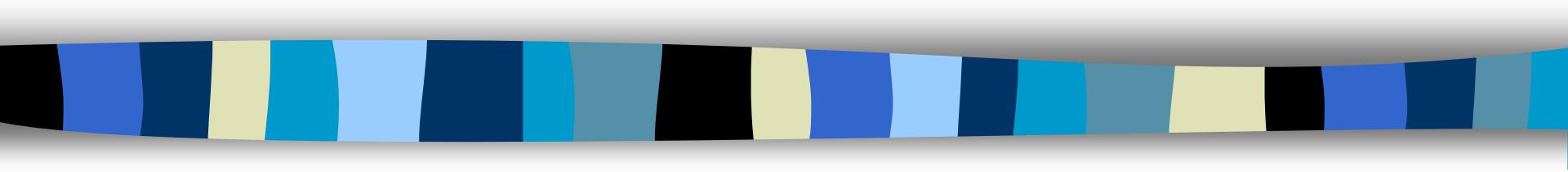
Data Security and Privacy



Topic 23: Publishing Marginals under Differential Privacy

What About High-Dimensional Data?

- Histogram publishing would not work
- It is infeasible to publish joint-distribution of all dimensions simultaneously
- Solutions:
  - Decompose the joint distribution into many smaller distribution
    - Similar in spirit to Probabilistic Graphical Models
  - Figure out which dimensions one cares about
    - As when mining frequent itemsets

Outline

- Background and motivation
- PriView: Practical Differentially Private Release of Marginal Contingency Tables
  - To appear in SIGMOD 2014
- PrivBasis: Mining Frequent Itemsets with Differential Privacy
  - In VLDB 2012.

Answering Marginal Queries: Problem Definition

- Given a d-dimensional binary dataset D, and a positive integer k < d, we want to differentially privately construct a synopsis of D, so that any k-way marginal table can be computed with reasonable accuracy
  - Assume d is large, e.g., between 30 and 200
  - And k is small, e.g., between 2 and 8

Relational Table

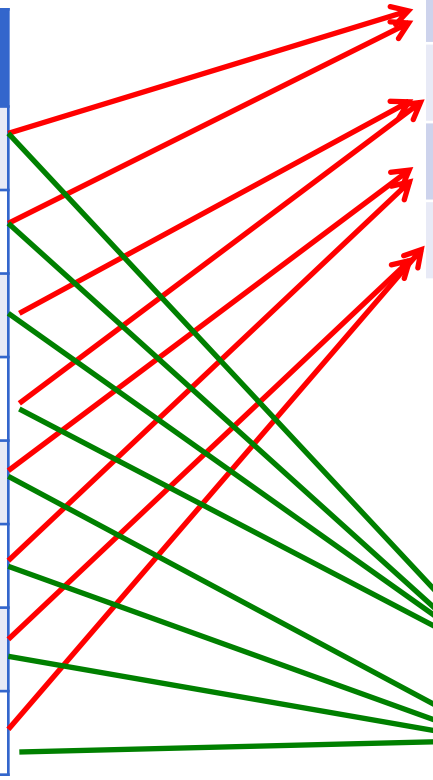
Name	Gender	Age	Income
Alice	Female	31	150k
Bob	Male	28	100k
Carlos	Male	30	110k
Dan	Male	45	200k
Eve	Female	19	50k
Frank	Male	24	40k

Contingency Table

Gender Male: 1 Female: 0	Age Larger than 25: 1 Otherwise : 0	Income More than 100k: 1 Otherwise: 0	Count
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	3

From Contingency Table to Marginal Tables

Gen	Age	Inc	Cnt
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	3



Gen	Age	Cnt
0	0	1
0	1	1
1	0	1
1	1	3

Gen	Cnt
0	2
1	4

Utility Metrics

- Utility: the generated k-way marginal should be close to the true values
  - Sum of Squared Errors (SSE)
  - Jensen-Shannon divergence

$$D_{\mathrm{KL}(P||Q)} = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i) \qquad M = \frac{P+Q}{2}$$

$$D_{\mathrm{JS}}(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M)$$

Direct method and Flat method

- Direct Method: Generate every k-way marginal tables
  - There are $\binom{d}{k}$ such tables, each requires a portion of the privacy budget
  - SSE is proportional to $\binom{d}{k}^2 2^k$
- Flat Method: Generate the complete contingency table
  - The complexity 2^d is infeasible
  - SSE is proportional to 2^d

