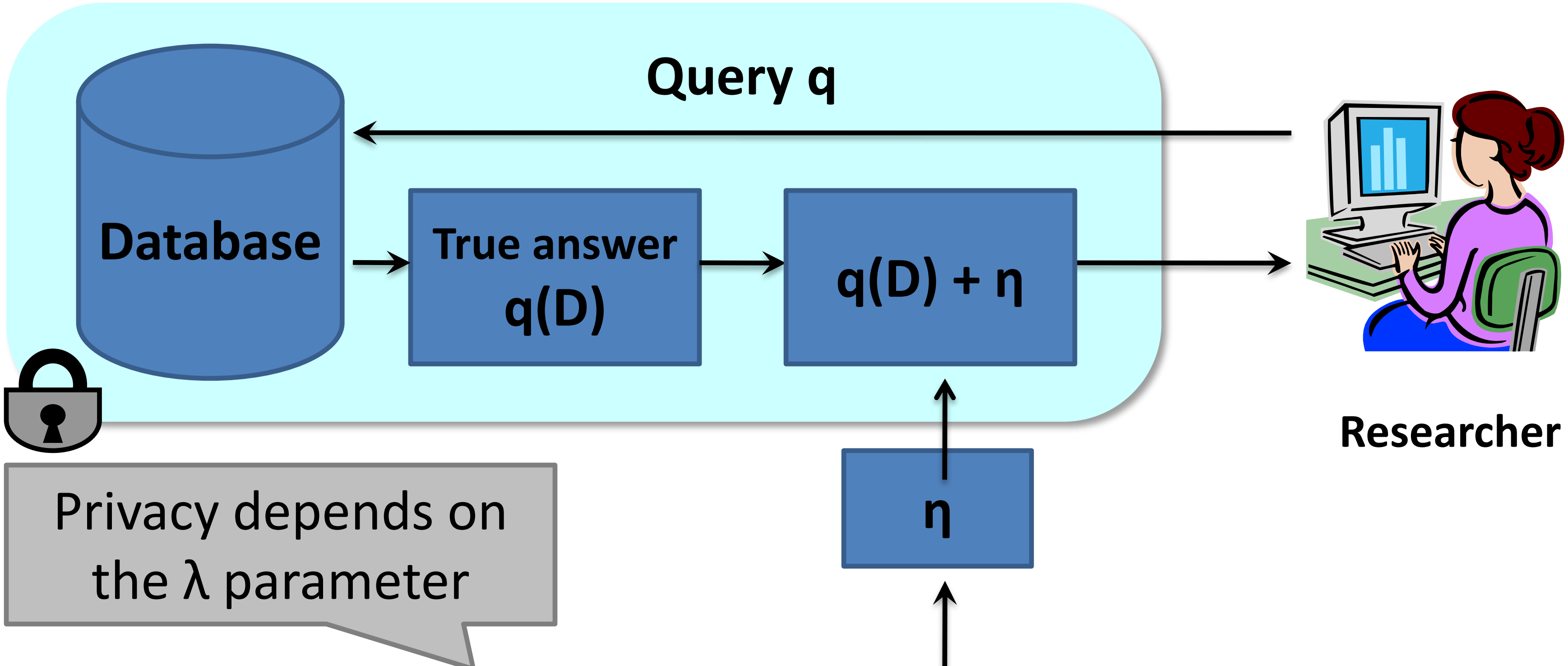
- In the original UE, 1 and 0 are treated symmetrically
  - $p_{1\to1} = p_{0\to0} = \frac{e^{\varepsilon/2}}{e^{\varepsilon/2}+1}, \quad p_{1\to0} = p_{0\to1} = \frac{1}{e^{\varepsilon/2}+1}$
- **Observation:** In the input, there are a lot more 0's than 1's when d is large.
- **Key Insight:** We can perturb 0 and 1 differently and should reduce $p_{0\to1}$ as much as possible
  - $p_{1\to1} = \frac{1}{2}, \quad p_{1\to0} = \frac{1}{2}$
  - $p_{0\to0} = \frac{e^{\varepsilon}}{e^{\varepsilon}+1}, \quad p_{0\to1} = \frac{1}{e^{\varepsilon}+1}$
    - $\frac{p_{1\to1}}{p_{0\to1}} \times \frac{p_{0\to0}}{p_{1\to0}} \leq e^{\epsilon}$

Optimized Local Hash (OLH)

- In original BLH, secret is **compressed** into a bit, **perturbed** and transmitted.
- Both steps cause information loss:
  - Compressing: loses much
  - Perturbation: information loss depends on ϵ
- **Key Insight**: We want to make a balance between the two steps:
  - By compressing into more groups, the first step carries more information
- Variance is optimized when $g = e^{\varepsilon} + 1$
- See our paper for details.

