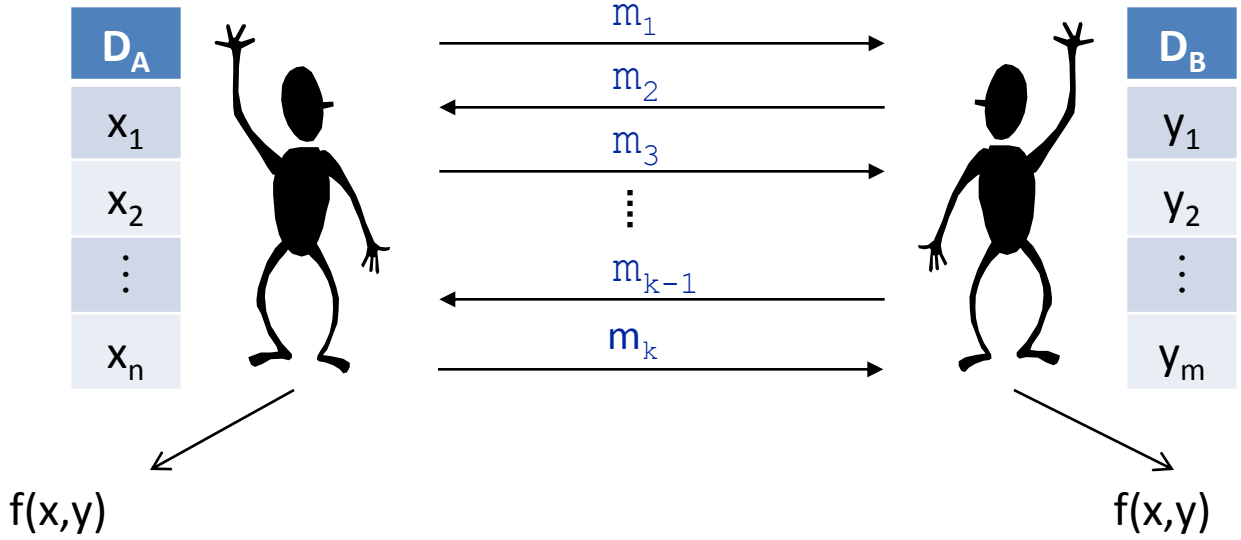
Privacy
Communication Complexity
and Information Complexity

Toniann Pitassi

2-Party Communication Complexity

2-party communication: each party has an input.

Goal is to compute a function f(x,y)



Communication complexity of a protocol for f is the number of bits exchanged between A and B.

Privacy: how much do Alice and Bob learn?

Two Different Notions of Privacy

- Cryptographic [deterministic protocols]
  (Perfect privacy, PRIV, PAR)
- Differential Privacy [randomized protocols]

All of our privacy applications require information complexity lower bounds (cc is not enough)

Two Different Notions of Privacy

- **Cryptographic**
  (Perfect privacy, PRIV, PAR)
- Differential Privacy

Cryptographic Privacy

Goal: Ideal protocol for f should reveal only f(x,y) and no other information

A deterministic protocol partitions M_f into disjoint rectangles (submatrices) until every rectangle is monochromatic (f is constant on all inputs in the submatrix)

Perfect Privacy

Perfect privacy

A protocol for 2-player function $f : X \times Y \to Z$ is **perfectly private** if every two inputs in the same **region** are partitioned into the same **rectangle**.

Characterizing perfect privacy (Kushilevitz '89)

The perfectly private functions of 2 inputs are fully characterized combinatorially. A private deterministic protocol for such functions is given by "decomposing" M_f.

Perfect privacy is unattainable for most functions.

Important Example: Vickrey Auction

Vickrey auction

The 2-player Vickrey auction is defined as $f : X \times Y \to Z$ where

$$X = Y = [2^n],\ Z = [2^{n+1}] \text{ and } f(x,y) = \begin{cases} (x,B), & \text{if } x \leq y \\ (y,A) & \text{if } y < x \end{cases}$$

	1	2	3	4	…	2^n-1	2^n
1	$(1,B)$	$(1,B)$	$(1,B)$	$(1,B)$	…	$(1,B)$	$(1,B)$
2	$(1,A)$	$(2,B)$	$(2,B)$	$(2,B)$	…	$(2,B)$	$(2,B)$
3	$(1,A)$	$(2,A)$	$(3,B)$	$(3,B)$	…	$(3,B)$	$(3,B)$
4	$(1,A)$	$(2,A)$	$(3,A)$	$(4,B)$	…	$(4,B)$	$(4,B)$
⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮
2^n-1	$(1,A)$	$(2,A)$	$(3,A)$	$(4,A)$	…	$(2^n-1,B)$	$(2^n-1,B)$
2^n	$(1,A)$	$(2,A)$	$(3,A)$	$(4,A)$	…	$(2^n-1,A)$	$(2^n,B)$

Vickrey (cont'd)

Vickrey auction

The 2-player Vickrey auction is defined as $f : X \times Y \to Z$ where

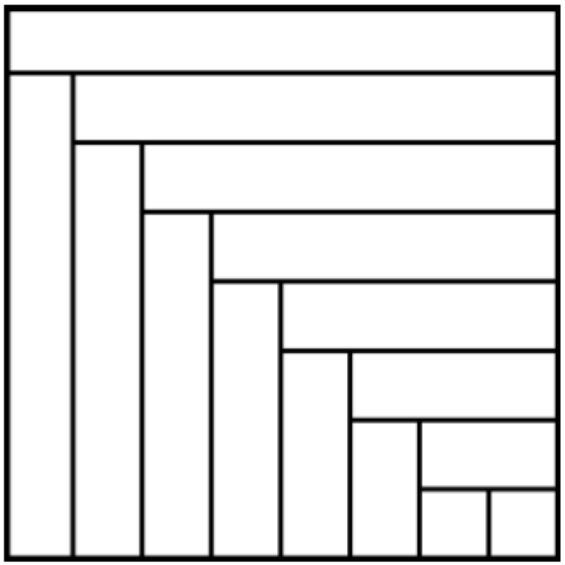
$$X = Y = [2^n],\ Z = [2^{n+1}] \text{ and } f(x,y) = \begin{cases} (x,B), & \text{if } x \leq y \\ (y,A) & \text{if } y < x \end{cases}$$

