Privacy preserving high-dimensional data mining

Raghav Bhaskar[1], Daniel Kifer[2], Srivatsan Laxman[1], Adam Smith[2], Abhradeep Thakurta[2]
[1]Microsoft Research India, [2]Pennsylvania State University

Research goals

- **Ranking** [BLST]
  - Consider a collection of elements $U = \{1, \cdots, u\}$
  - Each element i has a real valued score $q_T(i)$ bases on data set *T*
  - **Goal:** Output *k* elements with highest scores
- **Sparse linear regression (SLR)**
  - Consider the regression model $y = X\theta^* + w$ where $\theta^* \in \mathbb{R}^p, X \in \mathbb{R}^{n \times p}, w \in \mathbb{R}^n$
  - θ^*: regression vector, X:design matrix and w:noise vector
  - Sparsity assumption: θ^* has only *k* non-zero entries
  - **Goal:** Find $\hat{\theta}$ (*k*-sparse) which estimates θ^* given (y, X)
- **Privacy**
  - Differential privacy: Preserve the privacy of the entries in *T* (for ranking problem) and (y, X) for sparse regression

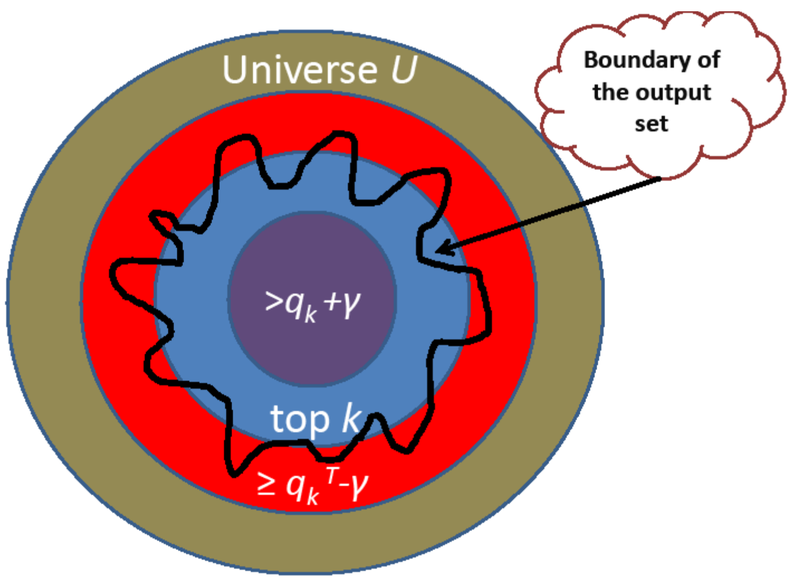
Differential privacy [DMNS]

A randomized algorithm $\mathcal{A}$ is ϵ-differentially private if for all data sets $T, T' \in \mathcal{D}^n$ differing in at most one entry and all events $\mathcal{O} \subseteq Range(\mathcal{A})$:

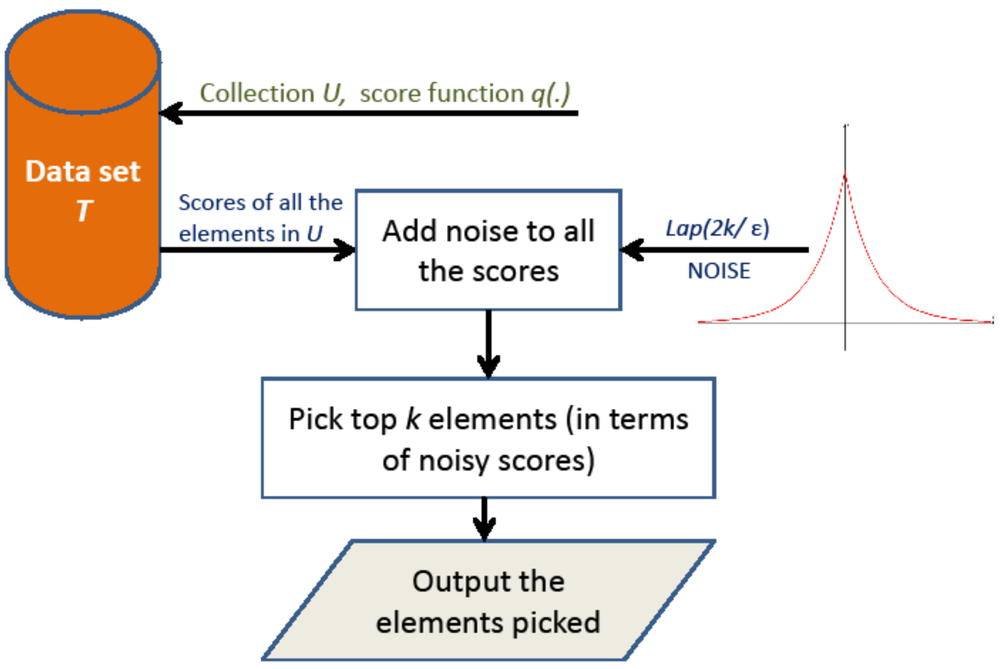
Approximate ranking

Let q_k^T be the k^{th} highest score based on data set T. An output list is γ-useful if:

- *(Soundness)* No element in the output has score less than $(q_k^T - \gamma)$.
- *(Completeness)* Every element with score greater than $(q_k^T + \gamma)$ is in the output.



