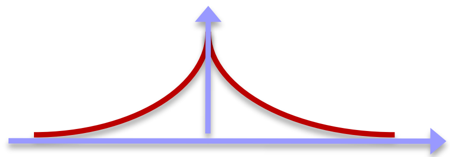
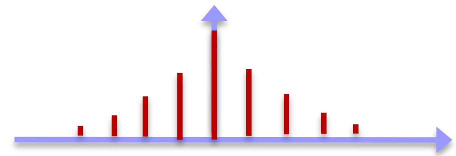




Differentially Private Mechanisms for Data Release

Graham Cormode

g.cormode@warwick.ac.uk

Joint work with

Magda Procopiuc (AT&T)

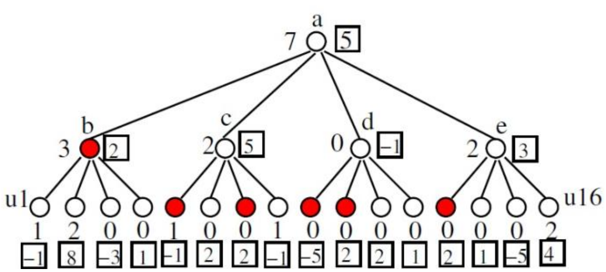
Divesh Srivastava (AT&T)

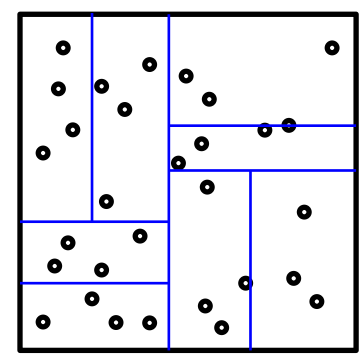
Entong Shen (NCSU)

Ting Yu (NCSU)

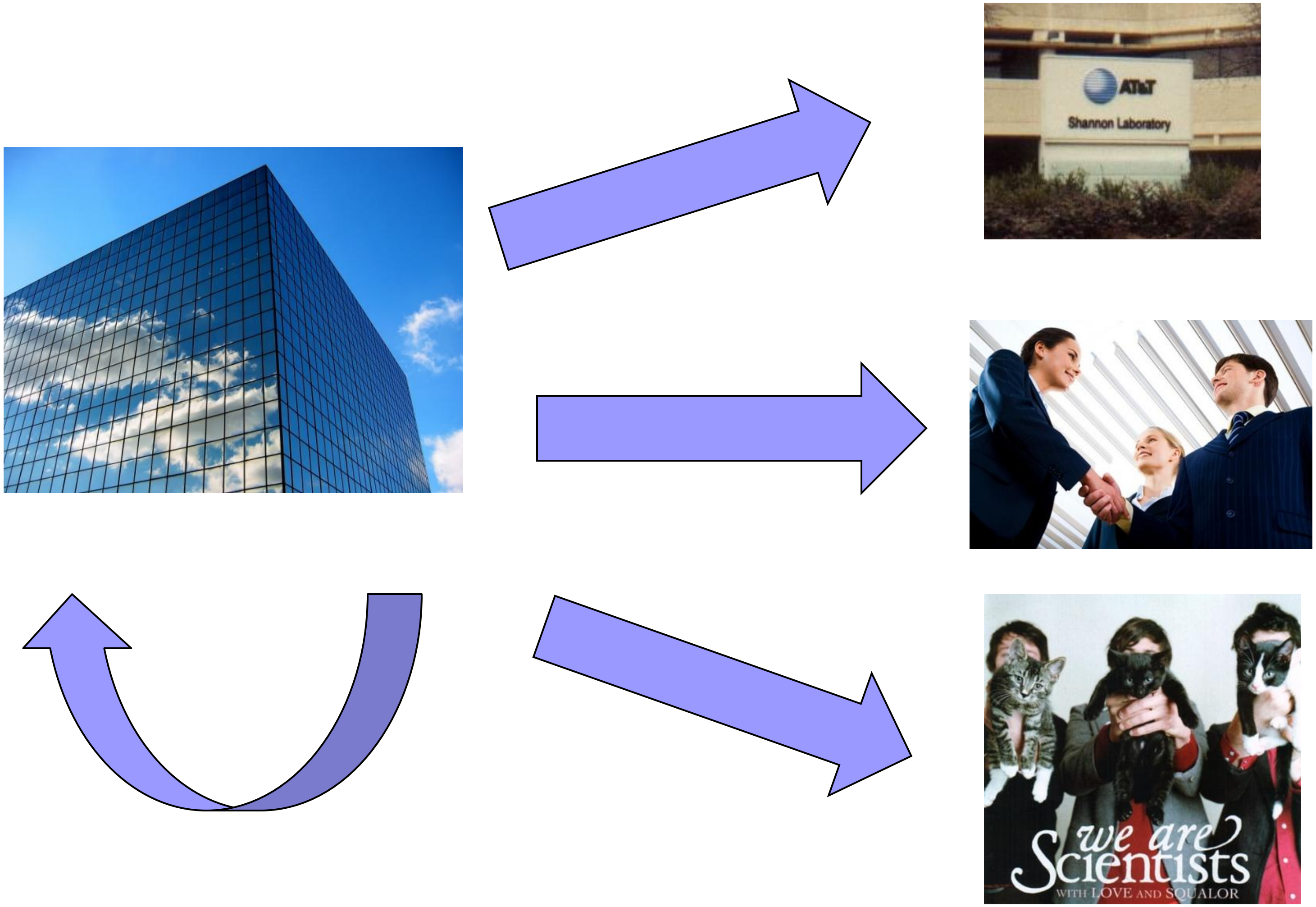
Xiaokui Xiao (NTU)