	1	2	3	4	…	2^n-1	2^n
1	$(1,B)$	$(1,B)$	$(1,B)$	$(1,B)$	…	$(1,B)$	$(1,B)$
2	$(1,A)$	$(2,B)$	$(2,B)$	$(2,B)$	…	$(2,B)$	$(2,B)$
3	$(1,A)$	$(2,A)$	$(3,B)$	$(3,B)$	…	$(3,B)$	$(3,B)$
4	$(1,A)$	$(2,A)$	$(3,A)$	$(4,B)$	…	$(4,B)$	$(4,B)$
⋮	⋮	⋮	⋮	⋮	…	⋮	⋮
2^n-1	$(1,A)$	$(2,A)$	$(3,A)$	$(4,A)$	…	$(2^n-1,B)$	$(2^n-1,B)$
2^n	$(1,A)$	$(2,A)$	$(3,A)$	$(4,A)$	…	$(2^n-1,A)$	$(2^n,B)$

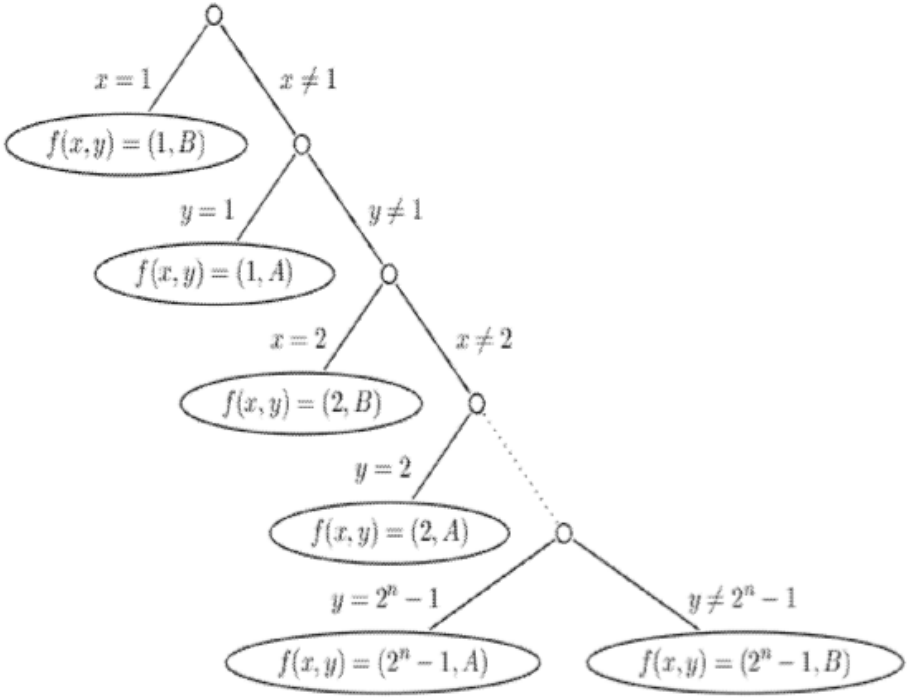
Vickrey (cont'd)

Vickrey auction

The 2-player Vickrey auction is defined as $f : X \times Y \to Z$ where $X = Y = [2^n]$, $Z = [2^{n+1}]$ and $f(x, y) = \begin{cases} (x, B), & \text{if } x \leq y \\ (y, A) & \text{if } y < x \end{cases}$



Perfectly Private Protocol for Vickrey



Ascending English bidding is the *only* perfectly private protocol. Lengthy!

Approximate Privacy (PAR)

Vickrey auction

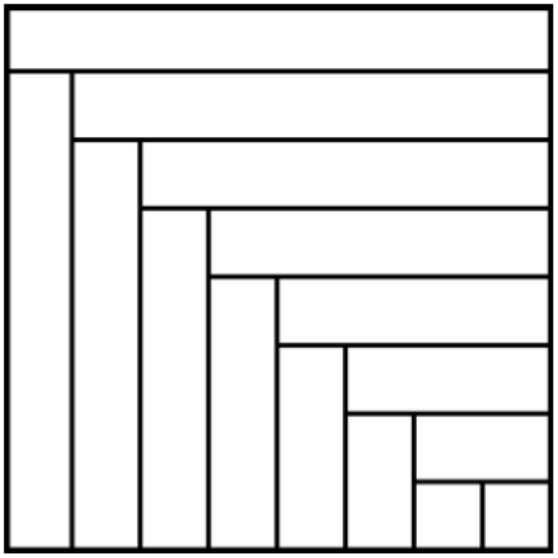
The 2-player Vickrey auction is defined as $f : X \times Y \to Z$ where

$X = Y = [2^n]$, $Z = [2^{n+1}]$ and $f(x, y) = \begin{cases} (x, B), & \text{if } x \leq y \\ (y, A) & \text{if } y < x \end{cases}$

Regions (preimages)

region $R_{x,y} = \{(x', y') \in X \times Y \mid f(x, y) = f(x', y')\}$

defined by **function** $\longrightarrow$



Rectangles

rectangle $P_{x,y} = \{(x', y') \in X \times Y \mid f(x, y) = f(x', y') \text{ and } \pi(x, y) = \pi(x', y')\}$

defined by **protocol**

Approximate Privacy (PAR)

Privacy approximation ratio (Feigenbaum Jaggard Schapira '10)

A protocol for f has **privacy approximation ratio**:

$$\text{worst-case PAR} = \max_{(x,y)} \frac{|R_{x,y}|}{|P_{x,y}|}$$

$$\text{average-case PAR} = \mathbb{E}_{(x,y)\sim\mu} \frac{|R_{x,y}|_\mu}{|P_{x,y}|_\mu} \text{ over distribution } \mu$$

Notes:

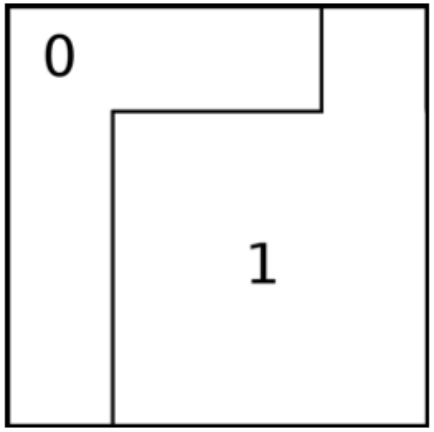
- Above is **External** PAR. Internal PAR can also be defined. For Vickrey, they are equal.
- Equivalent characterization of average-case PAR: $\Sigma_R\ |R|_D \times cut_P(R)$

