Nothing but Regrets — Privacy-Preserving Federated Causal Discovery

Osman Mian
Helmholtz Centre for Information Security

David Kaltenpoth
Helmholtz Centre for Information Security

Michael Kamp
Institute for AI in medicine IKIM, and Ruhr-University Bochum

Jilles Vreeken
Helmholtz Centre for Information Security

Abstract

In critical applications, causal models are the prime choice for their trustworthiness and explainability. If data is inherently distributed and privacy-sensitive, federated learning allows for collaboratively training a joint model. Existing approaches for federated causal discovery share locally discovered causal model in every iteration, therewith not only revealing local structure but also leading to very high communication costs. Instead, we propose an approach for privacy-preserving federated causal discovery by distributed min-max regret optimization. We prove that max-regret is a consistent scoring criterion that can be used within the well-known Greedy Equivalence Search to discover causal networks in a federated setting and is provably privacy-preserving at the same time. Through extensive experiments, we show that our approach reliably discovers causal networks without ever looking at local data and beats the state of the art both in terms of the quality of discovered causal networks as well as communication efficiency.

1 INTRODUCTION

Discovering causal dependencies from observational data is one of the most fundamental problems in science (Pearl, 2009). While a plethora of approaches for discovering causal networks are designed for single datasets (Spirtes et al., 2000; Chickering, 2002; Shimizu et al., 2006; Peters et al., 2014; Huang et al., 2018), in critical applications, such as healthcare, we cannot pool data due to privacy considerations. In such cases we have multiple sites each with their own private data. Learning causal networks over such data presents a challenging setting where we don't just want to discover the underlying causal network in a federated manner, we can also not compromise on privacy. We therefore consider the problem of discovering a global causal network over distributed datasets with a fixed set of variables — in a privacy preserving manner.

Most state-of-the-art causal discovery approaches that can work with multiple datasets require that we first pool the data (Mooij et al., 2016; Zhang et al., 2017), which makes them inapplicable to the current setting. Approaches that do not require data to be pooled, work only for a single target variable (Peters et al., 2016), or place strict assumptions on the causal mechanisms that are unlikely to hold in practice (Shimizu, 2012; Ghassami et al., 2017).

On the other hand, state of the art federated learning approaches allow to train models in a distributed manner without sharing any data, but their application to causal discovery is not straightforward. A naive approach is discovering individual causal models for each local dataset, pooling those models and computing the likely global causal model governing the process that generated all local datasets. Sharing models, however, is not privacy-preserving, since one can make inferences about local datasets from model parameters (Geiping et al., 2020; Lyu and Chen, 2021; Singhal et al., 2021). Another naive approach is to discover *local* causal networks for each dataset and compute their union. This has two major issues: (i) For finite dataset sizes, locally discovered causal models can vary substantially from the true network and may contain spurious edges, leading to a bad performance, and (ii) this still requires us to explicitly communicate the local causal networks for pooling, which may compromise privacy guarantees (Geiping et al., 2020; Wang et al., 2020).

In this work, we propose to discover the global causal network without sharing any data, model parameters, or even local causal networks— using regrets. Intuitively, the regret measures how much worse a given causal network is, compared to the best causal network for a given dataset. We show that minimizing the worst-case regret over these distributed datasets allows us to define a scoring criterion that, under mild assumptions, is guaranteed to be consistent. It can hence be employed as a score within Greedy Equivalence Search (GES) (Chickering, 2002) to discover

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

the global causal network by only using regrets obtained from local datasets: we first let each site discover the best network for its dataset using GES with any consistent scoring criterion (e.g., BIC), and then optimize the worst-case regret, once again using GES, with respect to the locally discovered causal networks. Throughout the entire learning process, the optimizing algorithm neither sees the data, nor knows the local model parameters. To ensure privacy of local data, we show that using the Laplace mechanism on the shared regrets guarantees ϵ-differential privacy.

To perform federated causal discovery, we instantiate our proposed approach, which we call PERI[1], using three well known consistent scoring criteria. Through extensive experiments we show that PERI discovers causal networks of higher quality than the state of the art on both synthetic and real-world data, scales upto 100 distributed environments while requiring orders of magnitude less communication.

We organize this paper as follows. We first discuss related work in Sec. 2. In Sec. 3 we describe preliminary details required for our proposed method. Sec. 4 and 5 describe theoretical guarantees resp. practical instantiation of our approach. Next, we provide privacy-preserving guarantees of our instantiation in Sec. 6 and report our empirical evaluation results in Sec. 7. We conclude with a discussion and future research directions in Sec. 8

2 RELATED WORK

Causal discovery, i.e., the discovery of causal networks from observational data, is perhaps the most important problem in causal inference as without a causal model, causal inference is impossible. Many methods have been proposed to discover causal networks given a single dataset (Spirtes et al., 2000; Chickering, 2002; Shimizu et al., 2006; Peters et al., 2014; Blöbaum et al., 2018; Huang et al., 2018; Zheng et al., 2018; Mian et al., 2021), much fewer for doing so given data collected from multiple environments (Zhang et al., 2017; Mooij et al., 2016), and only a small handful for doing so when the data cannot be gathered centrally (Ng and Zhang, 2022).

Methods that can consider only a single dataset are not applicable in our setting; even if we ignore all privacy aspects and were to centrally collect and pool all data, it is well known that naively pooling the data can introduce unwanted bias in estimation (Tillman, 2009). Methods that can consider multiple datasets, such as when data has been collected from different environments (Yang et al., 2018; Squires et al., 2020), come one step closer to the scenario we consider in this paper. The most prominent approaches still combine all data, adding one or more context variables to distinguish the rows of the combined datasets, and then perform causal discovery on the augmented data (Zhang et al., 2017; Magliacane et al., 2018). A very general such approach is the Joint Causal Inference (JCI) framework proposed by Mooij et al. (2016), which permits any constraint-based causal discovery algorithm to work with data from multiple environments. Each of these approaches require that all data is available at one site, which is prohibitive in our setting as this violates privacy.

Federated learning allows for learning without the need for centralized data. Rather than sharing data with other nodes, the key idea in federated learning is that we share (partial) local results. The topic of federated causal discovery is relatively young. Proposals for federated causal inference (Xiong et al., 2021) and federated causal discovery (Shimizu, 2012) require strong parametric assumptions. Recent approaches avoid these, either by sacrificing convergence guarantees (Gao et al., 2021) or by sharing additional learning parameters (Ye et al., 2022; Ng and Zhang, 2022). Although these methods do not directly share data, by sharing completely specified local causal models they can provide attackers sufficient information to reconstruct local data (Geiping et al., 2020; Singhal et al., 2021).

In this paper, we build upon a recent approach of (Mian et al., 2022), proposing a regret-based framework to federated learning and instantiating it using an approximate beam-search-based approach. However, it provides no theoretical or privacy guarantees of their solution. Furthermore, the proposed method does not scale beyond a few (10) variables. Our work, on the other hand, uses the the idea of regret-based learning to propose a theoretically sound score that comes with strong privacy guarantees and achieves lower communication costs while scaling up to 100 environments.

3 PRELIMINARIES

3.1 Notation and Assumptions

We consider data $\mathbf{X}$, consisting of m variables, split into d different datasets $X^{(1)}, \ldots, X^{(d)}$ of sizes $n^{(1)}, \ldots, n^{(d)}$. We assume that each $X^{(i)}$ is drawn i.i.d. from a distribution $P_i(X^{(i)})$, which are all are entailed by the same true causal network G^* but where the parameters associated with G^* may be different between $X^{(i)}$. Our goal is to solve the following problem.

