New Oracle-Efficient Algorithms for Private Synthetic Data Release

Giuseppe Vietri
University of Minnesota

Grace Tian
Harvard

Mark Bun
Boston University

Thomas Steinke
IBM Research-Almaden

Zhiwei Steven Wu
University of Minnesota

Privacy in Data Analysis

In many important cases, access to **data** is restricted due to **privacy** concerns.

Financial

Medical

Socioeconomic

Private Synthetic Data

Example: Q is a set of k-Way Marginal Queries.

$$q_\phi(D) = \frac{1}{D} \sum_{x \in D} \phi(x)$$

$$\phi(x) = \begin{cases} 1 & \text{if } x_a = 1 \wedge x_b = 0 \wedge x_c = 1 \\ 0 & \text{otherwise} \end{cases}$$

GOAL: Find such that $\max_i |q_i(\ ) - q_i(\ )| \leq \alpha$, subject to differential privacy.

Prior Work: MWEM

- **MWEM** is an **optimal** algorithm for generating differentially private synthetic data.
- It keeps track of a distribution over the data domain X.
- But **Runtime** is exponential in the dimension of the data. Intractable for high dimensions.
- This runtime in necessary in the works case [Ullman 2016], [Ullman & Vadhan 2011].

Our Contributions

	Error
FEM	$\widetilde{O}\left(\frac{d^{3/4}\log^{1/2}(\|Q\|)}{\sqrt{n\varepsilon}}\right)$ ⟵ **Best Empirically**
sepFEM	$\widetilde{O}\left(\frac{d^{5/8}\log^{1/2}(\|Q\|)}{\sqrt{n\varepsilon}}\right)$
DQRS	$\widetilde{O}\left(\frac{d^{1/5}\log^{3/5}(\|Q\|)}{n^{2/5}\varepsilon^{2/5}}\right)$
MWEM	$\widetilde{O}\left(\frac{d^{1/4}\log^{1/2}(\|Q\|)}{\sqrt{n\varepsilon}}\right)$

d = size of the data.

Oracle Efficient

Our algorithms are **Oracle Efficient**.

Computational efficient given access to an **optimization oracle** that can solve:

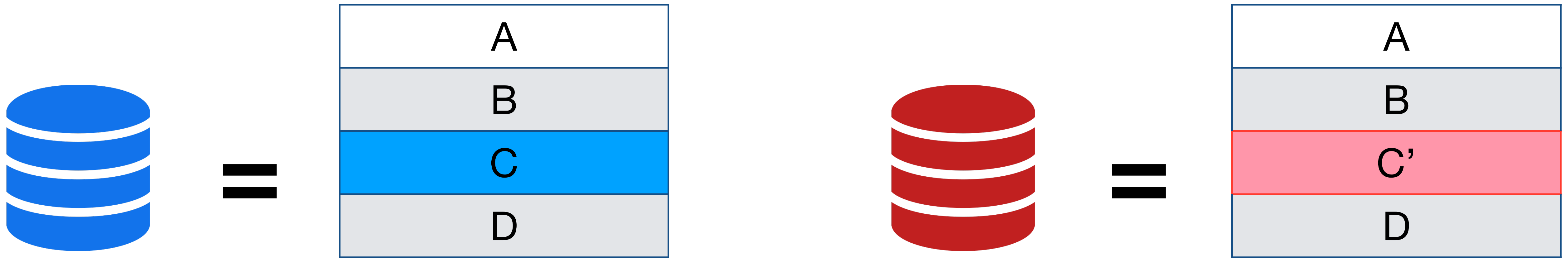
$$\arg\min_{x \in X} \sum_{i=1}^{t-1} \left(q_i(D) - q_i(x) \right) + \langle x, \sigma \rangle$$

We take advantage of **fast heuristics** like integer programming.

Differential Privacy (DP)

[Dwork et al., 2006]

Two datasets



are **neighbors** if the are different in only one row.

