Faster Private Release of Marginals on Small Databases

Karthekeyan Chandrasekaran[1], Justin Thaler[2], Jonathan Ullman[3], Andrew Wan[2]

[1] Simons Postdoc - Harvard University, [2] Postdoc - Simons Institute for the Theory of Computing,
[3] Postdoc - Harvard University

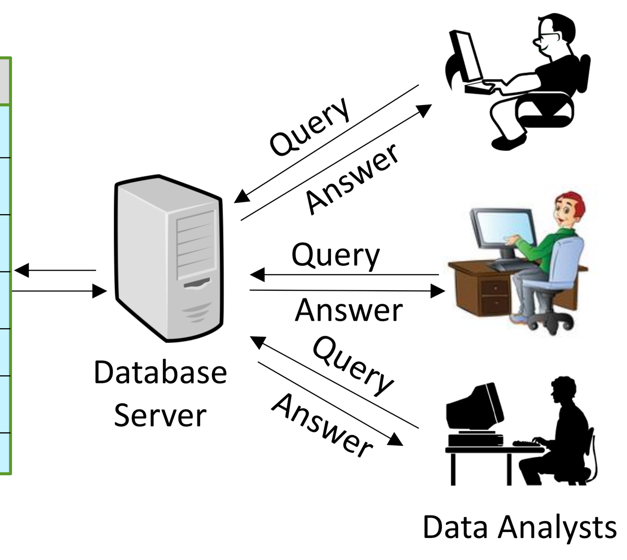
Privacy Tools for Sharing Research Data

A National Science Foundation Secure and Trustworthy Cyberspace Project

Introduction

Male	Asian	Age 25-30	Smoker	Married
1	1	1	0	0
0	1	0	0	0
1	1	0	1	0
1	1	1	1	1
1	0	0	1	1
0	0	0	1	1
1	0	1	1	1

Database with binary attributes



- Data analysts need statistical information about the database
- Analysts ask counting queries
  - E.g., What fraction of the patients are smokers?
- Most queries from the analysts focus on a small number of attributes
  - *k-way marginal queries*: What fraction of the records have a given set of values for a chosen set of k attributes
- Goal of the server: Return accurate answers without revealing information about individual records

Differential Privacy

- Database D contains n records, each containing d attributes, $D \in (\{0,1\}^d)^n$
- Query family Q
- Server employs a randomized algorithm $A: (\{0,1\}^d)^n \to [0,1]^{|Q|}$
- (ϵ, δ)-Privacy: For any database D, every record $x \in D, y \in \{0,1\}^d$, $D' = D - x + y$, for every subset $S \subseteq [0,1]^{|Q|}$,
  $\Pr(A(D) \in S) \le e^{\epsilon} \Pr(A(D') \in S) + \delta$
- $(0.01, 0.01)$-Accuracy: For every query q in the family Q of queries, $|A(D)[q] - q(D)| \le 0.01$
  with probability at least 0.99 for every database D

Related Work

- Accuracy and Privacy are conflicting goals
- Conflicting nature is especially perceived in databases with small number of records

		Minimum DB size needed for constant accuracy	Runtime per query
Algorithms designed to answer arbitrary counting queries	[DN03, BDMN05, DMNS06]	$\tilde{O}(d^{\frac{k}{2}})$	$O(1)$
	[BLR08]	$\tilde{O}(kd)$	$2^{\tilde{O}(kd)}$
	[HR10, GRU12]	$\tilde{O}(kd^{\frac{1}{2}})$	$2^{O(d)}$

[U13] Any algorithm that returns accurate and private answers for arbitrary counting queries in small DBs needs $2^{O(d)}$ run-time

$\cup$|

[BUV13] $\Omega(kd^{\frac{1}{2}})$ is a lower bound on the size of the database needed to guarantee constant accuracy for k-way marginals

		Minimum DB size needed for constant accuracy	Runtime per query
k-way Marginal queries	[HRS12, TUV12]	$d^{O(\sqrt{k})}$	$d^{O(\sqrt{k})}$

Results

Question: Can we exploit structure of k-way marginal queries to design *faster* private algorithms that are accurate on databases of size $\tilde{O}(kd^{\frac{1}{2}})$?