Problem Statement (Informal)**.** *Given data $\boldsymbol{X}$, discover the true causal network G^* in a federated (without pooling data) and privacy-preserving (without sharing any models fit over individual datasets) manner.*

To identify the underlying causal network, we need to assume that the distributions $P_i(X^{(i)})$ and the local causal network G^* satisfy the following common assumptions

[1] In astronomy, Peri is the point at which an orbiting object is closest to the center of mass of the body it is orbiting (such as a planet). In our approach, we aim to discover that network which is collectively closest to the local networks of all environments.

made in causal discovery. First is the *causal Markov condition*, stating that every variable in G^* is independent of its non-descendants conditional on its parents in G^* and second, the *Faithfulness* assumption is that if sets U, V are independent given a set W in P, then W d-separates U and V in G^*. Together, the Causal Markov Condition and Causal Faithfulness condition entail that the conditional independence relations in P correspond precisely to d-separation relations in G^*, which lets us identify the Markov equivalence class (MEC) of G^*, i.e. the set of all graphs entailing the exact same independence constraints as G. We write $G \sim H$ when G, H are in the same MEC, and $G \sqsupseteq H$ if $G \sim H'$ for some H' containing all edges in H. As we instantiate our proposed solution using Greedy Equivalence Search (Chickering, 2002) (Sec 3.2), we need to make the *causal sufficiency* assumption, telling us that there are no latent confounders. The sufficiency and faithfulness assumptions, however, are not always necessary and we provide a discussion in Sec 8 on how they can be avoided.

When all of the above assumptions hold, algorithms such as Greedy Equivalence Search can discover causal networks, for a single dataset, up to Markov equivalence (Glymour et al., 2019) i.e. partially oriented causal networks where all collider structures are correctly identified.

3.2 Greedy Equivalence Search

Greedy Equivalence Search (GES) (Chickering, 2002) is a score-based causal discovery approach that learns a causal network $\hat{G}$ from observational dataset $X^{(i)}$. To do so it uses a scoring criterion L to measure how well a network G describes $X^{(i)}$. Starting from an empty network, GES iteratively builds a causal network through repeated forward respectively backward-search. In each step of the forward search, GES chooses a single edge addition to the current best network such that the edge improves score the most and uses the new network as the best network for the next step. Similarly, in each step of the backward search, single edge deletions that improve score the most are chosen. Each phase ends when no modifications of the current network improve score anymore. GES is guaranteed to return the correct Markov equivalence class as $n \to \infty$ if the following two conditions hold. First, L is decomposable, meaning that L can be expressed as

$$L(X; G) = \sum_{j=1}^{m} l_j(X_j; pa_j^G)\,,$$

where pa_j^G are the parents of variable X_j in G and l_j is only a function of X_j and its parents. And second, L satisfies the consistency property which is formally stated as follows.

Definition 1 (Chickering (2002). Consistent Scoring Criterion)**.** *Let G, H be any pair of DAGs, X be a set of data consisting of m records that are i.i.d. samples from some distribution $P(\cdot)$. A (minimizing) scoring criterion L is consistent if in the limit $n \to \infty$, the following two properties hold:*

1. *If H contains P and G does not contain P, then $L(X; H) < L(X; G)$*
2. *If H and G both contain P, and G contains fewer parameters than H, then $L(X; G) < L(X; H)$,*

where *contains* means that G has the exact independence constraints implied by P.

Despite its greedy nature, if L is consistent, GES is guaranteed to find a graph in the MEC of the true G in the large sample limit, although (in the worst-case) this discovery could require runtime super-exponential in the number of variables. Examples of decomposable consistent scores include the Akaike's Information Criterion (AIC) (Akaike, 1974), Bayesian Information Criterion (BIC) (Schwarz, 1978) and scores defined using Minimum Description Length (MDL) (Grünwald, 2007; Mian et al., 2021).

GES, however, is limited to finding causal networks over a single dataset and can therefore only be used to learn individual networks $G^{(i)}$ for each $X^{(i)}$. To extend GES to a federated setting, we require that we can measure the score of a global network G relative to a locally learned $G^{(i)}$ *without* knowing what the local networks are. To do so, we introduce the concept of regret.

3.3 Regret

Given data X and some model M, from a model class $\mathcal{M}$ that explains the data, let $L(X; M)$ be a score function that is minimized when M is the true model for X. Regret $R(M)$ for a given model M with respect to data X is defined as the difference in scores when evaluating X using M instead of the best model M^* for X. Formally stated

$$R(M) \coloneqq L(X; M) - \min_{M^* \in \mathcal{M}} L(X; M^*)\,, \qquad (1)$$

where we drop the dependence on the data X to simplify notation. Simply put, regret measures how much worse the proposed model M is compared to the *best* model for the data. If both M and M^* are present in $\mathcal{M}$, $R(M)$ is lower bounded by 0, which is achieved when $M \equiv M^*$.

4 REGRET-BASED FEDERATED CAUSAL DISCOVERY

In this section we show that we can use the regret (Eq. (1)) to build a consistent score for GES. Using such a regret-based score, we provide a score-agnostic framework that can be used to perform federated causal discovery. We then show that the minimizer of the proposed score discovers the

correct underlying causal structure in the limit. We provide the proofs of our results in the Supplementary section.

For our model class $\mathcal{M}$ defined in Eq (1), we consider the space of all Directed Acyclic Graphs (DAGs), $\mathcal{G}$. Hence for our proposed setup we can write Eq. (1) as

$$R_i(G) \coloneqq L(X^{(i)}; G) - \min_{G^{(i)} \in \mathcal{G}} L(X^{(i)}; G^{(i)}) \,,$$

where $R_i(G)$ is the regret associated with dataset $X^{(i)}$ when using network G.

Now it becomes easy to see the merit of using regret from a federated learning perspective: Given a server S that aims to learn a global causal network using d different sites, each with their own private datasets $X^{(1)}, \ldots, X^{(d)}$, S can send a network G and a scoring criterion L to each site and optimize using the regrets that it receives back. To do so, S needs to consolidate these regret values received back from each site into a meaningful score. We propose this to be the worst-case regret calculated over *all* environments,

$$\begin{aligned} L_F(G) &\coloneqq \max_i R_i(G) \\ &= \max_i \left(L(X^{(i)}; G) - L(X^{(i)}; G^{(i)}) \right) \end{aligned} \tag{2}$$

where $G^{(i)}$ is the minimizer for $L(X^{(i)}; \cdot)$.

Using the aforementioned formulation, the goal of the server is to find that network G that minimizes the worst-case regret among all the networks. Formally stated

Problem Statement. *Given samples $X = \{X^{(1)}, \ldots, X^{(d)}\}$ over d environments that share a common underlying causal DAG, find $\widehat{G}$ such that*

$$\widehat{G} = \operatorname*{argmin}_{G \in \mathcal{G}} \max_i R_i(G) \,. \tag{3}$$

This obtained network $\widehat{G}$ is the one which trades off errors relative to one local network $G^{(i)}$ to another local network $G^{(j)}$ and tries to jointly minimize them. Such a $\widehat{G}$ is the least bad network relative to any of the local networks.

Next we show the conditions under which the minimizer for Eq. (3) finds the correct underlying causal network, up to Markov equivalence.

4.1 Consistency

To prove that the minimizer for Eq. (3) is the true causal network, we need to assume that $L(X^{(i)}; G)$ is of the form

$$L(X^{(i)}; G) = L(G) + L(X^{(i)}|G) \,,$$

where $L(G)$ is a function penalizing the complexity of the network G and the parameters associated with the class of generating functions e.g. linear or spline relationships between each variable and its parents, and $L(X^{(i)}|G)$ is the log-likelihood of the data given the G.