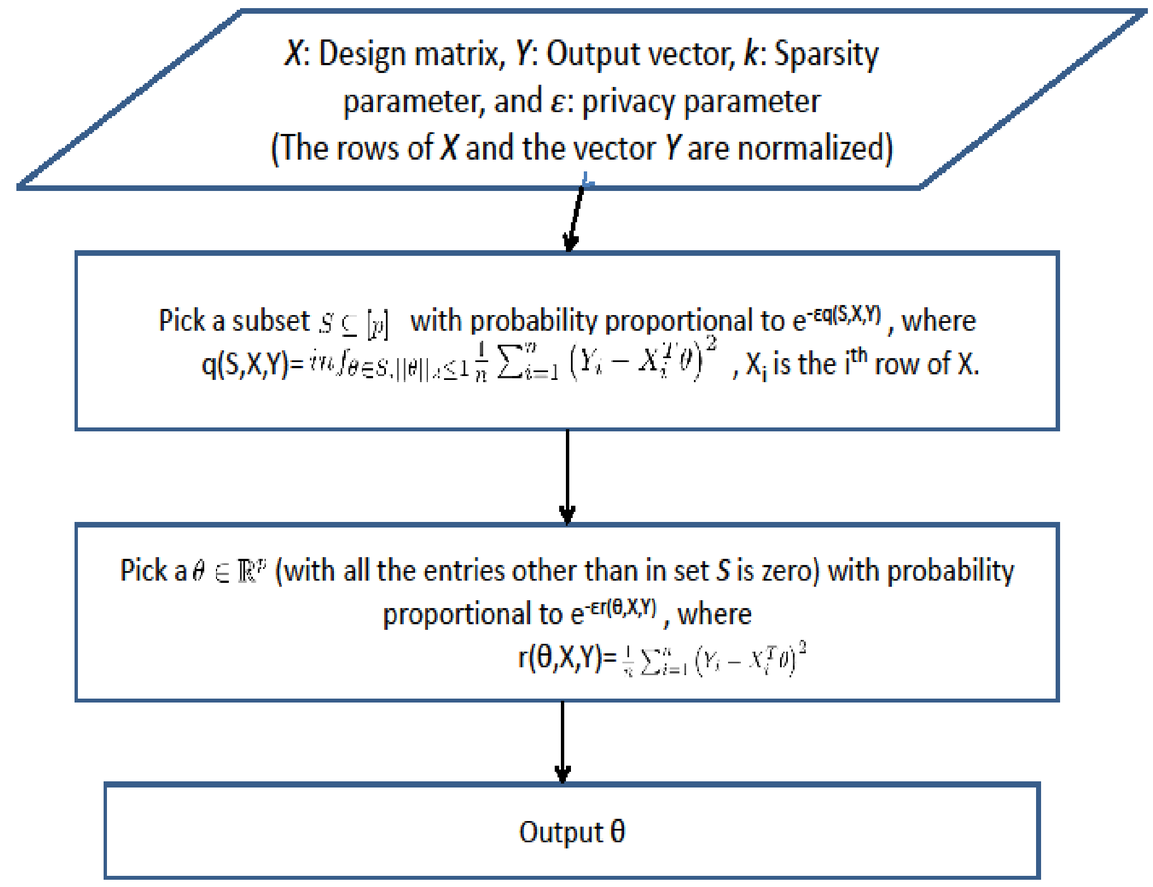
Private ranking algorithm



Privacy and utility

- **Theorem (Privacy):** *The algorithm is ϵ-differentially private.*
- **Theorem (Utility):** *For all $\rho > 0$: with probability at least $1 - \rho$, the output is γ-useful, where $\gamma = \frac{4k}{\epsilon}\left(\ln u + \ln(\frac{1}{\rho})\right)$.*
- **Theorem (Running Time):** *The algorithm runs in time $O(u)$.*

Private SLR



Privacy and utility

- ***Theorem(Privacy):*** *The algorithm is $\frac{16\epsilon}{n}$-differentially private*
- ***Theorem (Utility):***

References

[DMNS] C. Dwork, F. McSherry, K. Nissim, A. Smith: Calibrating Noise to Sensitivity in Private Data Analysis. TCC 2006.

[BLST] R. Bhaskar, S. Laxman, A. Smith: Discovering Frequent Patterns from Sensitive Data. KDD 2010.

[NRWY] S. Negahban, P. Ravikumar, M. Wainwright, B.Yu: A unified framework for high-dimensional analysis of *M*-estimators with decomposable regularizers. NIPS 2009.

[CMS] K. Chaudhuri, C. Monteleoni, A. Sarwate: Differentially private Emperical Risk Minimization.

Sponsors: Microsoft Research, India and National Science Foundation