		Minimum DB size needed for constant accuracy	Runtime per query
k-way Marginal queries	This work	$kd^{\frac{1}{2}+o(1)}$	$2^{O\left(\frac{d}{\log^{0.99} d}\right)}$
	This work	$\tilde{O}(kd^{\frac{1}{2}+0.01})$	$2^{d^{1-\Omega\left(\frac{1}{\sqrt{k}}\right)}}$

k-way Disjunction Query

- Specified by a subset $S \subseteq [d]$ of size at most k
- Answer is the fraction of records x in the database for which at least one of the attributes of the set S is $TRUE$

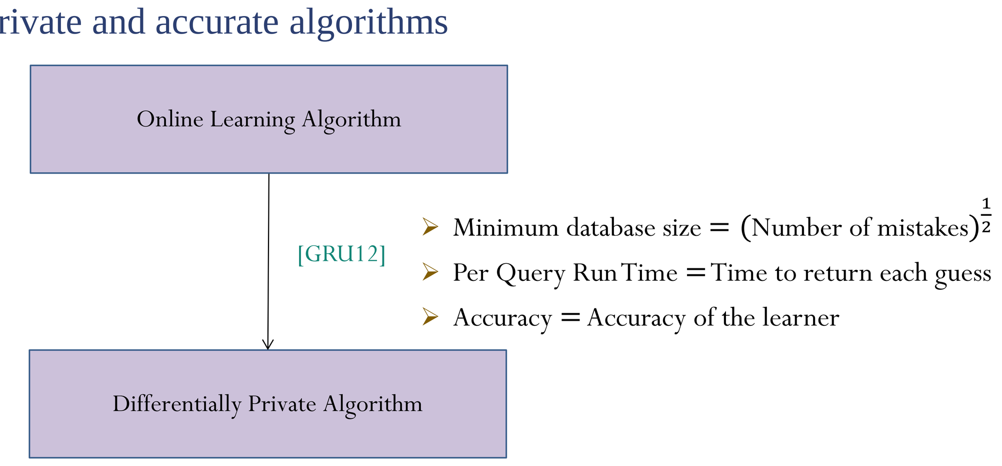
Approach

(1) Learning Approach to Designing Private Algorithms

- For each $D \in (\{0,1\}^d)^n$, there is an underlying function $f_D: \{0,1\}^d \to [0,1]$
- Input to the function is the indicator vector s of the query set S

k-way disjunction queries $\Rightarrow$ inputs to the function have at most k bits set to $TRUE$

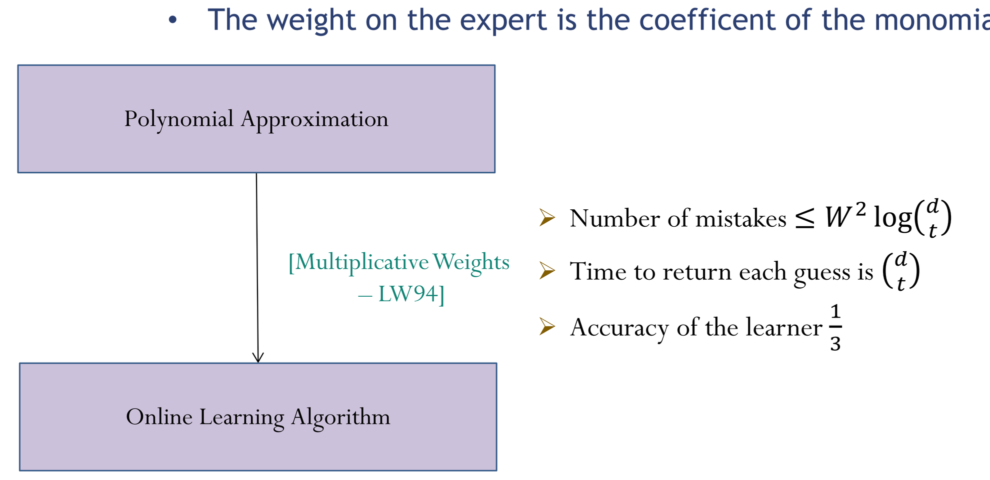
- The function value $f_D(s)$ is the answer to the query S on the database D
- [GRU12] Good online learning algorithm to learn a hypothesis that evaluates close to $f_D(s)$ on the queries of interest can be used to derive private and accurate algorithms



