Joint Selection: Adaptively Incorporating Public Information for Private Synthetic Data

Miguel Fuentes
University of Massachusetts Amherst

Brett Mullins
University of Massachusetts Amherst

Ryan McKenna
Google Research

Gerome Miklau
Tumult Labs

Daniel Sheldon
University of Massachusetts Amherst

Abstract

Mechanisms for generating differentially private synthetic data based on marginals and graphical models have been successful in a wide range of settings. However, one limitation of these methods is their inability to incorporate public data. Initializing a data generating model by pre-training on public data has shown to improve the quality of synthetic data, but this technique is not applicable when model structure is not determined a priori. We develop the mechanism JAM-PGM, which expands the adaptive measurements framework to jointly select between measuring public data and private data. This technique allows for public data to be included in a graphical-model-based mechanism. We show that JAM-PGM is able to outperform both publicly assisted and non publicly assisted synthetic data generation mechanisms even when the public data distribution is biased.

1 INTRODUCTION

A differentially private (DP) algorithm can extract valuable insights from sensitive data while provably limiting what can be learned about individuals (Dwork et al., 2006). However, when data is accessed repeatedly, the curator must track the accumulated privacy loss and add noise sufficient to protect the entire sequence of queries, which presents logistical challenges for many common settings, including exploratory data analysis. Therefore, there is considerable interest in releasing private synthetic datasets that can support a range of downstream analyses (Charest, 2011; Chen et al., 2015; Zhang et al., 2017; Xie et al., 2018; Jordon et al., 2019; Zhang et al., 2018; Asghar et al., 2020; Bowen and Liu, 2020; Vietri et al., 2020; Ge et al., 2021; McKenna et al., 2021a).

Much of the recent research for private synthetic data has its roots in the multiplicative-weights exponential mechanism (MWEM) for private query-answering (Hardt et al., 2012). These algorithms iteratively select workload queries to measure, then measure those queries (with noise) and use the results to update a model for the synthetic data. Finally they generate synthetic records from the model. The goal is to generate synthetic tabular data to accurately answer a set of workload queries. This general pattern has been referred to as the "select-measure-generate" paradigm (McKenna et al., 2021a). Recent work formalizes this pattern as the "adaptive measurements" framework (Liu et al., 2021b). Within this framework, research has focused on different model representations, estimation methods, selection mechanisms, and computational efficiency (Aydore et al., 2021; Zhang et al., 2021; Liu et al., 2021b,a; Cai et al., 2021; McKenna et al., 2022).

It is well known that public data, when available, can be used to boost accuracy of many differentially private algorithms (Ji and Elkan, 2013; Alon et al., 2019; Bassily et al., 2020; Amid et al., 2022; Zhou et al., 2020; Bassily et al., 2018; Kairouz et al., 2021; Wang and Zhou, 2020; Papernot et al., 2017). For example, public data can be used to pre-train models (Liu et al., 2021a; Yu et al., 2021), select hyperparameters or model structure (McKenna et al., 2021a), or even answer some queries directly to save privacy budget. Public data may come from releases that were done before DP restrictions were instituted or they may even come in the

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

form of previously released synthetic data. In general, if the public data and private data are "similar enough", performance gains can be very large. However, a significant issue is that one does not know in advance how similar the public and private data are. Some previous works assume public data and private data are from the same distribution, which is unrealistic (Bassily et al., 2020; Alon et al., 2019).

We address the problem of differentially private query answering and synthetic data generation with the assistance of public data. Unlike prior work, which uses public data for pre-training or to determine the support of the data distribution (Liu et al., 2021a,b), we integrate public data tightly into the selection process of the mechanism, which means we explicitly consider *when* and *how* to use public measurements as a proxy for private measurements. We focus primarily on workloads of marginals, and augment the selection step to allow measuring a marginal either from public or private data. Conceptually, measuring a private marginal is unbiased but requires privacy noise, while measuring a public marginal as an estimate of a private marginal is biased but noise-free: which of these is better depends on the data, workload, and privacy parameters. Therefore, the mechanism must (privately) decide which marginals to measure from public data and which marginals to measure from the private data.

In this paper, we develop joint adaptive measurements with PRIVATE-PGM (JAM-PGM), the first approach that incorporates public data *selection* into iterative methods for synthetic data. JAM-PGM is an adaptive measurements approach that uses PRIVATE-PGM (McKenna et al., 2019) to model the data distribution. JAM-PGM privately selects from both public and private measurements with scores that estimate the error reduction expected from each measurement. By automatically selecting which queries to answer with public data, JAM-PGM can benefit from public data that is accurate for some marginals but inaccurate for others. We show empirically that JAM-PGM can use public data to increase accuracy across a range of scenarios.

2 BACKGROUND

A private dataset D is a collection of n records each containing potentially sensitive information about one individual. Each record $r = (r_1, ..., r_m)$ has m attributes and each attribute r_i takes a value from the discrete finite set $\mathcal{X}_i$. Each record belongs to the data universe $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$. We also consider a public dataset D_{pub} that is a collection of $\hat{n}$ records which are not subject to differential privacy constraints. We assume that $D_{\text{pub}} \in \mathcal{X}^{\hat{n}}$.

2.1 Differential Privacy

Differential privacy is a formal model of privacy that bounds the effect of any individual record on the output of a randomized algorithm. We say that datasets $D, D' \in \mathcal{X}^n$ are neighboring, denoted $D \sim D'$, if D' can be obtained from D by modifying the values of at most one record. Note that all differentially private mechanisms considered in this paper use this notion of the neighboring relation.

Definition 2.1 (Differential privacy; DP)**.** A randomized mechanism $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{R}$ is said to be (ϵ, δ)-DP if, for all neighboring datasets $D \sim D' \in \mathcal{X}^n$ and all measurable subsets $S \subseteq \mathcal{R}$, we have $\Pr[\mathcal{M}(D) \in S] \leq e^\epsilon \cdot \Pr[\mathcal{M}(D') \in S] + \delta$.

A useful alternative notion of differential privacy for analyzing the composition of mechanisms is zero-Concentrated Differential Privacy.

Definition 2.2 (Zero-concentrated differential privacy; zCDP)**.** A mechanism $\mathcal{M}$ satisfies ρ-zCDP if for any neighboring datasets $D \sim D'$ and for all $\gamma \in (1, \infty)$, it holds that $D_\gamma(\mathcal{M}(D)||\mathcal{M}(D')) \leq \rho\gamma$, where D_γ is the γ-Renyi divergence between distributions $\mathcal{M}(D), \mathcal{M}(D')$.

Proposition 2.3 (zCDP to DP Conversion; Canonne et al. 2020)**.** *If mechanism $\mathcal{M}$ satisfies ρ-zCDP, then, it satisfies (ϵ, δ)-DP for any $\epsilon > 0$ and $\delta = \min_{\alpha>1} \frac{\exp((\alpha-1)(\alpha\rho-\epsilon))}{\alpha-1}\left(1-\frac{1}{\alpha}\right)^\alpha$.*

The synthetic data mechanisms considered in this paper utilize two building block mechanisms: the exponential mechanism for private selection and the Gaussian mechanism for private query measurement. To analyze the privacy of these mechanisms, an important quantity is sensitivity, the maximum change in function value on neighboring datasets. The L_p sensitivity of a function f is given by $\Delta_p(f) = \max_{D\sim D'} \|f(D) - f(D')\|_p$ where $f\colon \mathcal{X}^n \to \mathbb{R}^k$.

Proposition 2.4 (zCDP of Gaussian mechanism; Bun and Steinke 2016)**.** *Let $f : \mathcal{X}^n \to \mathbb{R}^k$ be a vector-valued function of the dataset. For dataset D, the Gaussian mechanism adds i.i.d. Gaussian noise to $f(D)$ with scale parameter σ^2 i.e., $\mathcal{M}(D) = f(D) + \sigma\Delta_2(f)\mathcal{N}(0, \mathbf{1})$, where $\mathbf{I}$ is the $k \times k$ identity matrix. The Gaussian Mechanism satisfies $\frac{1}{2\sigma^2}$-zCDP.*

Proposition 2.5 (zCDP of exponential mechanism; Cesar and Rogers 2021)**.** *Let $\epsilon > 0$ and $\text{Score} : \mathcal{R} \times \mathcal{X}^n \to \mathbb{R}$ be a function such that $\text{Score}(r, D)$ is the quality score of candidate $r \in \mathcal{R}$ for data set D. The exponential mechanism (McSherry and Talwar, 2007) outputs a candidate $r \in \mathcal{R}$ according to the following distribution: $\Pr[\mathcal{M}(D) = r] \propto \exp\left(\frac{\epsilon}{2\Delta_1}\text{Score}(r, D)\right)$, where $\Delta_1 = \sup_{r\in\mathcal{R}} \Delta_1(\text{Score}(r, D))$. The exponential mechanism satisfies $\frac{\epsilon^2}{8}$-zCDP.*

Later, we suppress the dependence on D and write $\text{Score}(r)$ when it is clear from context.

Our method adaptively selects which building block mechanisms to use and how much privacy budget to allocate at each round based on the output of previous rounds. Because of this, we use the following result for fully adaptive composition instead of more basic results for non-adaptive composition.

Proposition 2.6 (Fully adaptive composition for zCDP; Whitehouse et al. 2022)**.** *Let $(\mathcal{M}_i)_{i\geq 1}^{\ell}$ be a sequence of adaptively chosen mechanisms and $(\rho_i)_{i=1}^{\ell}$ be a sequence of adaptively chosen privacy parameters such that $\mathcal{M}_i$ satisfies ρ_i-zCDP for $1 \leq i \leq \ell$. Let $\mathcal{M}_{1:\ell}$ denote the mechanism releasing output $(\mathcal{M}_1, \ldots, \mathcal{M}_\ell)$. If it is always the case that $\Sigma_{i=1}^{\ell}\rho_i \leq \rho$ then the mechanism $\mathcal{M}_{1:\ell}$ satisfies ρ-zCDP.*

2.2 Marginals and Workloads

A marginal is a collection of linear queries that captures low-dimensional structure of the data distribution. Given a subset of attributes $\tau \subseteq \{1, \ldots, m\}$, the marginal on τ is a histogram over the possible values the attributes in τ can take.