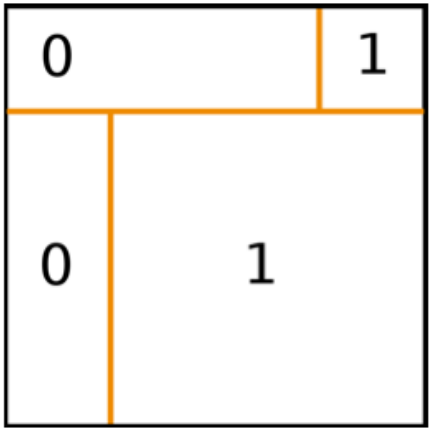
Privacy approximation ratio (Feigenbaum Jaggard Schapira '10)

A protocol for f has **privacy approximation ratio**:

$$\text{worst-case PAR} = \max_{(x,y)} \frac{|R_{x,y}|}{|P_{x,y}|}$$

$$\text{average-case PAR} = \mathbb{E}_{(x,y)\sim\mu} \frac{|R_{x,y}|_\mu}{|P_{x,y}|_\mu} \text{ over distribution } \mu$$



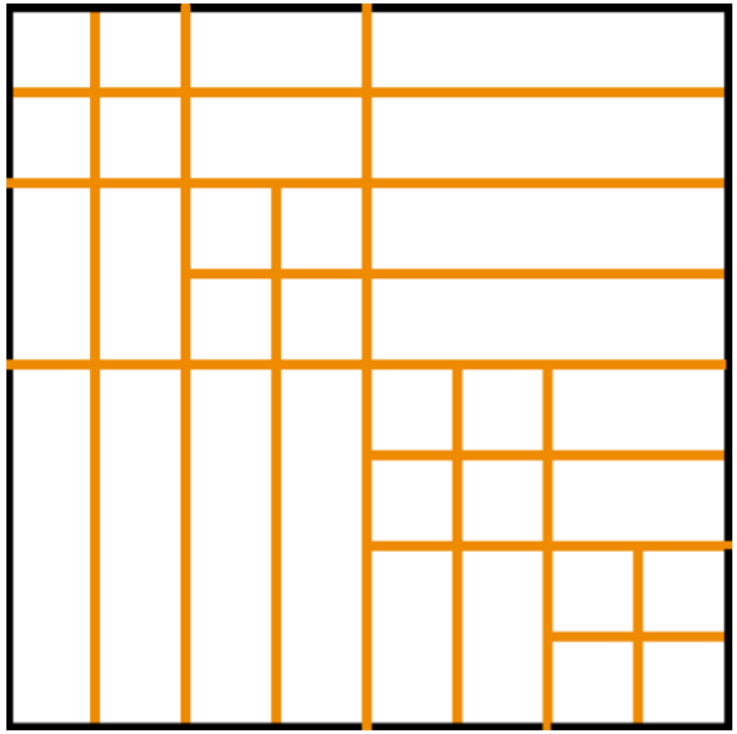


worst-case $\mathrm{PAR} = 10$
average-case $\mathrm{PAR} = 2$

Above matrices are 4-by-4. $|R_1|$=10, $|R_0|$=6
Avg case PAR = $|R_1|_U \times 2 + |R_0|_U \times 2 = 10/16 \times 2 + 6/16 \times 2 = 2$

Back to Vickrey: The Bisection Protocol

How short can we make a protocol for Vickrey auction?



Bisection protocol.

Tradeoffs between privacy and communication [ACCFKP]

Upper bounds for Vickrey auctions

	English bidding	bisection protocol
communication cost	2^n	$O(n)$
worst-case PAR	1	2^n
average-case PAR	1	$O(1)$

Worst-case lower bound

For all n, for all p, $2 \leq p \leq n/4$, any deterministic protocol for the n-bit two-player Vickrey auction obtaining PAR less than 2^{p-2} has length at least $2^{n/4p}$.

Average-case lower bound

For all $n < r$, any deterministic protocol of length $\leq r$ for the uniform n-bit Vickrey auction has average-case PAR greater than $\Omega(\frac{n}{\log(r/n)})$.

Information Complexity and Avg-case PAR

Definition (Klauck)

$$\mathrm{Priv}^{ext}_D(P) = I(XY; \pi_P(X,Y) \mid f(X,Y)),$$

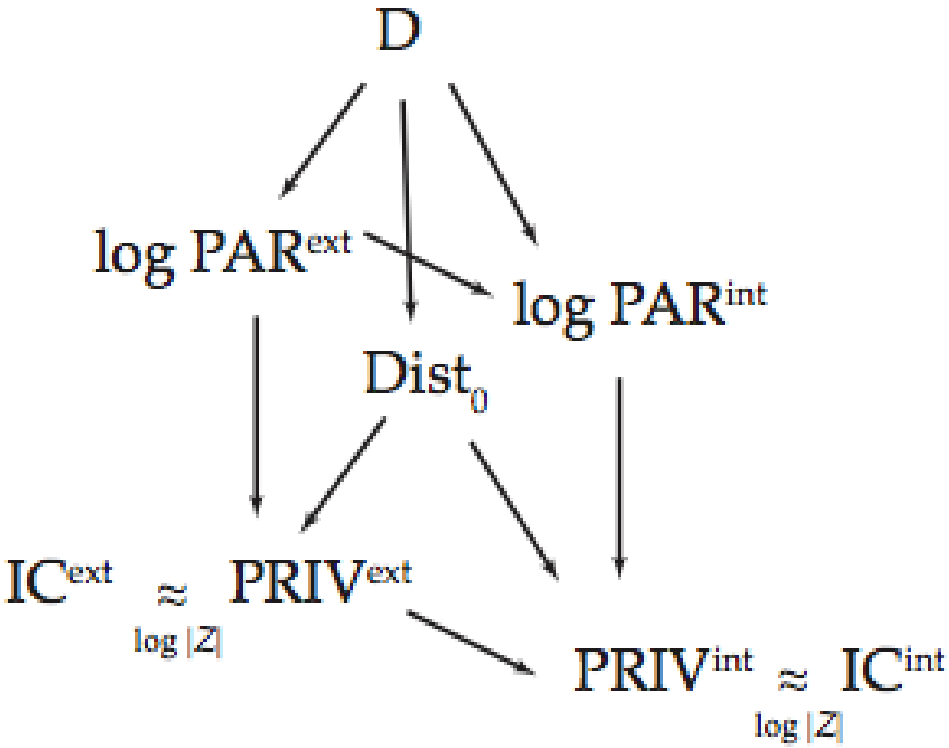
Proposition:

$$\mathrm{Priv}^{ext}_D(P) \le \mathrm{IC}^{ext}_D(P) \le \mathrm{Priv}^{ext}_D(P) + \log|Z|$$

Theorem: $\mathrm{Priv}^{ext}_D(P) \le \log(\mathrm{avg}_D\, \mathrm{PAR}^{ext}(P))$ (concavity)

- Analagous relationships hold for internal case.
- Thus IC lower bounds imply avg case PAR lower bounds!

