2024 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 22–25, 2024, LONDON, UK

ADAPTIVE PRIVACY FOR DIFFERENTIALLY PRIVATE CAUSAL GRAPH DISCOVERY

Payel Bhattacharjee, Ravi Tandon

Department of Electrical and Computer Engineering
University of Arizona, Tucson, AZ, USA.
E-mail: {*payelb, tandonr*}@arizona.edu

2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP) | 979-8-3503-7225-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/MLSP58920.2024.10734773

ABSTRACT

Causal Graph Discovery (CGD) enables the estimation of directed acyclic graph (DAG) that represents the joint probability distribution of observational data. To estimate DAGs, typical constraint-based CGD algorithms run a sequence of conditional independence (CI) tests, making the estimation process prone to privacy leakage. Now, *privacy affects utility*, and due to the high inter-dependency, initial CI tests need to be more accurate to avoid error propagation through subsequent iterations. Based on this key observation, we present *CURATE* (CaUsal gRaph AdapTivE privacy), a differentially private constraint-based CGD algorithm. In contrast to the existing works, in *CURATE* we propose a privacy preserving framework with adaptive privacy budgeting by minimizing error probability while keeping the cumulative leakage bounded. To validate our framework, we present comprehensive set of experiments on several datasets and show that *CURATE* achieves significantly higher utility compared to the existing DP-CGD algorithms.[1]

Index Terms— Causal Graph Discovery, Differential Privacy, Adaptive Privacy Budgeting.

1. INTRODUCTION

Causal graph discovery (CGD) is the method of estimating the underlying causal graph from observational data. CGD is an important part of causal inference [1], and is widely used in biology [2], genetics [3], finance, education and so on. CGD algorithms are broadly classified into two categories: (i) *Constraint-based algorithms* which run a sequence of conditional independence (CI) tests to estimate the causal graph, and (ii) *Score-based algorithms* which optimize a score function for the estimation.

Differentially Private CGD: Datasets used in CGD often contain sensitive information about the participants. Traditional constraint-based CGD algorithms need to run a sequence of interdependent statistical CI tests which makes the estimation process prone to privacy leakage. There is a line of work which incorporates Differential Privacy (DP) [4, 5, 6] that ensures the estimated DAG is *approximately* the same; irrespective of the presence/absence of a user in the observational dataset. For instance, EM-PC [7] uses *Exponential Mechanism*, Priv-PC and SVT-PC [8] adopt *Laplace Mechanism* and *Sparse Vector Technique* (SVT) for privatizing CI tests. *Score-based* DP-CGD algorithm, NOLEAKS [9] uses the *Gaussian Mechanism* to privatize the optimization process. For the scope of this paper, we focus on constraint-based DP-CGD algorithms, due to their low computational complexity.

Overview and Summary of Contributions: Existing constraint-based DP-CGD algorithms ensure privacy by perturbing every CI test with the *same amount of noise*. As we discuss in Section 3, the CI tests in CGD can be highly interdependent. If an edge between two vertices is erroneously deleted by a CI test, then the conditional interdependence between them (conditioned on any other subset of features) is never checked in later iterations. This issue also impacts the scalability of DP-CGD as the total privacy leakage blows up for datasets with a large number of features ($d >> 1$). This brings forth the important point that initial CI tests are more critical and motivates the idea of *adaptive privacy budgeting* in CGD. Specifically, given a total privacy budget, the initial CI tests need higher privacy budget to reduce the risk of error propagation to subsequent iterations. As *uniform privacy budget allocation* throughout any iterative optimization process affects the utility, several recent works [10, 11, 12] use the adaptive privacy budgeting in the context of DP. We adopt adaptive budgeting in the context of DP-CGD with our proposed framework, *CURATE* (CaUsal gRaph AdapTivE privacy). Based on the outcome of the previous order, *CURATE* optimizes privacy budgets for each order by minimizing the surrogate of total error probability. The adaptive budget allocation along with the error probability minimization enables *CURATE* to scale up utility while ensuring better privacy guarantees.

2. PRELIMINARIES ON CGD AND DP

Definition 1 (Probabilistic Graphical Model) *Given a joint probability distribution $\mathbb{P}(F_1, \ldots, F_d)$ of d random variables, the graphical model $\mathcal{G}^*$ with V vertices $(v_1, \ldots, v_d)$*