Definition 2.7. Let $\tau \subseteq \{1, \ldots, m\}$ be a subset of attributes, $\mathcal{X}_\tau = \prod_{i\in\tau} \mathcal{X}_i$, and $n_\tau = |\mathcal{X}_\tau|$. Define $r_\tau = (r_i)_{i\in\tau}$, the restriction of record r to τ. The marginal on τ is a vector of counts $\mathbf{x} \in \mathbb{R}^{n_\tau}$ indexed by $t \in \mathcal{X}_\tau$ such that $\mathbf{x}[t] = \sum_{r\in D} \mathbf{1}[r_\tau = t]$. We denote the function that computes the marginal on τ as q_τ.

For a marginal query q_τ, the L_1 sensitivity is 2 and the L_2 sensitivity is $\sqrt{2}$. To verify this, observe that neighboring datasets differ on the values of attributes τ for at most one record, increasing a count in the histogram by one and decreasing another by one. The sensitively of measuring one of the linear queries contained in the marginal is the same as the sensitivity of measuring the entire marginal. This useful property is sometimes called "the marginal trick" and it makes marginals an efficient class of measurements.

We define a workload W as a set of linear queries. In this paper, we focus on the class of workloads consisting of marginal queries. Many synthetic data generation algorithms take a workload as an input so that the distribution of the output data can be tailored to the given workload. A workload can be general, such as the set of all marginal queries for three or fewer attributes, or it can be specific, where it may be designed with particular dataset or application in mind.

The goal of synthetic data generation is to create a mechanism that will minimize error for any given workload and any input dataset. We define a notion of error for a given workload.

Definition 2.8. Let W be a workload of marginals. The workload error of synthetic dataset S on W is defined as follows for a fixed private data set D:

$$\text{Error}_W(S) = \frac{1}{n|W|} \sum_{\tau\in W} \|q_\tau(D) - q_\tau(S)\|_1 \quad (1)$$

We write $\text{Error}_\tau(S)$ if W contains a single marginal τ.

2.3 PRIVATE-PGM

PRIVATE-PGM (McKenna et al., 2019) is a general purpose and scalable approach to combining noisy measurements into a single representation of the data distribution from which records can be sampled. Mechanisms using PRIVATE-PGM such as MST and AIM are among the state-of-the-art methods for differentially private synthetic data generation (McKenna et al., 2021a, 2022). Given some marginal queries $\tau_1, \ldots, \tau_t$ and noisy query answers $y_1, \ldots, y_t$ PRIVATE-PGM produces an estimate of the data distribution p_θ where θ are the parameters of a probabilistic graphical model. In this paper, we will take for granted that p_θ can be used to answer marginal queries $q_\tau(p_\theta)$. PRIVATE-PGM solves an optimization problem to search the space of models $\theta \in \mathcal{P}$ for one that minimizes the loss function $\sum_{i=1}^{t} ||y_i - q_{\tau_i}(p_\theta)||_2^2$.

Since PRIVATE-PGM represents the data distribution as a graphical model, it is capable of scaling effectively to high-dimensional settings. However, as noted in prior work (McKenna et al., 2019, 2021a,b, 2022; Cai et al., 2021) the complexity of PRIVATE-PGM depends crucially on the set of marginals that have been measured. PRIVATE-PGM exposes a utility method IsTractable$(\tau_1, \ldots, \tau_t)$ that determines if PRIVATE-PGM is capable of efficiently handling a given set of marginals. Efficiency-aware mechanisms that use PRIVATE-PGM must utilize this function in order to prevent the mechanism from measuring marginals that PRIVATE-PGM cannot efficiently handle (Cai et al., 2021; McKenna et al., 2022).

3 JOINT ADAPTIVE MEASUREMENTS

Given a private data set D, a public data set D_{pub}, a workload W, and a privacy budget (ϵ, δ), our goal is to design a mechanism $\mathcal{M}$ that generates synthetic data S to minimize $\text{Error}_W(S)$ while satisfying (ϵ, δ)-DP. To solve this task we follow a design pattern called "adaptive measurements" (Liu et al., 2021b) which can be applied to most private synthetic data algorithms. A general version of this algorithmic pattern is provided in Algorithm 1. We augment this framework by extending the selection step and measurement step

Algorithm 1 Adaptive Measurements; Liu et al. (2021b)

Input: Private dataset D, Workload W, zCDP privacy budget ρ
Output: Synthetic dataset S
Initialize model p_{θ_0}
for $t = 0$ **to** $T-1$ **do**
  select τ_t where model p_{θ_t} poorly approximates D.

  measure let y_i be a private measurement of the marginal τ_t made by a noise-addition mechanism.

  update $p_{\theta_{t+1}}$ from noisy measured information.

$$p_{\theta_{t+1}} \leftarrow \arg\min_{p_\theta \in \mathcal{P}} L(p_\theta; y_1, \ldots, y_t)$$

generate synthetic data S from p_{θ_T} (or some function of the iterates $p_{\theta_0}, ..., p_{\theta_T}$)

to include the public data, we also presenting a novel budgeting strategy we call "frugal budgeting". We refer to this augmented framework as "Joint Adaptive Measurements" (JAM).

3.1 Public Proxy Estimator

When trying to estimate the value of a marginal τ on the private data, we can use the public data as a proxy for the private data. To do this, we evaluate the marginal query q_τ on D_{pub} and re-scale the result by $\frac{n}{\hat{n}}$ to account for the number of records in the data sets.

The public proxy estimator does not depend on the private data at all, so it can be used without expending any privacy budget. From a privacy standpoint, the public proxy is ideal. As a statistical estimator, the public proxy is unusual because it is deterministic and biased. In contrast, measurements made with the Gaussian mechanism incur a privacy cost but they provide an unbiased estimator that has a predictable error profile based on the noise scale.

Which estimator is expected to incur more error depends on the similarity between the public and private marginals, the dimensionality of the marginal vector, and the noise scale being used for the Gaussian mechanism. By carefully designing a score function that considers these factors, we use the exponential mechanism to select the best estimator for the situation.

3.2 Public/Private Measurement

For conciseness, we use one measurement function to capture the Gaussian mechanism and the public proxy estimator.

Definition 3.1. For public data D_{pub} and private data D, the function Measure : $2^{[m]} \times \{\mathsf{priv}, \mathsf{pub}\} \to \mathbb{R}^{n_q}$ takes a marginal query q_τ, public/private indicator i, and noise parameter σ^2 and is given by

$$\text{Measure}(\tau, i; \sigma^2) = \begin{cases} q_\tau(D) + \mathcal{N}(0, \sigma^2 \mathbf{I}) & i \text{ is } \mathsf{priv} \\ q_\tau(D_{\text{pub}}) \frac{n}{\hat{n}} & i \text{ is } \mathsf{pub}. \end{cases} \quad (2)$$

3.3 Joint Candidate Set

Adaptive measurement algorithms use the exponential mechanism to select from a pool of candidate queries. To allow for the possibility of measuring a marginal query with either the public proxy estimator or the Gaussian mechanism, we include public and private versions of queries in our candidate set. We refer to this as a "joint candidate set".

Definition 3.2. Given a set of candidate queries W, the joint extension $W^{\mathsf{priv},\mathsf{pub}}$ of W is $W \times \{\mathsf{priv}, \mathsf{pub}\}$.

The indicator element $i \in \{\mathsf{priv}, \mathsf{pub}\}$ indicates whether a candidate corresponds to a measurement on the private data with the Gaussian mechanism or a measurement using the public proxy.

In Algorithm 2 we utilize the joint extension of the downward closure candidate set. This allows lower-dimensional marginals to be selected and was first used in AIM (McKenna et al., 2022).

Definition 3.3. The joint downward closure candidate set for W is $W_{\downarrow}^{\mathsf{pub},\mathsf{priv}} = \{\tau' | \tau' \subseteq \tau, \tau \in W\} \times \{\mathsf{priv}, \mathsf{pub}\}$.

To select from this set, we construct a score function that applies to both public and private candidates. In practice, we filter this set of candidates $W_{\downarrow}^{\mathsf{pub},\mathsf{priv}}$ down to a set C containing marginals that can be added to PRIVATE-PGM without rendering it intractable.

3.4 Expected Improvement Score Function

JAM-PGM uses a score function that quantifies the expected improvement in the model after making a measurement. This score function has the goal of simultaneously considering which queries are being poorly approximated by the model and which marginals could be accurately measured (on the public or private data). This idea has been explored in the private data setting by evaluating the expected error of the Gaussian mechanism (McKenna et al., 2022). When measuring a private marginal with the Gaussian mechanism, the expected error is given by $\sqrt{2/\pi}\sigma n_\tau$ for a given noise scale σ and marginal size n_τ. When measuring a public marginal, the error is fixed and can be evaluated directly. Combining these gives the following function for measurement error:

Definition 3.4. For fixed public data D_{pub} and private data D, the predicted measurement error PredError : $2^{[m]} \times \{\mathsf{priv}, \mathsf{pub}\} \to \mathbb{R}$ of a marginal query τ and measurement indicator i with noise scale σ^2 is

$$\text{PredError}(\tau, i; \sigma^2) = \begin{cases} \sqrt{2/\pi}\sigma n_\tau & i = \mathsf{priv} \\ \left\| q_\tau(D) - q_\tau(D_{\text{pub}}) \frac{n}{\hat{n}} \right\|_1 & i = \mathsf{pub} \end{cases} \quad (3)$$

Note that for each round of the algorithm, D_{pub}, D, and σ^2 will be fixed so the measurement error is given as a function of just τ and i. To estimate the improvement in the model after making a measurement, we assume that the model will match the value of the measurement on the marginal $q_\tau(p_\theta) = \text{Measure}(\tau, i, \sigma^2)$. Under this assumption, the expected error in the next round of the algorithm would be $\text{PredError}(\tau, i; \sigma^2)$. So, the difference between the current error of the model and our estimate for the error after a measurement gives us the expected improvement score function:

Definition 3.5. For a fixed model p_θ, the expected improvement score function Score : $2^{[m]} \times \{\mathsf{priv}, \mathsf{pub}\} \to \mathbb{R}$ of a marginal query τ and public/private indicator i with noise parameter σ^2 is

$$\text{Score}(\tau, i; \sigma^2) = \text{Error}_\tau(p_\theta) - \text{PredError}(\tau, i; \sigma^2) \quad (4)$$

The L_1 sensitivity of this score function is 4 because changing one record can change the value of $\text{Error}_{p_\theta}(\tau)$ by 2 and the value of $\text{PredError}(\tau, i; \sigma^2)$ by 2.

