Faster Algorithms for Private Data Release

Anna Gavrilman†, John Ullman††, and Salil Vadhan†† annagav@cs.umb.edu, {jullman,salil}@seas.harvard.edu

† Class of 2014, Dept of Computer Science, University of Massachusetts Boston. Supported in part by an NSF REU Program.

††, School of Engineering and Applied Science Center for Research and Computation and Society, Harvard University.

Privacy Tools for Sharing Research Data

A National Science Foundation Secure and Trustworthy Cyberspace Project

Executive Summary

• Marginal queries enable a rich class of statistical analysis of a dataset, and designing **computationally efficient** algorithms for privately releasing marginal queries has become an important problem.

•The answer to a k-way marginal query is a fraction of records with at least one "1" among a set of up to k attributes.

• Building on the theoretical results in [1], our contributions are:

1. **Implementation** of an algorithm that outputs a private summary of the database that is capable of answering **all possible** k-way marginal queries.
2. Analysis of the algorithm's **computational limitations** and **statistical properties**, especially in the regime of where the dimension of the dataset is large.

Notion of Differential Privacy

A mechanism, $\mathcal{A}$, which maps database D to an output $\mathcal{A}(D)$, is ϵ-**differentially private**, if for *any* two neighbouring D and D' which differ by one row, the distributions of $\mathcal{A}(D)$ and $\mathcal{A}(D')$ are ϵ-close to each other. Formally, $\mathbb{P}\left(\mathcal{A}(D) \in S\right) \leq e^{\epsilon}\mathbb{P}\left(\mathcal{A}(D') \in S\right), \forall S$.

The Algorithm [1]

- D: database $D \in (\{0,1\}^d)^n$.
- n: number of records (rows) in the database
- d: number of attributes (columns) in the database
- k: maximum number of attributes that can appear in a marginal query

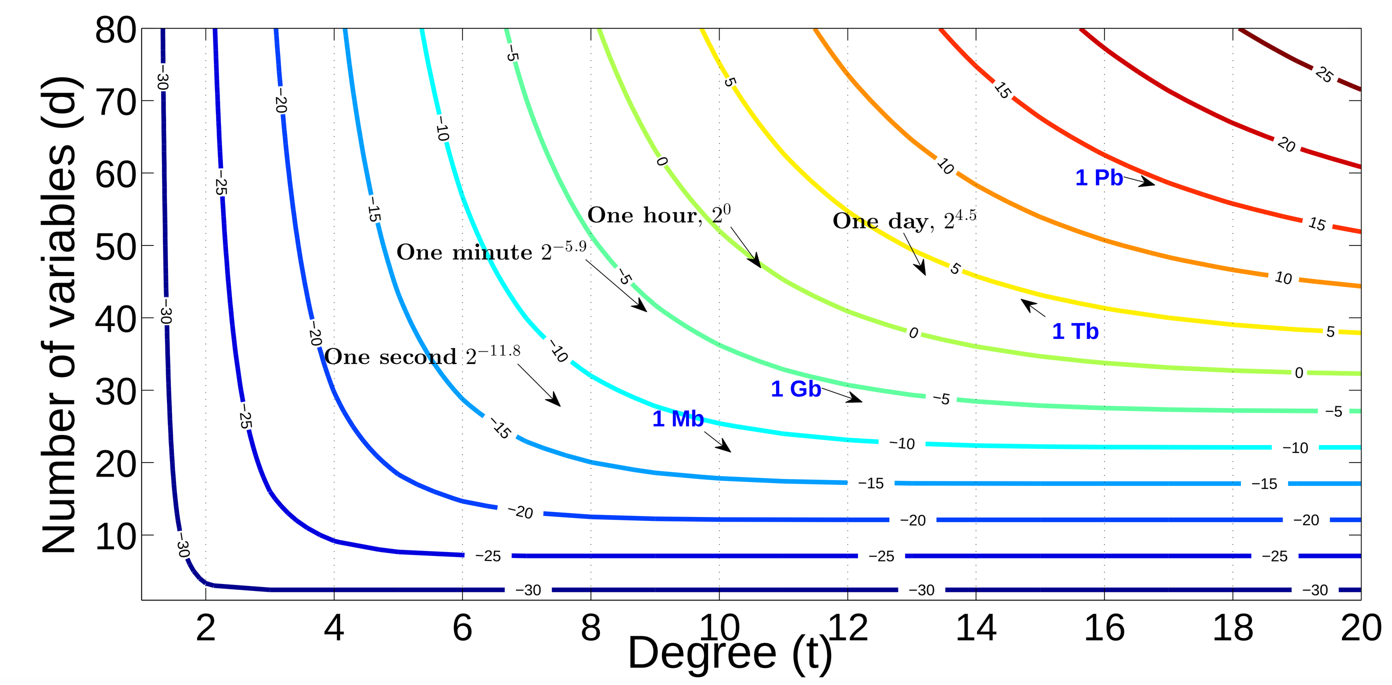
Key Idea: Using polynomial approximations to create a private summary of a database that gives approximate answers to all k-way marginals.

