Enhancing Causal Discovery in Federated Settings with Limited Local Samples

Xianjie Guo[*]
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, China
xianjieguo@mail.hfut.edu.cn

Liping Yi[*]
College of Computer Science
Nankai University
Tianjin, China
yiliping@nbjl.nankai.edu.cn

Xiaohu Wu
National Engineering Research Center of Mobile Network Technologies
Beijing University of Posts and Telecommunications
Beijing, China
xiaohu.wu@bupt.edu.cn

Kui Yu[†]
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, China
yukui@hfut.edu.cn

Gang Wang[†]
College of Computer Science
Nankai University
Tianjin, China
wgzwp@nbjl.nankai.edu.cn

Abstract

Causal discovery from observational data is crucial for understanding complex systems, but traditional methods often require centralized data, conflicting with growing privacy concerns. Although federated causal discovery (FCD) has emerged as a solution, existing methods struggle when individual clients possess limited local samples. This paper introduces `FedECD`, a novel approach addressing causal discovery in federated settings with limited local samples. `FedECD` comprises two phases: 1) Federated Causal Skeleton Optimization and 2) Federated Causal Structure Refinement, both leveraging Bootstrapping techniques to enhance robustness and accuracy across distributed clients. Both phases employ a two-layer aggregation strategy: client-layer aggregates results from Bootstrapped sub-datasets within each client, while server-layer aggregates across all clients. The first phase uses weighted aggregation to iteratively remove false causal edges based on conditional independence tests. In contrast, the second phase utilizes majority voting to determine edge directions, ensuring robust estimation of the true causal structure. Extensive experiments on eight benchmark Bayesian network datasets demonstrate the superiority of `FedECD` over existing FCD methods, particularly with limited

[*]These authors contributed equally to this work.
[†]Corresponding authors: Kui Yu and Gang Wang.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

sample sizes. `FedECD` achieves an average improvement of 7.53% in the Ar_F1 score compared to the best baseline, addressing a critical challenge in FCD.

1 Introduction

Background Causal discovery (CD) aims to uncover causal relationships between variables or events, often represented as causal structures, from observational data [1, 2, 3, 4, 5]. It is a promising approach to address the limitations of current machine learning techniques, particularly deep learning, which often lack robustness and interpretability [6, 7, 8]. Developing effective CD methods is crucial for realizing the "high-value" transformation of big data [9]. Numerous studies have sought to discover causal relationships in various fields, including medicine [10], computer science [11], and bioinformatics [12], to enable inference and analysis of events.

Traditionally, CD methods require large-scale datasets aggregated from multiple decentralized data centers to achieve satisfactory performance. However, with the development of big data and an increasing international emphasis on data privacy protection, data collection has become increasingly costly and often prohibited by privacy protection laws and regulations [13, 14]. To address these challenges, federated causal discovery (FCD) has emerged as a novel direction [15, 16, 17, 18, 19]. FCD leverages the privacy-preserving capabilities of the federated learning (FL) paradigm [20] to achieve satisfactory CD performance while protecting data privacy, allowing each client (institution or organization) to keep their data local. For example, in healthcare, multiple hospitals may wish to collaboratively discover causal relationships among various symptoms and diseases without sharing sensitive patient data. FCD enables these institutions to jointly learn a causal model while keeping their individual datasets confidential.

Motivation Although existing FCD algorithms have achieved satisfactory performance in various scenarios (e.g., data heterogeneity [18], nonlinear causal relationships [15], high-dimensional data [19]), they have primarily focused on data privacy issues while overlooking another critical challenge in federated learning scenarios: the limited sample size problem [21]. This problem arises when each FL client holds a very limited number of samples, which can severely degrade the performance of existing FCD methods. To visually illustrate this challenge, we conducted extensive experiments using a state-of-the-art FCD method, FedPC [17], on three benchmark Bayesian network (BN) datasets: Child, Insurance, and Alarm[3]. We generated multiple batches of datasets with varying sample sizes from these BNs. Setting the total number of clients to 10, we allocated an average sample size per client ranging from 150 to 1,000. We then ran the FedPC algorithm under different dataset scenarios and compared the learned causal structures with the ground truth using two common metrics: Ar_F1 and SHD [17]. The experimental results are shown in Figure 1, and we can observe that when the average sample size per client falls below 350 (in the range [150, 350]), the performance

[3]These benchmark BNs are publicly available at `http://www.bnlearn.com/bnrepository/`.