3.5 Frugal Budgeting

Besides accuracy, another advantage to measuring the public data is that it allows for what we call frugal budgeting. Each round, we allocate some portion of the remaining privacy budget to selection and some portion of the budget to measurement. If we select a public measurement, the budget allocated for measurement is unused. This allows us to take those budget savings and roll them over into the next round. Passing savings on to later rounds of the algorithm means that when a public measurement is selected, subsequent rounds will have higher budgets and lower noise scales.

3.6 Privacy Proof

The privacy analysis of Algorithm 2 is an application of Propositions 2.4 to 2.6.

Theorem 3.6. *For any number of rounds $T > 0$, budget split parameter $\alpha \in (0, 1)$, and privacy parameter $\rho > 0$, JAM-PGM satisfies ρ-zCDP.*

Proof. By construction, each round of JAM-PGM satisfies ρ^t_{used}-zCDP, where $\rho^t_{\text{used}} = \rho^t_{\text{select}} + \mathbf{1}[i_t = \mathsf{priv}] \cdot \rho^t_{\text{measure}}$ as defined in the algorithm: the selection step always satisfies ρ^t_{select}-zCDP, and the measurement step satisfies ρ^t_{measure}-zCDP if a private candidate is selected and 0-zCDP if a public candidate is selected. Also by construction, $\rho^t_{\text{used}} < \rho^t = (\rho - \sum_{s=0}^{t-1} \rho^s_{\text{used}})/(T-t)$, so we have the invariant that $\sum_{s=0}^{t} \rho^s_{\text{used}} \leq \rho$. Therefore, by Proposition 2.6, JAM-PGM satisfies ρ-zCDP. □

Algorithm 2 JAM-PGM

Input: Public dataset D_{pub}, Private dataset D, Workload W

Output: Synthetic dataset S

Hyperparameters: Privacy parameter ρ, number of rounds T, select-measure split α

Initialize $p_\theta \leftarrow \text{Uniform}[\mathcal{X}]$

for $t = 0$ **to** $T - 1$ **do**

$\rho^t \leftarrow \left(\rho - \sum_{s=0}^{t-1} \rho^s_{\text{used}}\right) / (T - t)$

$\rho^t_{\text{select}}, \rho^t_{\text{measure}} \leftarrow (1 - \alpha)\rho^t, \alpha\rho^t$

$\sigma^2_t \leftarrow 1/\rho^t_{\text{measure}}$

$C_t \leftarrow \{\tau, i \in W^{\mathsf{pub},\mathsf{priv}}_{\downarrow} | \text{Is-Tractable}(\tau, \tau_1, ..., \tau_{t-1})\}$

select τ_t, i_t from C_t using exponential mechanism with budget ρ^t_{select} and $\text{Score}(\tau, i; \sigma^2_t)$ from Equation (4)

measure τ_t publicly or privately with measurement function from Equation (2)

$$y_t = \text{Measure}(\tau_t, i_t; \sigma^2_t)$$

estimate the data distribution using PRIVATE-PGM

$$p_\theta \leftarrow \arg\min_{p'_\theta \in \mathcal{P}} \sum_{i=1}^{t} \|y_i - q_{\tau_i}(p'_\theta)\|_2^2$$

$\rho^t_{\text{used}} \leftarrow \rho^t_{\text{select}} + \mathbf{1}[i_t = \mathsf{priv}] \cdot \rho^t_{\text{measure}}$

generate synthetic data S from p_θ

4 PRIOR WORK

The field of DP synthetic data, also known as DP query release, has a rich history. The adaptive measurements framework provides useful language to describe a number of methods in a unified framework. Many of these algorithms draw inspiration from the MWEM algorithm (Hardt et al., 2012), the first to iteratively refine a data model by selecting poorly approximated queries in each round. Since MWEM, various data models, selection criteria, and optimization procedures have been explored in the literature (Hardt et al., 2012; Liu et al., 2021b; Gaboardi et al., 2014; Aydore et al., 2021; McKenna

et al., 2021a; Cai et al., 2021; McKenna et al., 2022).

The first method to make use of public data for synthetic data generation was PMW$^{\text{Pub}}$ (Liu et al., 2021a). This method is a version of the MWEM algorithm that incorporates the public data in two ways: initialization and domain restriction. The original MWEM algorithm represents the data distribution as a histogram where each entry corresponds to an element of the data universe $\mathcal{X}$. It then uses an adaptive measurement strategy along with a multiplicative weights update rule to update that data distribution. One problem with this algorithm is that the size of the histogram representation scales exponentially with the dimensionality of the data. To get around this, PMW$^{\text{Pub}}$ restricts the histogram to elements of the data universe that are present in the public data. It also initializes that histogram to match the distribution of the public data. When the public and private data distributions are similar, this greatly improves performance. But when the distributions are very different, the domain restriction can make it impossible to generate good synthetic data.

This domain restriction problem was addressed in subsequent work, which introduced the GEM$^{\text{Pub}}$ method (Liu et al., 2021b). This method represents the data distribution with a generator network and does not restrict the domain of the distribution. Instead, GEM$^{\text{Pub}}$ pre-trains on the public data to initialize the model. By doing this, GEM$^{\text{Pub}}$ realizes gains in performance without overly committing to public data, which may do a poor job of reflecting the private data distribution.

These works incorporate the public data into the initialization step of the adaptive measurements framework. This is fundamentally incompatible with methods that do not fix model structure a priori, such at PRIVATE-PGM. In a graphical model, the number and structure of parameters depend on the edges in the graph. In PRIVATE-PGM, edges are determined by the choice of marginal measurements so the model structure is not yet defined during the initialization step. In contrast, the joint selection framework incorporates the public data into the select and measure steps of the adaptive measurements framework.

5 EXPERIMENTS

In this section, we evaluate the performance of JAM-PGM against the baseline methods PMW$^{\text{Pub}}$ (Liu et al., 2021a) and GEM$^{\text{Pub}}$ (Liu et al., 2021b) which both incorporate public data. We also compare to AIM (McKenna et al., 2022) and GEM (Liu et al., 2021b) which do not use public data. Here we provide the key details of the experimental setup. Additional details on the compute environment and code are provided in the supplementary materials. Code for running our methods is provided on GitHub[1].

Task Specification The task is to minimize workload error while satisfying (ϵ, δ)-DP. We evaluate on the privacy parameters $\epsilon \in \{0.03, 0.10, 0.31, 1.00, 3.16, 10.00\}$ and $\delta = 1 \times 10^{-9}$. The workload provided was all 3-way marginals and 5 trials were run with 5 random seeds to provide standard error estimates.

Data Our first experiment is performed on the ADULT dataset (Kohavi et al., 1996). We split the dataset into public and private datasets using a stratified sampling strategy to ensure that the public data distribution is different from the private data distribution. We sampled a private dataset with 32,384 records such that 25% of those records are female. Then, we constructed various public datasets with 3,238 records each varying the percentage of female records.

We also experiment with the following datasets: SALARY (Hay et al., 2016), FIRE (Ridgeway et al., 2021), NIST-TAXI (Grégoire et al., 2021), and TITANIC (Harrell and Cason, 1994). We split the data by randomly selecting private records and iteratively adding them to the public dataset until the public data met an error target when used as a proxy for the private data. The TITANIC dataset contains about one thousand records total while the other datasets contain hundreds of thousands of records. Because of this, the behaviour of the various methods on TITANIC differ from the behaviour on the other datasets and the results are discussed separately. Information regarding the number of attributes, number of private data records, number of public data records, average 3-way marginal size, and total domain size of the dataset used in our experiments can be seen in appendix Table 1.

Hyperparameter Selection The main hyperparameter for each method is the number of rounds T (except for AIM which determines T adaptively). For all methods, we conducted limited preliminary experiments to non-privately select a value or values for this parameter. For PMW$^{\text{Pub}}$ and JAM-PGM it was relatively easy to find a single value that performed well across a range of epsilon values. However, this was not possible for GEM$^{\text{Pub}}$, which performed better with fewer rounds at low privacy budgets, and more rounds at high privacy budgets.