1. For each row x in D, compute a degree-t polynomial, p_x, which approximates the answers to all queries on x
2. Compute polynomial p_D by average over all rows:
   $P_D = \frac{1}{n}\sum_{x=1}^{n} p_x$
3. Form P'_D by adding i.i.d. noise to all coefficients of P_D to ensure ϵ-differential privacy
4. Output p'_D

Under theoretical assumptions in [1] the algorithm runs in time $d^{O(k)}$ realses a private summary capable of answering any k-way marginal query with at most ± 0.01 error as long as $n \geq d^{O(\sqrt{k})}$

Memory and Space

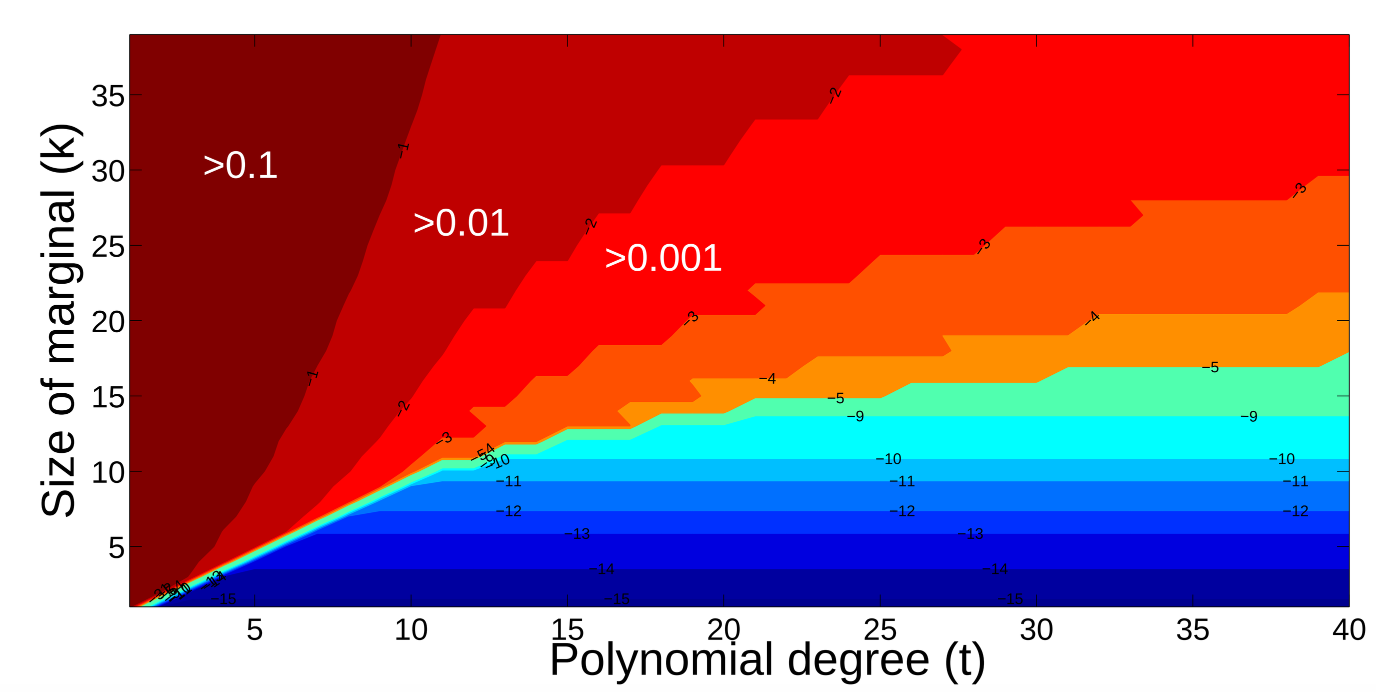
Space and time complexity both scale linearly with **number of coefficients**, which is at most $\binom{d+t}{t}$ for the original algorithm. We further reduce this by employing a multi-linear polynomial prepresentation, which yields $\sum_{1}^{min(d,t)} \binom{d}{i}$ coefficients.