A Middle Ground

- Publish a set of *Views*, each of size more than k and less than d
- E.g. d = 16, k = 2
  - Publish six 8-way marginal tables ensures that every pair is covered by one marginal
  - {1-4,5-8}, {1-4,9-12}, {1-4,13-16}, {5-8,9-12}, {5-8, 13-16}, {9-12,13-16}
  - Result in less noise than either direct or flat

PriView: The Steps

1. Choose the set of view coordinates
   - Each is a set of dimensions
2. Generate noisy marginals views
   - Add Laplacian noise to marginals
3. Consistency step
   - Make all noisy views consistent
4. Generate k-way marginals
   - Inferring from the noisy views

Step 1: Choosing the set of view coordinates

- We use covering design to choose a set of view coordinates such that any set of t dimensions is "covered" by at least one view
- Parameters:
  - t: balances noise errors and correlation errors
    - Optimal choice depends on dataset properties; t=2 works well empirically in our experiments
  - l: the number of dimensions in each view
  - t and l determines the number of views

Step 1: Choose the parameters

- We estimate the ESE to be on the order of

$$err = \frac{1}{N}\sqrt{\frac{\frac{2^{\ell+1}w^2}{\epsilon^2}}{\frac{w\ell(\ell-1)}{d(d-1)}}} = \frac{2^{\frac{\ell+1}{2}}}{N\epsilon}\sqrt{\frac{wd(d-1)}{\ell(\ell-1)}}$$

- Conclusions:
  - l should be around 8

Step 2: Generate Noisy Views

- For each set of dimensions, compute the corresponding marginal table, then add Laplace noise to all cells
  - Noise proportional to number of views
- The only step in PriView that needs direct access to the dataset.

Step 3: Consistency Step

- Perform constrained inference on the marginal tables
  - Ensures that any two noisy views are mutually consistent
  - Has two benefits: improve accuracy and enables query answering (step 4)
- In each step, ensures a set of views consistent on their intersection
- Use topological sort to decide ordering of steps

Step 3: Non-negativity and Consistency

- Negative number doesn't make sense
- Change negative numbers to zero introduces a bias and destroys consistency
- Our approach:
  - Ripple non-negativity: Turns negative counts into 0 while decreasing the counts for its neighbors to maintain overall count unchanged
  - Consistency + Non-negativity + Consistency

Step 4: Compute k-way Marginals

- Maximum Entropy
  - The probability distribution which best represents the current state of knowledge is the one with largest information-theoretical entropy

$$\begin{aligned} \text{maximize} \quad & -\sum_{a \in Q_A} \frac{\mathsf{T}_A(a)}{N_{\mathbf{V}}} \cdot \log\left(\frac{\mathsf{T}_A(a)}{N_{\mathbf{V}}}\right) \\ \text{subject to} \quad & \bigwedge_{a \in Q_A} \mathsf{T}_A(a) \geq 0 \\ & \bigwedge_{V_i \in \mathbf{V}} \bigwedge_{a' \in Q_{V_i \cap A}} \mathsf{T}_{V_i}(a') = \mathsf{T}_A(a') \end{aligned}$$

Experiment Datasets

Dataset	Dimension	Number of records
Kosarak	32	912,627
AOL	45	647,377
MSNBC	9	989,818
MCHAIN	64	1,000,000