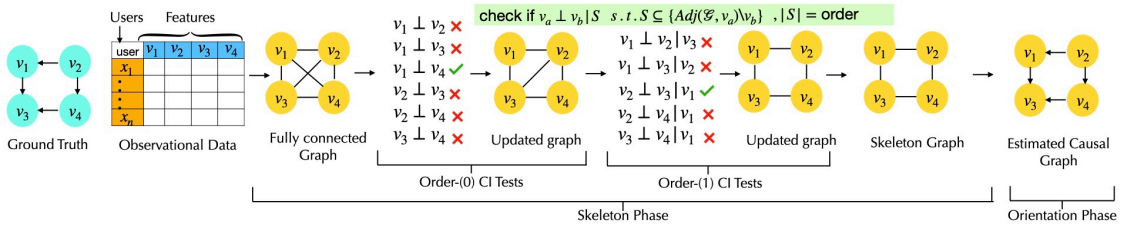

Fig. 1. Workflow of the skeleton-phase of non-private canonical PC algorithm for an observational dataset with 4 features. Through a sequence of interdependent CI tests the estimated graph is updated and the algorithm returns the skeleton graph.

and $E \subseteq V \times V$ edges is known as Probabilistic Graphical Model (PGM) if the joint distribution decomposes as:

$$\mathbb{P}(F_1, \ldots, F_d) = \prod_{F_a \in \{F_1, \ldots, F_d\}} \mathbb{P}(F_a | Pa(F_a)),$$

where, $Pa(F_a)$ represents the direct parents of the node F_a. It relies on the assumption of probability independence ($F_a \perp\!\!\!\perp_P F_b | S$) $\implies$ graphical independence ($v_a \perp\!\!\!\perp_G v_b | S$) [13].

Definition 2 (Causal Graph Discovery) *Given dataset $\mathcal{D}$ with the collection of n i.i.d. samples $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ drawn from a joint probability distribution $\mathbb{P}(F_1, \ldots, F_d)$ where $\mathbf{x}_i$ is a d-dimensional vector representing the d features of the i^{th} sample (user); the method of estimating the PGM $(\mathcal{G}^*)$ from $\mathcal{D}$ is known as Causal Graph Discovery (CGD)[8].*

Overview of PC algorithm: Canonical constraint-based CGD algorithms (such as the PC algorithm [1]) work in two phases: a *skeleton phase* followed by an *orientation phase*. In the *skeleton phase*, the algorithm starts with a fully connected graph ($\mathcal{G}$) and prunes it by conducting a sequence of conditional independence (CI) tests. The CI tests in PC are order dependent, and the order of a test represents the cardinality of the conditioning set S of features. In order-(i) tests, all the connected node pairs (v_a, v_b) in $\mathcal{G}$ are tested for statistical independence conditioned on the set S. The conditioning set S is chosen such that $S \subseteq \{Adj(\mathcal{G}, v_a) \backslash v_b\}$, where $Adj(\mathcal{G}, v)$ represents the adjacent vertices of the node v in the graph $\mathcal{G}$. Edge between the node pairs (v_a, v_b) gets deleted if they pass order-(i) CI test and never get tested again for statistical independence conditioned on set S with $|S| > i$. The remaining edges in $\mathcal{G}$ then get tested for independence in order-$(i+1)$ CI tests conditioned on a set S with $|S| = (i+1)$. This process of CI testing continues until all connected node pairs in $\mathcal{G}$ are tested conditioned on set S of size $(d-2)$. At the end of this phase, PC returns the skeleton graph. In the *orientation phase*, the algorithm orients the edges based on the separation set S of one independent node pair (v_a, v_b) without introducing cyclicity in $\mathcal{G}$ [1, 7] as shown in Figure 1. The privacy leakage in this two-step process is only caused in the *skeleton phase*, as this is when the algorithm directly interacts with the dataset $\mathcal{D}$. Therefore, the existing literature has focused on privatizing CI tests subject to the notion of differential privacy [4, 5, 6] which ensures the *presence/absence* of a user will not *significantly* change the estimated graph.

Definition 3 ((ϵ, δ)-Differential Privacy) *[4, 5, 6] For all pair of neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ by a single element, i.e., $||\mathcal{D} - \mathcal{D}'||_1 \leq 1$, a randomized algorithm $\mathcal{M}$ with an input domain of D and output range $\mathcal{R}$ is considered to be (ϵ, δ)-differentially private, if $\forall \mathcal{S} \subseteq \mathcal{R}$:*

$$\mathbb{P}[\mathcal{M}(\mathcal{D}) \in \mathcal{S}] \leq e^{\epsilon} \mathbb{P}[\mathcal{M}(\mathcal{D}') \in \mathcal{S}] + \delta.$$

Differentially private CGD algorithms have adopted *Exponential Mechanism* [7], *Laplace Mechanism, Sparse Vector Technique* [8], *Gaussian Mechanism* [9] to ensure DP.