We interpret this behavior as implicitly selecting "how much" to use the public data. With low epsilon (high privacy noise), it is beneficial to run for few rounds and remain close to the public data initialization, while

[1] https://github.com/Miguel-Fuentes/JAM_AiStats/

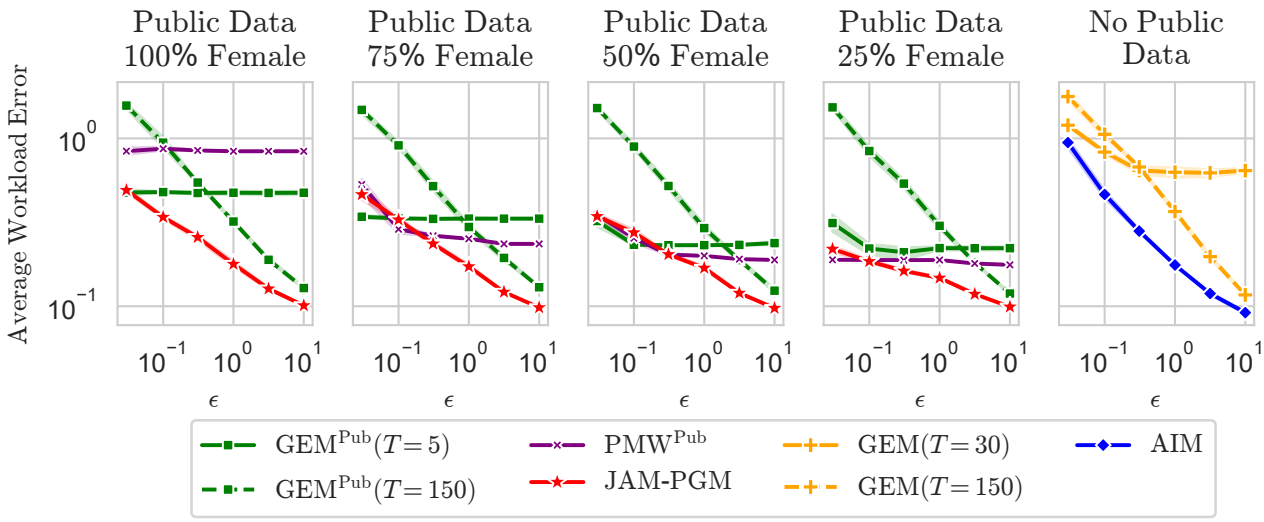

Figure 1: Average workload error (workload of all 3 way marginals) for ϵ in {0.03, 0.10, 0.31, 1.00, 3.16, 10.00} and $\delta = 1 \times 10^{-9}$ for the ADULT data set. Private dataset consists of 25% female records while public dataset consists of 100%, 75%, 50%, and 25% female records respectively. The plots are shown with the amount of public data bias decreasing from left to right.

with high epsilon (low privacy noise) it is beneficial to run for more rounds. To fairly represent the range of performance possible with GEM$^{\text{Pub}}$, we selected two values for T, the highest and lowest ones that performed reasonably in preliminary experiments. More details of hyperparameter selection appear in the appendix.

5.1 Varying Public Data Bias

The results of the ADULT experiment can be seen in Figure 1. For clarity, the methods that do not use public data are shown on a separate subplot.

Across all levels of public data bias and almost all values of ϵ, JAM-PGM achieved lower average workload error than the baseline methods. All of the methods that utilize the public data perform better when the public data distribution is more similar to the private data distribution. However, as the public data bias increases, the performance of JAM-PGM degrades more slowly than the other public data methods. Another change related to the public data bias is the relationship between the error curves of GEM$^{\text{Pub}}$ with $T = 5$ and GEM$^{\text{Pub}}$ with $T = 150$. When the public data are from the same distribution as the private data, it is beneficial to run GEM$^{\text{Pub}}$ for fewer rounds for a given value of ϵ but as the bias increases it becomes beneficial to run GEM$^{\text{Pub}}$ for more rounds for a given value of ϵ.

In the setting where the public data consists of only female records, we see that PMW$^{\text{Pub}}$ has much worse performance than in the settings where the public data includes male records. This highlights the risk described in prior work (Liu et al., 2021a,b): if the public and private data are not sufficiently compatible, PMW$^{\text{Pub}}$ will be unable to produce a good model of the private data regardless of privacy budget.

Note that while the performance of JAM-PGM and AIM are similar for high values of ϵ, the average workload error of AIM is slightly lower than that of JAM-PGM for $\epsilon = 10$ across all levels of public data bias.

5.2 Small Data Set

The results of the TITANIC experiment can be seen in Figure 2. In this setting, where the private data set consists of only one thousand records, there is a large separation between the methods that utilize public data and those that do not up to $\epsilon = 3.16$. PMW$^{\text{Pub}}$ performs best for $\epsilon \leq 0.31$, but it does not improve as the privacy budget increases. For larger values of ϵ, JAM-PGM and GEM$^{\text{Pub}}$ have very similar average workload error.

Notice that for most values of ϵ in this experiment, the number of rounds that performs best for the GEM$^{\text{Pub}}$ algorithm is 1. Increasing the number of rounds to 5 significantly decreases performance for $\epsilon < 3.16$. This suggests that in this setting, the model initialization is much more informative than the noisy measurements. In cases like these with few records and low privacy budget, it may be advantageous to not make any measurements at all and simply rely on the public data instead.

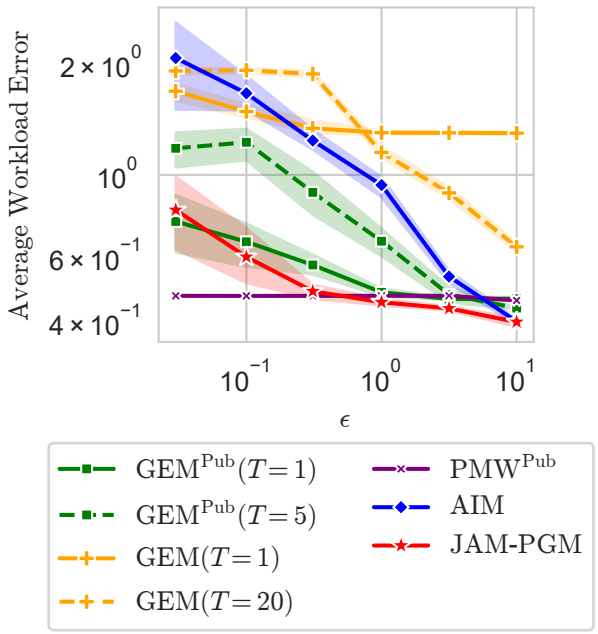

Figure 2: Average workload error (workload of all 3-way marginals) for ϵ in $\{0.03, 0.10, 0.31, 1.00, 3.16, 10.00\}$ and $\delta = 1 \times 10^{-9}$ for TITANIC data set.

5.3 Larger Data Sets

The results of the SALARY , FIRE , and NIST-TAXI experiments can be seen in Figure 3. In these experiments, JAM-PGM outperforms the other public-data utilizing mechanisms for all values of epsilon (except for $\epsilon = 10$ on the NIST-TAXI dataset). However, none of the methods that incorporate public data are optimal in these settings. AIM performs best in the SALARY experiment for $\epsilon > 1$ and in the FIRE experiment for $\epsilon > 0.1$ while GEM performs best in the NIST-TAXI experiment for $\epsilon \geq 3.16$. We might hope that having access to additional information would never hurt the performance of these synthetic data mechanisms but these experiments show that this is not the case.

6 DISCUSSION

Our development of the "joint adaptive measurements" framework incorporates public data into the *selection* and *measurement* steps of the "adaptive measurements" framework. This expands the design space for public-data assisted DP algorithms; using this technique, we develop JAM-PGM, which is able to use joint selection and PRIVATE-PGM in order to effectively select a combination of public measurements and private measurements and utilize them to generate high quality DP synthetic data.

Limitations One limitation of the joint selection framework is the lack of public data error estimates, a benefit of noise-addition mechanisms is that the distribution of noise is known and can be released publicly. PRIVATE-PGM (McKenna et al., 2019) uses this information to weight the measurements provided in its optimization objective based on the noise scale. AIM also uses this information to perform budget annealing which dynamically determines the number of rounds (McKenna et al., 2022). The error of a public measurement is sensitive information and budget would need to be spent to measure it. Because of this, JAM-PGM gives equal weight to all measurements in the PRIVATE-PGM loss function and uses a fixed number of rounds instead of a budget annealing strategy. Another limitation, inherited from PRIVATE-PGM, is that JAM-PGM can not be applied to attributes with continuous values without discretization this limitation is common in this area and also applies to the other methods we compare against.

Suboptimality of Publicly-Assisted Methods: Ideally, utilizing public data would strictly improve performance compared to not utilizing public data. Our experiments show that no public-data assisted mechanism currently achieves this. On all of the data sets we tested except for TITANIC , there were some values of ϵ for which none of the methods that incorporate public data outperform the baselines that do not use public data. This indicates that all of the techniques that have been developed to incorporate public data (domain restriction, pre-training, and joint selection) have some risk of performing worse than baseline methods that ignore public data. On the other hand, the public techniques clearly provide sizeable boosts in performance when the distribution of the public data closely matches the distribution or when ϵ is small. Navigating this trade-off in practice may be difficult for practitioners.

Benefits of JAM-PGM Across all our experiments, JAM-PGM achieved lower average workload error than the public data baselines for most values of ϵ. In addition, the adaptive nature of JAM-PGM alleviates some of the difficulties associated with the use of public data in practice. Users will not know a priori exactly how similar their public data and private data distributions are, so they may be hesitant to use PMW$^{\text{Pub}}$ because of the risks associated with domain restriction. Similarly but to a lesser extent, selecting the number of rounds for GEM$^{\text{Pub}}$ may be challenging because there is no setting for this parameter that works well across values of ϵ and for a given ϵ the optimal number of rounds depends on the data set. JAM-PGM is a safe choice for DP synthetic data generation. Unlike PMW$^{\text{Pub}}$, it does