We can now show that in the limit, when every site uses the same consistent score L and obtains arbitrarily much data then our method is guaranteed to find the correct MEC.

Theorem 1. *Let G^* be the true causal network for all $P(X^{(i)})$ and let $n^{(1)} \ldots, n^{(d)} \to \infty$. Further let L be a consistent score. Then*

$$\lim_{n^{(1)}, \ldots, n^{(d)} \to \infty} P\left(\widehat{G} \sim G^*\right) = 1 \,.$$

That is, $\max_i R_i(G)$ is consistent when all $n^{(i)} \to \infty$.

We can further relax Thm. 1 to not require that every site's amount of data grows over time. In fact, as long as even one of the datasets grows, we nevertheless find all edges.

Theorem 2. *Let G^* be the true causal network for all $P(X^{(i)})$ and let $N \coloneqq \max_i n^{(i)} \to \infty$. Further let L be a consistent score. Then*

$$\lim_{N \to \infty} P\left(\widehat{G} \sqsupseteq G^*\right) = 1 \,.$$

For scores L, like AIC, the correct MEC is generally impossible to recover precisely because the penalty for additional edges does not scale with the number of data points. In contrast, for the BIC score this is not an issue.

Corollary 3. *Let the assumptions of Thm. 2 hold and let L be the BIC score. Then*

$$\lim_{N \to \infty} P\left(\widehat{G} \sim G^*\right) = 1 \,.$$

That is, the score $\max_i R_i(G)$ is consistent when L incorporates a BIC-penalty for parameters and $N \to \infty$.

The proof of Cor. 3 applies equally to any other consistent criterion where the parameter-penalty grows strictly with sample size, e.g., MDL-based scores. In Sec. 8 we discuss how to extend our work to other types of scores.

These results imply that $R_i(G)$ is a consistent scoring criterion as long as L used within $R_i(G)$ is consistent. We can therefore define $R_i(G)$ using any consistent L and perform a search for the underlying causal network G by exhaustively evaluating all possible causal networks and choosing one that minimizes Eq (3). Exhaustive search, however, is super-exponential in the number of nodes and is only feasible for networks with very few variables. The problem of learning exact Bayesian network structure is NP-hard after all (Chickering et al., 2004). Using our consistency guarantees, however, we can instantiate our search more efficiently using GES. We show in the next section how we can instantiate an efficient regret-based causal learning framework, while maintaining privacy guarantees.

5 PRACTICAL ALGORITHM

We now describe PERI, a score-based federated causal discovery approach for distributed environments. Let L be

any consistent score used within GES, such as BIC, and let L_F be the composition that calculates the worst-case regret using L as defined in Eq.(2). Then L_F can be used as a consistent score within GES (Thm. 2, Cor. 3) to discover causal networks in a federated fashion.

As a result, we can perform federated causal discovery as shown in Algorithm 1. Given a server S and d different sites, each with their own private datasets $X^{(1)}, \ldots, X^{(d)}$, the server communicates L to each of the sites. Each site then learns a local network $G^{(i)}$ using GES (lines 1-2). The server then instantiates L_F as defined in Eq. (2) (l. 4) and runs its own GES using L_F. In the forward pass (l. 6), the server communicates the best discovered network G_t, at iteration t, to all sites. Each site converts G_t to the MEC $\mathcal{E}_t$ and calculates regret over all possible single edge extensions of $\mathcal{E}_t$. The list of these scores is communicated back to the server. Next, the server chooses the network G_{t+1} with the lowest worst-case regret among all these extensions and sets this network as the best network for the next iteration. The forward search ends when no extensions of G_t improve the score anymore. The backward search (l. 7) is analogous to the forward search except that the regret scores are calculated over single edge deletions of $\mathcal{E}_t$ at each iteration. We repeat the search process until convergence (l. 8). During the learning process, the server neither sees the data, nor knows the local models for any site. The only communication that takes place between server and sites is the list of regret values for networks in the MEC for the query DAG G_t.

This proposed approach has several advantages: First, the regret for a query network can be calculated locally at each site and returned back to the server, requiring no communication of model parameters — the job of the server is to choose the worst-case regret for a given network G. Second, PERI is *guaranteed* to converge. This is because regret is lower-bounded by 0, and we only take steps that reduce regret. Third, we do not need any additional assumptions except the ones required for L — to be used within GES we require L to be decomposable and consistent. PERI, in fact, can be viewed as a generalization of GES to multiple datasets as it is easy to see that running PERI using only a single site is akin to running GES on a single dataset.

With the algorithm explained, we now describe how we can guarantee differential privacy using PERI.

Algorithm 1: PERI for federated causal discovery

Input: Scoring criterion L
Output: Causal network G
1 **for** $i = 1 \ldots d$ **do**
2 site[i].GREEDYEQVSEARCH(L)
3 $G^* \leftarrow \emptyset$
4 Define $L_F(G) := \max_i \left[L(X^{(i)}, G) - L(X^{(i)}, G^{(i)})\right]$
5 **repeat**
6 $G^* \leftarrow$ server.FORWARDEQVSEARCH(G^*, L_F)
7 $G^* \leftarrow$ server.BACKWARDEQSEARCH(G^*, L_F)
8 **until** convergence;
9 return G^*

6 PRIVACY OF SHARING ONLY REGRETS

Intuitively, sharing only regrets reveals less about local data than sharing model parameters and causal networks: Attackers can infer membership in local datasets from model parameters (Shokri et al., 2017; Ma et al., 2020) and even reconstruct local datasets from model updates (Zhu and Han, 2020). Moreover, model parameters allow an attacker to craft poisoning and backdoor attacks (Sun et al., 2019). Sharing only causal graphs still does not fully protect local data, since "a causal graph can leak information about participants in the dataset" (Wang et al., 2020). PERI shares only regret values, but local causal networks can be reconstructed by optimizing Eq. (3) with respect to the target site, which is in principle NP-hard (Chickering et al., 2004).

By applying the Laplace mechanism (Dwork et al., 2006), i.e., adding appropriate noise to the regret values, we can guarantee that sensitive local data is protected in terms of ϵ-differential privacy. To prove this guarantee holds, it suffices to show that all regrets R_i have bounded sensitivity. For that, we assume that G corresponds at each site i to a parameter vector $\theta^{(i)}$ such that $X^{(i)}$ is modeled via $X_j^{(i)} = f(\mathrm{Pa}_j, \epsilon_j; \theta^{(i)})$ with independent noise ϵ_j. We assume that our score L is well-behaved in the following sense: when $X^{(i)}$ is of size n and $X'^{(i)}$ differs in one element from $X^{(i)}$ then the corresponding optimizers for L differ by $\left\|\theta^{(i)} - \theta'^{(i)}\right\|_1 \propto 1/n$. This assumption holds for many learning algorithms, e.g. convex empirical risk minimization with finite VC-dimension or Rademacher complexity (Von Luxburg and Schölkopf, 2011).

