Differential Privacy: What we Know and What we Want to Learn

Cynthia Dwork, Microsoft Research

Our Original Goal

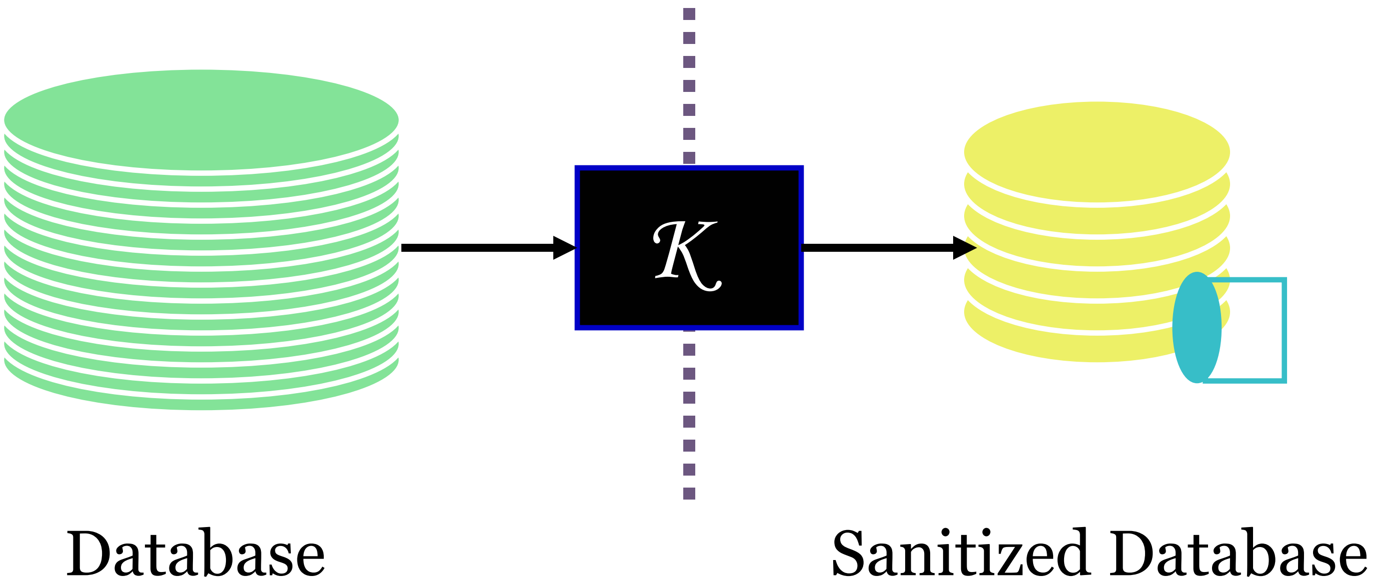
Privacy-Preserving Analysis of Confidential Data

- Mathematical Definition of Privacy
- Finding Statistical Correlations
  - Analyzing medical data to learn genotype/phenotype associations
  - Correlating cough outbreak with chemical plant malfunction
    - Can't be done with HIPAA safe-harbor sanitized data
- Noticing Events
  - Detecting spike in ER admissions for asthma
- Datamining Tasks
  - Clustering; learning association rules, decision trees, separators; principal component analysis
- Official Statistics
  - Contingency Table Release