Relationships between measures
(deterministic protocols, avg case PAR, boolean functions)



Also known that PAR^{ext} is at least rank(M_f). So under log-rank conjecture, D and log PAR^{ext} are polynomially related

Consequences of IC Connection: Lower Bounds for Avg PAR

Theorem (Braverman).

Let P be a randomized protocol for DISJn with error at most 1/3. Then $IC^{int}(P) = \Omega(n)$

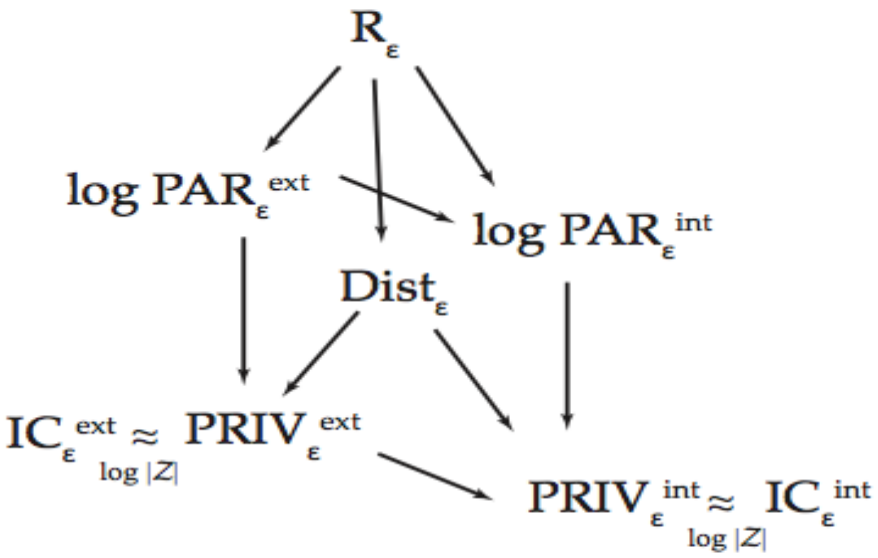
Theorem (ACCFKP). For all $n \geq 1$, and any protocol P for set intersection,

$$avg_U\ PAR(P) = \exp(\Omega(n))$$

Consequences of IC Connection [ACCFKP,KLX]

Problem	$\mathbf{PAR}_{\mu}^{\mathbf{ext}}$		$\mathbf{PRIV}_{\mu}^{\mathbf{ext}}$	$\mathbf{PAR}_{\mu,\epsilon}^{\mathbf{int}}$
	[FJS10] (for uniform μ)	**Our contribution** (for any μ)	(for some μ)	(for some μ)
Equality	-	2^n	$n-1$	$\Theta(1)$
Disjointness	$\left(\frac{3}{2}\right)^n$	2^n-1	$n-1$	$2^{\Theta(n)}$
Inner Product	-	2^n-1	$n-2-o(1)$	$2^{\Theta(n)}$
Greater Than	$2^n+\frac{1}{2^{n+1}}-\frac{1}{2}$	2^n-1	$n-1$	$2^{\Theta(\log n)}$

Randomized PAR [KLX]



Suggests that PAR could be helpful to study the randomized CC versus IC question.

Two Different Notions of Privacy

- Cryptographic
  (Perfect privacy, PRIV, PAR)
- **Differential Privacy**

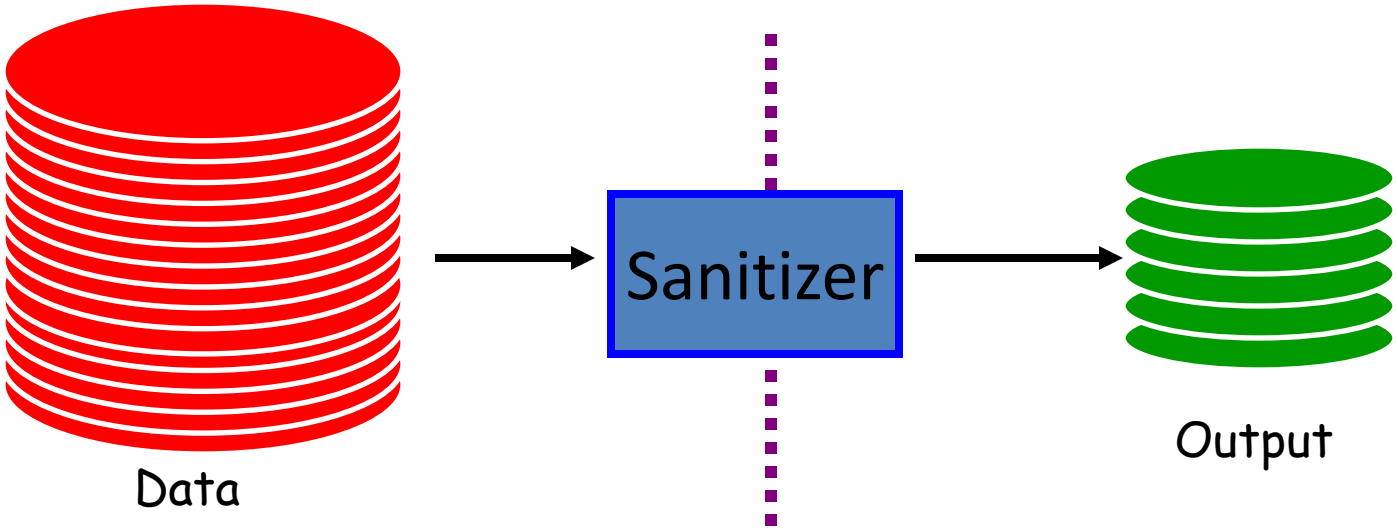
Differential Privacy [DwMcNiSm06]

Holy Grail: Get **utility** of statistical analysis (i.e., give reasonably accurate answers) while **protecting privacy** of every individual participant.

Guarantee: outcome of the analysis is nearly identical whether any user is in or out of the dataset. Thus participation does not increase risk of privacy violation.