Jun Zhang (NTU)





The anonymization scenario



Data-driven privacy

- Much interest in private data release
  - Practical: release of AOL, Netflix data etc.
  - Research: hundreds of papers
- In practice, many data-driven concerns arise:
  - Efficiency / practicality of algorithms as data scales
  - How to interpret privacy guarantees
  - Handling of common data features, e.g. sparsity
  - Ability to optimize for known query workload
  - Usability of output for general processing
- This talk: outline some efforts to address these issues

Differential Privacy [Dwork 06]

- Principle: released info reveals little about any individual
  - Even if adversary knows (almost) everything about everyone else!
- Thus, individuals should be secure about contributing their data
  - What is learnt about them is about the same either way
- Much work on providing differential privacy
  - Simple recipe for some data types e.g. numeric answers
  - Simple rules allow us to reason about composition of results
  - More complex for arbitrary data (exponential mechanism)
- Adopted and used by several organizations:
  - US Census, Common Data Project, Facebook (?)

Differential Privacy

The output distribution of a differentially private algorithm changes very little whether or not any individual's data is included in the input – so you should contribute your data

A randomized algorithm K satisfies ε-differential privacy if:

Given any pair of neighboring data sets,

D_1 and D_2, and S in Range(K):

$$\Pr[K(D_1) = S] \leq e^{\varepsilon} \Pr[K(D_2) = S]$$

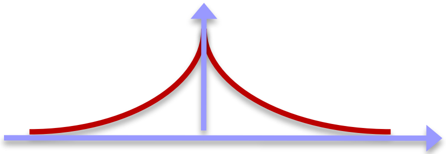
Achieving ε-Differential Privacy

(Global) Sensitivity of publishing:

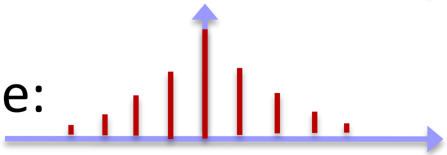
$s = \max_{x,x'} |F(x) - F(x')|$, x, x' differ by 1 individual

E.g., count individuals satisfying property P: one individual changing info affects answer by at most 1; hence $s = 1$

For every value that is output:



- Add Laplacian noise, $Lap(\varepsilon/s)$:
- Or Geometric noise for discrete case:



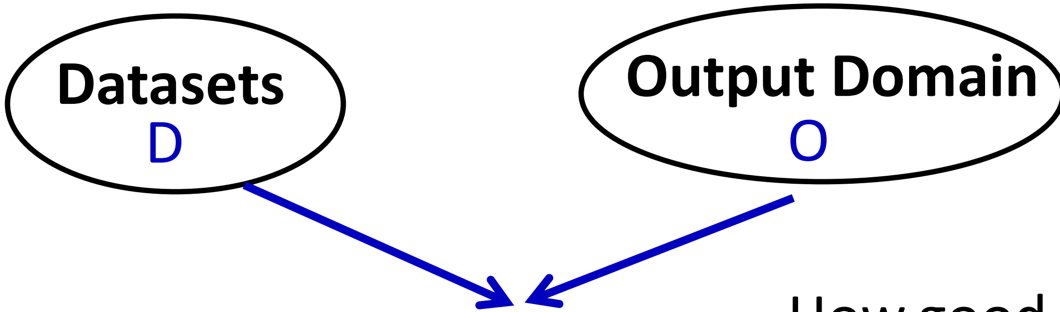
Simple rules for composition of differentially private outputs:

Given output O_1 that is ε_1 private and O_2 that is ε_2 private

- (Sequential composition) If inputs overlap, result is $\varepsilon_1 + \varepsilon_2$ private
- (Parallel composition) If inputs disjoint, result is $\max(\varepsilon_1, \varepsilon_2)$ private

Exponential Mechanism [MT07]

Given function F: Datasets $\rightarrow$ Outputs



Define score(D,O)$\in$R

How good O is as an answer to F(D)

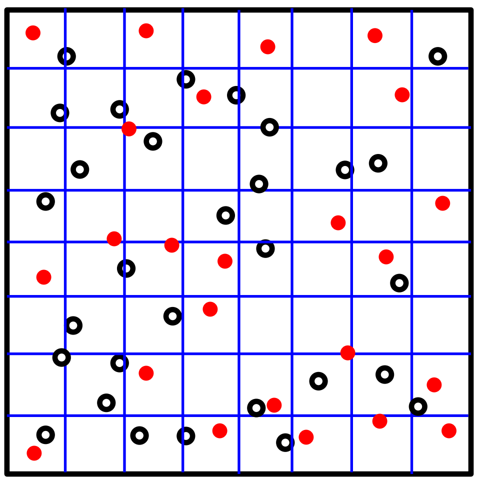
Exponential Mechanism: Return O with probability