Achieved Much

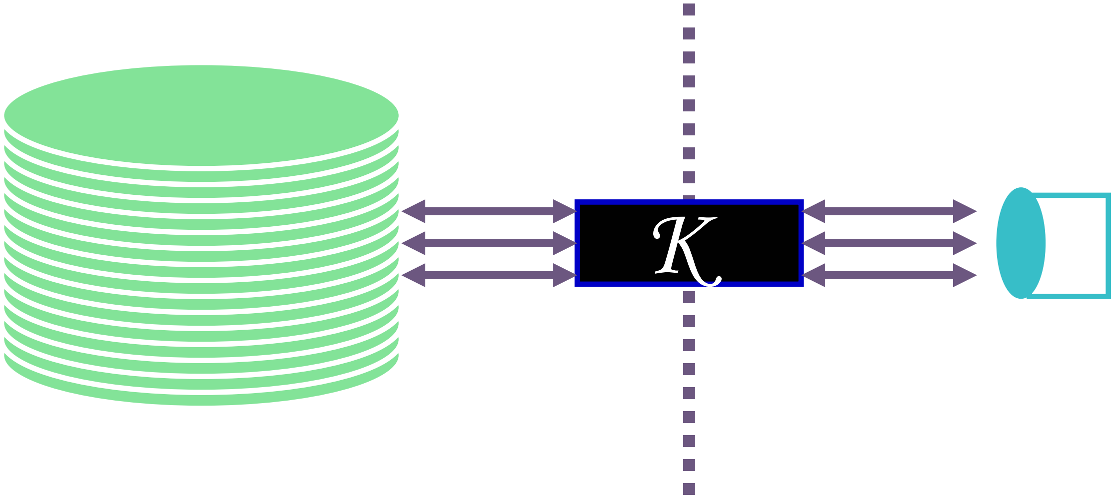
- Defined Differential Privacy
  - Natural goals unachievable
  - "Ad Omnia" definition; independent of linkage information
- General Approach; Rigorous Proof
  - Relates degree of distortion to the (mathematical) sensitivity of the computation needed for the analysis
    - "How much" can the data of one person affect the outcome?
  - Cottage Industry: redesigning algorithms to be insensitive
- Assorted Extensions
  - When noise makes no sense; when actual sensitivity is much less than worst-case; when the database is distributed; …
- Lower bounds on distortion

Two Models



Non-Interactive: Data are sanitized and released

Two Models



Database

Interactive: Multiple Queries, Adaptively Chosen

Privacy: Outputs vs. Processes

- Privacy comes from uncertainty.
- Differentially private mechanisms provide uncertainty.
- Probability space is the coin flips of the mechanism.

- Similar in spirit to randomized response:

  Are you now, or have you ever been, a member of the CP?

  Flip a coin. If heads, answer truthfully.

  If tails, flip again: say yes if heads, no if tails.
- This is a (ln 3)-differentially private mechanism.
  - If member, answer yes with probability ¾.
  - If never member, answer yes with probability ¼.
  - Ratio = 3, bounded by exp(ln 3).
  - Same possible answers in both cases, different distributions.

Privacy: Outputs vs. Processes

- Privacy comes from uncertainty.
- Differentially private mechanisms provide uncertainty.
- Probability space is the coin flips of the mechanism.

- Cf: traditional suppression of cells with low counts
  - Single datum can determine suppression/release of count.
  - NOT the same set of possible answers.