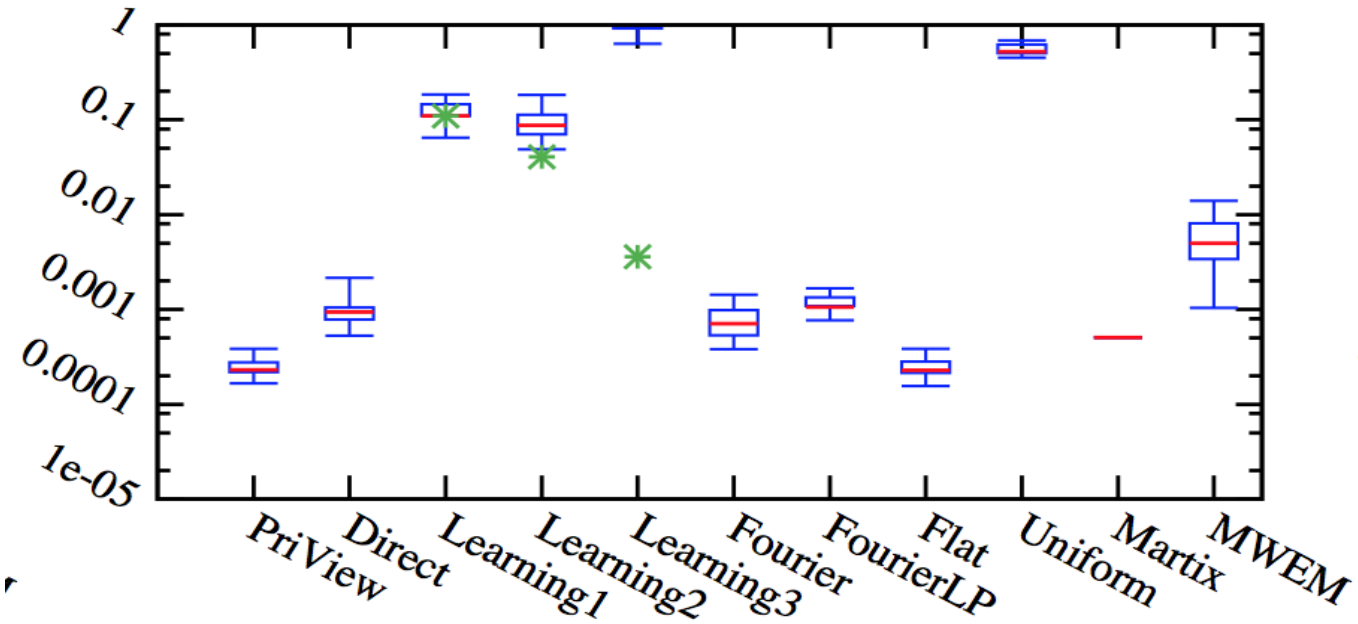
Other Methods

- Flat Method
- Direct Method
- Fourier Method[1]

- Data Cube[2]
- Matrix Mechanism[3]
- Multiplicative Weights Mechanism[4]
- Learning Based Approaches[5]

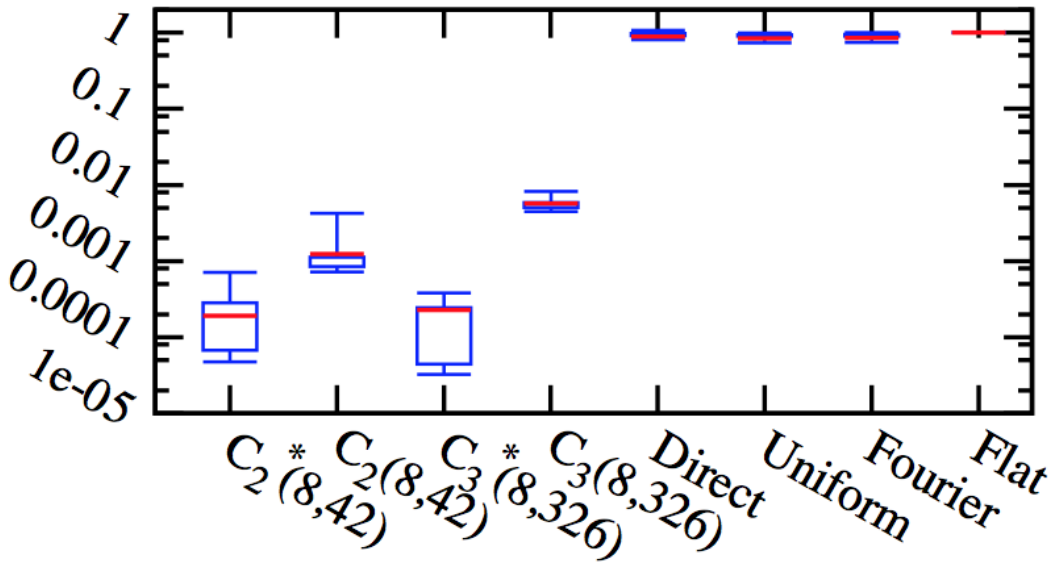
1. B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In PODS'07, pages 273–282, 2007.
2. B. Ding, M. Winslett, J. Han, and Z. Li. Differentially private data cubes: optimizing noise sources and consistency. In SIGMOD, pages 217–228, 2011.
3. C. Li and G. Miklau. An adaptive mechanism for accurate query answering under differential privacy. PVLDB, 5(6):514–525, Feb. 2012.
4. M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. In NIPS, pages 2348–2356, 2012.
5. J. Thaler, J. Ullman, and S. Vadhan. Faster algorithms for privately releasing marginals. In ICALP, pages 810–821, 2012.