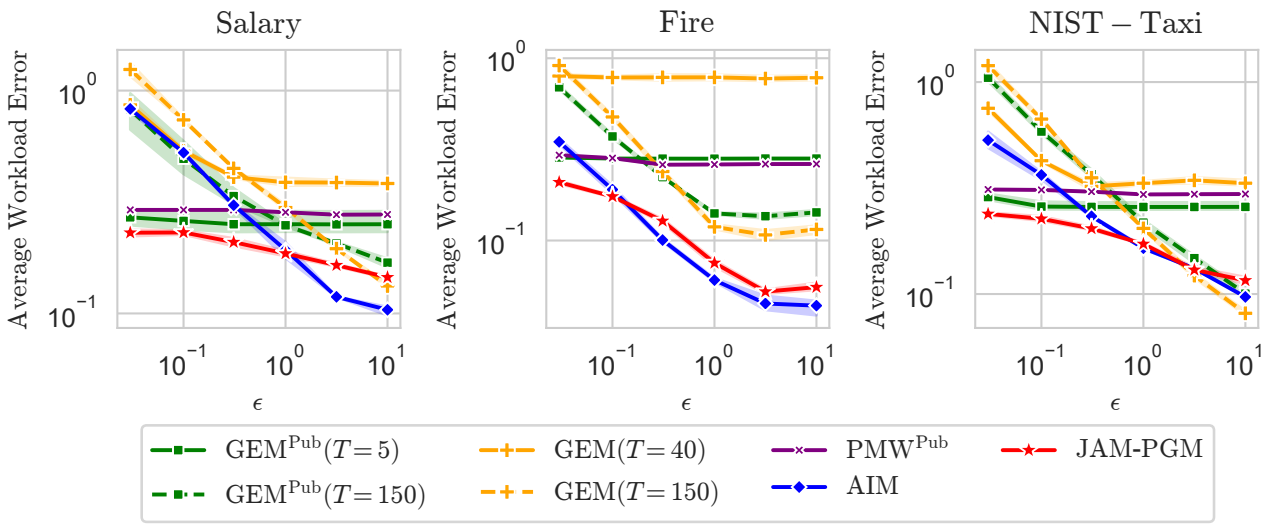

Figure 3: Average workload error (workload of all 3-way marginals) for ϵ in {0.03, 0.10, 0.31, 1.00, 3.16, 10.00} and $\delta = 1 \times 10^{-9}$ for the SALARY , FIRE , and NIST-TAXI datasets.

not have the risk associated with restricting the domain of the model based on public data. Additionally, there exist settings for the number of rounds parameter T that work well across a range of privacy budgets which may make it easier to use in practice than GEM$^{\text{Pub}}$.

Bounded vs Unbounded DP We adopt bounded DP, defining dataset neighbors as those differing by swapping a single record. Under this regime, the number of records in the private dataset is considered public information. This is useful when scaling public marginals to match the number of records in the private dataset. In contrast, unbounded DP considers neighbors based on adding or removing a single record, necessitating private estimation of the number of private records. While estimating this number is straightforward, for simplicity, we opt for bounded DP, omitting the need for a separate estimation step.

Acknowledgements

We would like to thank Cecilia Ferrando and Javier Burroni for their helpful comments on earlier drafts of this paper. This material is based upon work supported by the National Science Foundation under Grant No. 1749854.

References

Alon, N., Bassily, R., and Moran, S. (2019). Limits of private learning with access to public data. *Advances in neural information processing systems*, 32.

Amid, E., Ganesh, A., Mathews, R., Ramaswamy, S., Song, S., Steinke, T., Suriyakumar, V. M., Thakkar, O., and Thakurta, A. (2022). Public data-assisted mirror descent for private model training. In *International Conference on Machine Learning*, pages 517–535. PMLR.

Asghar, H. J., Ding, M., Rakotoarivelo, T., Mrabet, S., and Kaafar, D. (2020). Differentially private release of datasets using gaussian copula. *Journal of Privacy and Confidentiality*, 10(2).

Aydore, S., Brown, W., Kearns, M., Kenthapadi, K., Melis, L., Roth, A., and Siva, A. A. (2021). Differentially private query release through adaptive projection. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 457–467. PMLR.

Bassily, R., Cheu, A., Moran, S., Nikolov, A., Ullman, J., and Wu, S. (2020). Private query release assisted by public data. In *International Conference on Machine Learning*, pages 695–703. PMLR.

Bassily, R., Thakkar, O., and Guha Thakurta, A. (2018). Model-agnostic private learning. *Advances in Neural Information Processing Systems*, 31.

Bowen, C. M. and Liu, F. (2020). Comparative study of differentially private data synthesis methods. *Statistical Science*, 35(2):280–307.

Bun, M. and Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer.

Cai, K., Lei, X., Wei, J., and Xiao, X. (2021). Data synthesis via differentially private markov random fields.

Proceedings of the VLDB Endowment, 14(11):2190–2202.

Canonne, C. L., Kamath, G., and Steinke, T. (2020). The discrete gaussian for differential privacy. In *NeurIPS*.

Cesar, M. and Rogers, R. (2021). Bounding, concentrating, and truncating: Unifying privacy loss composition for data analytics. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 421–457.

Charest, A. (2011). How can we analyze differentially-private synthetic datasets? *Journal of Privacy and Confidentiality*, 2(2).

Chen, R., Xiao, Q., Zhang, Y., and Xu, J. (2015). Differentially private high-dimensional data publication via sampling-based inference. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 129–138. ACM.

Dwork, C., Nissim, F. M. K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284.

Gaboardi, M., Arias, E. J. G., Hsu, J., Roth, A., and Wu, Z. S. (2014). Dual query: Practical private query release for high dimensional data. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1170–1178. JMLR.org.

Ge, C., Mohapatra, S., He, X., and Ilyas, I. F. (2021). Kamino: Constraint-aware differentially private data synthesis. *Proceedings of the VLDB Endowment*, 14(10):1886–1899.

Grégoire, L., Task, C., Slavitt, I., Nicolas, G., Bhagat, K., Streat, D., and Howarth, G. (2021). SDNist: Benchmark Data and Evaluation Tools for Data Sythesizers.

Hardt, M., Ligett, K., and McSherry, F. (2012). A simple and practical algorithm for differentially private data release. In Bartlett, P. L., Pereira, F. C. N., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 2348–2356.

Harrell, F. and Cason, T. (1994). Encyclopedia titanica.

Hay, M., Machanavajjhala, A., Miklau, G., Chen, Y., and Zhang, D. (2016). Principled evaluation of differentially private algorithms using dpbench. In *Proceedings of the 2016 International Conference on Management of Data*, pages 139–154.

Ji, Z. and Elkan, C. (2013). Differential privacy based on importance weighting. *Machine learning*, 93(1):163–183.

Jordon, J., Yoon, J., and van der Schaar, M. (2019). PATE-GAN: generating synthetic data with differential privacy guarantees. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kairouz, P., Diaz, M. R., Rush, K., and Thakurta, A. (2021). (nearly) dimension independent private erm with adagrad rates via publicly estimated subspaces. In *Conference on Learning Theory*, pages 2717–2746. PMLR.

Kohavi, R. et al. (1996). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207.

Liu, T., Vietri, G., Steinke, T., Ullman, J. R., and Wu, Z. S. (2021a). Leveraging public data for practical private query release. In *ICML*, pages 6968–6977.

Liu, T., Vietri, G., and Wu, S. (2021b). Iterative methods for private synthetic data: Unifying framework and new methods. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

McKenna, R., Miklau, G., and Sheldon, D. (2021a). Winning the nist contest: A scalable and general approach to differentially private synthetic data. *Journal of Privacy and Confidentiality*, 11(3).

McKenna, R., Mullins, B., Sheldon, D., and Miklau, G. (2022). Aim: An adaptive and iterative mechanism for differentially private synthetic data. *Proc. VLDB Endow.*, 15(11):2599–2612.

McKenna, R., Pradhan, S., Sheldon, D. R., and Miklau, G. (2021b). Relaxed marginal consistency for differentially private query answering. *Advances in Neural Information Processing Systems*, 34.

McKenna, R., Sheldon, D., and Miklau, G. (2019). Graphical-model based estimation and inference for differential privacy. In *International Conference on Machine Learning*, pages 4435–4444.

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *FOCS*.

Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I. J., and Talwar, K. (2017). Semi-supervised knowledge transfer for deep learning from private training data. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Ridgeway, D., Theofanos, M., Manley, T., Task, C., et al. (2021). Challenge design and lessons learned

from the 2018 differential privacy challenges. Technical report, NIST.

Vietri, G., Tian, G., Bun, M., Steinke, T., and Wu, Z. S. (2020). New oracle-efficient algorithms for private synthetic data release. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9765–9774. PMLR.

Wang, J. and Zhou, Z.-H. (2020). Differentially private learning with small public data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(04), pages 6219–6226.

Whitehouse, J., Ramdas, A., Rogers, R., and Wu, Z. S. (2022). Fully adaptive composition in differential privacy. *arXiv preprint arXiv:2203.05481*.

Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. (2018). Differentially private generative adversarial network. *CoRR*, abs/1802.06739.

Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H. A., Kamath, G., Kulkarni, J., Lee, Y. T., Manoel, A., Wutschitz, L., et al. (2021). Differentially private fine-tuning of language models. In *International Conference on Learning Representations*.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. (2017). Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):25:1–25:41.

Zhang, X., Ji, S., and Wang, T. (2018). Differentially private releasing via deep generative model (technical report). *arXiv preprint arXiv:1801.01594*.

Zhang, Z., Wang, T., Li, N., Honorio, J., Backes, M., He, S., Chen, J., and Zhang, Y. (2021). Privsyn: Differentially private data synthesis. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 929–946. USENIX Association.

Zhou, Y., Wu, S., and Banerjee, A. (2020). Bypassing the ambient dimension: Private sgd with gradient subspace identification. In *International Conference on Learning Representations*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

1 Data

Here we describe the preprocessing and public/private splitting strategies applied to the ADULT , FIRE , SALARY , NIST-TAXI , and TITANIC data sets. We also provide a table showing additional information about those data sets.

1.1 Prepossessing

In order to be consistent with prior work, we follow the preprocessing steps described in (McKenna et al., 2022). The first step is attribute selection, again following the lead of McKenna et al. (2022), we keep all the attributes from the ADULT , SALARY , and TITANIC datasets but we remove the 15 attributes relating to incident times in the FIRE data set because after discretization, they contain redundant information. Next, we identify the domain of each data set. Normally a data provider would make this information public separately from the records in the data set, but this was not the case with these data sets. Therefore, we determine the domain by looking at the records in the data set. For each categorical attribute, we list the set of observed values and treat that as the set of possible values for that attribute. For each numerical attribute, we record the minimum and maximum observed value for that attribute. Finally, we discretize the continuous attributes by 32 equal-width bins, using the min/max values determined in the prior step.