Log-scaled space and running time, implemented in Java 1.7 on Dell XPS L502X (Intel Core i5 2.0GHz(Quad), 8GB DDR3, 7200RPM, Win7)

Approximation Error

The maximum degree of the approximating polynomials (t) determines number of coefficients as well as approximation erro. Higher values of $t \implies$ better approximation but more coefficients.



Log-scale maximum approximation error as a function of maximum degree and number of attributes

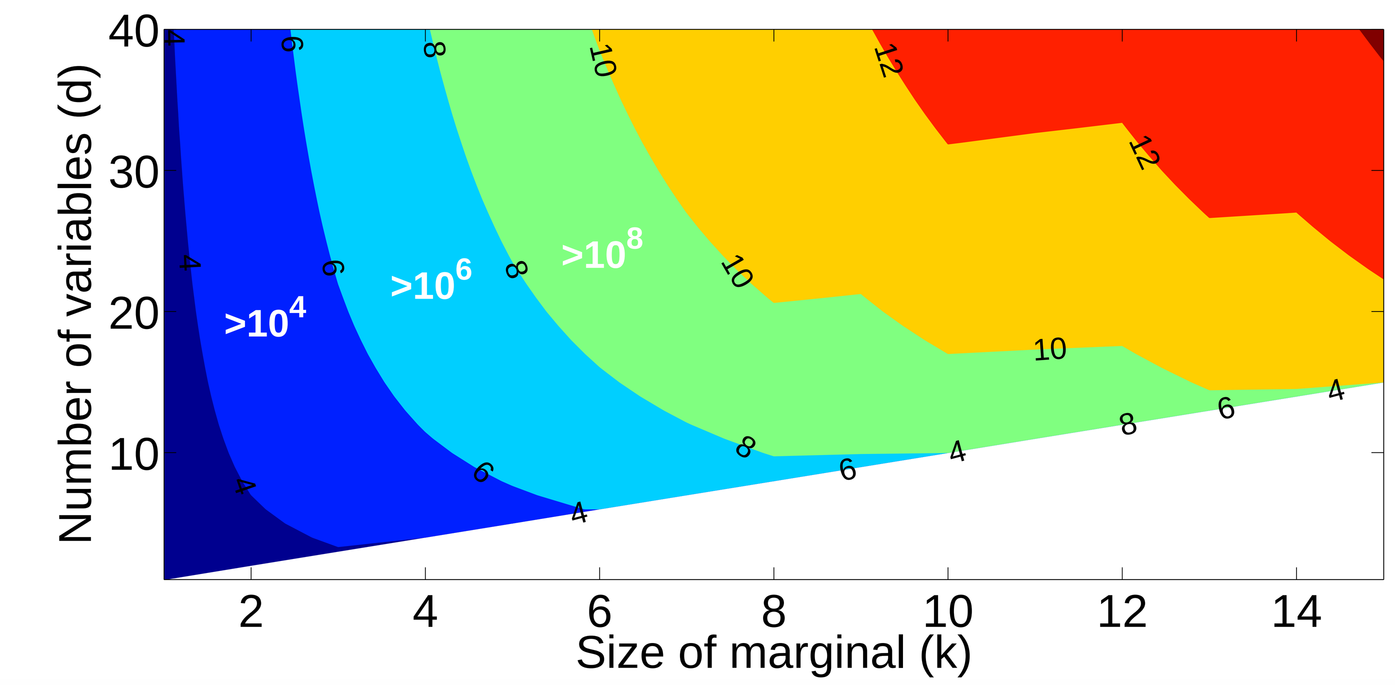
Properties of Noise

• Independent Laplace noises, $\{N_s\}$, are added to all coefficients to ensure differential privacy.
• Total error = approximation error + error due to noise
• **Proposition**. To ensure 1-differential privacy, it suffices to add noise terms, such that

$$\text{STD}\left(\text{Total noise for } k\text{-way marginal}\right) = \frac{1}{n} S_{d,t} \sqrt{\sum_{1}^{min(k,t)} \binom{k}{i}},$$