Experimental Result: d=9



(c) $\epsilon = 0.1, k = 2$

Experimental Result: d=45



(i) $L_2, \epsilon = 1.0, k = 6$

Fourier Method

- Starting from Direct Method; tries to solve the inconsistency problem
- Instead of publishing noisy marginals, publishes noisy Fourier coefficients for constructing the marginals
- Any set of Fourier coefficients correspond to a full contingency table; however, the one corresponding to perturbed Fourier coeffcients may contain non-negative/non-integral values
- Linear programming can be used to find an integral contingency table close to perturbed coefficients

1. B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In PODS'07, pages 273–282, 2007.

Limitations of the Fourier Method

- Same accuracy for marginal queries compared with Direct method
- Linear program step solves for 2^d variables, can be carried out only when d is small
- When d is small, using Flat is pretty good

Learning Based Method

- Given a set of counting queries indexed using $\{0,1\}^L$, a dataset D can be viewed as a function f_D on L binary variables
  - And f_D can be viewed as sum of f_x for each $x \in D$
- For each tuple $x \in D$, compute a polynomial f'_x approximating f_x using the Chebyshev polynomials.
- Compute f'_D by summing up all f'_x and add Laplace noise to coefficients

J. Thaler, J. Ullman, and S. Vadhan. Faster algorithms for privately releasing marginals. In ICALP, pages 810–821, 2012.

Limitation of Learning-Based Methods

- Most mathematically interesting among the methods
- Becomes computationally expensive when d>9
- While asymptotic analysis shows that its accuracy is better than Direct, this occurs only when k>60 (recall that we are issing k-way marginal queries)

Multiplicative Update

- Starts with an initial full contingency table that is uniform, and then selects, using the exponential mechanism, a k-way marginal that is most incorrectly answered by the current distribution.
- One then obtains a noisy answer to the selected marginal, and updates the distribution to match the current state of knowledge.
- This process is repeated T times.

M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. In NIPS, pages 2348–2356, 2012.

Limitations of PriView

- Requiring covering every pair of dimensions does not scale to higher dimensions
  - With higher number of dimensions, one needs to choose which sets of dimensions to focus on
  - Or decides not to cover all pairs, e.g., using randomly generated views
- Utility depends on nature of dataset

Membership Privacy: A Unifying Framework For Privacy Definitions

Ninghui Li
Department of Computer Science and CERIAS
Purdue University

Joint work with
Wahbeh Qardaji, Dong Su, Yi Wu, Weining Yang

Two Versions of Differential Privacy

- Unbounded differential privacy
  - Two datasets are neighboring if one is obtained by adding one tuple to the other dataset
    - One has n tuples, one has n-1 tuples
- Bounded differential privacy
  - Two datasets are neighboring if one is obtained by replacing a tuple with another tuple
    - Both have the same number of tuples

Need for a General Privacy Framework

- Desire to relax differential privacy
  - Differential privacy under sampling [Li et al. 2012]
- A series of papers challenging differential privacy
  - Differential privacy is not robust to arbitrary background knowledge (Kifer and Machanavajjhala 2012)
  - Difficult to choose ε in DP, propose differential identifiability (Lee and Clifton 2012)
  - Differential privacy does not prevent attribute disclosure (Cormode 2012)

Why Membership Privacy?

- Privacy incidents demonstrate
  - Privacy violation = positive membership disclosure
- No membership disclosure means no attribute disclosure and no re-identification disclosure
- Membership privacy = Protect any tuple t's membership in input dataset

Formalizing Membership Privacy

- Adversary has some prior belief about the input dataset (modeled by a prob. dist. over all possible datasets)
  - Gives the prior probability of any t's membership
- Adversary updates belief after observing output of the algorithm, via Bayes rule
  - Obtains posterior probability of t's membership
- For any t, posterior belief should not change too much from prior
  - Whether this holds may depend on the prior distribution
- Membership privacy is relative to the family of prior distributions the adversary is allowed to have

Positive Membership Privacy

Definition (Positive Membership Privacy ($(\mathbb{D}, \gamma)$-PMP))

We say that a mechanism $\mathcal{A}$ provides γ-positive membership privacy under a family $\mathbb{D}$ of distributions over $2^{\mathcal{U}}$, i.e., $((\mathbb{D}, \gamma)\text{-PMP})$, where $\gamma \geq 1$, if and only if for any $S \subseteq \text{range}(\mathcal{A})$, any distribution $\mathcal{D} \in \mathbb{D}$, and any entity $t \in \mathcal{U}$, we have

$$\Pr_{\mathcal{D},\mathcal{A}}[t \in \mathbf{T} \mid \mathcal{A}(\mathbf{T}) \in S] \leq \gamma \Pr_{\mathcal{D}}[t \in \mathbf{T}] \quad (1)$$

$$\text{and } \Pr_{\mathcal{D},\mathcal{A}}[t \notin \mathbf{T} \mid \mathcal{A}(\mathbf{T}) \in S] \geq \frac{\Pr_{\mathcal{D}}[t \notin \mathbf{T}]}{\gamma} \quad (2)$$

where $\mathbf{T}$ is a random variable drawn according to the distribution $\mathcal{D}$.