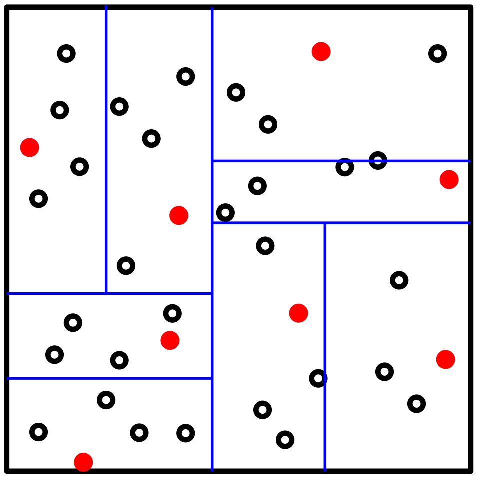
$$Pr[O] \propto exp(\frac{\varepsilon}{2\Delta} score(D,O))$$

where Δ = max |score(D,O) – score(D', O)| ,
taken over outputs O, neigbouring datasets D, D'

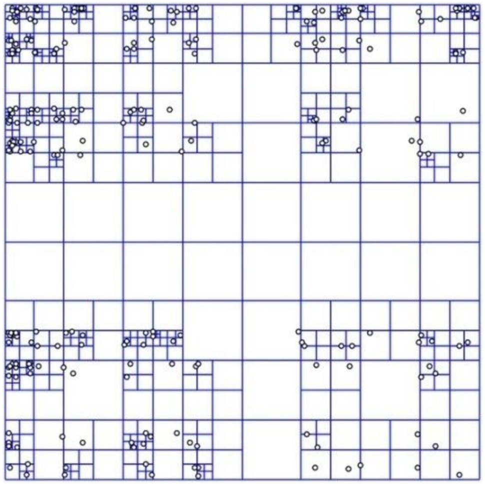
Sparse Spatial Data [ICDE 2012]

- Consider location data of many individuals
  - Some dense areas (towns and cities), some sparse (rural)
- Applying DP naively simply generates noise
  - lay down a fine grid, signal overwhelmed by noise
- Instead: compact regions with sufficient number of points

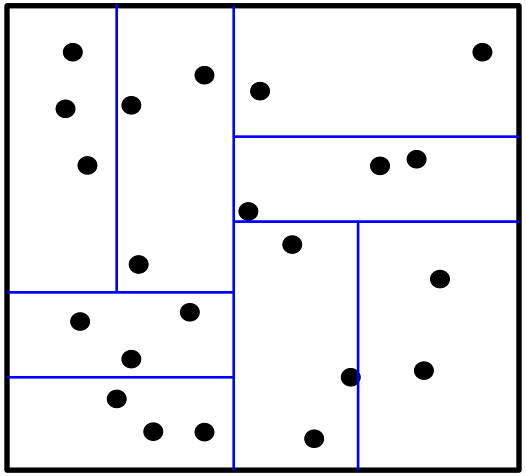




Private Spatial decompositions



quadtree



kd-tree

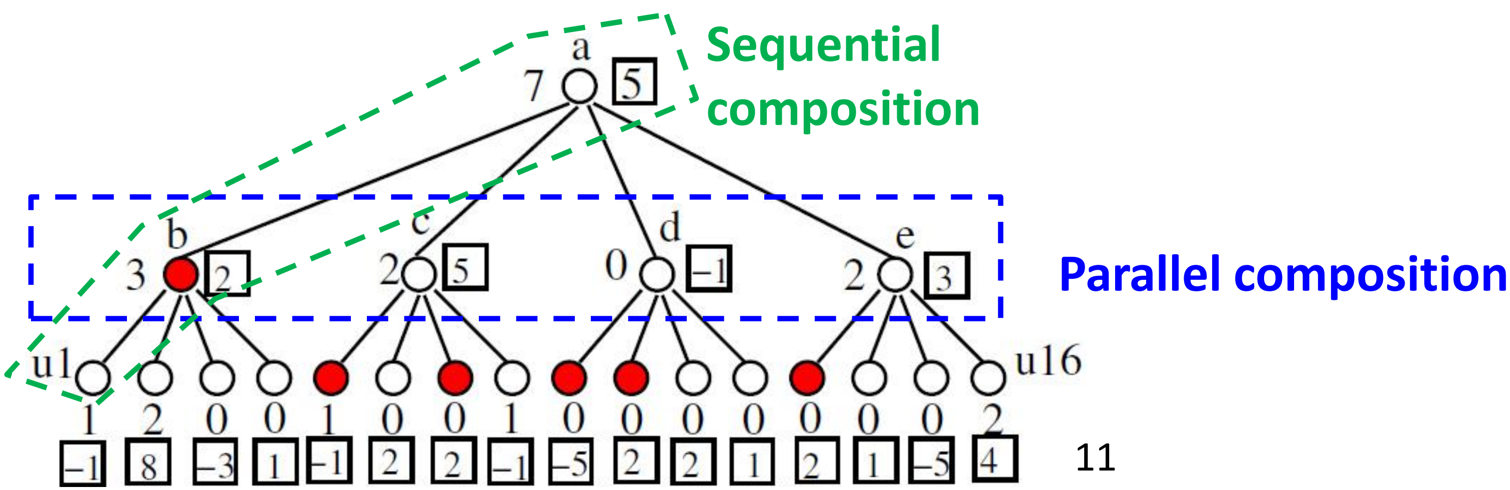
- Build: adapt existing methods to have differential privacy
- Release: a private description of data distribution (in the form of bounding boxes and noisy counts)