Definition: Mechanism M satisfies ε-**DP** if, for all neighboring datasets and for all $r \in \text{range}(M)$

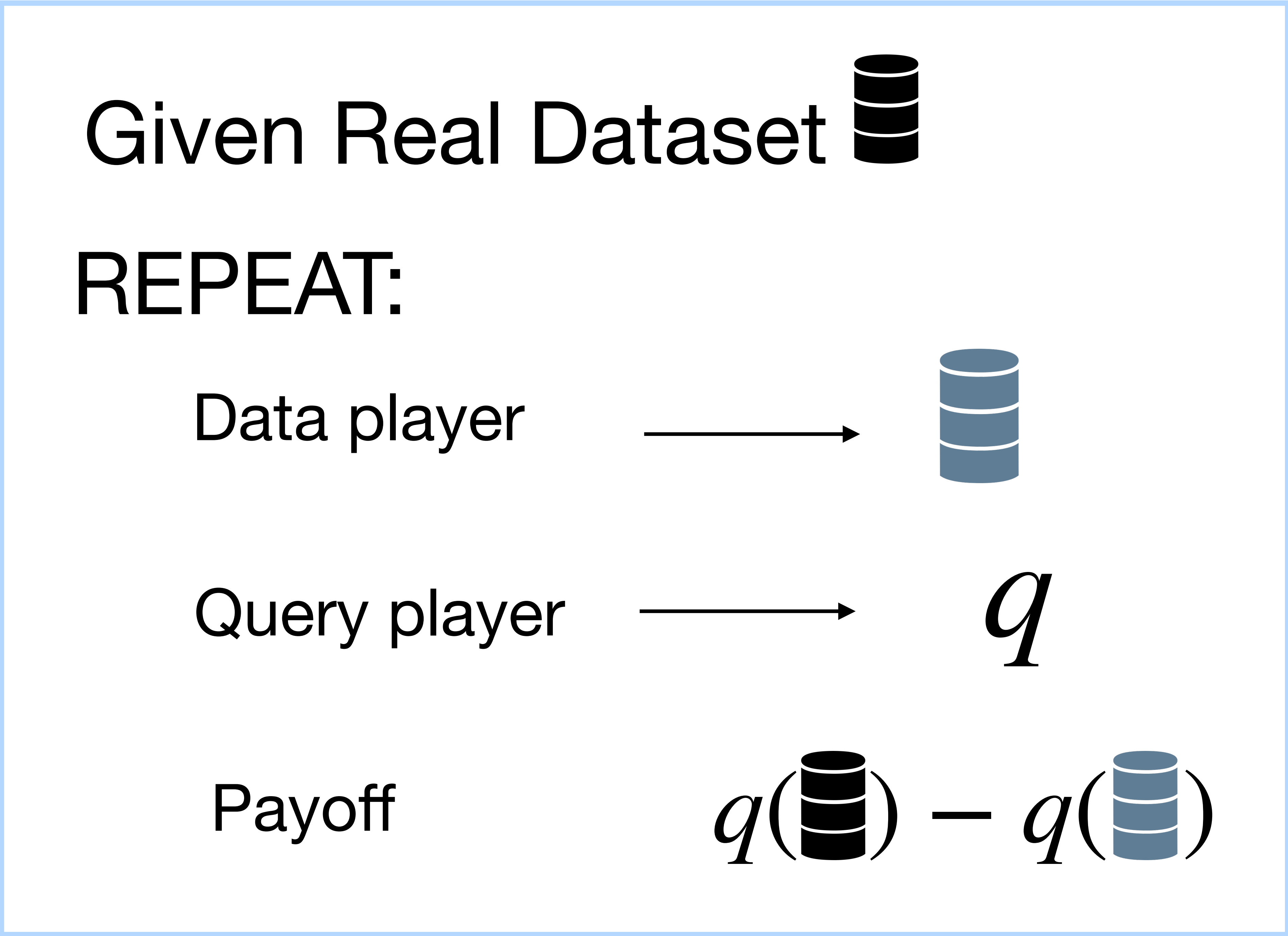
Pr[M() $= r] \leq e^{\varepsilon}$ Pr[M() $= r]$

Exponential Mechanism (EM)

- Suppose we have a set of low sensitivity Queries $Q = \{q_1, \ldots, q_m\}$
- And some score function $s : X \times Q \rightarrow [0,1]$, where s has sensitivity Δs
- The EM chooses $q \in Q$ with the maximum possible score.
- Satisfies ε-differential privacy, with error $\approx \frac{\Delta s}{\varepsilon} \ln(|Q|)$

A Framework for Private Synthetic Data

- Zero-sum game.
- No-regret Dynamics -> Game equilibrium.
  (Freud & Schapire, 1997)
- Game equilibrium -> Accurate synthetic data.
  (Hsu et al., 2013; Gaboardi et al., 2014).
- Similar to GANs.