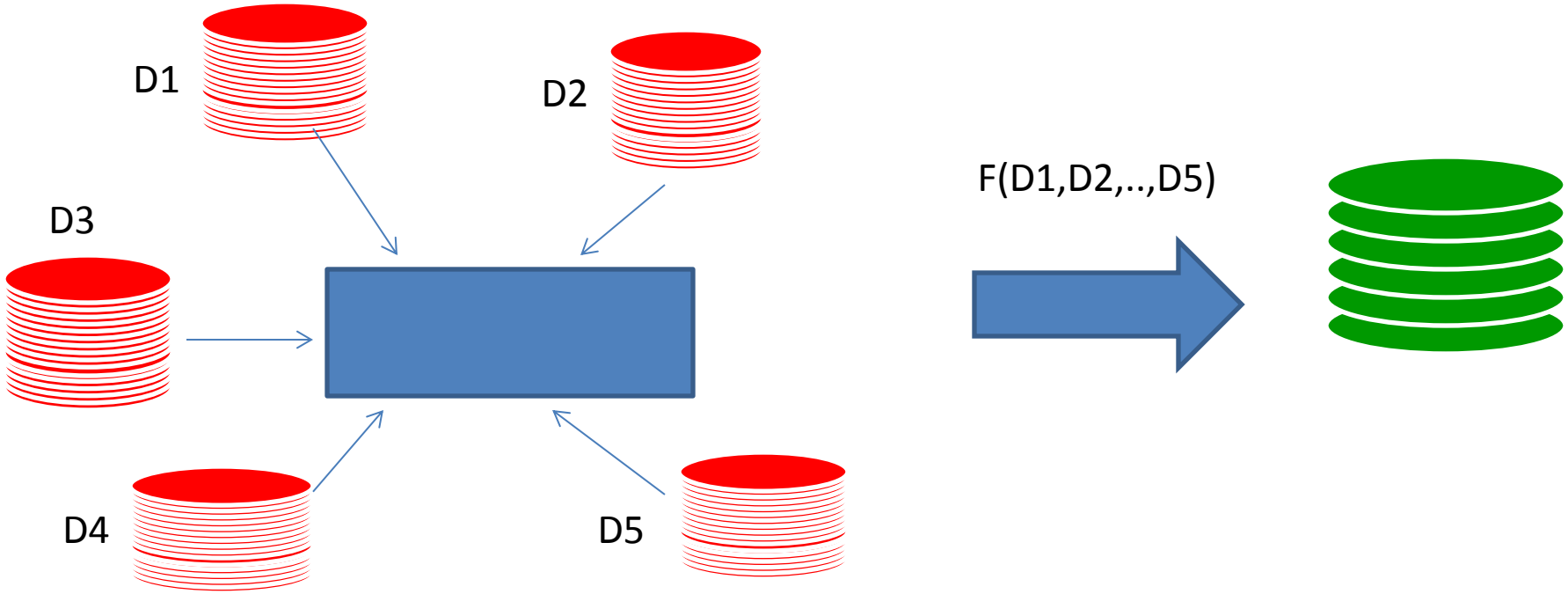
The Basic DP Scenario
(Client-Server Setting)

- Data is owned by a curator.
- Statistical queries are made. The curator computes the query releases a sanitized version .



Differential Privacy: The Distributed Setting

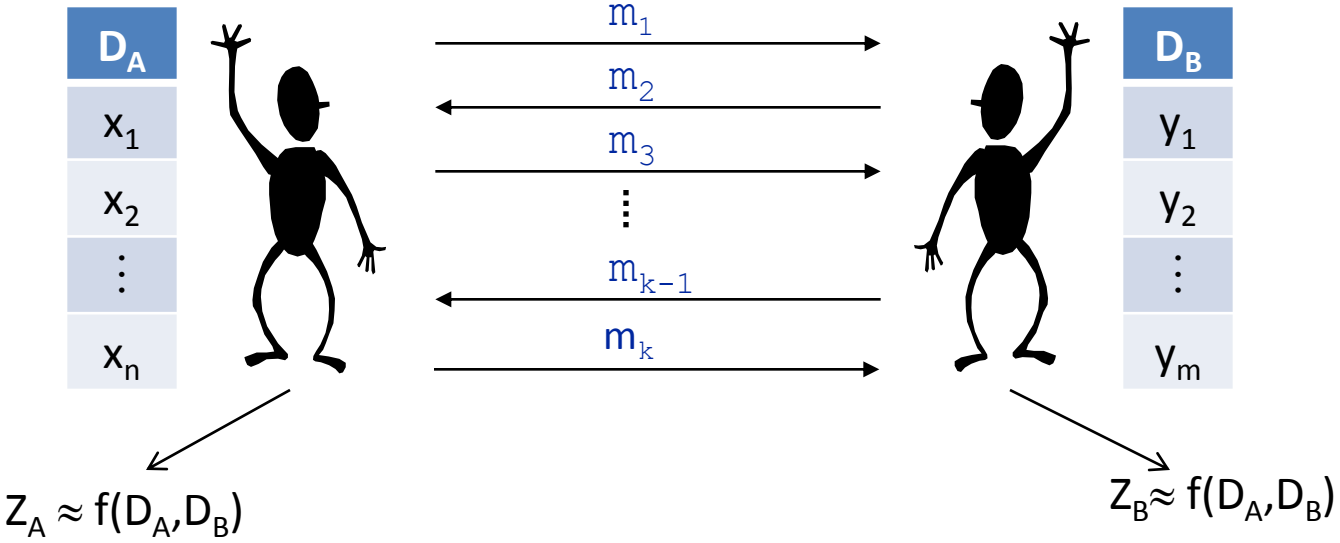
Multiple databases, each with private data.



Goal: compute a joint function while maintaining privacy for any individual, with respect to both the outside world and the other database owners.

2-Party Setting: Differentially Private CC [MMPRTV]

Each party has a dataset; want to compute a joint function $f(D_A, D_B)$



A's view should be a **differentially private** function of D_B (even if A deviates from protocol), and vice-versa

Two-Party Differential Privacy

Let P(x,y) be a 2-party protocol. **P is ε-DP** if:

(1) for all y, for every pair x, x′ that are neighbors, and for every transcript π,

$$\Pr[P(x,y) = \pi] \le \exp(\varepsilon) \Pr[P(x',y) = \pi]$$

(2) symmetrically, for all x, for every pair of neighbors y,y′ and for every transcript π

$$\Pr[P(x,y) = \pi] \le \exp(\varepsilon) \Pr[P(x,y') = \pi]$$

Question: Can anything that can be done DP-ly in client-server setting also be done DP-ly in multiparty setting?