- Minimum database size = (Number of mistakes)$^{\frac{1}{2}}$
- Per Query Run Time = Time to return each guess
- Accuracy = Accuracy of the learner

(2) Polynomial Approximations for Learning

- Suppose we have a polynomial p_D for each function $f_D: \{0,1\}^d \to [0,1]$
  - $|p_D(s) - f_D(s)| \le 0.01$ for every input s
  - Sum of absolute values of the coefficients of $p_D(s)$ is at most W
  - Degree$(p_D) \le t$
- Can derive a learning algorithm:
  - Given samples $(s, f_D(s))$, need to learn a hypothesis h satisfying $|h(s) - f_D(s)| \le 0.01\ \forall s$
  - Find a hypothesis among the possible p_D
    - i.e., learn the coefficients of the polynomial
  - Use Multiplicative Weights Method
    - Each monomial is an expert
    - The weight on the expert is the coefficent of the monomial



- Number of mistakes $\le W^2 \log\binom{d}{t}$
- Time to return each guess is $\binom{d}{t}$
- Accuracy of the learner $\frac{1}{3}$

Polynomial Approximations

Let $H_k \subset \{0,1\}^d$ be the subset of inputs with at most k bits set to TRUE

Goal: Find a polynomial p_D for each function $f_D: H_k \to [0,1]$

- $|p_D(s) - f_D(s)| \le 0.01$ for every input $s \in H_k$
- Sum of absolute values of the coefficients of $p_D(s)$ is at most W
- Degree$(p_D) \le t$

Question: What is the least possible W and t?

Simplifying the problem

- Sufficient to find approximating polynomials for $OR_x: H_k \to \{0,1\}$

$$OR_x(s) = \bigvee_{i \in S} x_i$$

  - This is because $f_D(s) = \frac{1}{n}\sum_{x \in D} OR_x(s)$
  - Average of approximating polynomials gives the required approximating polynomial for f_D
- In fact, sufficient to find *one* approximating polynomial for $OR: H_k \to \{0,1\}$

$$OR(s) = \begin{cases} 1, & s \ne \emptyset \\ 0, & o.w. \end{cases}$$

  - Can express the OR_x functions in terms of OR and vice-versa
- Seeking low-weight low-degree polynomial to approximate disjunction over low-Hamming weight inputs