Prior Work: MWEM

Input: Private Dataset $D \in X^n$, target privacy parameters ε, δ.

$\widehat{D}_1 =$ uniform distribution over X

For $t = 1\ldots, T$**:**

1. $s(q) = q(D) - q(\widehat{D}) \quad \forall q \in Q$
2. Sample q_t from the Exponential Mechanism with score function s and privacy $\frac{2T}{\varepsilon}$
3. $\boxed{\forall_{x \in X} \widehat{D}_{t+1}(x) \propto \widehat{D}_t(x) \exp\left(q_t(x)\left(q_t(D) - q_t(\widehat{D})\right)/2n\right)}$

Output $\frac{1}{T}\sum_{t=1}^{T} \widehat{D}_t$

Computational Bottleneck

Our First Algorithm: FEM

- FEM: Follow-the-Perturbed-Leader with Exponential Mechanism
- Replace the MWEM's computational bottleneck (MW) by $FTPL$
- We use the $FTPL$ algorithm from [Suggala et al. 2019].

The FEM Algorithm

Input: Real Dataset , target privacy parameters ε. Algorithms $FTPL$ and M_E.

Set: Number of steps T, and Exponential Mechanism privacy parameter ε_0

For $t = 1\ldots, T$**:**

1. $_t \sim FTPL$ # Data player gets a sample from FTPL.
2. $s(q) = q(\,) - q(\,_t) \quad \forall q \in Q$
3. Sample q_t from the Exponential Mechanism with score function s and privacy ε_0
4. $FTPL$ incurs loss $q_t(\,) - q_t(\,_t)$

Output $\cup \{\,_\mathbf{1}, \ldots, \,_\mathbf{T}\}$

Follow the Perturbed Leader

Input: Previous Queries $q_1, \ldots, q_{t-1}$, Number of samples s

For $i = 1\ldots, s$**:**

1. Sample σ from exponential distribution.
2. $x_i = \text{arg}\min_{x \in X} \sum_{i=1}^{t-1} \left(q_i(D) - q_i(x)\right) + \langle x, \sigma \rangle$

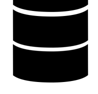
Output $\widehat{D} = \{x_1, \ldots, x_s\}$

FTPL has sub-linear regret.

Benchmark: HDMM

- **H**igh **D**imensional **M**atrix **M**echanism (McKenna et al., 2018).
- Considered the state-of-the-art in practice.
- Used by US Census Bureau.
- Empirical evaluation benchmark.