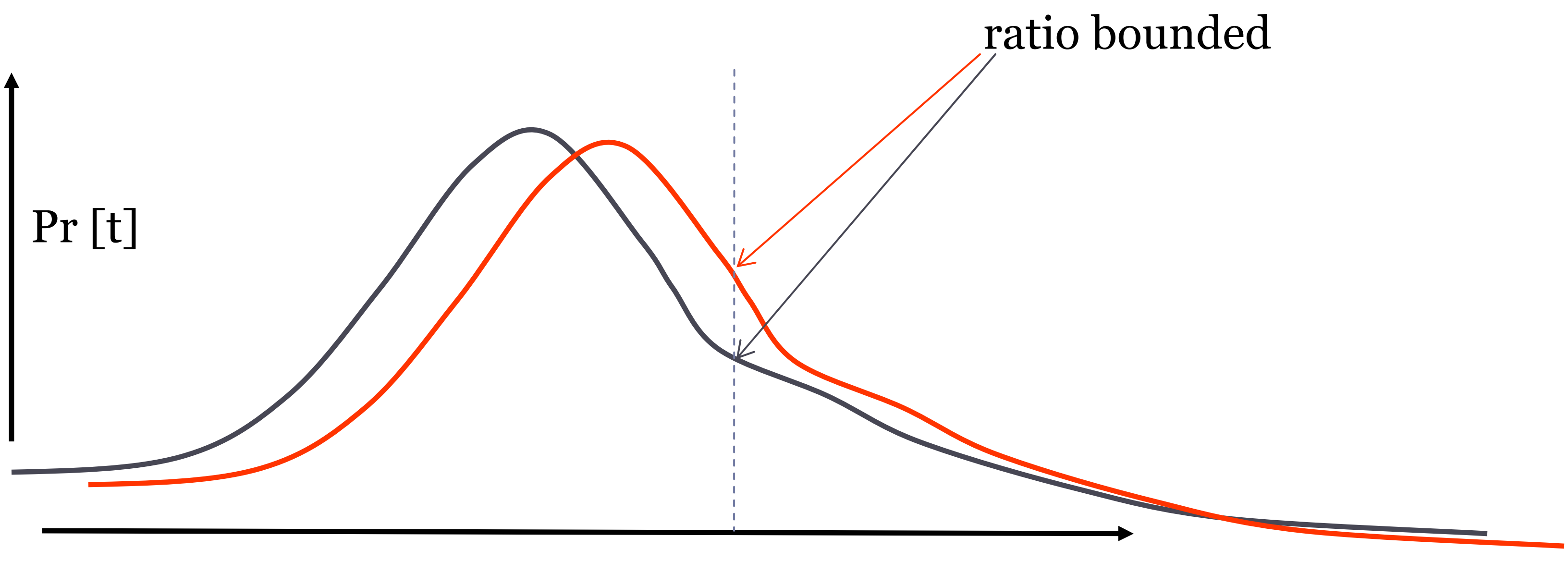
Semantic Security for Statistical Databases?

- Dalenius, 1977
  - Anything that can be learned about a respondent from the statistical database can be learned without access to the database.
- Unachievable
  - Auxiliary Info/Linkage Data is the stumbling block.
  - Fun proof; can be told as a parable.
- Suggests new criterion: risk incurred by joining DB
  - Before/After interacting vs Risk when in/notin DB

Differential Privacy

$\mathcal{K}$ gives ε-differential privacy if for all values of DB, DB'
differing in at most one row, and all S ⊆ Range($\mathcal{K}$)