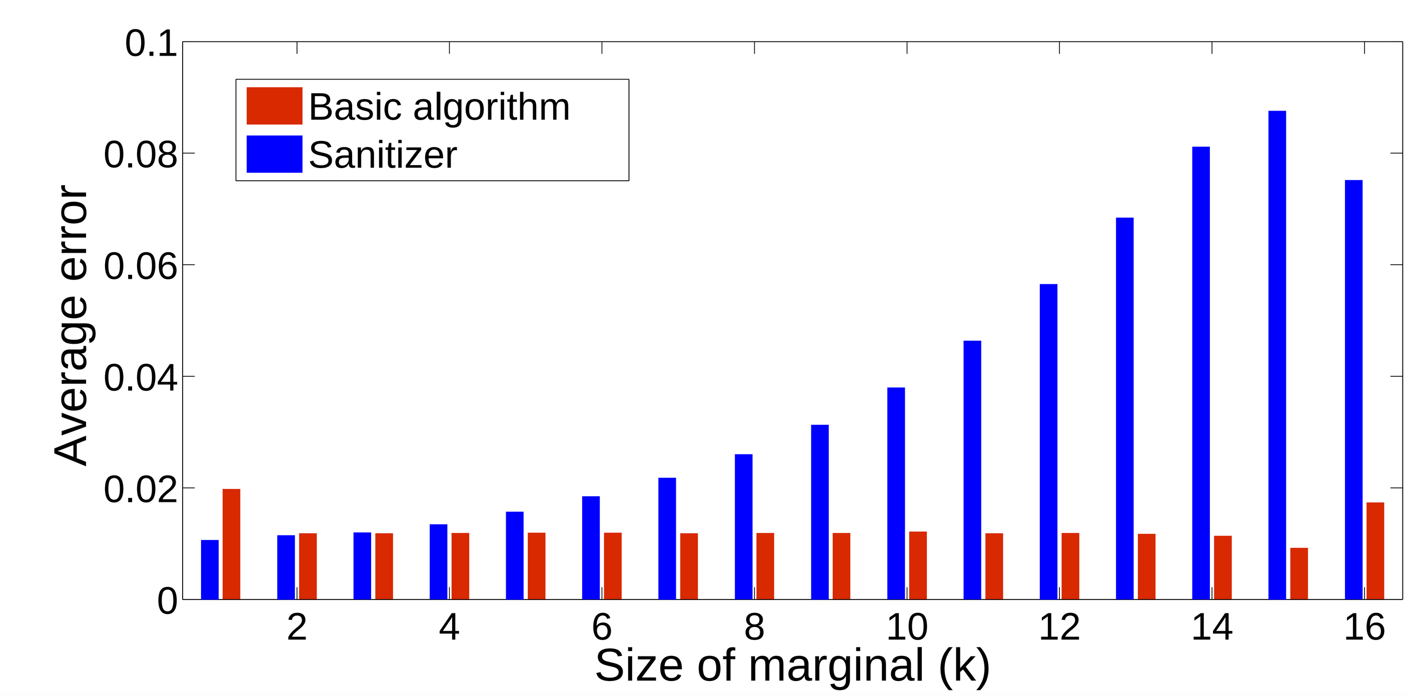
where $S_{d,t}$ is the sum of the absolute values of all coefficients.

Noise Accuracy



Minimum number of rows (n) required, such that the error due to noise for a k-way marginal is within ± 0.01 with at least 99% probability.

Experiments



Average total error based on the Adult census data set [2]. The error is compared to that of a the naive private algorithm which outputs the answers to 2^d queries.

Conclusions

• Our results demonstrate that the algorithm is able to provide efficient private summaries for dataset with large number of attributes ($d \approx 100$), for moderate-sized marginal queries ($k \approx 15$).
• It remains challenging to maintain a fixed level of error as the number of attributes, d, becomes large.

Acknowledgement

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

References

[1] J. Thaler, J. Ullman, and S. Vadhan, Faster algorithms for privately releasing marginals, *Proceedings of ICALP*,810–821, 2012.
[2] UCI Adult Data Set. http://archive.ics.uci.edu/ml/datasets/Adult