Results for the polynomial approximation problem

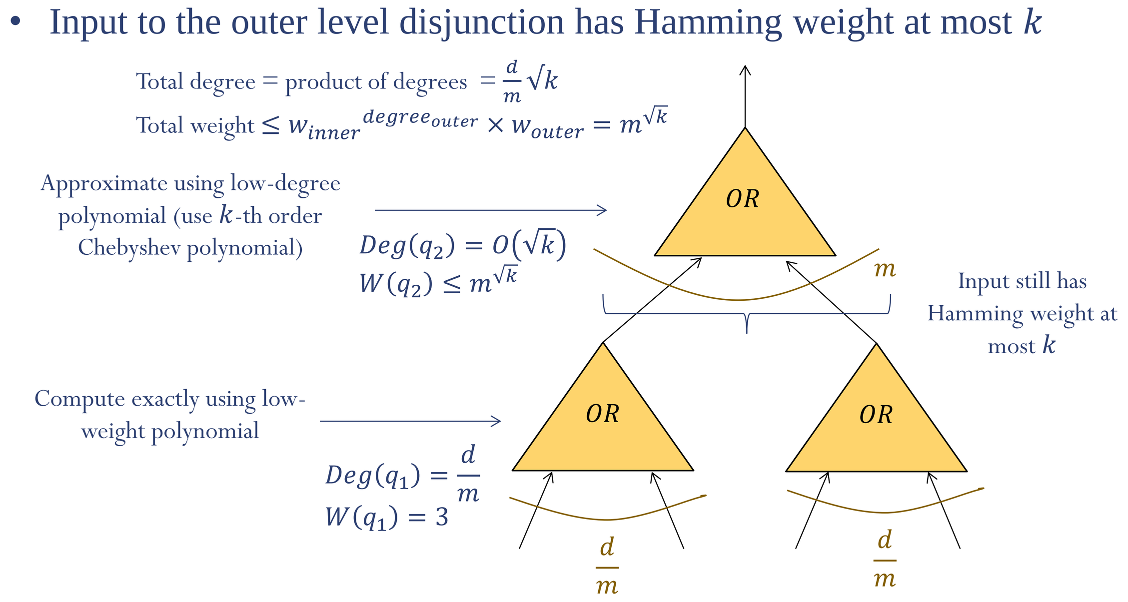
Explicit Construction to achieve

- $W = poly(d)$
- $t = \min\left\{d^{1-\frac{1}{\sqrt{k}}}, \frac{d}{\log^{0.99} d}\right\}$

Lower bound for $k = o(\log d)$

- If $W = poly(d)$, then $t \ge d^{1-\frac{1}{\sqrt{k}}}$

- Construction: View disjunction as a disjunction of m disjunctions (choose $m = d^{1-\frac{1}{\sqrt{k}}}$ to optimize the final parameters)
- Input to the outer level disjunction has Hamming weight at most k



- Lower bound: express the existence of the polynomial as a linear program
- Show infeasibility by constructing a feasible solution to the dual
- Dual construction by combining dual solutions witnessing:
  - Any Polynomial approximating OR has high degree
  - Any low-degree polynomial that approximates OR on inputs of Hamming weight at most "1" has large weight

Future Directions

- Used polynomial approximations for $(OR_x)_{x \in \{0,1\}^d}$ to derive good online algorithms and in turn *faster* private and accurate algorithms
- The polynomial approximations can be viewed as linear combination of monomials
- A linear combination of some other small set of functions with similar properties can be used by the same approach to improve run-time
- Is there a collection of functions $\Gamma_1, \Gamma_2, \ldots, \Gamma_r: H_k \to [-1,1]$ such that
  - For each $x \in \{0,1\}^d$, there exists a linear combination $p_x(s) = \sum_{j=1}^{r} \Gamma_j(s).c_j^x$
  - $|p_x(s) - OR_x(s)| \le 0.01\ \forall s \in H_k, \forall x \in \{0,1\}^d$
  - $\sum_{j=1}^{r} |c_j^x| \le W\ \forall x \in \{0,1\}^d$
  - $r = O(d^{\sqrt{k}}), W = O(kd^{\frac{1}{2}})$?

References

[BDMN05] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In PODS, pages 128–138, 2005.

[BLR08] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In STOC, pages 609–618, 2008.

[BUV13] Mark Bun, Jonathan Ullman, Salil P. Vadhan. Fingerprinting Codes and the Price of Approximate Differential Privacy. arXiv: 1311.3158.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In TCC '06, pages 265–284, 2006.

[DN03] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In PODS, pages 202–210, 2003.

[GRU12] Anupam Gupta, Aaron Roth, and Jonathan Ullman. Iterative constructions and private data release. In TCC, pages 339–356, 2012.

[HR10] Moritz Hardt and Guy N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In FOCS, pages 61–70, 2010.

[HRS12] Moritz Hardt, Guy N. Rothblum, and Rocco A. Servedio. Private data release via learning thresholds. In SODA, pages 168–187, 2012.

[LW94] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. Information and Computation, 108(2): 212–261, 1994.

[TUV12] Justin Thaler, Jonathan Ullman, and Salil P. Vadhan. Faster algorithms for privately releasing marginals. In ICALP, pages 810–821, 2012.

[U13] Jonathan Ullman. Answering $n^{2+o(1)}$ counting queries with differential privacy is hard. In STOC, pages 361–370, 2013.

Contact

{karthe,jthaler, jullman, atw12}@seas.harvard.edu

- Karthekeyan Chandrasekaran – supported by Simons Fellowship
- Justin Thaler – supported by NSF Graduate Research Fellowship and NSF grants CNS-1011840 and CCF-0915922.
- Jonathan Ullman – supported by NSF grant CNS-1237235 and Siebel Scholarship
- Andrew Wan – supported by NSF grant CCF-0964401 and NSFC grant 61250110218

CRCS Center for Research on Computation and Society at Harvard School of Engineering and Applied Sciences

BERKMAN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

The Institute for Quantitative Social Science at Harvard University

DATA PRIVACY LAB