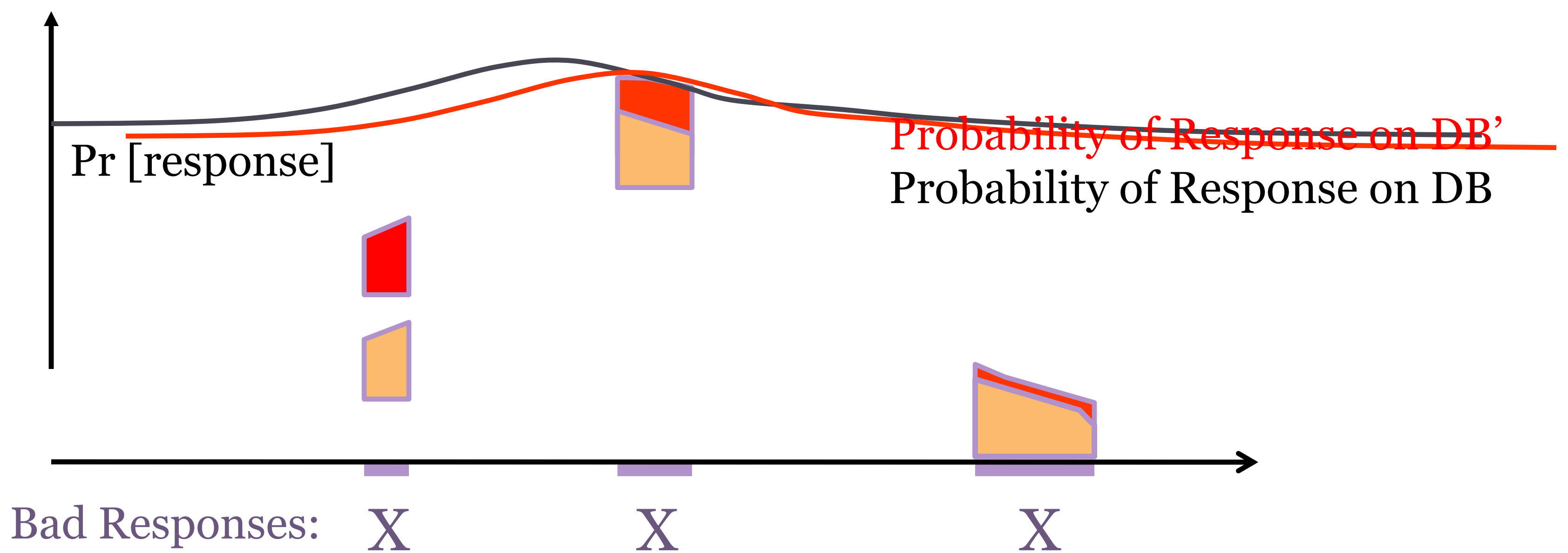
$$\frac{\Pr[\mathcal{K}(\mathrm{DB}) \subseteq S]}{\Pr[\mathcal{K}(\mathrm{DB'}) \subseteq S]} \leq e^{\varepsilon} \approx (1+\varepsilon)$$



Same set of possible answers; different probability distributions

Differential Privacy: An Ad Omnia Guarantee

- $\mathcal{K}$ behaves essentially the same way, independent of whether any individual opts in or opts out
- No perceptible risk is incurred by joining DB
- Holds independent of aux info, comp power



A Natural Relaxation: (ε, δ)-Differential Privacy

For all DB, DB' differing in at most one element,
for all $S \subseteq \text{Range}(\mathcal{K})$,

$$\Pr[\mathcal{K}(DB) \subseteq S] \leq e^{\varepsilon} \Pr[\mathcal{K}(DB') \subseteq S] + \delta$$

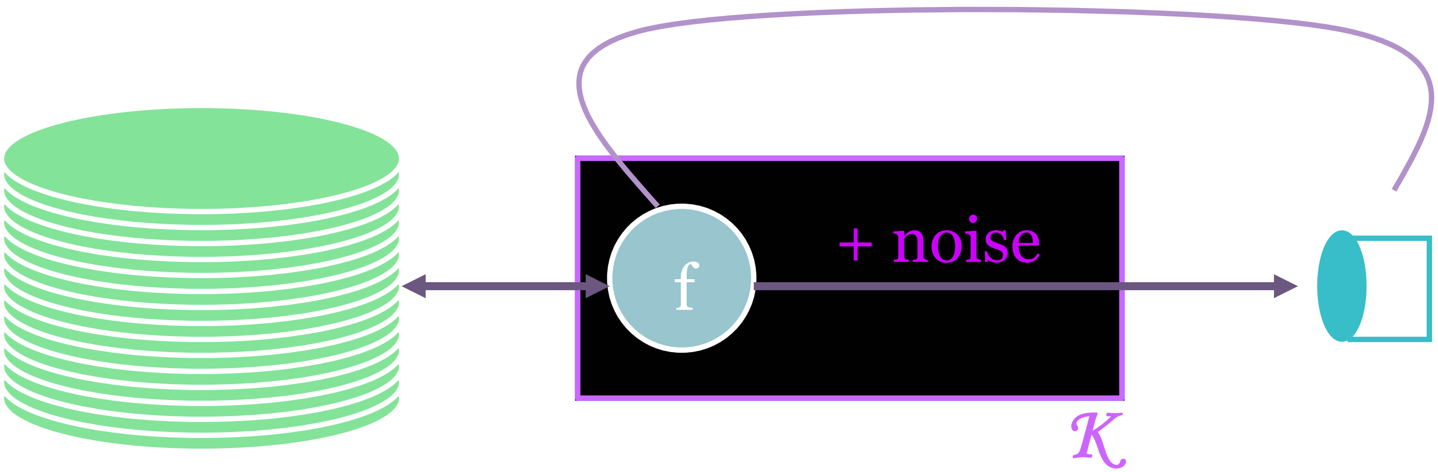
where $\delta = \delta(n)$ is negligible.