Building a Private kd-tree

- Process to build a private kd-tree
  - Input: maximum height *h*, minimum leaf size *L*, data set
  - Choose dimension to split
  - Get (private) median in this dimension
  - Create child nodes and add noise to the counts
  - Recurse until:
    - Max height is reached
    - Noisy count of this node less than *L*
    - Budget along the root-leaf path has used up
- The entire PSD satisfies DP by the composition property

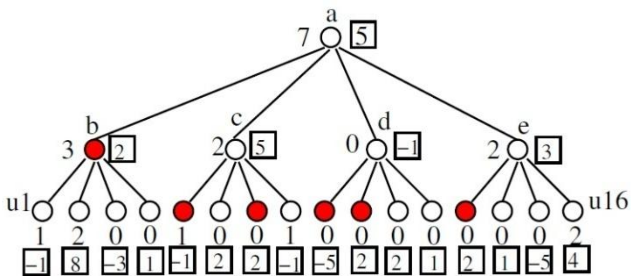
Building PSDs – privacy budget allocation

- Data owner specifies a total budget reflecting the level of anonymization desired
- Budget is split between medians and counts
  - Tradeoff accuracy of division with accuracy of counts
- Budget is split across levels of the tree
  - Privacy budget used along any root-leaf path should total ε



Privacy budget allocation

- How to set an ε_i for each level?
  - Compute the number of nodes touched by a 'typical' query
  - Minimize variance of such queries
  - Optimization: min $\sum_i 2^{h-i} / \varepsilon_i^2$ s.t. $\sum_i \varepsilon_i = \varepsilon$
  - Solved by $\varepsilon_i \propto (2^{(h-i)})^{1/3}\varepsilon$: more to leaves
  - Total error (variance) goes as $2^h/\varepsilon^2$
- Tradeoff between noise error and spatial uncertainty
  - Reducing h drops the noise error
  - But lower h increases the size of leaves, more uncertainty



Post-processing of noisy counts

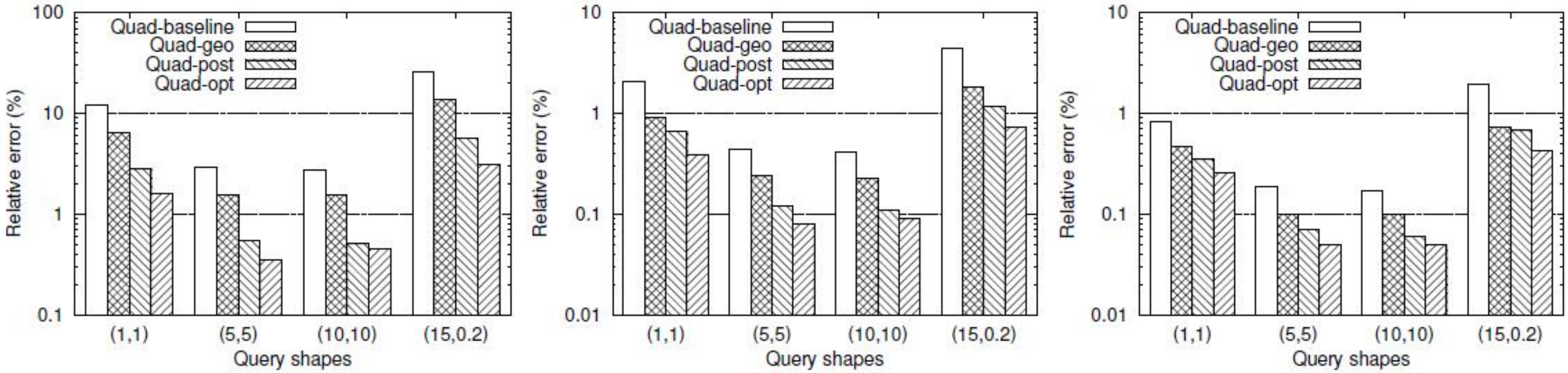
- Can do additional post-processing of the noisy counts
  - To improve query accuracy and achieve consistency
- Intuition: we have count estimate for a node and for its children
  - Combine these independent estimates to get better accuracy
  - Make consistent with some true set of leaf counts
- Formulate as a linear system in n unknowns
  - Avoid explicitly solving the system
  - Expresses optimal estimate for node v in terms of estimates of ancestors and noisy counts in subtree of v
  - Use the tree-structure to solve in three passes over the tree
  - Linear time to find optimal, consistent estimates