[DMNS 06]

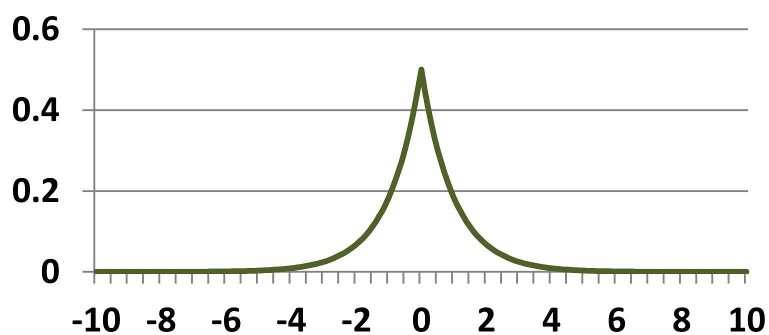
Laplace Mechanism



$$h(\eta) \propto \exp(-\eta / \lambda)$$

Mean: 0,
Variance: $2\lambda^2$

Laplace Distribution – Lap(λ)



How much noise for privacy?

[Dwork et al., TCC 2006]

Sensitivity: Consider a query q: *I* → R. S(q) is the smallest number s.t. for any neighboring tables D, D',

$$| q(D) - q(D') | \leq S(q)$$

Thm: If **sensitivity** of the query is **S**, then the following guarantees ε-differential privacy.

$$\lambda = S/\varepsilon$$

Sensitivity: COUNT query

- Number of people having disease
- Sensitivity = 1

- Solution: $3 + \eta$,
  where η is drawn from $Lap(1/\varepsilon)$
  - Mean = 0
  - Variance = $2/\varepsilon^2$

D

Disease (Y/N)
Y
Y
N
Y
N
N

More on Sensitivity

- Suppose all the n values x are in [a,b]
- Quiz (3 min break):
  - Sensitivity for sum?
  - Sensitivity for mean
  - Sensitivity for median

More on Sensitivity

- Suppose all values x are in [a,b]
- Sensitivity for sum: b
  - One record can increase sum up to b
- Sensitivity for mean: (b-a)/(n+1)
  - Change the total from na to na+b
  - Thus mean: na/n->(na+b)/(n+1)
- Sensitivity for median: (b-a)/2
  - Consider a,a,b->a,a,b,b

Privacy of Laplace Mechanism

- Consider neighboring databases D and D'
- Consider some output O

Laplace Distribution:

$$f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

$$= \frac{1}{2b} \begin{cases} \exp\left(-\frac{\mu-x}{b}\right) & \text{if } x < \mu \\ \exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \geq \mu \end{cases}$$

$$\frac{\Pr\left[A(D) = O\right]}{\Pr\left[A(D') = O\right]} = \frac{\Pr\left[q(D) + \eta = O\right]}{\Pr\left[q(D') + \eta = O\right]}$$

$$= \frac{e^{-|O-q(D)|/\lambda}}{e^{-|O-q(D')|/\lambda}}$$

$$\leq e^{\left|q(D)-q\left(D'\right)\right|/\lambda} \leq e^{S(q)/\lambda} = e^{\varepsilon}$$