Examples

1. **Ones(x,y)** = the number of ones in xy
   Ones(00001111,10101010) = 8.

   CC(Ones) = logn.
   There is a low error DP protocol.

2. **Hamming Distance HD(x,y)** = the number of positions i where $x_i \neq y_i$.
   HD(00001111, 10101010) = 4

   CC(HD)=n.
   No low error DP protocol ?

Is this a coincidence? Is there a connection between low cc and low-error DP protocols?

What is the smallest (additive) error of any protocol for computing the Hamming distance between x,y that is differentially private for both sides?

Lower Bounds for DP Protocols and Information Cost

$I(X;Y) = H(X) - H(X \mid Y)$

measures the average amount of info Y reveals about X.

$IC^{ext}_{\mu}(P) = I(XY; \pi(X,Y))$

measures the average amount of information the transcript reveals about XY

$IC^{ext}_{\mu}(F) = \min_P \ (IC^{ext}_{\mu}(P))$

Theorem. If P has ε-DP, then for every distribution μ on X x Y, $IC^{ext}_{\mu}(P) \leq 3\varepsilon n$

Note: over uniform distribution, bound improved to $\varepsilon^2 n$

Lower Bounds for DP Protocols via IC

Theorem. If P has ε-DP, then for every distribution μ on X x Y,

$$IC^{ext}_{\mu}(P) \leq 3\varepsilon n$$

Proof. Let z=(x,y). For every z, z′ differential privacy implies

$$\exp(-2\epsilon n) \leq \frac{\Pr[\Pi(z) = \pi]}{\Pr[\Pi(z') = \pi]} \leq \exp(2\epsilon n).$$

$$\exp(-2\epsilon n) \leq \frac{\Pr[\Pi(z) = \pi]}{\Pr[\Pi(Z') = \pi]} \leq \exp(2\epsilon n).$$

where Z′ is an independent sample. Thus by definition of IC we have:

$$\begin{aligned} I(\Pi(Z); Z) &= H(\Pi(Z)) - H(\Pi(Z)|Z) \\ &= \mathrm{E}_{(z,\pi)\leftarrow(Z,\Pi(Z))} \log \frac{\Pr[\Pi[Z] = \pi | Z = z]}{\Pr[\Pi(Z) = \pi]} \\ &\leq 2(\log_2 e)\epsilon n. \end{aligned}$$

Lower Bounds for Hamming Distance

Theorem. The Gap-Hamming Problem (distinguishing inputs with distance at most $n/2-c\sqrt{n}$ from those with distance at least $n/2+c\sqrt{n}$) has information complexity $\Omega(n)$.

Corollary. There exists an ε such that any ε-DP protocol for Hamming Distance must incur an additive error $\Omega(\sqrt{n})$

Notes:

- Our lower bound for Hamming distance is tight since there is an $O(\sqrt{n})$ error ε-DP protocol.
- There are other functions (with sensitivity 1) where any ε-DP must incur an additive error $\Omega(n)$

Implications of DP Lower bounds:

I. Separation beteen computational and info-theoretic DP

[MPRV] defined **computational ε-DP protocols**.

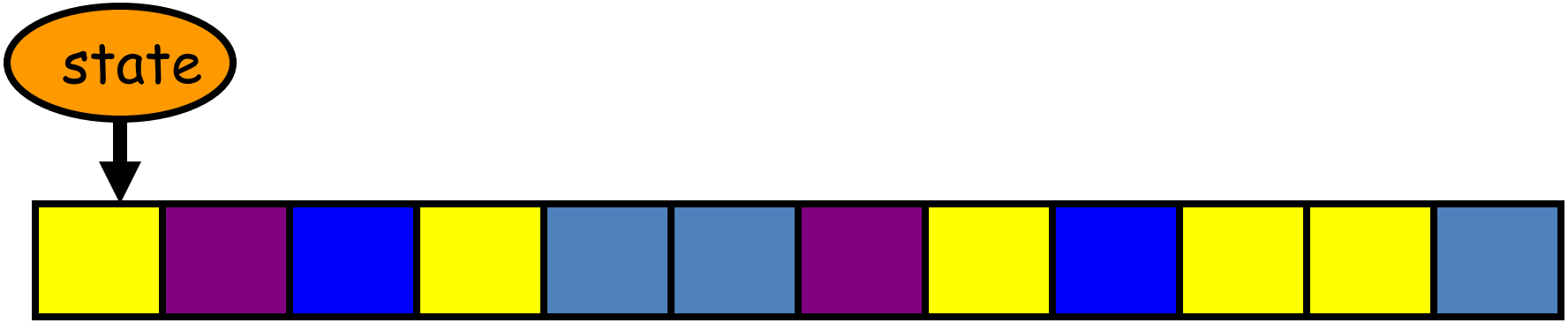
- Loosely speaking, now the probability distribution over the transcripts for neighboring x,x' is e^{ε}-indistinguishable to a polytime algorithm.

- Via fully homomorphic encryption, any low sensitivity $f(x,y)$ has a $O(1)$ error computational ε-DP protocol, including Hamming distance.

- **Thus our lower bound shows that in the context of distributed protocols, there can be a huge gain by relaxing DP to computational DP.**

Implications of DP Lower Bounds: II. Pan-Private Streaming Model [DPRNY]

- Data is a **stream** of items; each item belongs to a user. Sanitizer sees each item and updates internal state. Generates output at end of the stream (**single pass**).



Pan-Privacy: For every two **adjacent streams**, at any **single point in time**, the **internal state** (and final output) are differentially private.

Pan-private algorithms exist for many statistics!

- Stream density / number of distinct elements
- t-cropped mean: mean, over users, of min(t, #appearances)
- Fraction of users appearing exactly k times
- Fraction of users appearing exactly 0 times modulo k
- Fraction of heavy-hitters, users appearing at least k times

DP Lower Bounds imply lower bounds for Pan Private Protocols

Lower Bounds for ε-DP communication protocols imply pan privacy lower bounds for density estimation (via Hamming distance lower bound).

Lower bounds also hold for multi-pass pan-private models

Analogy: 2-party communication complexity lower bounds imply lower bounds in streaming model.

DP Protocols and Compression

So back to Ones(x,y) and HD(x,y)...is DP the same as compressible?

Theorem. [BBCR] **(Low Icost implies compression)**
For every product distribution μ, and protocol P, there exists a protocol Q (β-approximating P) with comm. complexity ~ $Icost_\mu(P)$ x polylog(CC(P))/β

Corollary. **(DP protocols can be compressed)**
Let P be an ε-DP protocol P. Then there exists a protocol Q of cost 3εn polylog(CC(P))/β and error β.