Lemma 4. *Assume that $P_i(x;\theta)$ is uniformly lower-bounded bounded by r, i.e., $\forall x \in \mathcal{X}\, \forall \theta \in \Theta : P_i(x;\theta) \geq r$, that $\|\theta\| \leq M$ for all local model parameters $\theta \in \Theta$, and that the score L is partially differentiable with respect to θ. Let $X^{(i)}$ and $X'^{(i)}$ be datasets that differ in a single element, i.e. $X^{(i)} \setminus X'^{(i)} = x_k$, θ and θ' the respective local parameters, and $\widehat{R_i}(G)$ and $\widehat{R'_i}(G)$ the respective regrets. Assume that $\|\theta - \theta'\|_1 \leq 2M/n$. Then the sensitivity $\Delta\widehat{R_i}$ of the regret is bounded by*

$$\max \left|\widehat{R_i}(G) - \widehat{R'_i}(G)\right| \leq (4M+1)\log r + \mathcal{O}\left(\frac{\log n}{n}\right) .$$

With this, it follows from the Laplace mechanism (Dwork et al., 2006) that adding Laplacian noise to regrets before sending them to the server guarantees ϵ-differential privacy.

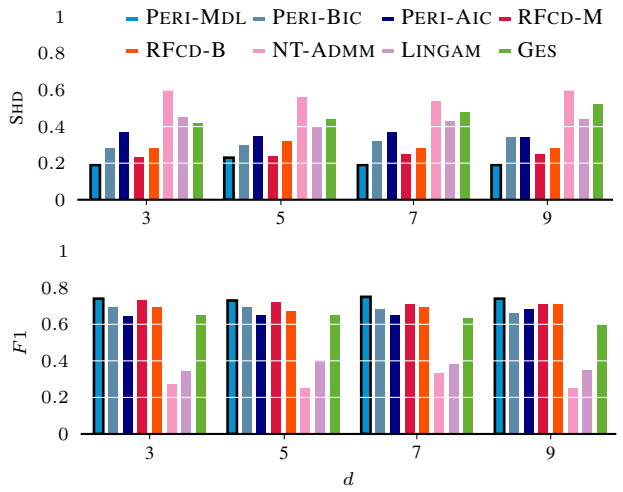

Figure 1: [Top, Lower is better] SHD and [Bottom, Higher is better] $F1$ over environment size $d = \{3, 5, 7, 9\}$. PERI-MDL performs the best overall.

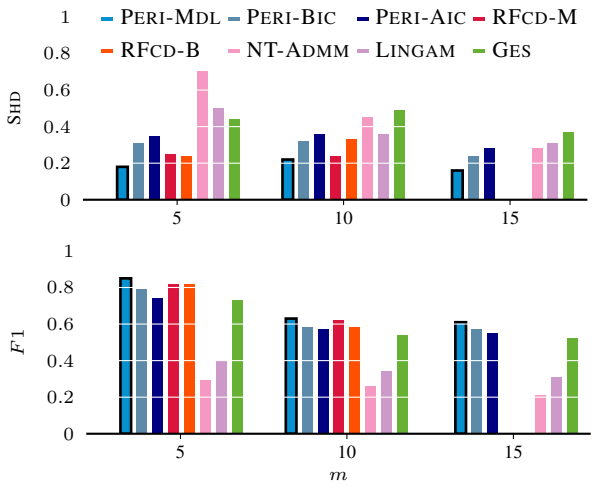

Figure 2: [Top, Lower is better] SHD and [Bottom, Higher is better] $F1$ over networks with variable count $m = \{5, 10, 15\}$. PERI-MDL consistently performs the best overall. RFCD does not terminate within 24 hours for any 15 variable networks.

Proposition 5. *Assume that each local regret $\widehat{R}_i$ has sensitivity $\leq Q$. Then* PERI *with i.i.d. Laplace noise with scale $\lambda = Q/\epsilon$ added to each $\widehat{R}_i$ is ϵ-differentially private.*

In practice, adding noise can deteriorate the training process, but we show in Sec. 7 that the practical performance of PERI is robust against noise added to local regret values and that it performs well under privacy requirements.

7 EVALUATION

Setup We instantiate PERI using three consistent scoring criteria, which are: the AIC (Sakamoto et al., 1986), BIC (Schwarz, 1978) and spline-based MDL score (Mian et al., 2021). We refer to these instantiations as PERI-AIC, PERI-BIC and PERI-MDL respectively. Since GES could get stuck in local-optima when discovering local causal networks with limited sample sizes (Lu et al., 2021), for practical reasons we run PERI in two rounds to prevent it from being misled due to incorrectly discovered local networks: first we use PERI to learn $\widetilde{G}$ using the learned $G^{(i)}$ for each environment. Next, we learn the actual G^* using PERI by enforcing $\widetilde{G}$ as the local model for all environments.

We compare to RFCD (Mian et al., 2022) as representative score-based approach. As representative ANM based method we compare to Direct-LINGAM (Shimizu, 2012), which is a modified version of the original LINGAM (Shimizu et al., 2006) for causal discovery over multiple groups. We compare to the nonlinear version of NOTEARS-ADMM (NT-ADMM) (Ng and Zhang, 2022) as continuous optimization based federated causal discovery approach. Both of the above approaches require that the model parameters be communicated between server and sites. As baseline, we use GES (Chickering, 2002) to locally discover causal networks within each environment and take a union over the discovered networks to predict the global causal network. While no parameter exchange takes place, the local causal networks are still shared with the server. We cannot compare to CDNOD (Zhang et al., 2017) or to JCI (Mooij et al., 2016) as they first pool all data and are therefore not applicable to our setting.

We evaluate the predicted networks in terms of structural similarity using the Structural Hamming Distance (SHD) (Tsamardinos et al., 2006) — which counts the number of edges where two networks differ. For comparability across multiple experiments, we normalize SHD to be in the range $[0, 1]$. To measure correctness of edge orientations in the predicted networks, we use the $F1$ score. For synthetic data, we terminate all experiments that do not finish within 24 hours. We standardize all data to have zero mean and unit variance to avoid practical issues like varsortability (Reisach et al., 2021) and make all code and data available for research purposes[2].

Results Next, we provide empirical results of our work on PERI. We extensively test PERI using both synthetic and real-world data and evaluate PERI's performance on five distinct aspects: 1) causal discovery in our intended setting 2) causal discovery when only a subset of environments are available at each learning iteration 3) communication efficiency, 4) performance under privacy considerations, and 5) causal discovery on real-world data.

Causal discovery in our intended setting We start with the simplest setting where we generate multiple datasets

[2]https://eda.rg.cispa.io/prj/peri/

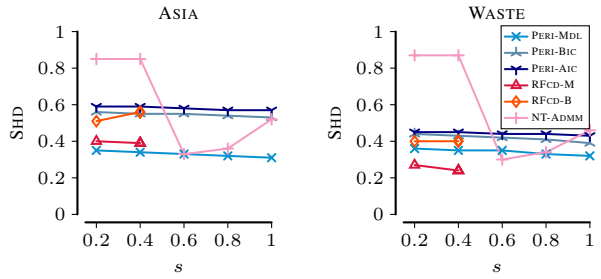

Figure 3: [Lower is better] Averaged SHD over ASIA (Left) and WASTE (Right) networks when querying only a subset s of environments. PERI-MDL performs best. Results for PERI progressively improve as more environments are allowed to be queried. RFCD-B and RFCD-M do not finish within 24 hours for any experiments with $s > 0.4$.