1.2 Public/Private Split

Here, we describe how we split the data into a private data sets and public data sets. The public data that was generated by variable stratification is intentionally biased with respect to the private data, because of this, the public proxy estimator error will not approach 0 as the number of public records increases. To generate public data without variable stratification we sampled very small public data sets so that the sampling error would be significant enough that the algorithms would still need to access the private data.

1.2.1 Variable Stratification

We performed variable stratification on theADULT dataset based on the sex attribute (Kohavi et al., 1996). To do this, we split male records from the female records, then we sampled a private dataset with 32,384 records such that 25% of those records are female. The next step was to sample four public datasets with 3,238 records each such that 100%, 75%, 50%, and 25% records were female. To generate a public data set with p% female records, we would randomly sample $0.01 * p * 3,238$ of the remaining female records and $0.01 * (1 - p) * 3,238$ of the remaining male records.

1.2.2 Public Error Targeting

To split the SALARY , FIRE , NIST-TAXI , and TITANIC data sets, we set a target for public proxy error. Public proxy error is given by $\frac{1}{|W|} \sum_{\tau \in W} ||q_\tau(D) - \frac{n}{\hat{n}} q_\tau(D_{\text{pub}})||_1$. To create a split with the desired public proxy error, we started with one record in public data set. Then, we evaluated the public proxy error, if it was greater than the public error target we would double the size of the public data set by removing records from the private data set. We repeated this process until the public data met the error target. This doubling process gave a rough estimate for the number of public records but would give public error that was too far below the target. To resolve this, we would round down the number or records to the nearest hundred or thousand until we got closer to the error target. The original error target for all four data sets was 0.3. This target was reached for the SALARY , FIRE , and NIST-TAXI data sets. The TITANIC data set is smaller than the others so we could not find a split that achieved the 0.3 target; because of this, we changed the target to 0.5.

Dataset Name	Columns	n	$\hat{n}$	Avg 3-way Marginal Size	Total Domain Size
ADULT	15	32,384	3,238	5.82×10^3	4.09×10^{16}
TITANIC	9	1,004	300	2.07×10^3	8.92×10^7
SALARY	9	131,727	4,000	1.64×10^5	1.34×10^{13}
FIRE	15	304,249	870	3.50×10^3	4.21×10^{15}
NIST-TAXI	10	223,551	2,500	2.76×10^4	1.87×10^{13}

Table 1: Additional information for the datasets used in our experiments: Number of columns, number of private data records, nuber of public data records, average 3-way marginal size, and total domain size

1.2.3 Additional Dataset Information

2 Additional Experimental Details

Code: To run AIM and use PRIVATE-PGM in the context of JAM-PGM, we used the code provided by the authors at https://github.com/ryan112358/private-pgm. The AIM code assumed unbounded DP, whereas JAM-PGM assumes bounded DP. In order to compare fairly, we modified the sensitivity values in the AIM code to use bounded DP sensitivities. The code to run JAM-PGM and the version of AIM with bounded DP will be available publicly at the time of publication. To run GEM$^{\text{Pub}}$, we use the code provided by the authors at https://github.com/terranceliu/dp-query-release. To run PMW$^{\text{Pub}}$, we use the code provided by the authors at https://github.com/terranceliu/pmw-pub.

Compute Environment: All experiments were run on internal compute clusters. The GEM$^{\text{Pub}}$ code is compatible with GPU acceleration, so it was run with on various types of NVIDIA GPUs, mostly (GeForce GTX TITAN X GPUs).

Size Limit: The PRIVATE-PGM based methods require setting a size limit. We used a size limit of 80MB for both AIM and JAM-PGM across all experiments. When applying a size limit to a PRIVATE-PGM based method. Our experiments ran with an adaptive size limit; during each round t a size limit s^t is determined based on the amount of privacy budget used so far $s^t = s_{\text{total}}(\frac{\rho_{\text{used}}^{t-1}}{\rho})$. This adaptive size limit ensures that the model grows slowly over the course of the rounds, which also speeds up inference in the early rounds because the model is guaranteed to be smaller.

3 Hyperparameters

The effect of changing the number of rounds hyperparameter T for JAM-PGM, GEM$^{\text{Pub}}$, GEM, and PMW$^{\text{Pub}}$ are shown in visually in Figures 1-10. JAM-PGM has one other hyperparameter that was not searched over and that was $\alpha = 0.8$. GEM$^{\text{Pub}}$ and GEM have several other hyperparameters that were not searched over in (Liu et al., 2021b) so we followed their lead and kept those constant. Those hyperparameters were hidden layer sizes of $(512, 1024, 1024)$, learning rate 0.0001, $B = 1000$, and $\alpha = 0.67$.

3.1 GEM$^{\text{Pub}}$ and GEM Round Sensitivity

Across many of our experiments, we see that the version of GEM that does not incorporate public data does have a relationship between the privacy budget and the optimal number of rounds. However, running for a relatively high number of rounds tends to lead to relatively good performance across all privacy budgets. The same is not true for GEM$^{\text{Pub}}$, once public data is incorporated the gap between high numbers of rounds and low numbers of rounds widens when the privacy budget is small. This is especially true when the public data are very representative of the private data.

3.1.1 ADULT Data Set

The sensitivity in performance of GEM$^{\text{Pub}}$ with respect to the number of rounds hyperparameter on the ADULT datasets are given in Figure 5. The sensitivity in performance of GEM with respect to the number of rounds hyperparameter on the private ADULT data is given in Figure 4.

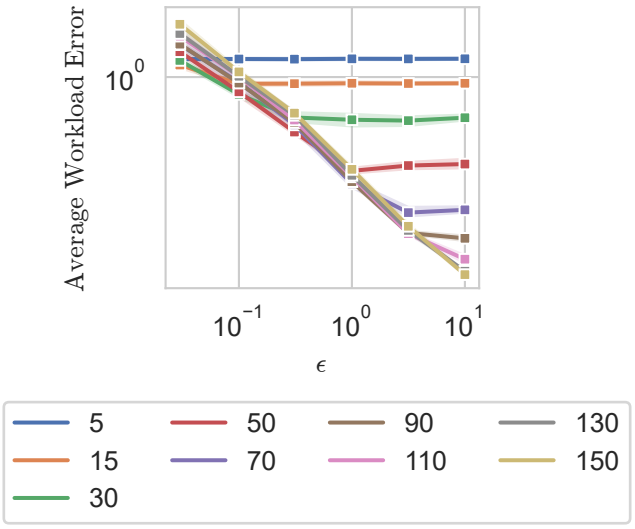

Figure 4: Average workload error for GEM on the private ADULT data set. The colors indicate the setting of the rounds hyperparameter T.

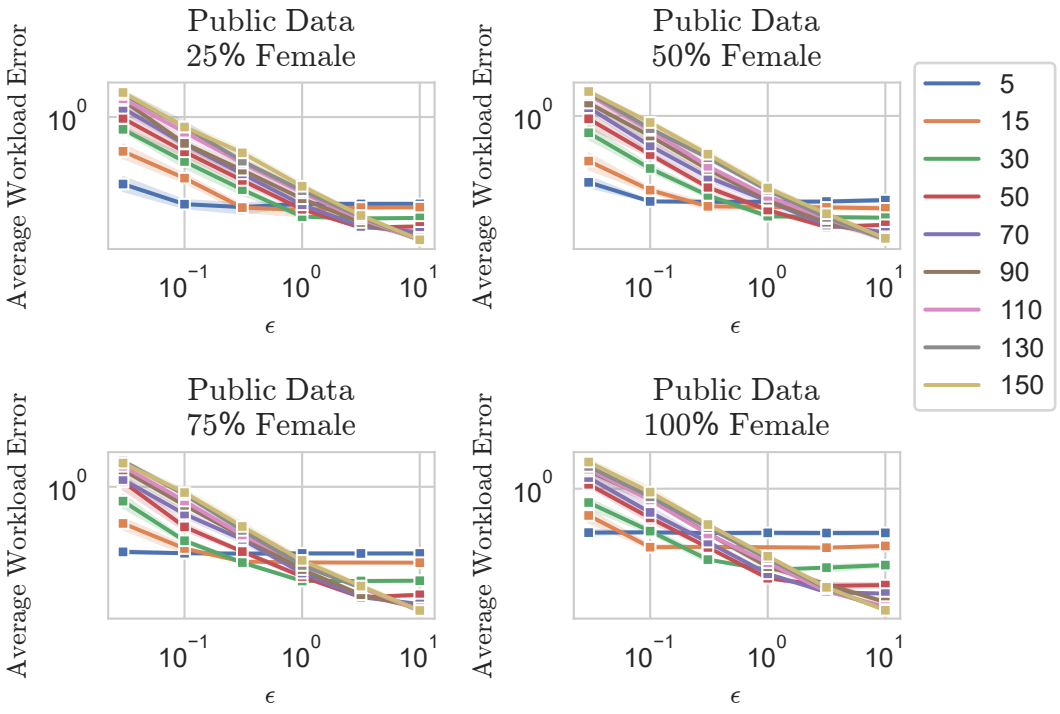

Figure 5: Average workload error for GEM$^{\text{Pub}}$ on the ADULT data sets. The colors indicate the setting of the rounds hyperparameter T.

3.1.2 TITANIC Data Set

The sensitivity in performance of GEM$^{\text{Pub}}$ and GEM with respect to the number of rounds hyperparameter on the TITANIC data set are given in Figure 6.