Sensitivity of CI tests & Composition of DP: For the class of constraint-based algorithms, an edge between the nodes (v_a, v_b) from estimated graph $\mathcal{G}$ gets deleted conditioned on set S if $(f_{v_a, v_b | S}(\mathcal{D}) > T)$, where $f_{v_a, v_b | S}(\cdot)$ is the test statistic, and T is the test threshold. Thus the structure of the estimated causal graph depends on the nature of $f(\cdot)$ and the threshold (T). Also, in DP-CGD, the amount of added noise is proportional to the l_k-sensitivity (Δ_k) of the test statistic $f_{v_a, v_b | S}(\cdot)$. Therefore, to maximize the predictive performance, test statistics with lower sensitivity with respect to sample size n of the dataset $\mathcal{D}$ are preferred. Through analysis we observed the l_1-sensitivity of the *Kendall's τ* test statistic can be bounded as $\Delta_1 \leq \frac{C}{\sqrt{n}}$ (C is a constant obtained from the analysis presented in *Supplementary document*). However, any other CI test statistics mentioned in Section 1 can be readily adopted in the framework of *CURATE*. As *Composition* is a critical tool in DP-CGD, the total leakage can be calculated by *Basic Composition* [4, 5, 14, 15], *Advanced Composition* [15, 6], *Optimal Composition* [16], *Adaptive Composition* [17], *Moments Accountant* [18].

3. MAIN RESULTS

Overview and Key Idea Behind CURATE: In this Section, we present the main proposed idea of this paper, *CURATE*, that enables adaptive privacy budgeting while minimizing the error probability. As, the CI tests in constraint-based CGD algorithm are highly interdependent, predicting the total number of CI tests in CGD before the execution of the tests is difficult. The number of order-(i) CI tests (t_i) enables the framework to have an approximation of per-order privacy budgets for later iterations $(\epsilon_i, \ldots, \epsilon_{d-2})$ based on the total remaining privacy budget $(\epsilon^{(i)}_{\text{Total}})$. One naive data agnostic way to upper bound t_i is: $t_i \leq \binom{d}{2} \cdot \binom{d-2}{i}$, where

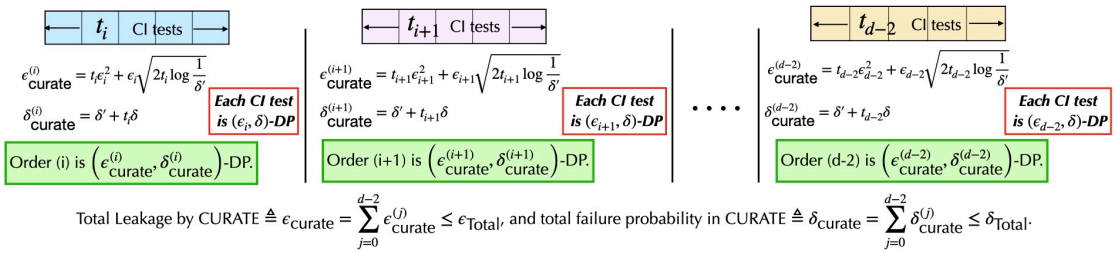

Fig. 2. The composition mechanism in *CURATE* across all order of CI tests. For every order-(i), total privacy leakage is calculated with *Advanced Composition*, and across all orders the total leakage by *CURATE* is calculated with *Basic Composition*.

$\binom{d}{2}$ represents the number of ways to select an edge from the edges of a fully connected graph (the way of selecting an edge between 2 connected nodes out of d nodes), and $\binom{d-2}{i}$ refers to the selection of conditioning set (S) with cardinality $|S| = i$. However, this upper bound is too large and does not depend on the outcome of the previous iteration. A better approximation of t_i is always possible given the outcome of the previous iteration. As, DP is immune to *post-processing* [5], releasing the number edges (e_{i+1}) after executing order-(i) differentially private CI tests will preserve differential privacy. For instance, the possible number of order-$(i+1)$ CI tests can always be upper-bounded as $t_{i+1} \leq e_{i+1} \cdot \binom{d-2}{i+1}$ where e_{i+1} represents the remaining edges after order-(i) tests. We have studied both of the methods and observed that $t_{i+1} \leq e_{i+1} \cdot \binom{d-2}{i+1}$ is a better estimate of t_{i+1} as $e_i \leq \binom{d}{2}, \forall i \in \{0, d-2\}$. Given the outcome of order-$(i-1)$ tests graph $\mathcal{G}$ with edges e_i and a total (remaining) privacy budget of $\epsilon^{(i)}_{\text{Total}}$, we assign a privacy budgets $(\epsilon_i, \ldots, \epsilon_{d-2})$. As every order-$(i)$ CI test in *CURATE* is (ϵ_i, δ)-DP, with DP failure probabilities $\delta, \delta' > 0$, the total leakage in order-(i) is calculated with Advanced Composition [6] as: $\epsilon^{(i)}_{\text{curate}} = t_i\epsilon_i^2 + \sqrt{2\log(\frac{1}{\delta'})t_i\epsilon_i^2}$, and the total failure probability in DP as: $\delta^{(i)}_{curate} = (\delta' + t_i\delta)$. However, as different orders have different privacy budgets, the total privacy leakage by *CURATE* is calculated with Basic Composition [6] as: $\sum_{j=0}^{d-2} \epsilon^{(j)}_{\text{curate}} = \sum_{j=0}^{d-2} \left(t_j\epsilon_j^2 + \sqrt{2t_j \log(\frac{1}{\delta'})\epsilon_j^2}\right)$, and the cumulative failure probability of *CURATE* is $\sum_{j=0}^{d-2} \delta^{(j)}_{curate}$ (refer Figure 2). Therefore, given the outcome of order-$(i-1)$ tests, the total leakage in *CURATE* must satisfy: $\sum_{j=i}^{d-2} \left(t_j\epsilon_j^2 + \sqrt{2t_j \log(\frac{1}{\delta'})\epsilon_j^2}\right) \leq \epsilon^{(i)}_{\text{Total}}$, where $t_j = e_j \cdot \binom{d-2}{j}$, and $\sum_{j=0}^{d-2} \delta^{(j)}_{curate} \leq \delta_{\text{Total}}$. We enforce $\epsilon_i \geq \ldots \geq \epsilon_{d-2}$, so that the initial CI tests get a higher privacy budget.