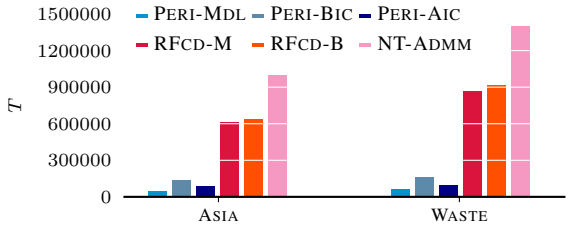

Figure 4: [Lower is better] Average number of parameter values, T, communicated to infer causal structures over ASIA and WASTE networks at $d = 100$. For baseline GES, the number of parameters is always 6400 and 8100 for ASIA resp. WASTE networks. Neither RFCD-M nor RFCD-B produce any results within 24 hours for $d = 100$. We therefore report their results for $d = 40$.

using the same underlying distribution. We have number of environments $d \in \{3, 5, 7, 9\}$, number of variables $m \in \{5, 10, 15\}$, and samples per environment $n = 5000$ as our experimental setting. We perform a total of 52 experiments for each m. We simulate DAGs using the Erdős-Rényi model and generate each effect, X_i from its parents Pa_i using functions of the form $X_i = f(\text{Pa}_i) + \epsilon_i$, where f is a non-linear function defined over Pa_i, and ϵ_i is independent additive noise Gaussian noise with zero mean. We generate complex causal relationships by defining f to be a randomly initialized 2-layer neural network, using the causal discovery toolbox (Kalainathan and Goudet, 2019).

We report the results across varying number of environments in Fig. 1 and for different sized networks in Fig. 2. We see that overall PERI-MDL outperforms all other approaches in terms of both SHD as well as orientation-$F1$. One reason for this is that spline-based MDL score uses non-parmetric regression to model causal relationships and is therefore able to identify the causal parents with higher accuracy. This is in contrast to PERI-BIC and RFCD-B, both of which use the BIC score with a lenient parameter penalty which could support inclusion of spurious edges. We see in Fig. 2 that both RFCD variants, despite their competitive performance, fail to scale to networks with $m = 15$. Moreover we find that baseline GES has better $F1$-scores than LINGAM and NT-ADMM..

Discovering networks when only a subset of environments are available As our next experiment, we generate data using two well known causal structures, namely the ASIA (Lauritzen and Spiegelhalter, 1988) and WASTE (Lauritzen, 1992) networks. We generate a total of 10 experiments, each containing 100 unique environments. At each round of update, we allow the methods to only query a fraction $s \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ of randomly chosen environments. We average the results over 30 iterations of each experiment for PERI-MDL, PERI-BIC and PERI-AIC whereas for NT-ADMM,RFCD-M and RFCD-B we average over 10 iterations due to longer run times. We omit LINGAM as it does not contain a mechanism to query a subset of environments. We show the results in Fig 3 where we see that PERI-MDL performs the best overall. All of the PERI approaches show improvement in results as the available number of environments increase. Surprisingly, NT-ADMM shows inconsistent performance which initially improves with increasing environment, but subsequently worsens even when all of the sites are available.

Communication efficiency To measure communication efficiency between server and sites, we investigate the total rounds of communications required by each approach to infer a causal network. Overall PERI-MDL is able to discover the causal network on average 15 rounds of communications for the ASIA network, with PERI-BIC and PERI-AIC following closely with 21 resp. 23 rounds. This is much less than NT-ADMM which always terminates after the max iteration cap of 176 rounds set by Ng and Zhang (2022). This means that the number of parameters that PERI exchanges during the course of learning for both ASIA and WASTE networks are significantly less than NT-ADMM and RFCD as we show in Fig. 4.

Performance under privacy considerations We test the effect of adding Laplacian noise with 0 mean and increasing scale λ over the range $[0.01-100]$ to the values of regret before communicating the regret values to the server. The results in Fig. 5 indicate that PERI is robust to Laplacian noise. Indeed, the performance of PERI does not change significantly with λ up to 1; and not even with $\lambda = 10$ when we use MDL. Since the larger noise corresponds to stronger privacy guarantees, this implies that PERI performs well under privacy requirements. We find that neither RFCD-M nor RFCD-B produced any output after 24 hours for any of the settings in this experiment.

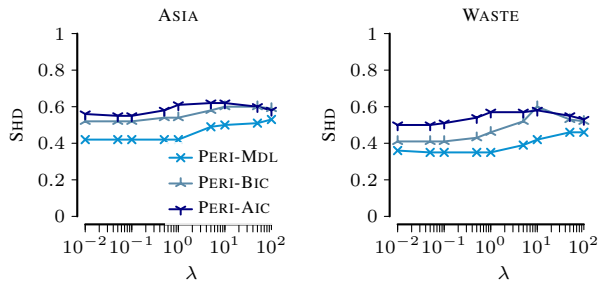

Figure 5: [Lower is better] Averaged SHD over ASIA (Left) and WASTE (Right) networks with $d = 100$ and Laplace noise on regret values with scale parameter $\lambda \in [0.0, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 100]$. PERI-MDL deteriorates the slowest. RFCD is omitted as it does not produce any output after 24 hours for this experiment.

Real world Data To see how well PERI performs on real-world data, we consider three distinct real-world networks. We consider two non-overlapping networks of sizes $\{5, 15\}$ from the Lung cancer gene-expression dataset (REGED) (Statnikov et al., 2015). For each of the REGED networks we generate 9 distinct environments without any sample overlap, each with 2000 samples per environment. Third, we consider the SACHS protein signaling network (Sachs et al., 2005) consisting of 11 variables, already measured over 9 distinct environments. The SACHS dataset provides a challenging setting since each environment has its data generated from a *different* intervened-upon causal network. This violates our assumption of a common, shared ground truth network.

We see from the results in Table 1 that PERI-MDL discovers the exact ground truth for both REGED5 and REGED15 networks and is marginally outperformed by LINGAM on the assumption-breaking case of SACHS dataset. We find that RFCD-M, which also uses a spline-based MDL score, recovers the correct causal network for REGED5 but fails to do the same for REGED15.

Table 1: [Lower is better] SHD for multiple real-world networks. PERI-MDL discovers the exact ground truth for both REGED5 and REGED15.

	REGED5	REGED15	SACHS
PERI-MDL	**0**	**0**	18
PERI-BIC	1	25	18
RFCD-M	**0**	5	**17**
RFCD-B	1	37	18
LINGAM	4	26	**17**
NT-ADMM	6	23	23
GES	2	55	25

8 DISCUSSION AND CONCLUSION

We considered the problem of discovering causal networks in a federated setup. We have proposed a new method PERI that allows us to discover causal networks in a privacy-preserving manner. Extensive experiments on diverse settings show that PERI outperforms the state of the art in federated causal discovery, both in quality of the discovered causal networks and communication efficiency while providing privacy guarantees on top of it.

We considered three different scores to instantiate PERI: AIC (Akaike, 1974), BIC (Schwarz, 1978), and spline-based MDL score (Mian et al., 2021). We found that while all three work well, the MDL score overall works best in practice. One of the reasons for the superiority is the ability of the proposed MDL score to model causal relationships non-parametrically in combination with an adaptive penalty for the parameters, rendering the method robust even when large noise values are added to regret.

We discover the global DAG by sharing only regrets, but we do not obtain the global models for each causal relationship; methods that share local model parameters do obtain them, at the cost of privacy and communication. We could additionally measure the regret with respect to global parameters θ. In such a scenario, the server proposes both G and θ to each site, instead of sending G alone. The conditions under which the parameter space Θ can be efficiently searched remains an open question.

We have used orientation-$F1$ to measure the correctness of edge orientation. Alternatively, one could consider the use of Structural Intervention Distance SID (Peters and Bühlmann, 2015), which measures the number of intervention distributions where two networks differ. It is, however, not straightforward to interpret SID between two Markov equivalence classes. This is because, as opposed to SHD, the SID of the ground-truth Markov equivalence class with itself is almost always non-zero. This makes interpretation of SID dependent on the underlying Markov equivalence class and prevents comparison across experiments.