Experimental study

- 1.63 million coordinates from US TIGER/Line dataset
  - Road intersections of US States
- Queries of different shapes, e.g. square, skinny
- Measured median relative error of 600 queries for each shape

Experimental study

- Effectiveness of geometric budget and post-processing

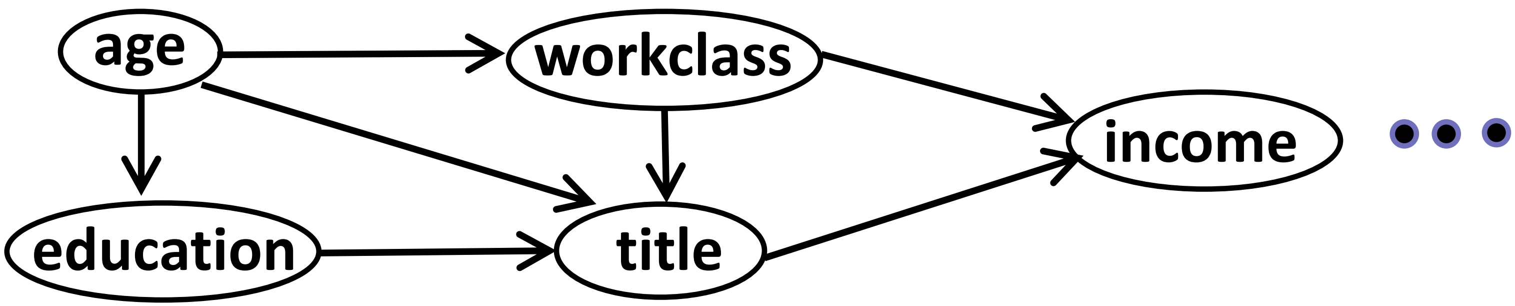


(a) $\varepsilon = 0.1$ (b) $\varepsilon = 0.5$ (c) $\varepsilon = 1.0$

  - Relative error reduced by up to an order of magnitude
  - Most effective when limited privacy budget

PrivBayes [SIGMOD14]

- Directly materializing a full distribution: low signal, high noise
- Use a **Bayesian network** to approximate the full-dimensional distribution by lower-dimensional ones:



$$\begin{aligned}\mathbf{Pr[H]} \quad \approx \quad & \mathbf{Pr[age]} \cdot \mathbf{Pr[education|age]} \cdot \mathbf{Pr[workclass|age]} \cdot \\ & \mathbf{Pr[title|age,education,workclass]} \cdot \mathbf{Pr[income|workclass,title]} \cdot \\ & \mathbf{Pr[marital\ status|age,income]} \cdots\end{aligned}$$

low-dimensional distributions: high signal-to-noise

PrivBayes (SIGMOD14)

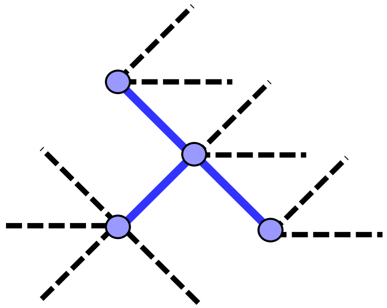
- **STEP 1:** Choose a suitable Bayesian Network BN
  - in a differentially private way
- **STEP 2:** Compute distributions implied by edges of BN
  - straightforward to do under differential privacy (Laplace)
- **STEP 3:** Generate synthetic data by sampling from the BN
  - post-processing: no privacy issues
- Evaluate utility of synthetic data for variety of different tasks

STEP 1: 1-degree BN [Chow-Liu'68]

- Optimal 1-degree BN maximizes $\sum_{(A_i,A_j):\ edge} MI(A_i, A_j)$ (MI: mutual information)
- Follows Prim's MST algorithm:
  - Pick arbitrary starting point $S = \{A_1\}$
  - For $i = 1 \ldots d-1$:

    Pick $e_i = (A_i, A_{i+1})$ to maximize $MI(e_i)$ where $A_i \in S$ and $A_{i+1} \notin S$

    Add e_i to BN, $S \leftarrow S \cup \{A_{i+1}\}$



- Use exponential mechanism to pick edge with high mutual information at each step
- For higher-degree BNs, pick a k'th order distribution at each step
  - Pick a set of parents A_i for A_{i+1} with high mutual information
- Problem solved?

Choosing an edge in BN

First attempt: define score(edge) = MI(edge)

- We prove $\Delta(MI) = \Theta(\log n / n)$, where $n = |D|$, size of the data
- Applying exponential mechanism, the MST algorithm chooses $e_i = (A_i \in S, A_{i+1} \notin S)$ with probability

$$Pr[e_i] \propto exp\left(\frac{\varepsilon n}{2} \frac{MI(e_i)}{\log n}\right)$$

- Problem: sensitivity $\Delta(MI)$ can be large compared to MI
  - Gives high chance of sampling an edge with low information
  - Can we find a better quality function for exponential mechanism?