Utility of Laplace Mechanism

- Laplace mechanism works for **any function** that returns a real number

- Error: $E(\text{true answer} – \text{noisy answer})^2$

$$= Var(\ Lap(S(q)/\varepsilon)\ )$$

$$= 2*S(q)^2 / \varepsilon^2$$

Exponential Mechanism

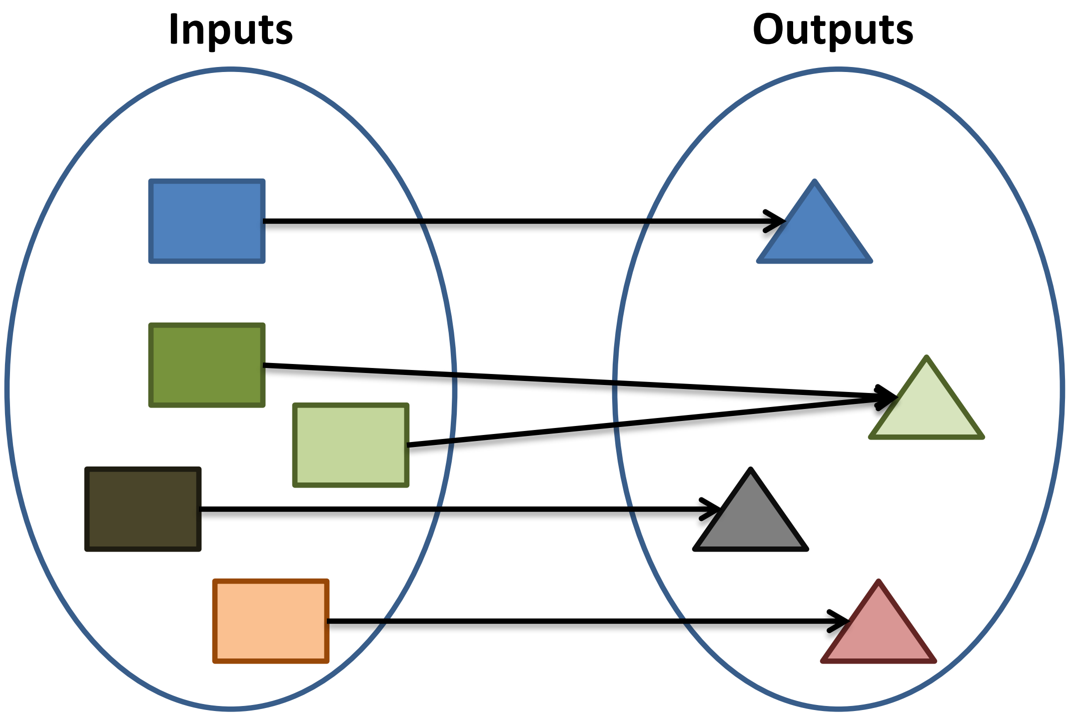
- For functions that do not return a real number …
  - “what is the most common nationality in this room”: Chinese/Indian/American…

- When perturbation leads to invalid outputs …
  - To ensure integrality/non-negativity of output

[MT 07]

Exponential Mechanism

Consider some function f (can be deterministic or probabilistic):



How to construct a differentially private version of f?

Exponential Mechanism

- Scoring function *w: Inputs x Outputs* → *R*

- D: nationalities of a set of people
- #(D, O): # people with nationality O
- f(D): most frequent nationality in D
- w(D, O) = |#(D, O) - #(D, f(D))|

Exponential Mechanism

- Scoring function *w: Inputs x Outputs* → *R*

- Sensitivity of w

$$\Delta_w = \max_{O\&D,D'} |w(D,O) - w(D,O')|$$

where D, D’ differ in one tuple

Exponential Mechanism

Given an input D, and a scoring function w,

Randomly sample an output O from *Outputs* with probability

$$\frac{e^{\frac{\varepsilon}{2\Delta} \cdot w(D,O)}}{\sum_{Q \in Outputs} e^{\frac{\varepsilon}{2\Delta} \cdot w(D,Q)}}$$

- Note that for every output O, probability O is output > 0.

Randomized Response (a.k.a. local randomization)

D

Disease (Y/N)
Y
Y
N
Y
N
N

With probability p,
Report true value

With probability 1-p,
Report flipped value

O

Disease (Y/N)
Y
N
N
N
Y
N

Differential Privacy Analysis

- Consider 2 databases D, D' (of size M) that differ in the j^{th} value
  - D[j] ≠ D'[j]. But, D[i] = D'[i], for all i ≠ j

- Consider some output O

$$\frac{P(D \rightarrow O)}{P(D' \rightarrow O)} \leq e^{\varepsilon} \Leftrightarrow \frac{1}{1+e^{\varepsilon}} < p < \frac{e^{\varepsilon}}{1+e^{\varepsilon}}$$

Randomized Response (a.k.a. local randomization)

D		O
Disease (Y/N)		**Disease (Y/N)**
Y		Y
Y		N
N		N
Y		N
N		Y
N		N

With probability p,
Report true value

With probability 1-p,
Report flipped value

Differential Privacy Analysis

- Consider 2 databases D, D' (of size M) that differ in the j^{th} value
  - D[j] ≠ D'[j]. But, D[i] = D'[i], for all i ≠ j