While in this work we instantiate PERI using GES, regret-based federated causal discovert framework is agnostic of the underlying causal discovery algorithm: For any score-based causal discovery algorithm $\mathcal{A}$ and a consistent score $\widetilde{L}$ with respect to $\mathcal{A}$, if L_F defined in Eq. (3) can be proven to be consistent for $\widetilde{L}$, we can simply replace GES in Algorithm. 1 with $\mathcal{A}$ and perform federated causal learning using $\widetilde{L}$ as the score. This implies that, unlike GES, if $\mathcal{A}$ does not require the faithfulness assumption as in the case of GSP (Solus et al., 2017), we can perform causal discovery without the latter. How to entail consistency guarantees for such score-based approaches, as well as for the ones that consider a mixture of observational and interventional data (Yang et al., 2018; Squires et al., 2020; Brouillard et al., 2020) is an engaging line of future work.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE TAC*, 19(6):716–723. 3, 8

Blöbaum, P., Janzing, D., Washio, T., Shimizu, S., and Schölkopf, B. (2018). Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, pages 900–909. PMLR. 2

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. (2020). Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*. 8

Chickering, D. M. (2002). Optimal structure identification with greedy search. *JMLR*, 3:507–554. 1, 2, 3, 6

Chickering, M., Heckerman, D., and Meek, C. (2004). Large-sample learning of bayesian networks is np-hard. *JMLR*, 5. 4, 5

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer. 5

Gao, E., Chen, J., Shen, L., Liu, T., Gong, M., and Bondell, H. (2021). Federated causal discovery. *arXiv preprint arXiv:2112.03555*. 2

Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M. (2020). Inverting gradients - how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, volume 33, pages 16937–16947. Curran Associates, Inc. 1, 2

Ghassami, A., Salehkaleybar, S., Kiyavash, N., and Zhang, K. (2017). Learning causal structures using regression invariance. *arXiv preprint arXiv:1705.09644*. 1

Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*. 3

Grünwald, P. (2007). *The Minimum Description Length Principle*. MIT Press. 3

Huang, B., Zhang, K., Lin, Y., Schölkopf, B., and Glymour, C. (2018). Generalized score functions for causal discovery. In *KDD*. ACM. 1, 2

Kalainathan, D. and Goudet, O. (2019). Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278*. 7

Lauritzen, S. L. (1992). Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87(420):1098–1108. 7

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194. 7

Lu, N. Y., Zhang, K., and Yuan, C. (2021). Improving causal discovery by optimal bayesian network learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8741–8748. 6

Lyu, L. and Chen, C. (2021). A novel attribute reconstruction attack in federated learning. *arXiv preprint arXiv:2108.06910*. 1

Ma, C., Li, J., Ding, M., Yang, H. H., Shu, F., Quek, T. Q., and Poor, H. V. (2020). On safeguarding privacy and security in the framework of federated learning. *IEEE network*, 34(4):242–248. 5

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. (2018). Domain adaptation by using causal inference to predict invariant conditional distributions. In *NIPS*, volume 31. 2

Mian, O., Kaltenpoth, D., and Kamp, M. (2022). Regret-based federated causal discovery. In *The KDD 22 Workshop on Causal Discovery*, pages 61–69. PMLR. 2, 6

Mian, O., Marx, A., and Vreeken, J. (2021). Discovering fully oriented causal networks. In *AAAI*. 2, 3, 6, 8

Mooij, J. M., Magliacane, S., and Claassen, T. (2016). Joint causal inference from multiple contexts. *JMLR*, 21. 1, 2, 6

Ng, I. and Zhang, K. (2022). Towards federated bayesian network structure learning with continuous optimization. In *International Conference on Artificial Intelligence and Statistics*. 2, 6, 7

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition. 1

Peters, J. and Bühlmann, P. (2015). Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799. 8

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Statist. Soc. B*, pages 947–1012. 1

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *JMLR*, 15. 1, 2

Reisach, A., Seiler, C., and Weichwald, S. (2021). Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34. 6

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529. 8

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, 81(10.5555):26853. 6

Schwarz, G. (1978). Estimating the dimension of a model. *Annals Stat.*, 6(2):461–464. 3, 6, 8

Shimizu, S. (2012). Joint estimation of linear non-gaussian acyclic models. *Neurocomputing*, 81. 1, 2, 6

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *JMLR*, 7. 1, 2, 6

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE. 5

Singhal, K., Sidahmed, H., Garrett, Z., Wu, S., Rush, J., and Prakash, S. (2021). Federated reconstruction: Partially local federated learning. *Advances in Neural Information Processing Systems*, 34. 1, 2

Solus, L., Wang, Y., Matejovicova, L., and Uhler, C. (2017). Consistency guarantees for permutation-based causal inference algorithms. *arXiv preprint arXiv:1702.03530*. 8

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, prediction, and search*. MIT Press. 1, 2

Squires, C., Wang, Y., and Uhler, C. (2020). Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1039–1048. PMLR. 2, 8

Statnikov, A., Ma, S., Henaff, M., Lytkin, N., Efstathiadis, E., Peskin, E. R., and Aliferis, C. F. (2015). Ultra-scalable and efficient methods for hybrid observational and experimental local causal pathway discovery. *JMLR*, 16:3219–3267. 8

Sun, Z., Kairouz, P., Suresh, A. T., and McMahan, H. B. (2019). Can you really backdoor federated learning? *arXiv preprint arXiv:1911.07963*. 5

Tillman, R. E. (2009). Structure learning with independent non-identically distributed data. In *ICML*, pages 1041–1048. 2

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78. 6

Von Luxburg, U. and Schölkopf, B. (2011). Statistical learning theory: models, concepts, and results. In *Inductive Logic*, volume 10 of *Handbook of the History of Logic*, pages 651–706. Elsevier. 5, 12

Wang, L., Pang, Q., and Song, D. (2020). Towards practical differentially private causal graph discovery. *Advances in Neural Information Processing Systems*, 33:5516–5526. 1, 5

Xiong, R., Koenecke, A., Powell, M., Shen, Z., Vogelstein, J. T., and Athey, S. (2021). Federated causal inference in heterogeneous observational data. *arXiv preprint arXiv:2107.11732*. 2

Yang, K., Katcoff, A., and Uhler, C. (2018). Characterizing and learning equivalence classes of causal dags under interventions. In *ICML*, pages 5541–5550. PMLR. 2, 8

Ye, Q., Amini, A. A., and Zhou, Q. (2022). Distributed learning of generalized linear causal networks. *arXiv preprint arXiv:2201.09194*. 2

Zhang, K., Huang, B., Zhang, J., Glymour, C., and Schölkopf, B. (2017). Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI*. 1, 2, 6

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31. 2

Zhu, L. and Han, S. (2020). Deep leakage from gradients. In *Federated learning*, pages 17–31. Springer. 5

Nothing but Regrets — Privacy-Preserving Federated Causal Discovery Appendix

A PROOFS AND FORMAL RESULTS

In this section we provide the proofs for the formal results in Sec. 4 and Sec. 6.

Theorem 1. *Let G^* be the true causal network for all $P(X^{(i)})$ and let $n^{(1)} \dots, n^{(d)} \to \infty$. Further let L be a consistent score. Then*

$$\lim_{n^{(1)},\dots,n^{(d)}\to\infty} P\left(\widehat{G} \sim G^*\right) = 1\,.$$

That is, $\max_i R_i(G)$ is consistent when all $n^{(i)} \to \infty$.