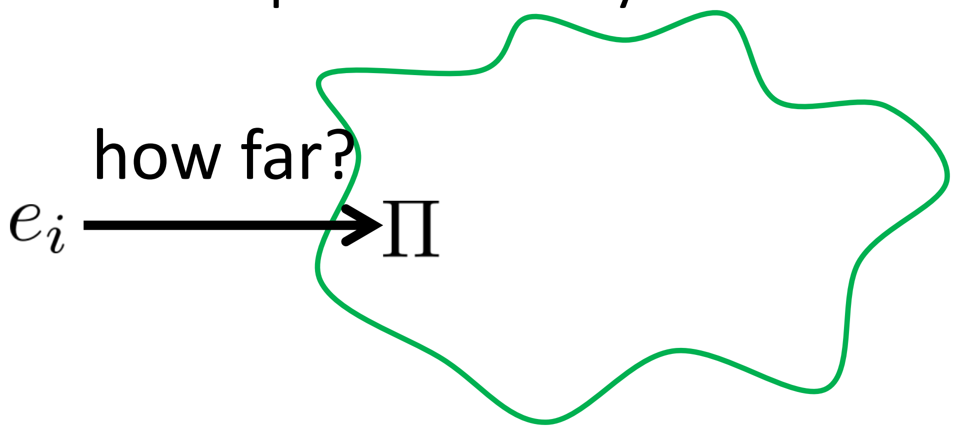
Defining a new score function

GOAL: $Pr[e_i] \propto exp\left(\frac{\varepsilon n}{2} score(e_i)\right)$

and large scores should correspond to large MI's

IDEA: define score to agree with MI at maximum values

and interpolate linearly in-between

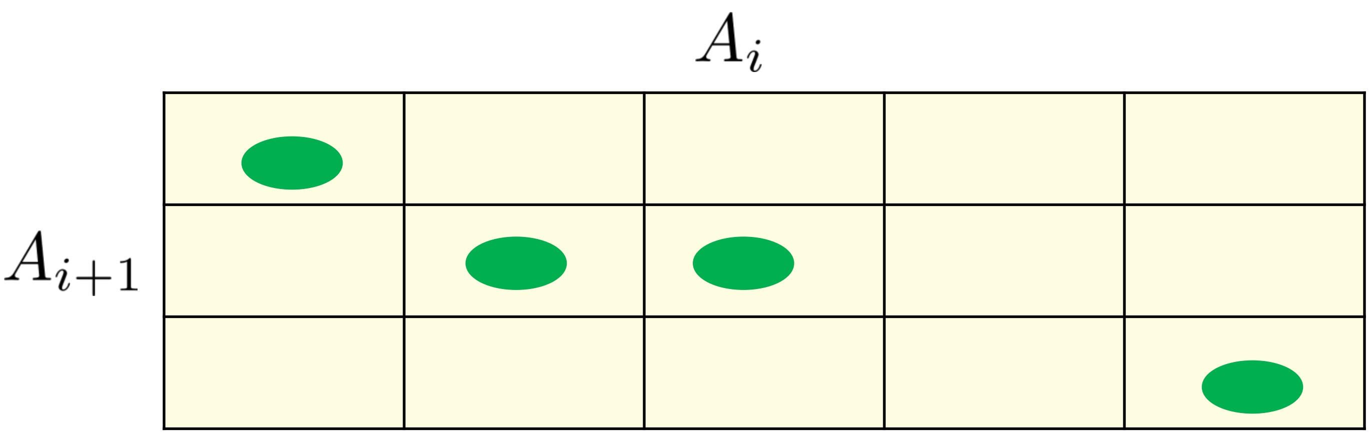


Π: "optimal" dbns over (A_i, A_{i+1}) that maximize $MI(A_i, A_{i+1})$

We define: $score(e_i) = -\frac{1}{2} \min_{\Pi:\ optimal} \|e_i - \Pi\|_1$

Δ(score) = 1/n by triangle inequality

Optimal Distributions



Can prove that necessary conditions for optimality are:

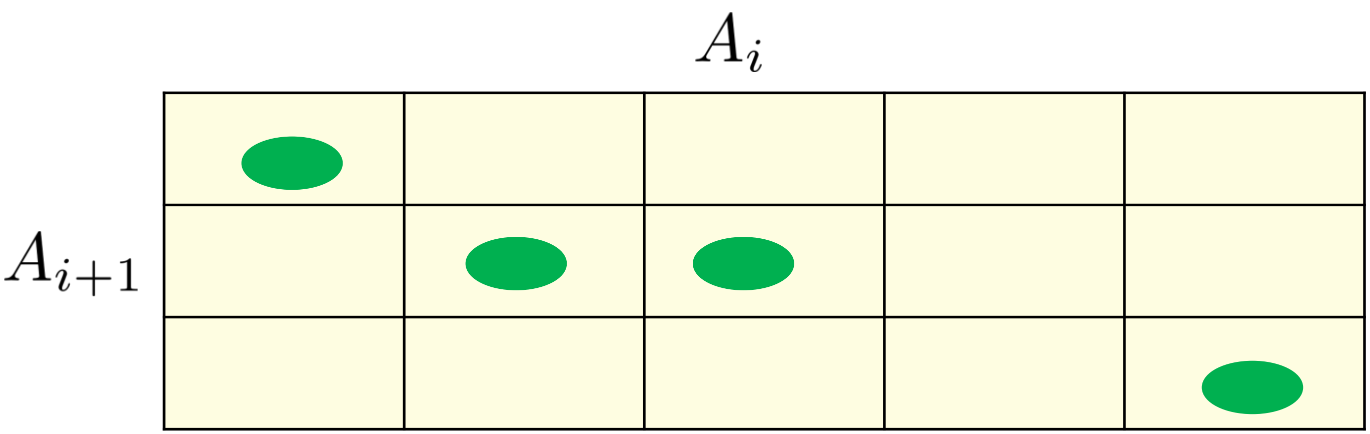
1. Uniform marginal:

$$\sum(\quad) = \frac{1}{|A_{i+1}|}$$ in each row

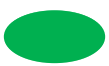
2. Sparse: At most one per column

Infinitely many such distributions!

Optimal Distributions



However, can show that

- size of the  does not matter;
- only their *position* matters

But still (doubly) exponentially many possibilities…

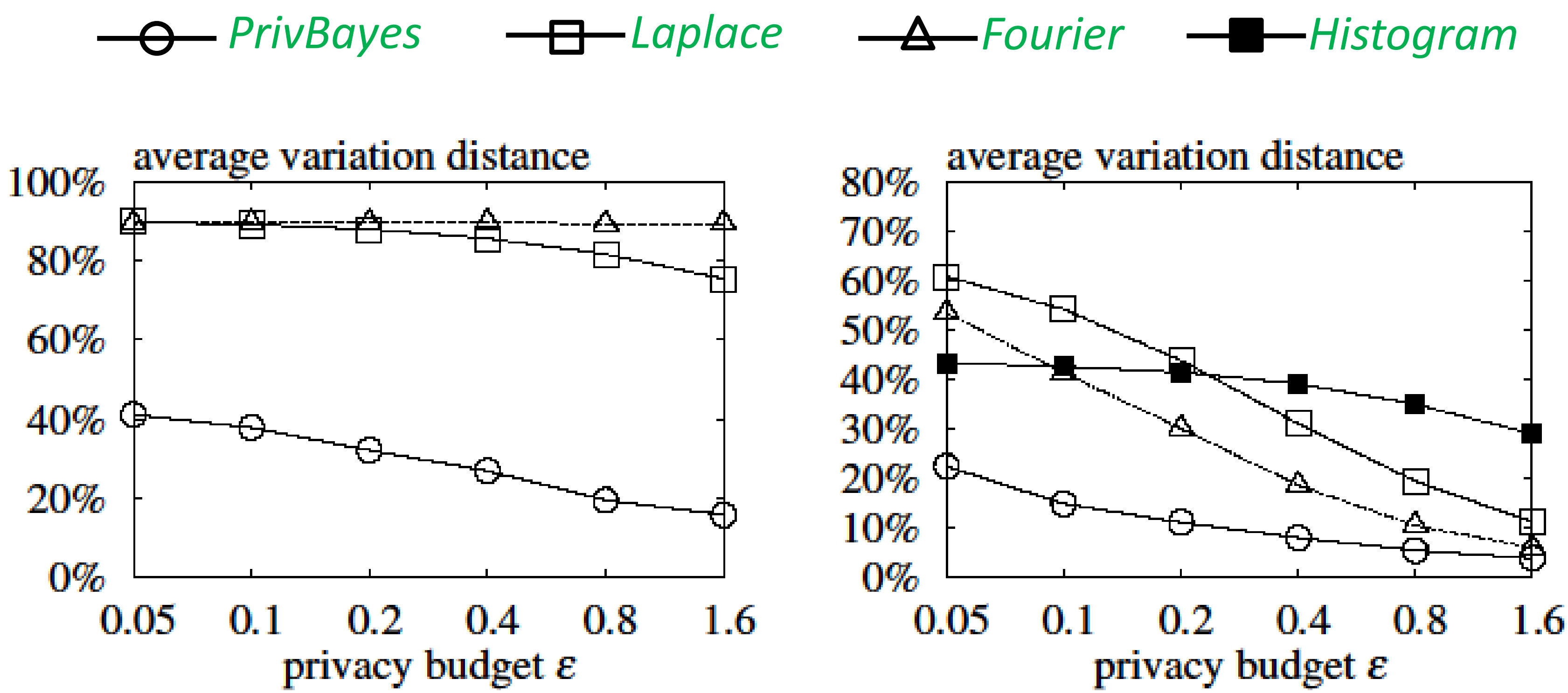
Define score(e_i) by discrete optimization over layouts