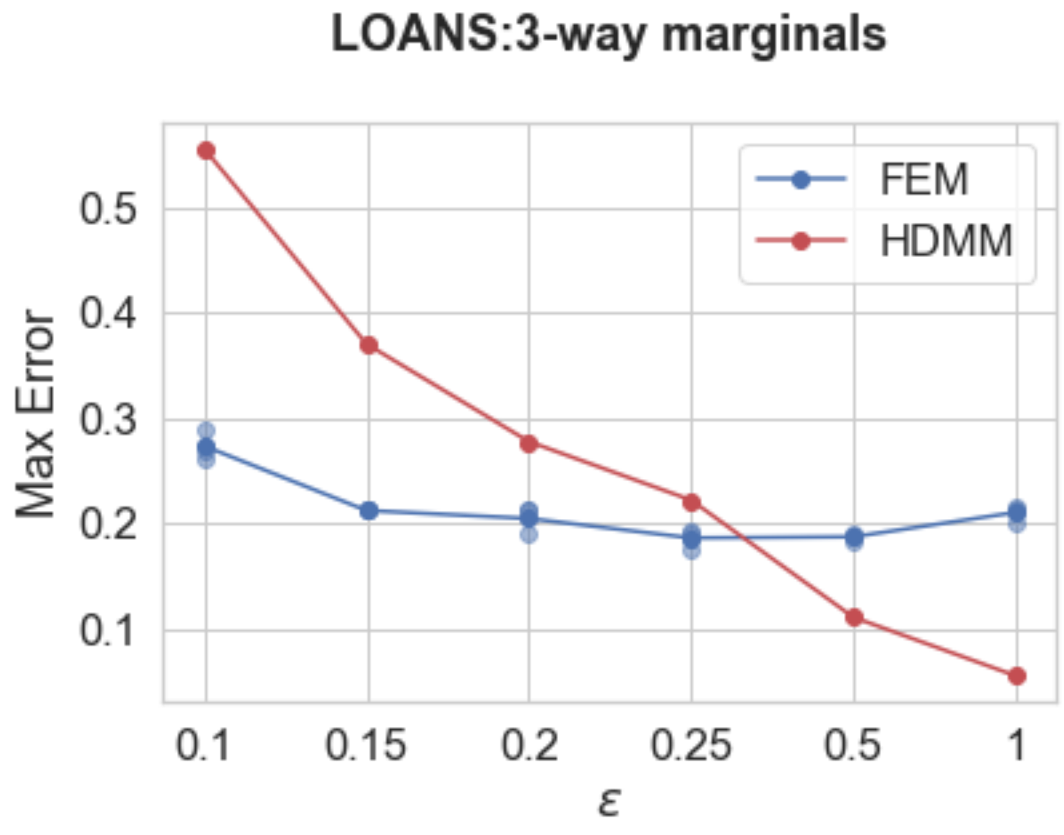
Experiments

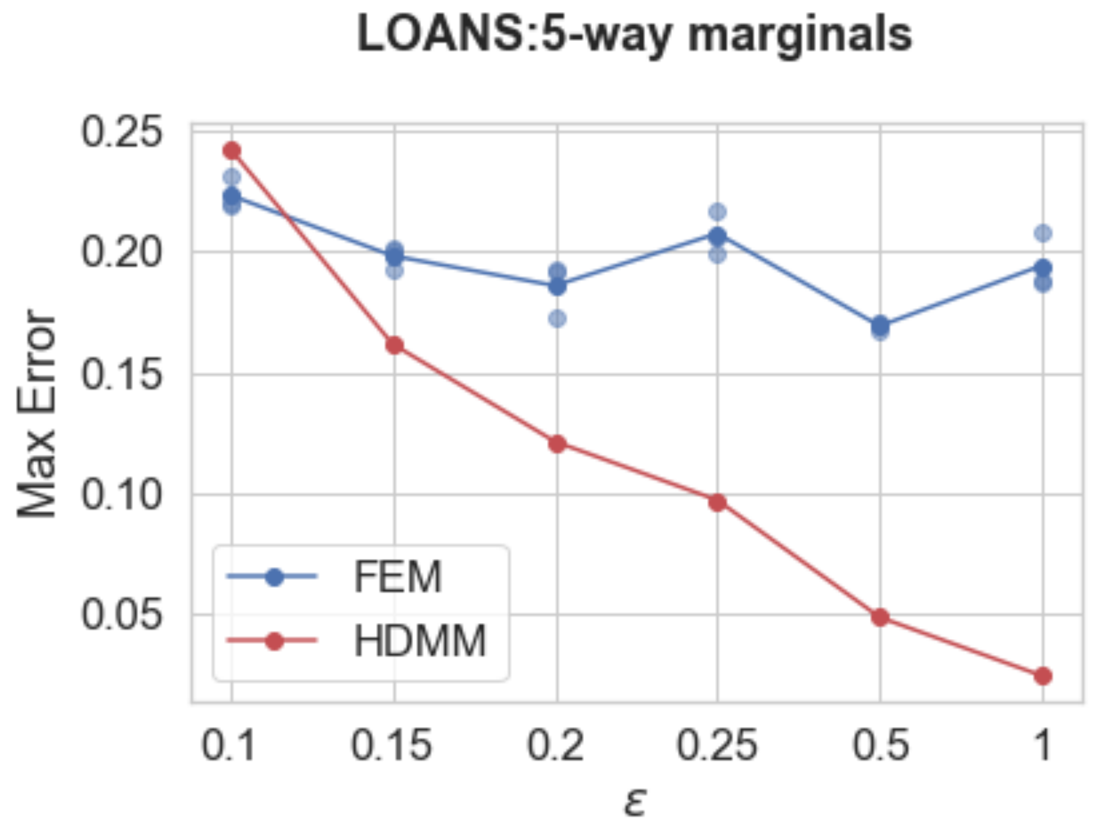
- Two high dimensional Datasets: ADULT and LOANS.
- We compared FEM against HDMM.
- Focus on large workloads and low privacy budget.
- Workload consists of k-way marginals.
- Basis for comparison is the maximum error: $\max_i | q_i($  $) - q_i($  $) |$