- Consider some output O

$$\frac{P(D \rightarrow O)}{P(D' \rightarrow O)} \leq e^{\varepsilon} \Leftrightarrow \frac{1}{1+e^{\varepsilon}} < p < \frac{e^{\varepsilon}}{1+e^{\varepsilon}}$$

Laplace Mechanism vs Randomized Response

Privacy

- Provide the same ε-differential privacy guarantee

- Laplace mechanism assumes data collected is trusted
- Randomized Response does not require data collected to be trusted
  - Also called a *Local* Algorithm, since each record is perturbed

Laplace Mechanism vs Randomized Response

Utility

- Suppose a database with N records where μN records have disease = Y.
- Query: # rows with Disease=Y

- Std dev of Laplace mechanism answer: $O(1/\varepsilon)$
- Std dev of Randomized Response answer: $O(\sqrt{N})$

Why Composition?

- Reasoning about privacy of a complex algorithm is hard.

- Helps software design
  - If building blocks are proven to be private, it would be easy to reason about privacy of a complex algorithm built entirely using these building blocks.

Sequential Composition

- If M_1, M_2, ..., M_k are algorithms that access a private database D such that each M_i satisfies ε_i -differential privacy,

  then running all *k* algorithms sequentially satisfies ε-differential privacy with $\varepsilon=\varepsilon_1+\ldots+\varepsilon_k$

Privacy of Sequential Composition

- Consider neighboring databases D and D'
- Consider some output O

$$\frac{\Pr[A(D) = O, O']}{\Pr[A(D') = O, O']} = \frac{\Pr[q(D) + \eta = O]\Pr[q'(D) + \eta' = O']}{\Pr[q(D') + \eta = O]\Pr[q'(D') + \eta' = O']}$$

$$= \frac{e^{-|O-q(D)|/\lambda} \times e^{-|O'-q'(D)|/\lambda}}{e^{-|O-q(D')|/\lambda} \times e^{-|O'-q'(D')|/\lambda}}$$

$$\leq e^{|q(D)-q(D')|/\lambda} \times e^{|q'(D)-q'(D')|/\lambda} \leq e^{\varepsilon}$$

Parallel Composition

- If $M_1, M_2, \ldots, M_k$ are algorithms that access disjoint databases $D_1, D_2, \ldots, D_k$ such that each M_i satisfies ε_i -differential privacy,

  then running all k algorithms in "parallel" satisfies ε-differential privacy with $\varepsilon = \max\{\varepsilon_1,\ldots,\varepsilon_k\}$

Postprocessing

- If M_1 is an ε-differentially private algorithm that accesses a private database D,

  then outputting $M_2(M_1(D))$ also satisfies ε-differential privacy.

Advanced Composition

[DRV10]

- Composing k algorithms, each satisfying ϵ-DP ensures ϵ_g-DP with probability $1-\delta$
$$\epsilon_g = O\left(\epsilon\sqrt{k\ln\frac{1}{\delta}} + k\epsilon^2\right)$$

- Analyze privacy loss as a random variable: given output o and neighbors (D, D')
$$PL(o) = \ln\frac{\Pr[M(D)=o]}{\Pr[M(D')=o]}$$

Advanced Composition

[DRV10]

- Composing k algorithms, each satisfying ϵ-DP ensures ϵ_g-DP with probability $1 - \delta$

$$\epsilon_g = O\left(\epsilon\sqrt{k \ln\frac{1}{\delta}} + k\epsilon^2\right)$$

- Each algorithm has privacy loss $PL(o)$
  - Worst case (DP): $\Pr[|PL(o)| \leq \epsilon] = 1$
  - Expected loss: $\mathrm{E}[PL(o)] \leq \epsilon(e^{\epsilon} - 1)$
  - Total privacy loss ϵ_g is bounded by Azuma's inequality

What Can Be Achieved Under Centralized DP?

- Possible to publish high-quality statistical information for low-dimensional data
- For high-dimensional data (data with hundreds or more attributes), achieving privacy while preserving arbitrary statistical information is hard
  - Possible to perform specific tasks, such as learning a classifier, learning frequent itemsets (and association rules)

Summary

- Motivation to use DP
- LDP Mechanisms
- DP Mechanisms
  - Laplace
  - Exponential
  - Random Response
- DP Properties
  - Sequential/parallel/advanced composition
  - Postprocessing is free