DP almost implies low cc, except for this annoying polylog(CC(P)) factor

Moreover, the low cc protocol can often be made DP (if the number of rounds is bounded.)

Differential Privacy and Compression

- We have seen that DP protocols have low information cost
- By BBCR this implies they can be compressed (and thus have low comm complexity)

What about the other direction? Can functions with low cc be made DP?

Yes! (with some caveats..the error is proportional not only to the cc, but also the number of rounds.)

Proof uses the exponential mechanism [MT]

Open Problems

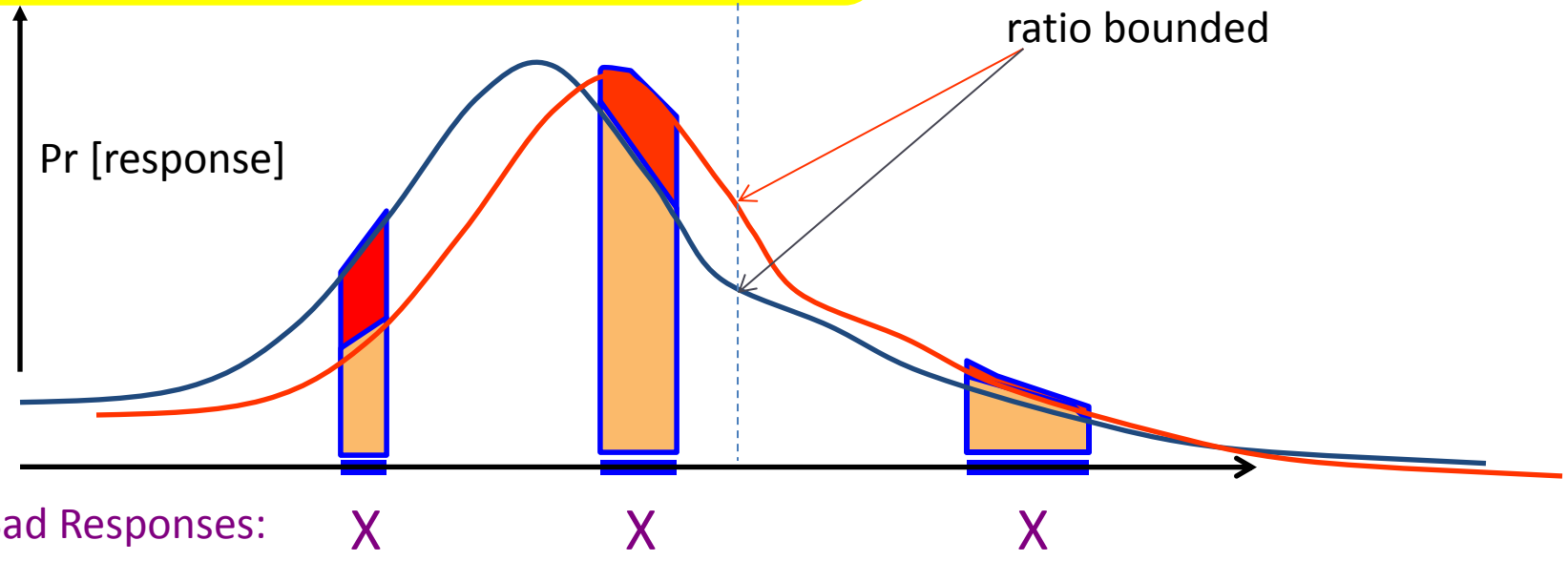
- PAR for randomized protocols
- Better relationship between DP protocols, IC, and CC (Can small cc protocols with unbounded rounds be made DP?)
- Approximate DP
  - No connection known between approx DP and IC
  - Approx DP in client-server setting versus 2-party setting
  - Approx DP in info-theoretic versus computational setting
- Unifying view of all of these concepts

Thanks!

Differential Privacy [DMNS 2006]

M is ε-differentially private if $\forall$ adjacent $D_1 D_2$,
$\forall C \subseteq \text{range}(M)$: $\Pr[M(D1) \in C] \leq e^{\varepsilon} \Pr[M(D2) \in C]$

Neutralizes all linkage attacks.
Composes automatically: $\Sigma_i \varepsilon_i$



Sensitivity of a Function

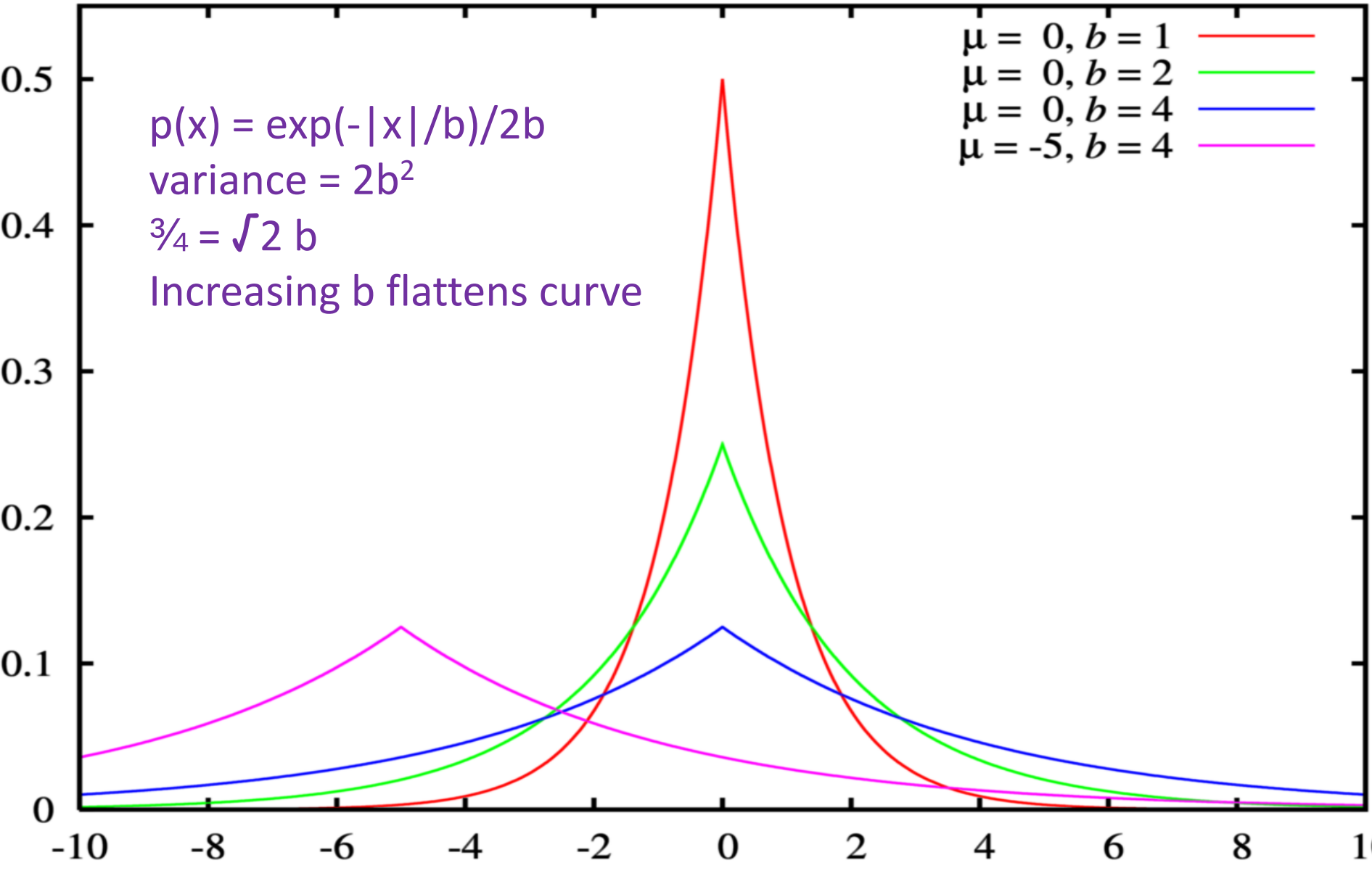
$$\Delta f = \max_{\text{adj } D,D'} |f(D) - f(D')|$$