DP-CI Test in CURATE: The differentially private order-(i) CI test with privacy budget ϵ_i, for variables $(v_a, v_b) \in \mathcal{G}$ conditioned on a set of variables S is defined as follows:

- if $\hat{f} > T(1+\beta_2) \implies$ delete edge (v_a, v_b)
- else if $\hat{f} < T(1-\beta_1) \implies$ keep edge (v_a, v_b)
- else keep the edge with probability $\frac{1}{2}$,

where $\hat{f} := f_{v_a,v_b|S}(\mathcal{D}) + \text{Lap}(\frac{\Delta}{\epsilon_i})$, $\text{Lap}(\frac{\Delta}{\epsilon_i})$ is *Laplace noise*, Δ denotes the l_1-sensitivity of the test statistic, T denotes the threshold, and (β_1, β_2) denote margins. In order to keep the utility high, one would ideally like to pick $(\epsilon_i, \epsilon_{i+1}, \cdots, \epsilon_{d-2})$ that minimize the error probability $\mathbb{P}[E] = \mathbb{P}[\mathcal{G} \neq \mathcal{G}^*]$, where $\mathcal{G}^*$ is the true causal graph, and $\mathcal{G}$ is the estimated causal graph. Unfortunately, we do not have access to $\mathcal{G}^*$; in this paper, we instead propose to use a *surrogate* for error by considering Type-I and Type-II errors relative to the unperturbed (non-private) statistic. Type-I error relative to the unperturbed CI test occurs when the private algorithm keeps the edge given that the unperturbed test statistic deletes the edge $(f_{v_a,v_b|S}(\mathcal{D}) > T)$, and relative Type-II error occurs when the algorithm deletes an edge given that the unperturbed test statistic keeps that edge $(f_{v_a,v_b|S}(\mathcal{D}) < T)$. The next *Lemma* gives upper bounds on relative Type-I and Type-II error probabilities in *CURATE*.

Lemma 1 *For some $c_1, c_2 \in (0,1)$, and non-negative test threshold margins (β_1, β_2), the relative Type-I ($\mathbb{P}[E_1^i]$) and Type-II ($\mathbb{P}[E_2^i]$) errors in order-(i) CI tests in CURATE with privacy budget ϵ_i and l_1-sensitivity Δ can be bounded as:*

$$\mathbb{P}[E_1^i] \leq \underbrace{\frac{c_1}{2} + \frac{1}{2}e^{(-\frac{T\beta_1\epsilon_i}{\Delta})}}_{q_i^{(1)}}, \quad \mathbb{P}[E_2^i] \leq \underbrace{\frac{c_2}{2} + \frac{1}{2}e^{(-\frac{T\beta_2\epsilon_i}{\Delta})}}_{q_i^{(2)}}.$$

The proof of *Lemma 1* is presented in the *Supplementary document*. The main objective of *CURATE* is to allocate privacy budgets adaptively for order-(i) CI tests by minimizing the total relative error. The leakage in DP-CGD depends on the number of CI tests and the number of CI tests depend upon the number of edges in the estimated graph $\mathcal{G}$. As, the number of edges in the true graph is not known, we use $\mathbb{P}[E_1^i] + \mathbb{P}[E_2^i]$ as a surrogate for the total error probability $\mathbb{P}[E]$. Given the outcome of order-$(i-1)$ tests, the algorithm can make Type-I error by preserving an edge which is not present in the true graph till order-$(d-2)$. If such an edge is present after order-$(i-1)$ tests, the probability of Type-I error at the end of the order-$(d-2)$ can be represented as: $\prod_{j=i}^{d-2} q_j^{(1)}$ since independent noise addition to each CI test enables the framework to bound the probability of error in each order independently and at the end of order-$(d-2)$ the total probability of error is the cumulative error made by the algorithm in every