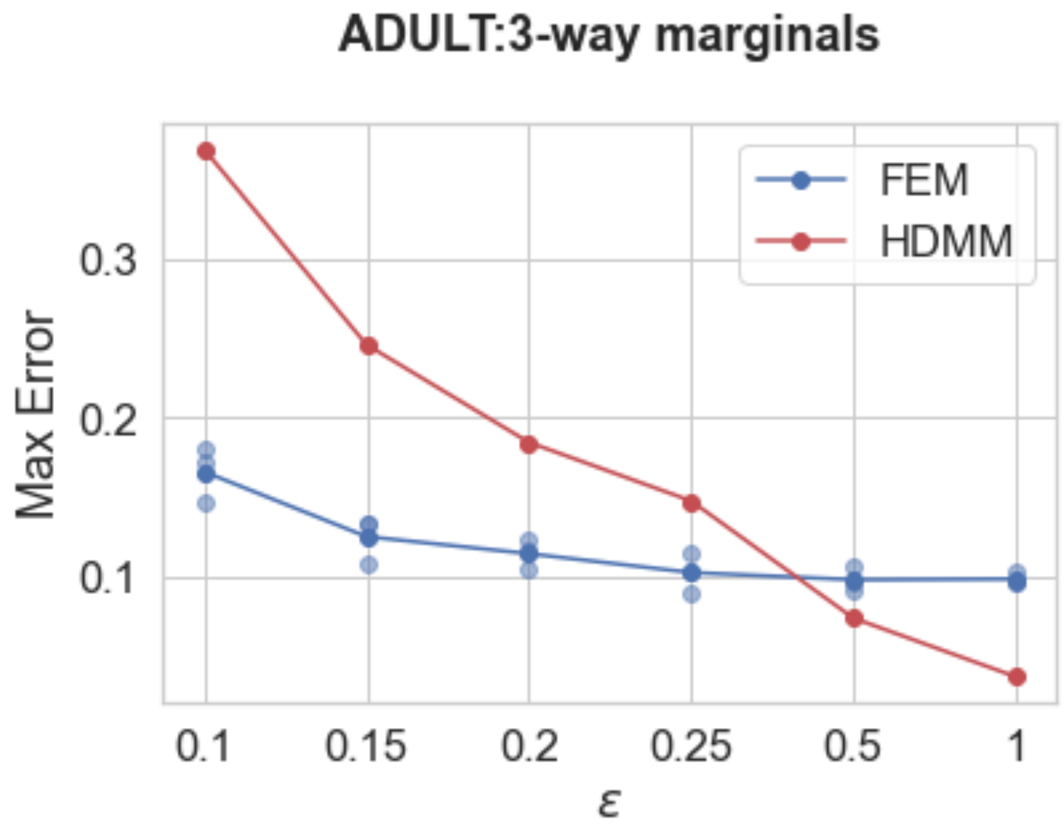
Datasets

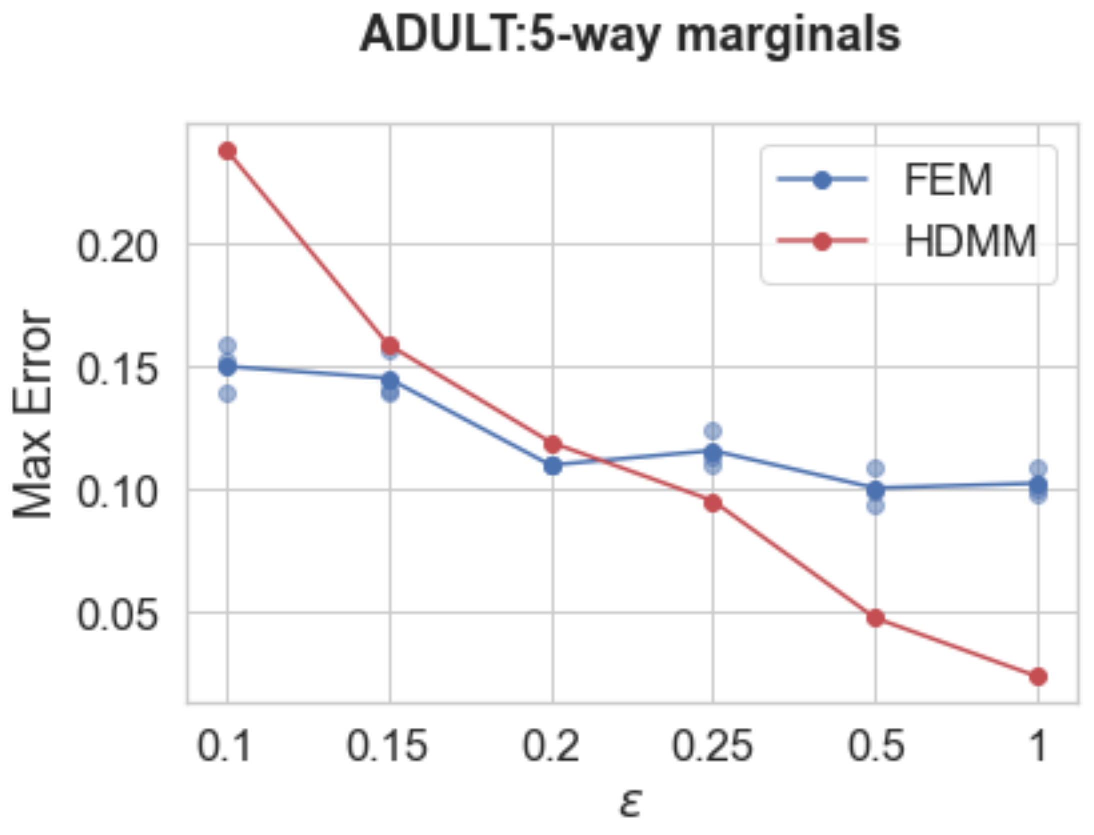
	Rows	Attributes	K-way marginals	Total Queries
ADULT	48,842	15	5	1,213,952
LOANS	42,535	48	5	588,584