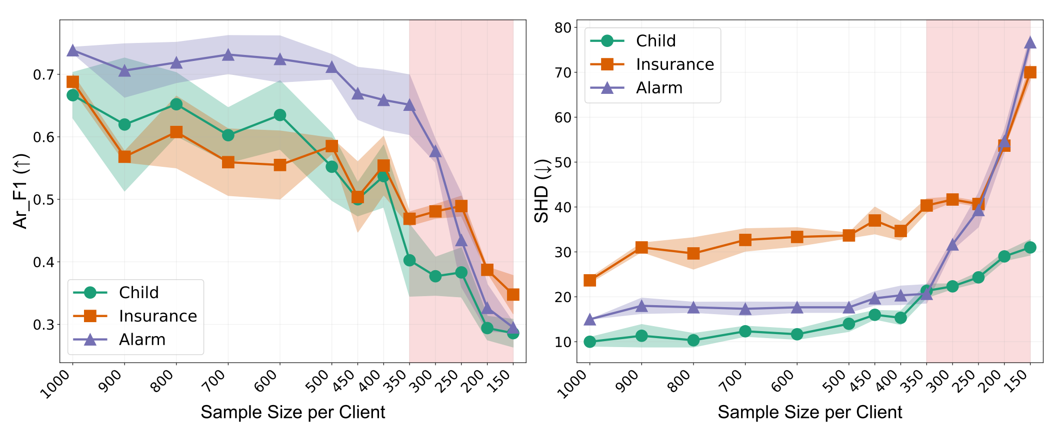

Figure 1: Performance of the FedPC algorithm under varying sample sizes per client.

of FedPC drops dramatically. This indicates that the limited sample sizes per client in federated scenarios significantly impact the performance of existing FCD algorithms.

Inspired by existing studies [22, 23] that have demonstrated the efficacy of Bootstrapping [24] in scenarios with limited sample sizes, we pose a critical question: can we leverage Bootstrapping techniques to resample local client data, thereby addressing the challenges posed by limited sample sizes and ultimately enhancing the performance of FCD in these demanding settings?

Contributions In this paper, we focus on enhancing FCD performance in scenarios where each FL client holds very limited samples. We propose the `FedECD` algorithm, which consists of two phases:

- Phase 1: *Federated Causal Skeleton Optimization Using Bootstrapping*. `FedECD` employs a two-layer causal skeleton aggregation strategy. In the first layer, clients internally aggregate causal skeletons learned from resampled sub-datasets generated via Bootstrapping. In the second layer, the server aggregates the causal skeletons returned by each FL client. This strategy helps prevent the loss of true causal edges that might occur due to the small sample sizes of individual clients.
- Phase 2: *Federated Causal Structure Refinement Leveraging Bootstrapping*. `FedECD` utilizes a two-layer causal structure aggregation strategy. In the first layer, clients internally aggregate causal structures learned from resampled datasets generated via Bootstrapping. In the second layer, the server aggregates the causal structures returned by each client. This strategy helps mitigate issues of incorrect edge orientations in causal structures that might arise due to the small sample sizes of individual clients.

To the best of our knowledge, `FedECD` is the first algorithm to tackle the challenge of limited local samples in FCD scenarios. Extensive experiments on eight benchmark Bayesian network datasets demonstrate the superiority of `FedECD` over seven state-of-the-art baselines, with an average improvement of 7.53% in the Ar_F1 score compared to the best baseline.

2 Related Work

Federated causal discovery (FCD) has emerged as a critical research direction, addressing the need to uncover causal relationships between variables from decentralized data while preserving privacy. In this section, we provide a comprehensive overview of existing approaches in FCD.

2.1 Continuous Optimization-based Methods

Early FCD algorithms derived primarily from continuous optimization-based CD methods. These methods became pioneers in FCD largely due to their foundation in gradient descent techniques, which allowed them to directly leverage the various optimization strategies already well-developed in the field of federated learning (FL). This natural compatibility made the extension of CD methods to federated learning settings relatively straightforward and effective.