E.g., γ=1.25, when $\Pr[t \in \mathbf{T}]$=0.8, $\Pr[t \in \mathbf{T} \mid A(\mathbf{T}) \in S] \leq \min(0.8*1.25, 1-0.2/1.25)$
$= \min(1, 1-0.16) = 0.84$

when $\Pr[t \in \mathbf{T}]$=0.2, $\Pr[t \in \mathbf{T} \mid A(\mathbf{T}) \in S] \leq \min(0.2*1.25, 1-0.8/1.25)$
$= \min(0.25, 1-0.64) = 0.25$

Negative Membership Privacy

- Positive membership privacy bounds the ability to conclude a tuple t is in the input dataset; it is allowed to conclude that a tuple t is not in the dataset

- Negative membership privacy is analogously defined to bound the increase in posterior probability that a tuple is not in the dataset

Results from Membership Privacy

- Membership privacy for the family of all possible distributions is infeasible
  - Requires publishing similar output distributions for two completely different datasets
  - Output has (almost) no utility

- Moral: One has to make some assumptions about the adversary's prior belief
  - Assumptions need to be clearly specified and reasonable

Differential Privacy as Membership Privacy

- Unbounded differential privacy is equivalent to (positive + negative) membership privacy under the family of all Mutually Independent (MI) distributions
  - Each MI distribution can be written as
    $\Pr[T] = \prod_{t \in T} p_t \prod_{t \notin T} (1-p_t)$ where there is p_t for each t

- Bounded differential privacy is equivalent to (positive + negative) membership privacy under the family of all distributions obtained by restricting MI distributions to allow only distributions of a fixed length

- Differential privacy insufficient for membership privacy without independence assumption

Differential Identifiability [Lee & Clifton 2012] as Membership Privacy

DEFINITION 5.4. *(ρ-differential identifiability (DI) [20]) A mechanism $\mathcal{A}$ is said to satisfy ρ-DI if for any dataset T, any entity $t \in T$, let $T' = T \setminus \{t\}$, for any output event S*

$$\Pr[t \in \mathbf{T} | \mathcal{A}(\mathbf{T}) \in S, T'] \leq \rho,$$

where $\mathbf{T}$ is a random variable drawn from the following distribution: each dataset $T' \cup \{t'\}$, where $t' \in \mathcal{U} \setminus T'$ is equally likely.

- Intuition: For each t, posterior membership prob bounded by ρ, assuming prior is such that t can be replaced with any tuple not already in T
- Equivalent to positive membership privacy for a sub-family of that corresponding to bounded differential privacy

A New Privacy Notion

- Membership privacy under the single uniform distribution (each tuple has prob 0.5) is a new notion that enables better utility
  - max can now be answered with high accuracy

Membership Privacy Notions We Considered

All distributions:
Privacy with Almost no Utility

Mutually Independent (MI) dist:
- **Pr[T] = $\Pi_{t \in T} p_t \ \Pi_{t \notin T} (1-p_t)$**
- **Unbounded Differential Privacy**

Bounded Mutually Independent dist.:
- **MI distributions conditioned on all datasets have the same size**
- **Bounded Differential Privacy**

- **MI distributions where each p_t is either 0 or β**
- **Differential Privacy Under Sampling**

- **MI distributions where each p_t is either 1 or 1/m, where m is number of t's have probability ≠1**
- **Differential Identifiability**

- **MI distributions where each p_t is 1/2**
- **New privacy notion**

Other Related Work

- Pupperfish privacy [Kifer & Machanavajjhala 2012]
  - Require specifying a set of potential secrets, set of discriminative pair of DBs
  - Generalizes DP to allow defining which pairs of DBs result in close output distribution
- Coupled-world privacy [Bassily et al. 2013]
  - Require D and $D_{\_t}$ result in close output distribution, where D is drawn from some distribution

Conclusions

- We have introduced membership privacy framework
  - Motivated by real-world privacy incidents
  - Captures what the society views as privacy violations
- Membership privacy framework improves analyzing/understanding existing notions
- Membership privacy framework provides a principled way to define new privacy notions for better privacy/utility tradeoff

Next Lecture

- Full Homomorphic Encryption