Cf : ε –Differential Privacy is unconditional, independent of n

Advantage: Can permit improved accuracy.

See also, *eg*, Abowd *et al*., 2008

An Interactive Mechanism: $\mathcal{K}$



f: DB → R

Eg, CountP(DB) = # rows in DB with Property P

$\mathcal{K}$(f, DB) = f(DB) + Noise

Sensitivity of a Function f

Assume DB and DB' differ only in one row (Me).

How Much Can f(DB) Exceed f(DB')?

- Recall: $\mathcal{K}$(f, DB) = f(DB) + noise
- Question Asks: What difference must noise obscure?

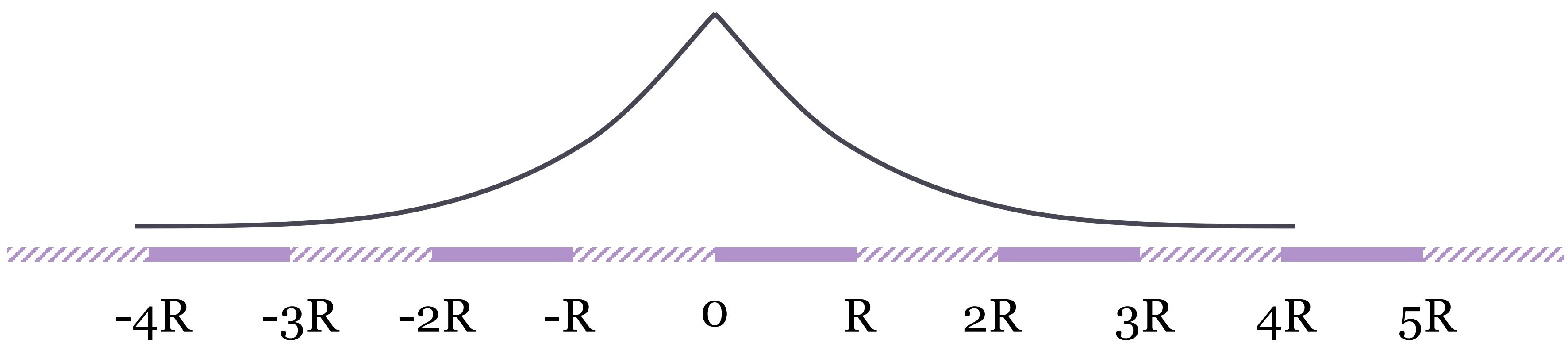
$$\Delta f = \max_{d(DB,\, DB')=1} |f(DB) - f(DB')|$$

eg, ΔCount = 1

Calibrate Noise to Sensitivity

$$\Delta f = \max_{d(DB,\ DB')=1} |f(DB) - f(DB')|$$

Theorem: Can achieve ε-differential privacy by adding scaled symmetric noise ~ Lap(Δf/ε).