order-(j). Probability of keeping an edge which is present in the ground truth after order-$(i-1)$ tests can be represented as $\prod_{j=i}^{d-2}(1-q_j^{(2)})$, therefore, the total Type-II error can be represented as: $\left(1-\left(\prod_{j=i}^{d-2}(1-q_j^{(2)})\right)\right)$. This leads to the construction of the main objective function of this paper given the outcome of order-$(i-1)$ CI tests $\mathcal{G}$. The minimization objective function is given as:

$$\prod_{j=i}^{d-2} q_j^{(1)} + \left(1-\left(\prod_{j=i}^{d-2}(1-q_j^{(2)})\right)\right). \tag{1}$$

Since the number of edges in true graph are unknown, we propose to minimize (1) as a surrogate for the error probability.

Optimization for Privacy Budget Allocation: By observing the differentially private outcome of order-$(i-1)$ CI tests (remaining edges e_i in graph $\mathcal{G}$), *CURATE* optimizes for $\bar{\epsilon} = \{\epsilon_i, .., \epsilon_{d-2}\}$ (privacy budgets for subsequent order-(i) tests and beyond) while minimizing the objective function as described in (1). Formally, we define the optimization problem in *CURATE*, denoted as $OPT(\epsilon_{\text{Total}}^{(i)}, e_i, i)$:

$$\underbrace{\arg\min_{\bar{\epsilon}} \prod_{j=i}^{d-2} q_j^{(1)} + \left(1-\left(\prod_{j=i}^{d-2}(1-q_j^{(2)})\right)\right)}_{OPT(\epsilon_{\text{Total}}^{(i)}, e_i, i)}$$

$$\begin{cases} \sum_{j=i}^{d-2} \underbrace{\left(t_j\epsilon_j^2 + \sqrt{2\log(\frac{1}{\delta'})t_j\epsilon_j^2}\right)}_{\text{total leakage in order-(j)}} \leq \epsilon_{\text{Total}}^{(i)} \\ \epsilon_j \geq \epsilon_{j+1}. \end{cases} \tag{2}$$

Given the outcome of order-$(i-1)$ tests, the above optimization function $OPT(\epsilon_{\text{Total}}^{(i)}, e_i, i)$ takes the following inputs: (a) remaining total budget ($\epsilon_{\text{Total}}^{(i)}$), (b) remaining edges ($e_i$) in the output graph $\mathcal{G}$ after all order-$(i-1)$ tests, (c) the index of order, i.e., i. The function then optimizes and outputs the privacy budgets $(\epsilon_i, \ldots, \epsilon_{d-2})$ for remaining order tests, while satisfying the two constraints mentioned in (2). As the optimization problem in (2) is difficult to solve in a closed form, in our experiments we have used Sequential Least Squares Programming (SLSQP) for optimizing the objective function.

CURATE Algorithm: Now we present the algorithm *CURATE* that enables adaptive privacy budget allocation by solving the optimization problem in (2). In *CURATE*, we use the optimization function $OPT(\cdot)$ recursively to observe adaptively chosen per-order privacy budgets. Given remaining privacy budget for order-(i) tests $\left(\epsilon_{\text{Total}}^{(i)}\right)$, $OPT(\cdot)$ calculates the remaining budget for order-$(i+1)$ CI tests $\epsilon_{\text{Total}}^{(i+1)}$ as:

$$\underbrace{\epsilon_{\text{Total}}^{(i+1)}}_{\text{budget for order(i+1)}} = \underbrace{\epsilon_{\text{Total}}^{(i)}}_{\text{budget before order(i)}} - \underbrace{\left(t_i\epsilon_i^2 + \epsilon_i\sqrt{2t_i\log(\frac{1}{\delta'})}\right)}_{\epsilon_{\text{curate}}^{(i)}\text{: total leakage order-}i}.$$