- General case: solved by Integer Program
- When $|A_{i+1}| = 2$: can solve by Dynamic Program

Naïve Bayes Summary

- To choose next distribution to materialize:
  - For each possible next child A_{i+1}
    - Find optimal distribution via discrete optimization (DP or IP)
    - Find score as L1 distance of $Pr[A_{i+1}, A_i]$ from optimal
  - Use exponential mechanism to pick next based on score
- Can pick the degree of the Bayesian network based on estimated noise (independent of data)
- Generate data from the released (private) Bayesian network
  - Plug into any desired application, e.g. classification, regression

Experiments: Counting Queries

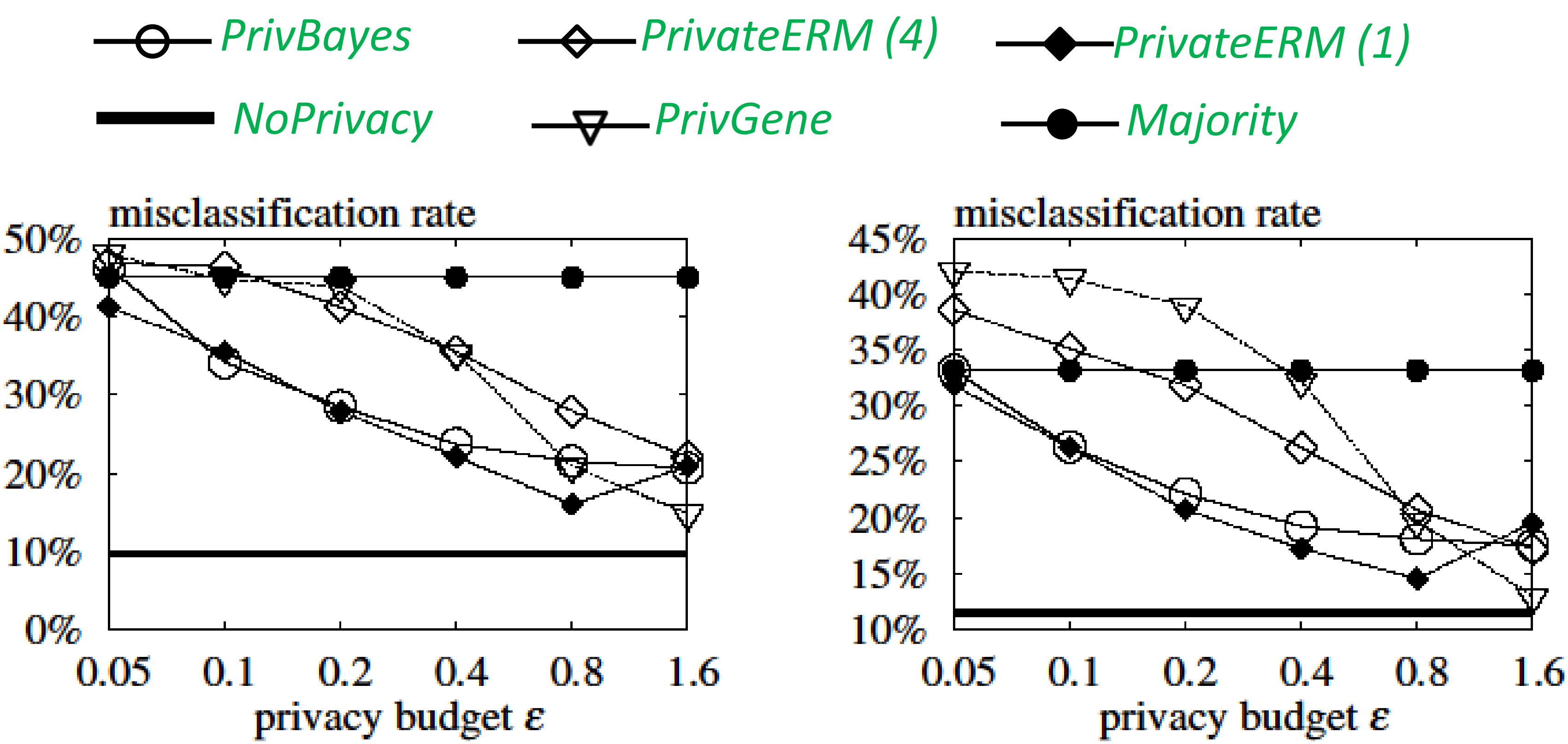


Adult dataset

NLTCS dataset

Query load = Compute all 3-way marginals

Experiments: Classification



Y = education: post-secondary degree?

Y = marital status: never married?

Adult dataset, build 4 classifiers

Concluding Remarks

- Differential privacy can be applied effectively for data release
- Solutions: classical techniques (e.g., sampling, kd-tree, BN) adapted to provide differentially privacy
  - With a different trade off: minimize the privacy cost
- Many open problems remain:
  - Transition these techniques to tools for data release
  - Extend to other forms of data: mobility data, graph data
  - Allow joining anonymized data sets accurately
  - Obtain alternate (workable) privacy definitions

Thank you!