Pr[x] proportional to exp(-|x|ε/Δf)

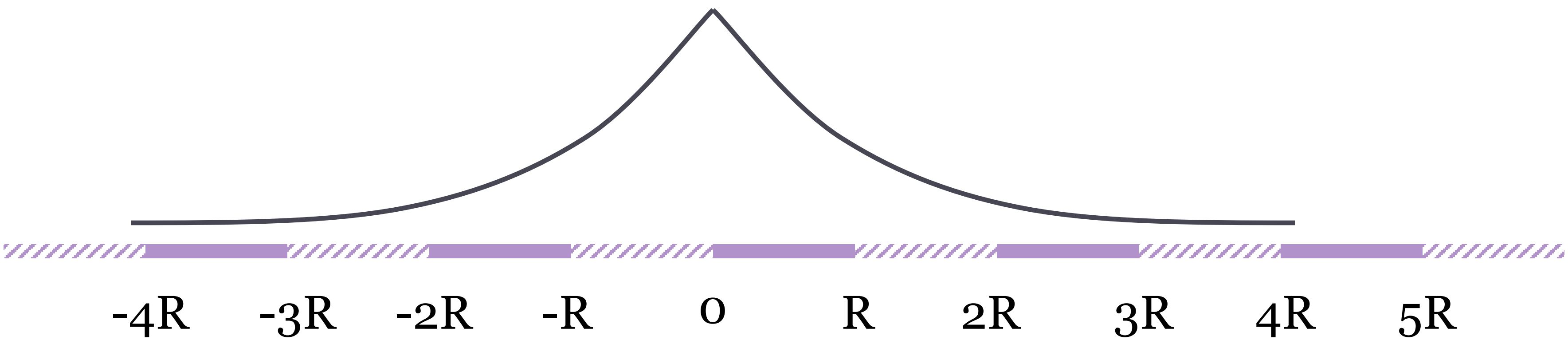
Increasing R= Δf/ε flattens curve; more privacy

Noise depends on f and ε, not on the database

Multiple/Complex Queries f: DB → R^k

$\Delta f = \max_{d(DB,\, DB')=1} ||f(DB) - f(DB')||_1$

Theorem: Can achieve ε-differential privacy by adding scaled symmetric noise ~ $[\mathrm{Lap}(\Delta f/\varepsilon)]^k$.



Noise Grows (and must grow!) with Total Number of Queries

T Counting Queries: Δ = T

Multiple/ Complex Queries f: DB $\rightarrow$ R^k

$$\Delta f = \max_{d(DB,\, DB')=1} ||f(DB) - f(DB')||_2$$

Theorem: Can achieve (ε, δ) -differential privacy by adding noise $\sim \mathcal{N}(0, 2 \ln(2/\delta) (\Delta f/\varepsilon)^2)^k$.

T Counting Queries: $\Delta = \sqrt{T}$

Examples

- Simple Counting Queries
  - Extremely Powerful Computational Primitive

    Data inference, singular value decomposition, principal component analysis, k-means clustering, perceptron learning, association rules, ID3 decision tree, SQ learning model, approximate halfspaces, density estimation, …
- Histograms
  - A histogram looks like many queries, has low sensitivity!

    Data of any one person can change only 2 cells, each by 1.
- Contingency Tables
  - Each table is a histogram
  - Each marginal is a histogram
  - Can even get consistency across multiple marginals…

Release of Contingency Table Marginals

Privacy, Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release

- Barak, Chaudhuri, Dwork, Kale, McSherry, and Talwar, 2007

Release of Contingency Table Marginals

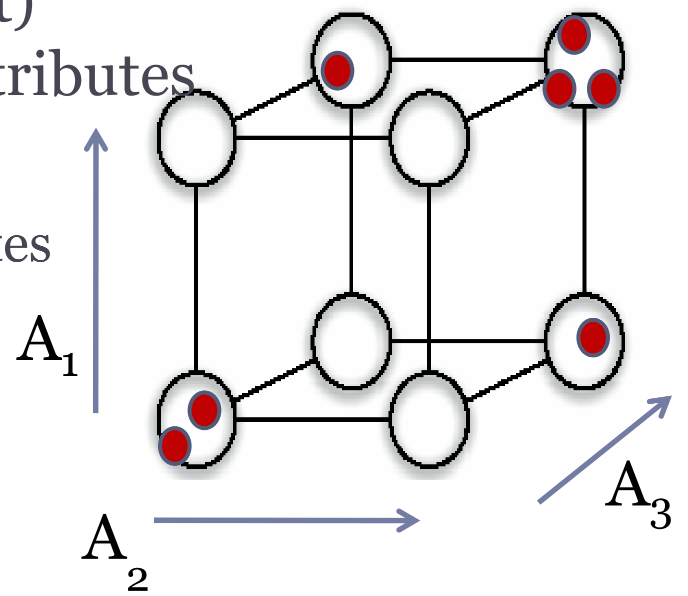
- Simultaneously ensure:
  - Consistency
  - Accuracy
  - Differential Privacy