Proof. Since L is a consistent score, we know that $\lim_{n^{(i)}\to\infty} P(G^{(i)} = G^*) = 1$ for all i. Thus $P\left(\widehat{G} = \operatorname{argmin}_G \max_i \left(L(X^{(i)};G) - L(X^{(i)};G^*)\right)\right) = 1$, which is clearly minimized when $\widehat{G} \sim G^*$. □

Theorem 2. *Let G^* be the true causal network for all $P(X^{(i)})$ and let $N \coloneqq \max_i n^{(i)} \to \infty$. Further let L be a consistent score. Then*

$$\lim_{N\to\infty} P\left(\widehat{G} \sqsupseteq G^*\right) = 1\,.$$

Proof. When *all* $n^{(i)} \to \infty$, Thm. 1 applies.

We therefore consider the case where some $n^{(i)}$ remain bounded. Let $I = \left\{i : n^{(i)} < \infty\right\}$ and $M = \max\left\{\limsup n^{(i)} : i \in I\right\}$. Then we have $\max_G \max_{i\in I} R_i(G) \leq cM < \infty$ for some $c > 0$. Meanwhile for all i with $n^{(i)} \to \infty$ we have for all $G \subsetneq G^*$ that $L(X^{(i)};G) - L(X^{(i)};G^{(i)}) \approx L(X^{(i)};G) - L(X^{(i)};G^*) \propto n^{(i)} \to \infty$. Hence any smaller $G \sqsubset G^*$ achieves strictly worse minmax regret than any $G \sqsupseteq G^*$ as $N \to \infty$. □

Corollary 3. *Let the assumptions of Thm. 2 hold and let L be the BIC score. Then*

$$\lim_{N\to\infty} P\left(\widehat{G} \sim G^*\right) = 1\,.$$

That is, the score $\max_i R_i(G)$ is consistent when L incorporates a BIC-penalty for parameters and $N \to \infty$.

Proof. When L is the BIC score then for any dataset i such that $n^{(i)} \to \infty$ we have $R_i(G) \propto \log(n^{(i)}) \to \infty$ when $G \sqsupset G^*$ is too large. This grows larger than any finite penalty incurred from any of the datasets j with $n^{(j)} \leq M$ bounded, so that picking $\widehat{G} \sim G^*$ will be the best choice as $N \to \infty$.

□

Lemma 4. *Assume that $P(x;G)$ is uniformly lower-bounded bounded by r, i.e., $\forall x \in \mathcal{X}\ \forall G : P(x;G) \geq r$, that $\|\theta\| \leq M$ for all local model parameters θ, and that the score L is partially differentiable wrt. θ. Let $X^{(i)}$ and $X'^{(i)}$ be datasets that only differ in a single element, i.e., $X^{(i)} \setminus X'^{(i)} = x_k$, θ and θ' the respective local parameters, and $\widehat{R_i}(G)$ and $\widehat{R'_i}(G)$ the respective regrets. Assume that $|\theta - \theta'| \leq 2M/n$. Then the sensitivity $\Delta\widehat{R_i}$ of the regret is bounded by*

$$\max\left|\widehat{R_i}(G) - \widehat{R'_i}(G)\right| \leq (4M+1)\log r + \mathcal{O}\left(\frac{\log n}{n}\right)\,.$$

Proof. Removing a single element from a local dataset $X^{(i)}$ changes also the local causal model, both in terms of the DAG $G^{(i)}$ and the local model parameters $\theta^{(i)}$. There fore, the local score changes for two reasons: (i) the dataset the score is computed on changes, and (ii) the local causal model changes. That is, the sensitivity is

$$\begin{aligned}\max \left|\widehat{R_i}(G) - \widehat{R_i'}(G)\right| = &\Big|L(X^{(i)}, G) - L(X^{(i)}, G^{(i)}) \\ &-L(X'^{(i)}, G) + L(X'^{(i)}, G'^{(i)})\Big| \\ = &\left|L(X'^{(i)}, G'^{(i)}) - L(X^{(i)}, G^{(i)})\right| \quad .\end{aligned}$$

Thus, it suffices to bound $|L(X', G') - L(X, G)|$ for datasets X and X' that only differ in a single element and corresponding different DAGs G, G' and local model parameters θ, θ'. This difference encompasses both the difference in DAGs and local model parameters. Since the difference in DAGs is determined by the difference of θ and θ', we for convenience write $L(X, G) = L(X, \theta)$ and show that the difference $|L(X, \theta) - L(X', \theta')|$ is bounded. Since

$$\begin{aligned}|L(X, \theta) - L(X', \theta')| \leq &|L(X', \theta) - L(X, \theta)| \\ &+ \|\theta - \theta'\| \, |L(X, \theta') - L(X, \theta)| \quad ,\end{aligned}$$

we can use the linearization of L and get

$$\begin{aligned}|L(X, \theta) - L(X', \theta')| \leq &\underbrace{|L(x_k, \theta)|}_{\leq \log r} \\ &+ \|\theta - \theta'\| \underbrace{|L(X, \theta) - L(X, \theta')|}_{\propto n} \\ &+ \underbrace{\|\theta - \theta'\|}_{\leq 2M/n} |L(\theta) - L(\theta')| + \mathcal{O}\left(\frac{\log n}{n}\right) \\ \leq &\log r + 2M \log r + 2M \log r + \mathcal{O}\left(\frac{\log n}{n}\right) \\ = &(4M + 1) \log r + \mathcal{O}\left(\frac{\log n}{n}\right) \quad .\end{aligned}$$

It follows that the sensitivity is bounded by $(4M + 1) \log r + \mathcal{O}(\log n/n)$. Note that the assumption $|\theta - \theta'| \leq 2M/n$ for θ, θ' optimized on datasets that only differ in a single element holds for most learning algorithms, e.g., convex empirical risk minimization with finite VC-dimension or Rademacher complexity Von Luxburg and Schölkopf (2011). □

Proposition 5. *Assume that $P(x; G)$ is uniformly lower-bounded bounded by r and that the regret has sensitivity $(4M + 1) \log r + \mathcal{O}(\log n/n)$. Then* PERI *with Laplacian noise with scale $\lambda = \epsilon^{-1}((4M + 1) \log r + \mathcal{O}(\log n/n))$ added to local regret values is ϵ-differentially private.*

Proof. The Laplace mechanism guarantees that adding noise with mean 0 and scale λ to a function f with sensitivity δf is $\delta f/\lambda$-differentially private. Since the regret has sensitivity $(4M + 1) \log r + \mathcal{O}\left(\frac{\log n}{n}\right)$, choosing $\lambda = \epsilon^{-1}((4M + 1) \log r + \mathcal{O}(\log n/n))$ results in a sensitivity of

$$\frac{\delta R}{\lambda} = \frac{(4M + 1) \log r + \mathcal{O}\left(\frac{\log n}{n}\right)}{\epsilon^{-1}((4M + 1) \log r + \mathcal{O}(\log n/n))} = \epsilon \quad .$$

□

B ALL EVALUATION RESULTS

B.1 Causal Discovery in our intended setting

Table 1: [Lower is better] Normalized SHD for methods over varying numbers of environments d, averaged over network sizes $\{5, 10, 15\}$.

d	PERI-MDL	PERI-BIC	PERI-AIC	RFCD-B	RFCD-M	NT-ADMM	GES	LINGAM
3	**0.185**	0.281	0.371	0.277	0.230	0.599	0.429	0.445
5	**0.232**	0.304	0.353	0.317	0.243	0.563	0.438	0.400
7	**0.186**	0.322	0.370	0.277	0.245	0.542	0.478	0.434
9	**0.194**	0.342	0.337	0.277	0.255	0.603	0.520	0.440