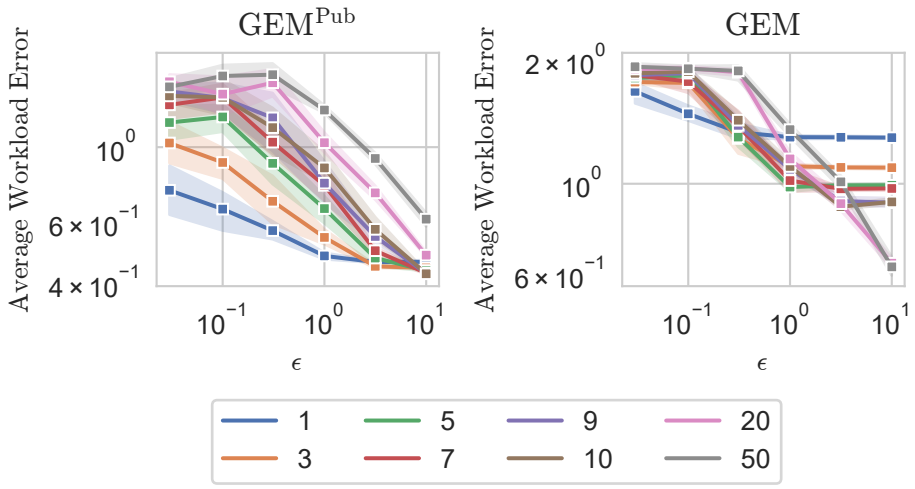

Figure 6: Average workload error for GEM$^{\text{Pub}}$ on the TITANIC data set. The colors indicate the setting of the rounds hyperparameter T.

3.1.3 Large Data Sets

The sensitivity in performance of GEM$^{\text{Pub}}$ and GEM with respect to the number of rounds hyperparameter on the SALARY , FIRE , and NIST-TAXI data sets are given in Figure 7.

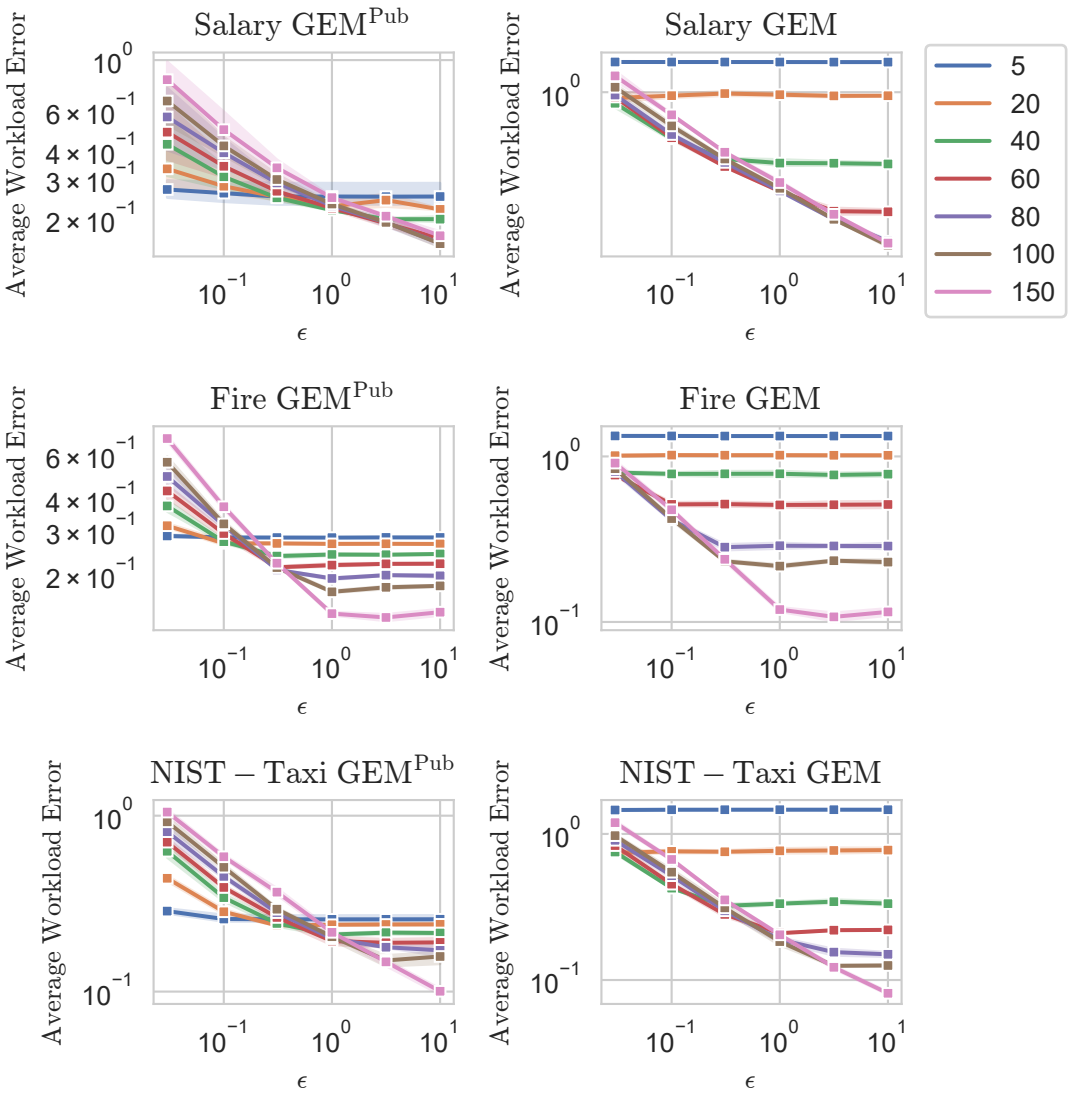

Figure 7: Average workload error for GEM$^{\text{Pub}}$ on the SALARY , FIRE , and NIST-TAXI data sets. The colors indicate the setting of the rounds hyperparameter T.

3.2 PMW$^{\text{Pub}}$ Round Sensitivity

The average workload error of PMW$^{\text{Pub}}$ is very insensitive to the number of rounds, notice that the y axes of the figures in this section have a very small range.

3.2.1 ADULT Data Set

The sensitivity in performance of PMW$^{\text{Pub}}$ with respect to the number of rounds hyperparameter on the ADULT data sets are given in Figure 8.

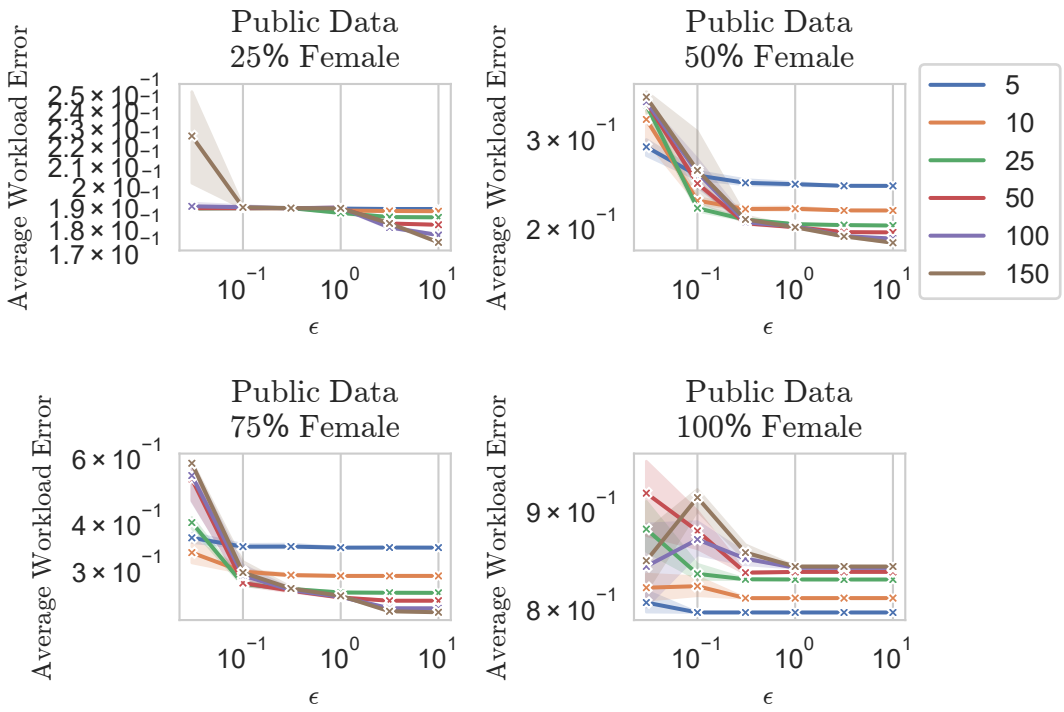

Figure 8: Average workload error for PMW$^{\text{Pub}}$ on the ADULT data sets. The colors indicate the setting of the rounds hyperparameter T.

3.2.2 TITANIC Data Set

The sensitivity in performance of PMW$^{\text{Pub}}$ with respect to the number of rounds hyperparameter on the TITANIC data set is given in Figure 9.

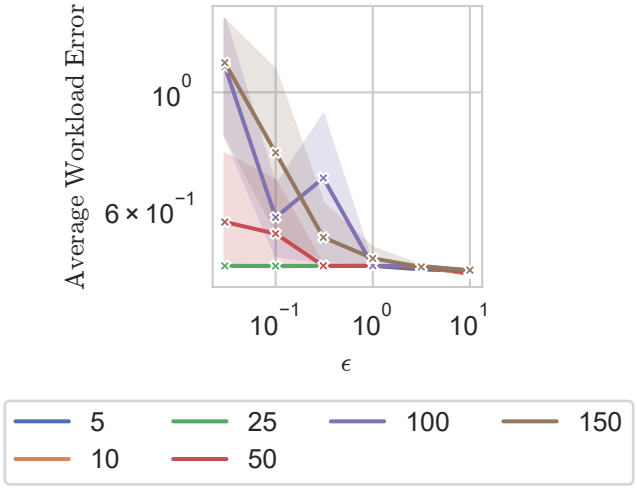

Figure 9: Average workload error for PMW$^{\text{Pub}}$ on the TITANIC data set. The colors indicate the setting of the rounds hyperparameter T.