Release of Contingency Table Marginals

- Simultaneously ensure:
  - Consistency
  - Accuracy
  - Differential Privacy
- Terms To Define:
  - Contingency Table
  - Marginal
  - Consistency
  - Accuracy

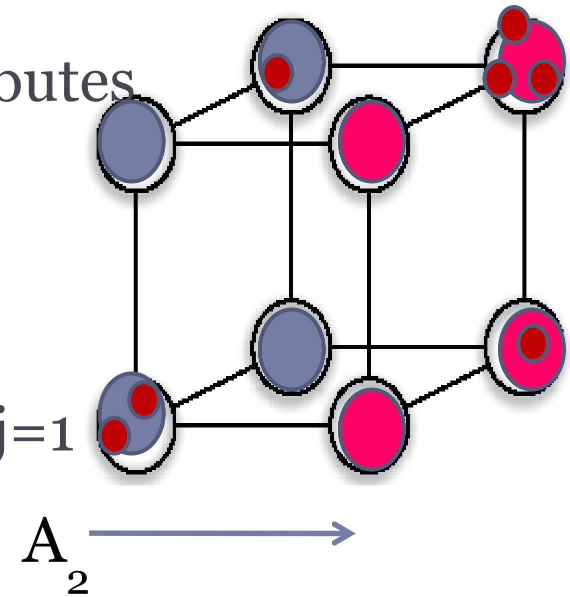
Contingency Tables and Marginals

- Contingency Table: Histogram / Table of Counts
  - Each respondent (member of data set) described by a vector of k (binary) attributes
  - Population in each of the 2^k cells
    - One cell for each setting of the k attributes



Contingency Tables and Marginals

- Contingency Table: Histogram / Table of Counts
  - Each respondent (member of data set) described by a vector of k (binary) attributes
  - Population in each of the 2^k cells
    - One cell for each setting of the k attributes
- Marginal: sub-table
  - Specified by a set of $j \leq k$ attributes, eg, j=1
  - Histogram of population in each of 2^j (eg, 2) cells
    - One cell for each setting of the j selected attributes
    - $A_2 = 0$: 3, $A_2 = 1$: 4, so the A_2 marginal is (3,4)



Consistency Across Reported Marginals

There exists a fictional contingency table T* whose marginals equal the reported marginals

- Marginals(T*) = Reported Marginals(T)
- Who cares about consistency?
  - Not we.
  - Software?

Release of Set M of Marginals

- Release noisy contingency table; compute marginals?
  - Consistency among marginals; differential privacy
  - Noise per cell of T: $Lap(1/\varepsilon)$
  - Noise per cell of M: about $2^{k/2}/\varepsilon$ for low order marginals

- Release noisy versions of all marginals in M?
  - Noise per cell of M: $Lap(|M|/\varepsilon)$
  - Differential privacy and better accuracy
  - Inconsistency among marginals

Consistency Across Reported Marginals

There exists a fictional contingency table T* whose marginals equal the reported marginals

- Marginals(T*) = Reported Marginals(T)
  - Can view T* (and its marginals) as synthetic data
  - T*, M(T*) may have negative and/or non-integral counts
- Who cares about integrality, non-negativity?
  - Not we.
  - Software?
  - See the paper.