Table 2: [Higher is better] $F1$ score for all methods over varying numbers of environments d, averaged over network sizes $\{5, 10, 15\}$.

d	PERI-MDL	PERI-BIC	PERI-AIC	RFCD-B	RFCD-M	NT-ADMM	GES	LINGAM
3	**0.741**	0.694	0.647	0.697	0.739	0.273	0.657	0.340
5	**0.726**	0.694	0.653	0.670	0.722	0.258	0.653	0.402
7	**0.750**	0.680	0.646	0.697	0.712	0.330	0.632	0.381
9	**0.740**	0.659	0.678	0.711	0.712	0.248	0.608	0.350

Table 3: [Lower is better] Normalized SHD for all methods over varying numbers of variables m, averaged over environment sizes $\{3, 5, 7, 9\}$.

m	PERI-MDL	PERI-BIC	PERI-AIC	RFCD-M	RFCD-B	LINGAM	NT-ADMM	GES
5	**0.180**	0.310	0.350	0.250	0.240	0.500	0.710	0.440
10	**0.218**	0.325	0.368	0.240	0.330	0.363	0.450	0.495
15	**0.163**	0.240	0.275	-	-	0.310	0.280	0.375

Table 4: [Higher is better] $F1$ score for all methods over varying numbers of variables m, averaged over environment sizes $\{3, 5, 7, 9\}$.

m	PERI-MDL	PERI-BIC	PERI-AIC	RFCD-M	RFCD-B	LINGAM	NT-ADMM	GES
5	**0.850**	0.790	0.740	0.820	0.820	0.400	0.290	0.730
10	**0.628**	0.578	0.570	0.623	0.575	0.343	0.263	0.545
15	**0.613**	0.568	0.550	-	-	0.305	0.208	0.518

B.2 Discovering networks when only a subset of environments are available

Table 5: [Loweris better] Averaged SHD over ASIA and WASTE networks when querying only a subset s of environments. PERI-MDL performs best. Results for PERI progressively improve as more environments are allowed to be queried. RFCD-B and RFCD-M do not finish within 24 hours for any experiments with $s > 40\%$.

s	Dataset	PERI-MDL	PERI-BIC	PERI-AIC	RFCD-M	RFCD-B	NT-ADMM
20%		**0.35**	0.56	0.59	0.4	0.51	0.85
40%		**0.34**	0.55	0.59	0.39	0.56	0.85
60%	ASIA	**0.33**	0.55	0.58	-	-	0.33
80%		**0.32**	0.54	0.57	-	-	0.36
100%		**0.31**	0.53	0.57	-	-	0.52
20%		0.36	0.44	0.45	**0.27**	0.4	0.87
40%		0.35	0.43	0.45	**0.24**	0.4	0.87
60%	WASTE	**0.35**	0.42	0.44	-	-	0.3
80%		**0.33**	0.41	0.44	-	-	0.34
100%		**0.32**	0.39	0.43	-	-	0.46

Table 6: [Higher is better] Orientation $F1$ over ASIA and WASTE networks when querying only a subset s of environments. PERI-MDL performs best. Results for PERI progressively improve as more environments are allowed to be queried. RFCD-B and RFCD-M do not finish within 24 hours for any experiments with $s > 40\%$.

s	Dataset	PERI-MDL	PERI-BIC	PERI-AIC	RFCD-M	RFCD-B	NT-ADMM
20%		**0.35**	0.56	0.59	0.4	0.51	0.85
40%		**0.34**	0.55	0.59	0.39	0.56	0.85
60%	ASIA	**0.33**	0.55	0.58	-	-	0.33
80%		**0.32**	0.54	0.57	-	-	0.36
100%		**0.31**	0.53	0.57	-	-	0.52
20%		0.36	0.44	0.45	**0.27**	0.4	0.87
40%		0.35	0.43	0.45	**0.24**	0.4	0.87
60%	WASTE	**0.35**	0.42	0.44	-	-	0.3
80%		**0.33**	0.41	0.44	-	-	0.34
100%		**0.32**	0.39	0.43	-	-	0.46

B.3 Performance under privacy considerations

Table 7: [Lower is better] Averaged SHD over ASIA and WASTE networks with number of environments $d = 100$ and Laplace noise on regret values with scale parameter $\lambda \in [0.0, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 100]$. RFCD is omitted as it does not produce any output after 24 hours for this experiment.

λ	Dataset	PERI-MDL	PERI-BIC	PERI-AIC
0	ASIA	**0.43**	0.52	0.55
0.01		**0.42**	0.52	0.56
0.05		**0.42**	0.52	0.55
0.1		**0.42**	0.52	0.55
0.5		**0.42**	0.54	0.58
1		**0.42**	0.54	0.61
5		**0.49**	0.58	0.62
10		**0.50**	0.60	0.62
50		**0.51**	0.60	0.60
100		**0.53**	0.59	0.58
0	WASTE	**0.35**	0.41	0.50
0.01		**0.36**	0.41	0.50
0.05		**0.35**	0.41	0.50
0.1		**0.35**	0.41	0.51
0.5		**0.35**	0.43	0.54
1		**0.35**	0.46	0.57
5		**0.39**	0.52	0.57
10		**0.42**	0.60	0.58
50		**0.46**	0.53	0.55
100		**0.46**	0.52	0.53

Table 8: [Higher is better] Averaged F1 over ASIA and WASTE networks with number of environments $d = 100$ and Laplace noise on regret values with scale parameter $\lambda \in [0.0, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 100]$. RFCD is omitted as it does not produce any output after 24 hours for this experiment.

λ	Dataset	PERI-MDL	PERI-BIC	PERI-AIC
0	ASIA	**0.41**	0.38	0.34
0.01		**0.42**	0.38	0.34
0.05		**0.43**	0.38	0.34
0.1		**0.42**	0.38	0.34
0.5		**0.42**	0.37	0.32
1		**0.4**	0.34	0.3
5		**0.34**	0.3	0.29
10		**0.34**	0.29	0.29
50		**0.31**	0.27	0.28
100		**0.31**	0.27	0.27
0	WASTE	0.4	**0.44**	0.36
0.01		0.41	**0.44**	0.36
0.05		0.41	**0.44**	0.37
0.1		0.41	**0.43**	0.37
0.5		**0.41**	0.4	0.32
1		**0.4**	0.35	0.31
5		**0.35**	0.3	0.28
10		**0.35**	0.28	0.29
50		**0.33**	0.27	0.27
100		**0.34**	0.27	0.26

B.4 Real-world data

Table 9: [Lower is better] SHD for multiple real-world networks. PERI-MDL discovers the exact ground truth for both REGED5 and REGED15.

	REGED5	REGED15	SACHS
PERI-MDL	**0**	**0**	18
PERI-BIC	6	25	18
PERI-AIC	6	70	28
RFCD-M	**0**	5	**17**
RFCD-B	1	37	18
LINGAM	4	26	**17**
NT-ADMM	6	23	23
GES	2	56	25

Table 10: [Higher is better] Orientation $F1$ for multiple real-world networks. PERI-MDL discovers the exact ground truth for both REGED5 and REGED15.

	REGED5	REGED15	SACHS
PERI-MDL	**1.0**	**1.0**	0.38
PERI-BIC	0.59	0.58	**0.50**
PERI-AIC	0.59	0.30	0.43
RFCD-M	**1.0**	0.89	0.37
RFCD-B	0.89	0.41	0.28
LINGAM	0.22	0.26	0.33
NT-ADMM	0.36	0.13	0.38
GES	0.84	0.39	0.42