Initially, the remaining budget for order-(0) CI tests is equal to the assigned total privacy budget, i.e., $\epsilon_{\text{Total}}^{(0)} = \epsilon_{\text{Total}}$ and the edges in the complete graph $\mathcal{G}$ can be expressed as $e_0 = \binom{d}{2}$. In order-(0), *CURATE* solves for $(\epsilon_0, \ldots, \epsilon_{d-2})$ by using the function $OPT(\epsilon_{\text{Total}}^{(0)}, e_0, 0)$. After completion of all order-(0) CI tests, the algorithm calculates the remaining budget for order-(1) CI tests as $\epsilon_{\text{Total}}^{(1)} = \epsilon_{\text{Total}}^{(0)} - \left(t_0\epsilon_0^2 + \epsilon_0\sqrt{2t_0\log(\frac{1}{\delta'})}\right)$ and by observing the remaining edges e_1, it solves for the next set of privacy budgets $(\epsilon_1, \ldots, \epsilon_{d-2})$. We recursively apply this process to get $\epsilon_i\ \forall i \in \{0, 1, \ldots, d-2\}$ CI tests. *CURATE* ensures $(\epsilon_{\text{Total}}, \delta_{\text{Total}})$-differential privacy. Notably, *CURATE* inherits the low computational complexity characteristic from constraint-based algorithms [1], which makes *CURATE* readily applicable for large datasets. As *sub-sampling* amplifies differential privacy [19], we can readily incorporate sub-sampling parameters within the optimization framework of *CURATE*.

Algorithm 1 *CURATE Algorithm*

Data: Dataset $\mathcal{D}$, total privacy budget (ϵ_{Total}), DP-failure probabilities $(\delta, \delta' > 0)$, total failure probability (δ_{Total}), test statistic $f(\cdot)$, threshold T, margins (β_1, β_2), l_1-sensitivity Δ, fully connected graph $\mathcal{G}$

Result: Partially connected graph $\mathcal{G}$

Perform sub-sampling: $\mathcal{D}'' \leftarrow_{\frac{m}{n}} \mathcal{D}, n = |\mathcal{D}|, m = |\mathcal{D}''|$

Initiation: $i = 0$, $\epsilon_{\text{Total}}^{(0)} = \epsilon_{\text{Total}}$, $\delta \leq 10^{-1.5m}$, $e_0 = \binom{d}{2}$

for $i = \{0, 1, \ldots, d-2\}$ **do**

  Initiate number of order-i CI tests as: $t_i = 0$

  $(\epsilon_i, \ldots, \epsilon_{d-2}) = OPT(\epsilon_{\text{Total}}^{(i)}, e_i, i)$

  $\forall$ connected node pairs (v_a, v_b) in $\mathcal{G}$ that has not been tested on S s.t. $S \subseteq \{Adj(\mathcal{G}, v_a) \backslash v_b\}, |S| = i$

  Evaluate $\hat{f} := f_{v_a, v_b|S}(\mathcal{D}'') + \text{Lap}(\frac{\Delta}{\epsilon_i})$

  - if $\hat{f} > T(1+\beta_2)$ then delete edge (v_a, v_b)
  - else if $\hat{f} < T(1-\beta_1)$ then keep edge (v_a, v_b)
  - else keep the edge with probability $\frac{1}{2}$

  Update $\mathcal{G}$, $t_i = t_i + 1$

  $\epsilon_{\text{Total}}^{(i+1)} = \epsilon_{\text{Total}}^{(i)} - \left(t_i\epsilon_i^2 + \epsilon_i\sqrt{2t_i\log(\frac{1}{\delta'})}\right)$

  $\epsilon_{\text{curate}}^{(i)} = \left(t_i\epsilon_i^2 + \epsilon_i\sqrt{2t_i\log(\frac{1}{\delta'})}\right)$

  $\delta_{\text{curate}}^{(i)} = \delta' + (t_i \cdot \delta)$

  e_{i+1} = edges in updated graph $\mathcal{G}$

  if $\sum_{j=0}^{i} \delta_{curate}^{(j)} < \delta_{Total}$ **then**

    Continue

  end

end

return Skeleton $\mathcal{G}$, Total Leakage $(\sum_{j=0}^{d-2} \epsilon_{\text{curate}}^{(j)}, \sum_{j=0}^{d-2} \delta_{\text{curate}}^{(j)})$

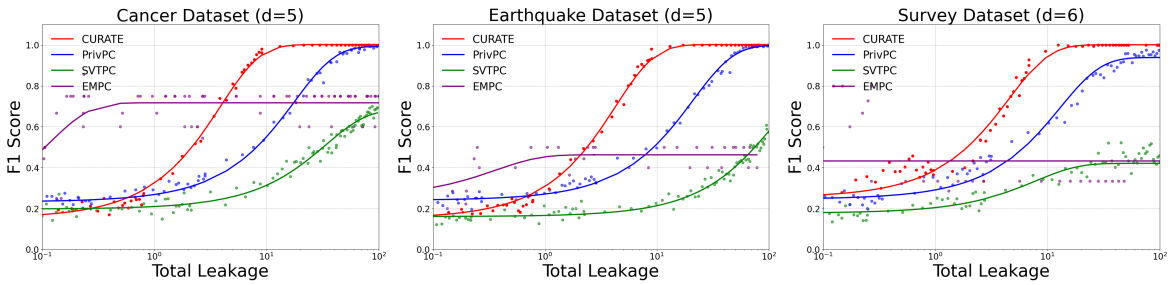

Fig. 3. Performance evaluation of private CGD algorithms EM-PC [7], SVT-PC, Priv-PC [8] and *CURATE* in terms of total leakage vs F1 score on 3 public CGD datasets: Cancer, Earthquake, and Survey with Test threshold $(T) = 0.05$.