Adjacent databases differ in at most one row.

Counting queries have sensitivity 1.

Sensitivity captures how much one person's data can affect output.

Laplacian Distribution Lap(b)



$p(x) = \exp(-|x|/b)/2b$

variance = $2b^2$

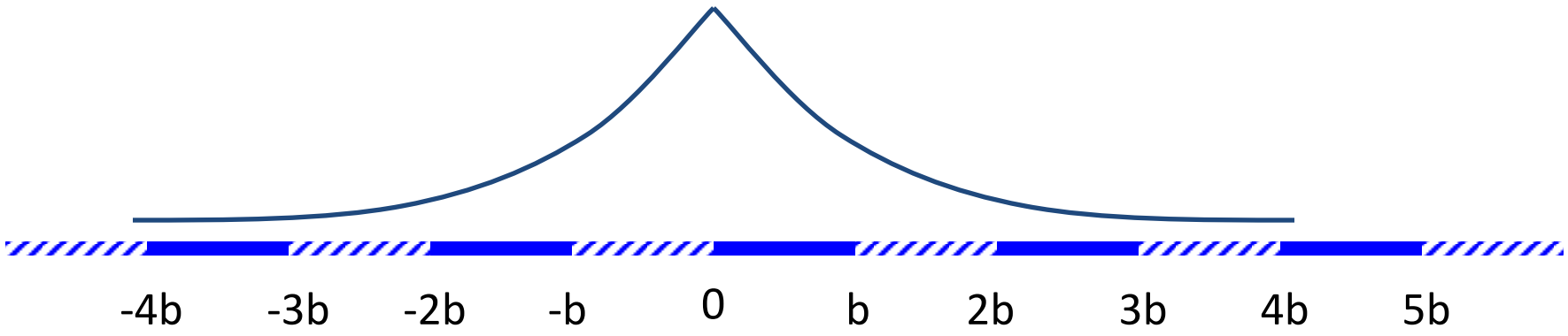
¾ = √2 b

Increasing b flattens curve

Calibrate Noise to Sensitivity

$$\Delta f = \max_{\text{adj } D,D'} |f(D) - f(D')|$$

Theorem: To achieve ε-differential privacy for f, use scaled symmetric noise [Lap(b)] with $b = \Delta f/\varepsilon$.



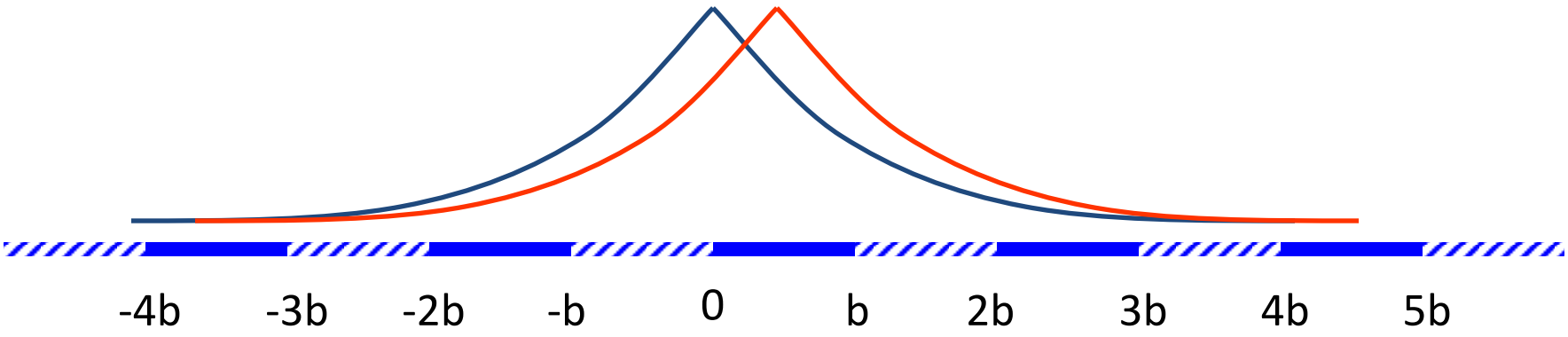
Noise depends on f and ε, not on the database

Smaller sensitivity (Δf) means less distortion

Why Does it Work ?

$$\Delta f = \max_{D,D'} \; |f(D) - f(D')|$$

Theorem: To achieve ε-differential privacy, add scaled symmetric noise [Lap(b)] with $b = \Delta f/\varepsilon$.



$$\frac{\Pr[\mathcal{K}(f, D) = t]}{\Pr[\mathcal{K}(f, D') = t]} = \exp(-(||t - f(D)|| - ||t - f(D')||)/b) \leq \exp(\Delta f/b)$$