Move to the Fourier Domain

- Just a change of basis. Why bother?
  - T represented by 2^k Fourier coefficients (it has 2^k cells)
  - To compute j-ary marginal only need 2^j coefficients
  - For any M, expected noise/cell depends on number of coefficients needed to compute M(T)
    - Independent of n and k
  - For M3 (all 3-way marginals): E[noise/cell] ≈ (k choose 3)/ε.

- The Algorithm for R(M(T)):
  - Compute set of Fourier coefficients of T needed for M(T)
  - Add noise; gives Fourier coefficients for M(T*)
    - 1-1 mapping between set of Fourier coefficients and tables ensures consistency
  - Convert back to obtain M(T*)
  - Release R(M(T))=M(T*)

Accuracy of Reported Values

- Roughly, described by $E[||R(M(T)) - M(T)||_1]$
  - Expected error in each cell: proportional to $|M|/\varepsilon$
  - A little worse
  - Probabilistic guarantees on size of max error

- Key Point: Error is Independent of n (and k)
  - Depends on the "complexity" of M
  - Depends on the privacy parameter ε

Improving Accuracy

- Gaussian noise, instead of Laplacian
  - E[noise/cell] for M3 looks more like $O((\log(1/\delta)^{1/2} k^{3/2}/\varepsilon)$
  - (ε,δ) -differential privacy

- Use Domain-Specific Knowledge
  - We have, so far, avoided this!
  - If most attributes are considered (socially) insensitive, can add less noise, and to fewer coefficients
    - Eg, ΔM3 with 1 sensitive attribute $\approx k^2$ (instead of k^3)
    - Reduce further using Gaussian noise: $\log(1/\delta)^{1/2} k$

What We Want to Learn

- Noise Reduction for Counting Queries
  - Is it necessary?
    - Can safely release answers to almost-linear number of counting queries with noise o(square root of population size). When is this too noisy? M3?
  - What is the correct interpretation of DiNi+ results?
    - Can't answer "too many" (weighted) subset sum queries "too accurately. But in M3 can't "zoom in" on a small subset of users and launch DiNi-style attacks.
    - There is a reasonable noise generation model for which, if want to bound even just over than half the queries to a small error p, and the coefficients can be as large as 2.1 p, then can attack any row using p-1 queries and $O(p^4)$ computation.

What We Want to Learn

- Noise Reduction for General Queries
  - Eg: Nissim, Raskhodnikova, Smith '07
    - Smoothed Sensitivity can be hard to work with
    - Subsample and Aggregate seems easier; powerful
  - Test-estimate-release [DL, in progress]
    - Use differentially private test for "nice" data; proceed iff nice

  - Not counting against sensitivity, or perturbing answers to, queries on non-sensitive data?
    - If, in a hypothetical world, sensitive data are *always* handled in a differentially private manner, maybe don't need to worry about insensitive fields being sufficient to identify an individual. That is, these can be used as a key, but so what?

What We Want to Learn

- Understand what it means *not* to provide ε-DiffeP
  - When is it a problem?
    - Failure to provide ε-DiffeP might result in 2ε-DiffeP
    - How bad is this?
  - Can this suggest a useful weakening?
    - Finite Differential Privacy?
    - How much residual uncertainty is enough?
  - (ϵ,δ) Differential Privacy when δ is non-negligible?
    - E.g, $1/n^2$ is very small when n is internet scale

What We Want to Learn

- Understand the relationship between robust statistics and Differential Privacy
  - Adam will say more about this
  - Understand what it means for statistical distributional assumptions to be false

- Differentially Private Algorithms for Statistical Tasks
  - Parameter estimation, regression, R, SAS?

What We Want to Learn

- Differential Privacy for Social Networks
- What Can Be Computed Insensitively?
- When can the Exponential Mechanism be efficient?
- Synthetic Data
  - Low-quality, low-sensitivity generation of synthetic set that will tell where to spend your privacy budget?