Error vs ε

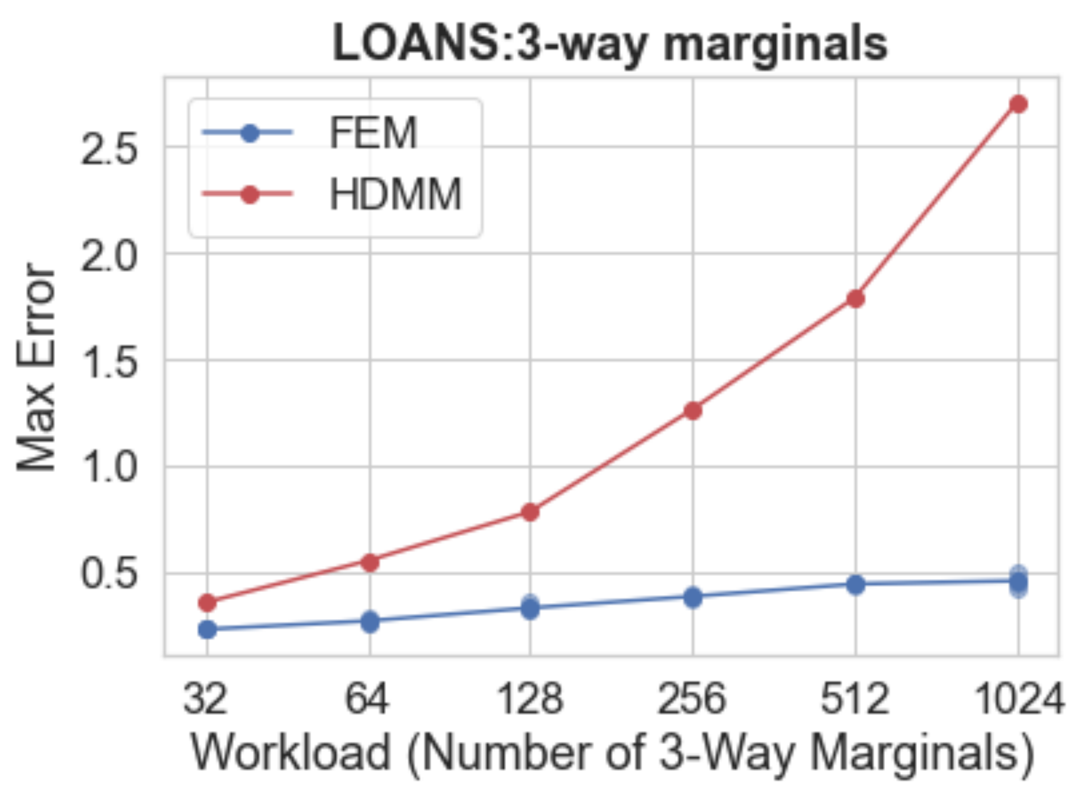


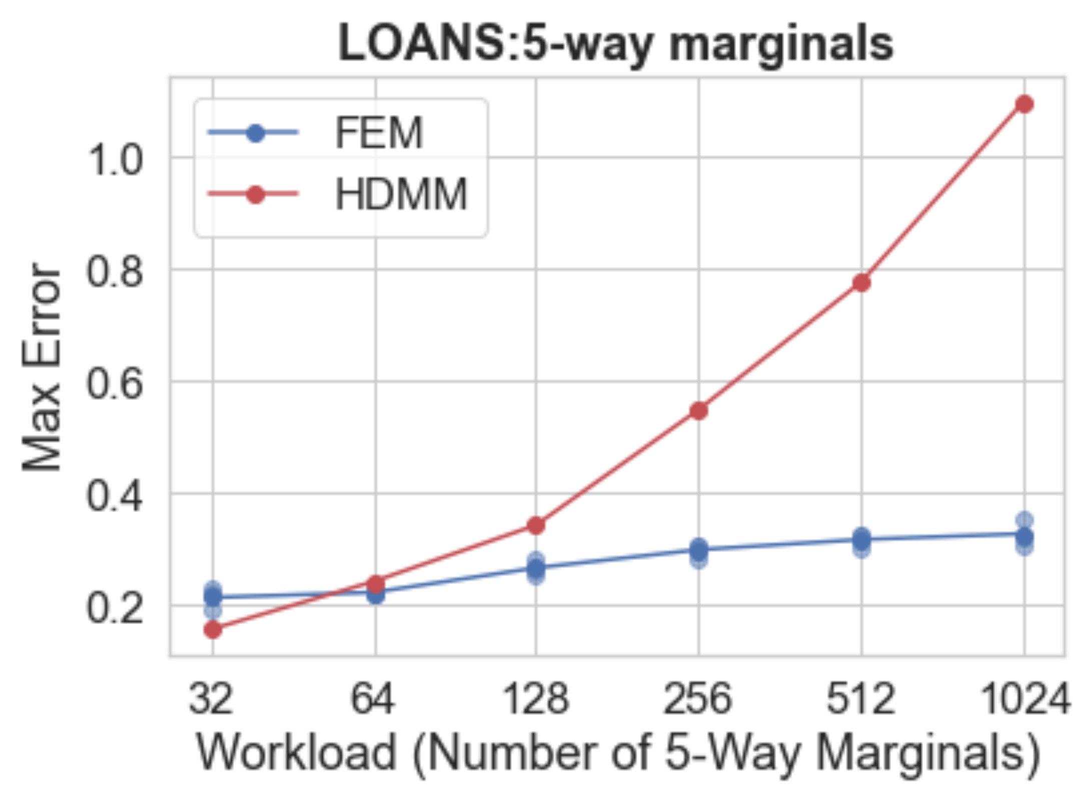


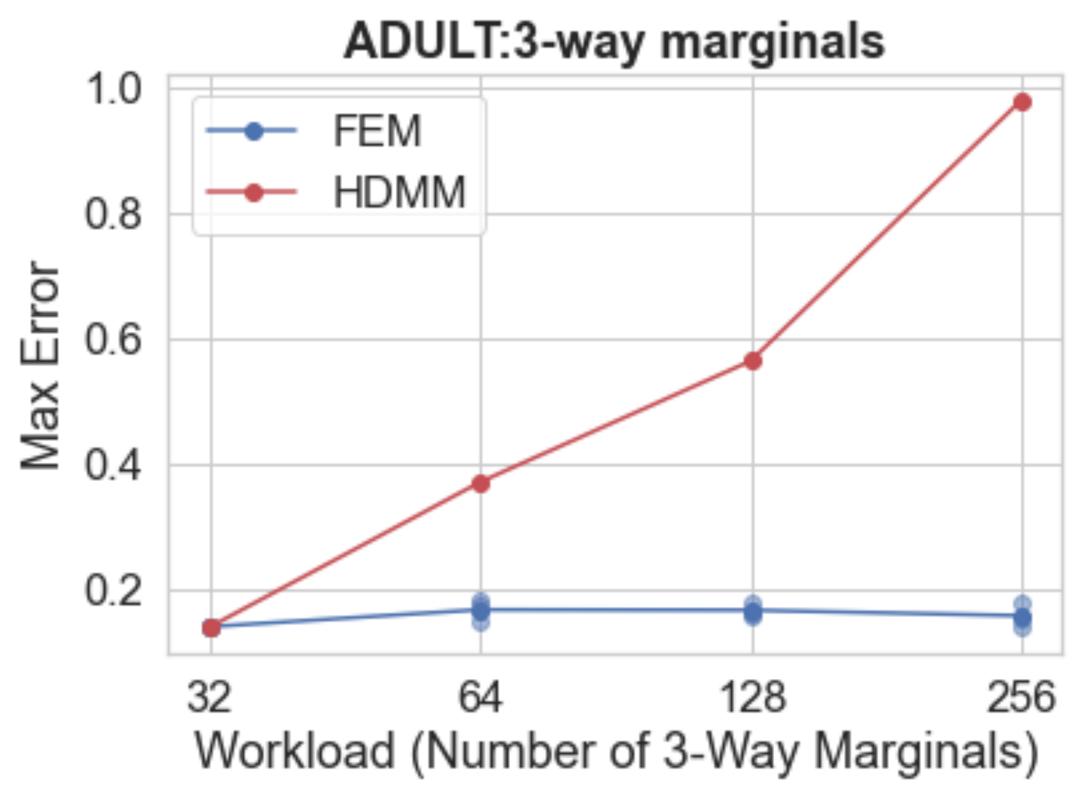


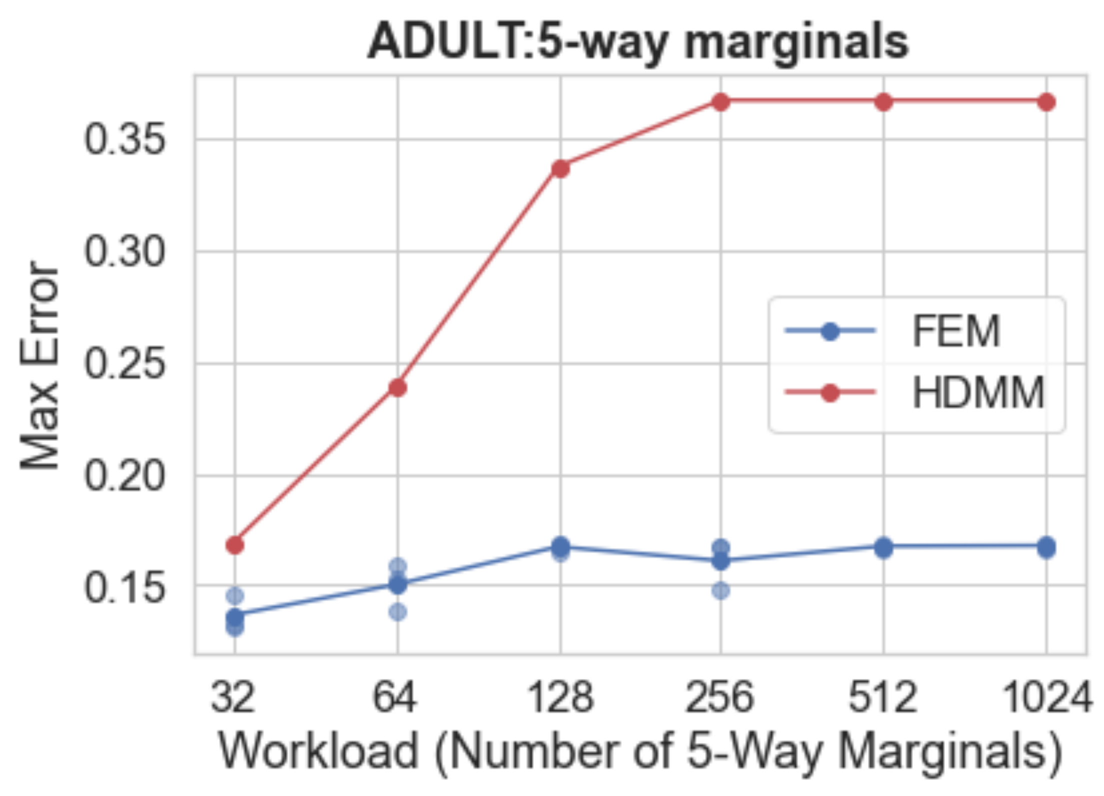


Error vs Workload









Conclusion

- We introduced three new algorithms for private synthetic data release.
- Our algorithms are close to the theoretical optimal.
- Computationally efficient for high dimensional settings.
- Our algorithm performs better in practice than the state-of-the art in
  - the high-privacy and large workload setting.