Dataset	Algorithm	F1-score		
		$\epsilon_{\text{Total}} = 1.0$	$\epsilon_{\text{Total}} = 5.0$	$\epsilon_{\text{Total}} = 10.0$
Cancer	***CURATE***	0.27 ± 0.17	0.72 ± 0.22	0.96 ± 0.10
	SVTPC	0.16 ± 0.14	0.22 ± 0.11	0.24 ± 0.18
	EMPC	0.60 ± 0.00	0.75 ± 0.00	0.75 ± 0.00
	PrivPC	0.32 ± 0.14	0.43 ± 0.19	0.62 ± 0.17
Earthquake	***CURATE***	0.24 ± 0.17	0.72 ± 0.20	0.93 ± 0.15
	SVTPC	0.17 ± 0.15	0.13 ± 0.13	0.20 ± 0.15
	EMPC	0.25 ± 0.00	0.5 ± 0.00	0.5 ± 0.00
	PrivPC	0.31 ± 0.17	0.44 ± 0.14	0.62 ± 0.20
Survey	***CURATE***	0.36 ± 0.19	0.68 ± 0.22	0.92 ± 0.14
	SVTPC	0.22 ± 0.20	0.30 ± 0.21	0.32 ± 0.21
	EMPC	0.33 ± 0.00	0.33 ± 0.00	0.33 ± 0.00
	PrivPC	0.32 ± 0.21	0.52 ± 0.22	0.61 ± 0.22

(a)

Algorithm	Cancer	Earthquake	Survey
PC (non-private)	**6**	**6**	**11**
CURATE	**22**	**23**	**36**
SVTPC	26	24	38
Priv-PC	58	68	48
EMPC	60	68	50

(b)

Fig. 4. Predictive performance of private CGD algorithms on 3 public CGD datasets: Cancer, Earthquake and Survey: Table (a) represents the mean and standard deviation of F1-score in three privacy regimes, ϵ_{Total}=1.0, ϵ_{Total}=5.0, $\epsilon_{\text{Total}} = 10.0$. Table (b) Average CI tests required to achieve the maximum F1 score with comparatively large amount of total leakage ($\epsilon_{\text{Total}} = 1.0$).

4. EXPERIMENTAL RESULTS AND DISCUSSION

Specifications	Cancer	Earthquake	Survey
Features (nodes)	5	5	6
Edges	4	4	6
Samples	100K	100K	100K
F1 Score (PC [1])	1.0	1.0	1.0
CI tests (PC [1])	6	6	11

Table 1. Dataset description and causal graph discovery results of non-private PC algorithm on 3 public CGD datasets Cancer, Earthquake and Survey.

In this Section, we present experimental results on 3 public CGD datasets: Cancer [20], Earthquake [20], and Survey [21] that demonstrates the utility-privacy trade-off achieved by *CURATE* compared to existing differentially private CGD algorithms: Priv-PC [8], SVT-PC [8], and EM-PC [7]. Utility of the CGD algorithms are measured using F1-score[2]. The implementation code and supplementary document of *CURATE* is available [3].

[2]Given the ground truth graph $\mathcal{G}^* = (\mathcal{V}, \mathcal{E}^*)$ and the estimated graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, if Precision= $|\frac{\mathcal{E} \cap \mathcal{E}^*}{\mathcal{E}}|$, and Recall= $|\frac{\mathcal{E} \cap \mathcal{E}^*}{\mathcal{E}^*}|$, the F1-score is defined as:$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

[3]https://github.com/PayelBhattacharjee14/CURATE

Dataset Description and Utility with PC Algorithm [1]: Table 1 presents the data description along with the F1-score obtained by the execution of non-private PC algorithm [1]. The F1-score is obtained using Kendall's τ test statistic with sub-sampling rate $(q = 1.0)$, test threshold $(T) = 0.05)$.

Privacy vs Utility Trade-offs: The experimental result presented in Figure 3 shows that with adaptive privacy budgeting and minimization of the relative total probability of error, *CURATE* outperforms the existing DP-CGD algorithms including EM-PC [7], SVT-PC, Priv-PC [8]. We can observe that for Cancer and Earthquake in moderate privacy regime ($\epsilon_{\text{Total}} \geq 1$) and for Survey dataset in comparatively high privacy regime ($\epsilon_{\text{Total}} \geq 0.1$), *CURATE* outperform all the existing methods. *CURATE* achieves the same F1-score as PC [1] with less leakage compared to the existing algorithms. EM-PC (corresponding to the Exponential Mechanism; and is computationally demanding algorithm) performs better in the high privacy (very low ϵ_{Total}) regime; however it's F1-score saturates and does not always converge to the F1-score of non-private PC [1]. Total failure probability in *CURATE* is $\delta_{\text{Total}} = 10^{-10}$ for all experiments. Figure 4 (Table (a)) represents the mean and standard deviation of the F1-score for each algorithm (averaged over 50 trials).