3.2.3 Large Data Sets

The sensitivity in performance of PMW$^{\text{Pub}}$ with respect to the number of rounds hyperparameter on the SALARY , FIRE , and NIST-TAXI data sets are given in Figure 10.

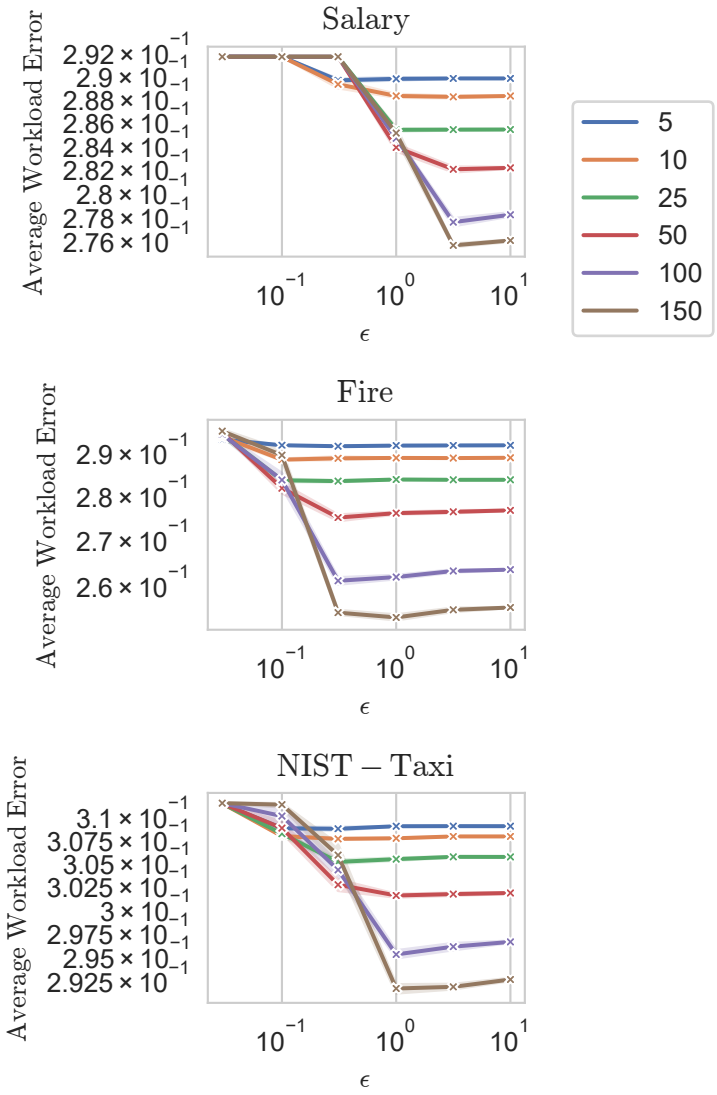

Figure 10: Average workload error for PMW$^{\text{Pub}}$ on the SALARY , FIRE , and NIST-TAXI data sets. The colors indicate the setting of the rounds hyperparameter T.

3.3 JAM-PGM Round Sensitivity

The performance of JAM-PGM is not very sensitive to the number of rounds, it seems that if the number of rounds is too low performance suffers but as long as the number of rounds is high enough the performance does not vary much.

3.3.1 ADULT Data Set

The sensitivity in performance of JAM-PGM with respect to the number of rounds hyperparameter on the ADULT data sets are given in Figure 11.

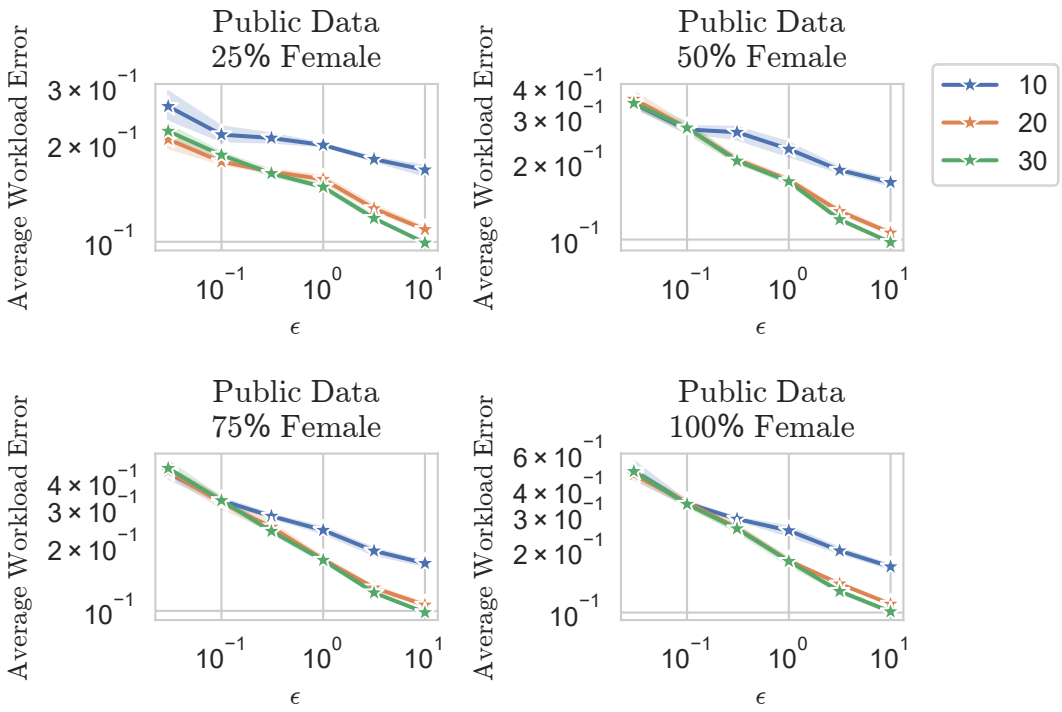

Figure 11: Average workload error for JAM-PGM on the ADULT data sets. The colors indicate the setting of the rounds hyperparameter T.

3.3.2 TITANIC Data Set

The sensitivity in performance of JAM-PGM with respect to the number of rounds hyperparameter on the TITANIC data set is given in Figure 12.

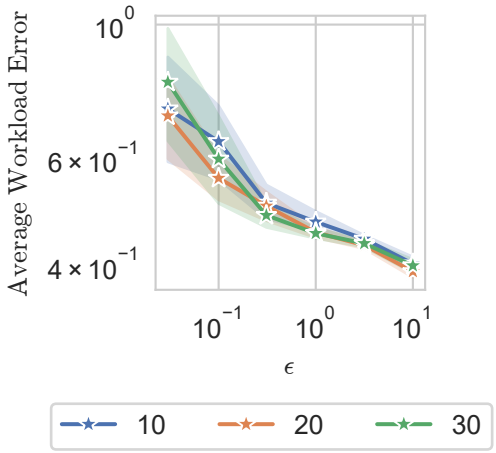

Figure 12: Average workload error for JAM-PGM on the TITANIC data set. The colors indicate the setting of the rounds hyperparameter T.

3.3.3 Large Data Sets

The sensitivity in performance of JAM-PGM with respect to the number of rounds hyperparameter on the SALARY , FIRE , and NIST-TAXI data sets are given in Figure 13.

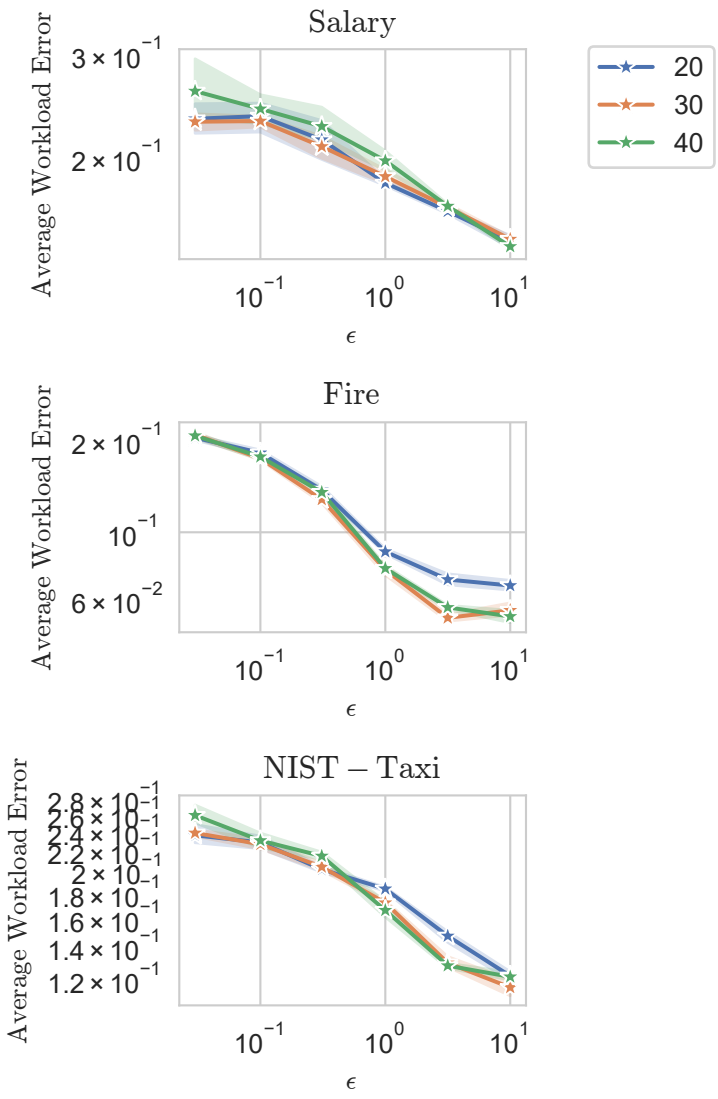

Figure 13: Average workload error for JAM-PGM on the SALARY , FIRE , and NIST-TAXI data sets. The colors indicate the setting of the rounds hyperparameter T.