Comparison of Number of CI Tests: Total number of CI tests in constraint-based DP-CGD algorithms directly influence the total amount of leakage. The privacy leakage can be provably reduced by efficient and accurate CI testing. Intuitively, in *CURATE*, the total leakage decreases as the adaptive choice of privacy budgets makes the initial CI tests more accurate, therefore, *CURATE* tends to run less number of overall CI tests compared to other DP-CGD algorithms. We confirm this intuition in the results presented in Figure 4 (Table (b)).

5. CONCLUSIONS

In this paper, we presented *CURATE*, a differentially private causal graph discovery framework that improves the privacy-utility trade-off by adaptive privacy budgeting. *CURATE* is based on the idea of minimizing a surrogate for error probability while ensuring that initial CI tests get higher privacy budget. Our experimental results validate the proposed approach and show that adaptivity can help in significantly improving utility subject to differential privacy. There are several interesting directions for future work: (i) adaptive privacy budget designing for score based algorithms (such as [9]); (ii) use the outcomes of previous noisy CI tests to adaptively design other hyper-parameters, including CI test thresholds and margins.

6. ACKNOWLEDGEMENT

This work was supported by NSF grants CAREER 1651492, CCF-2100013, CNS-2209951, CNS-1822071, CNS-2317192, and by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing under Award Number DESC-ERKJ422, and NIH Award R01-CA261457-01A1.

7. REFERENCES

[1] Peter Spirtes et Al., *Causation, Prediction, and Search*, Jan. 1993.

[2] Karen Sachs et Al., "Causal protein-signaling networks derived from multiparameter single-cell data," *Science (New York, N.Y.)*, Apr. 2005.

[3] Bin Zhang et Al., "Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease," *Cell*, 2013.

[4] Cynthia Dwork et Al., "Our Data, Ourselves: Privacy Via Distributed Noise Generation," 2006, Springer.

[5] Cynthia Dwork et Al., "Calibrating Noise to Sensitivity in Private Data Analysis," in *Theory of Cryptography*, Berlin, Heidelberg, 2006, Springer.

[6] Cynthia Dwork et Al., "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends® in Theoretical Computer Science*, , no. 3-4, 2013.

[7] Depeng Xu et Al., "Differential privacy preserving causal graph discovery," in *2017 IEEE PAC*.

[8] Lun Wang et Al., "Towards practical differentially private causal graph discovery," in *Advances in Neural Information Processing Systems*, 2020.

[9] Pingchuan Ma et Al., "NoLeaks: Differentially Private Causal Discovery Under Functional Causal Model," *IEEE Transactions on Information Forensics and Security*, 2022.

[10] NhatHai Phan et Al., "Adaptive laplace mechanism: Differential privacy preservation in deep learning," in *2017 IEEE international conference on data mining (ICDM)*.

[11] Jaewoo Lee et Al., "Concentrated differentially private gradient descent with adaptive per-iteration privacy budget," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[12] Lin Chen et Al., "Differentially private deep learning with dynamic privacy budget allocation and adaptive optimization," *IEEE TIFS*, 2023.

[13] Alessio Zanga et Al., "A survey on causal discovery: theory and practice," *International Journal of Approximate Reasoning*, 2022.

[14] Cynthia Dwork et Al., "Differential privacy and robust statistics," in *Proceedings of the 41 annual ACM symposium on Theory of computing*, 2009.

[15] Cynthia Dwork et Al., "Boosting and Differential Privacy," in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, 2010.

[16] Peter Kairouz et Al., "The Composition Theorem for Differential Privacy," *IEEE Transactions on Information Theory*, , no. 6, June 2017.

[17] Ryan M Rogers et Al., "Privacy Odometers and Filters: Pay-as-you-Go Composition," in *Advances in Neural Information Processing Systems*, 2016.

[18] Martín Abadi et Al., "Deep Learning with Differential Privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[19] Borja Balle et Al., "Privacy Amplification by Subsampling: Tight Analyses via Couplings and Divergences," in *Advances in Neural Information Processing Systems*, 2018.

[20] Kevin B Korb et Al., *Bayesian artificial intelligence*, CRC press, 2010.

[21] Marco Scutari Denis, Jean-Baptiste, *Bayesian Networks: With Examples in R*, Chapman and Hall/CRC, New York, June 2014.