Specifically, NOTEARS-ADMM [15] adapted the NOTEARS algorithm [25] to a federated setting using the ADMM [26] optimization method. Importantly, the authors proposed two versions: NOTEARS-ADMM for linear causal relationships, and NOTEARS-MLP-ADMM for nonlinear causal relationships. The latter employs multilayer perceptrons (MLPs) to model complex, nonlinear interactions between variables, significantly extending the applicability of the method to more realistic scenarios. Building on this foundation, FedDAG [16] introduced a two-level structure in local models, separating graph structure learning from causal mechanism approximation. This design allowed each client to adapt to its local data characteristics while contributing to the learning of a global causal structure through gradient-based optimization. FedCausal [27] proposed a global optimization formula to aggregate causal graphs from client data while constraining the acyclicity of the global graph. Its flexible optimization objective enables adaptive handling of both homogeneous and heterogeneous data, further advancing FCD's capability in dealing with complex data environments.

2.2 Constraint-based Methods

Constraint-based methods in FCD focus on learning causal structures through conditional independence tests performed across distributed datasets. FedPC [17] adapted the PC algorithm [28] to a federated setting, introducing a layer-wise aggregation strategy for skeleton learning and a consistent separation sets identifying strategy for skeleton orientation. This approach showed promise in handling larger-scale problems but faced data heterogeneity challenges. Very similar to FedPC, FedC^{2}SL [29] developed a federated conditional independence test protocol to minimize privacy leakages and address client heterogeneity. To further address the challenge of data heterogeneity in FCD, FedCDH [18] introduced a surrogate variable to account for distribution differences across clients. It proposed a federated conditional independence test (FCIT) for skeleton discovery and a federated independent change principle (FICP) for causal direction determination, making no assumptions about specific functional forms.

2.3 Hybrid and Novel Optimization Methods

Some approaches have combined elements from multiple categories or introduced novel optimization techniques. Specifically, DARLS [30] proposed a method simulating an annealing process to search over the space of topological sorts. It used distributed optimization to find the optimal graphical structure, providing theoretical guarantees of convergence to an Oracle solution. FedCSL [19] introduced a federated local-to-global learning strategy to improve scalability for high-dimensional data. It also proposed a novel weighted aggregation strategy to handle uneven sample allocation across clients without compromising privacy. PERI [31] introduced an approach based on distributed min-max regret optimization. By sharing only regret information, PERI achieves federated causal discovery while minimizing the amount of information exchanged between clients and the central server. FED-CD [32] introduced a framework for inferring causal structures from both observational and interventional data in a privacy-preserving manner. It proposed a knowledge aggregation method based on proximity to interventions within the global causal structure.

Although existing FCD approaches have addressed various challenges, including data heterogeneity, scalability, and privacy concerns, a critical issue remains largely unexplored: the performance degradation in scenarios with limited local samples. This challenge is particularly acute in real-world federated learning applications, where individual clients often possess small sample sizes. In this paper, we propose `FedECD` to tackle this overlooked problem by introducing a novel Bootstrapping-based approach, enhancing causal discovery in federated settings with limited local samples.

3 Preliminaries

3.1 Notations and Assumptions

Let $\mathcal{X} = \{X_1, X_2, ..., X_d\}$ be a set of d variables, and $\mathcal{C} = \{c_1, c_2, ..., c_m\}$ be a set of m clients. We consider a horizontal FL setting, where clients have large overlaps in the variable space but little overlap in the sample space. For each client c_k ($k \in \{1, 2, ..., m\}$), $\mathcal{D}^{c_k} \in \mathbb{R}^{n_{c_k} \times d}$ represents its local dataset, where n_{c_k} is the number of samples. Each sample in $\mathcal{D}^{c_k}$ is independently drawn from the probability distribution $\mathbb{P}^{c_k}(\mathcal{X})$. We define $\mathcal{D}^{\mathcal{C}} = \{\mathcal{D}^{c_1}, \mathcal{D}^{c_2}, ..., \mathcal{D}^{c_m}\}$ as the decentralized dataset and $\mathbb{P}^{\mathcal{C}}(\mathcal{X}) = \{\mathbb{P}^{c_1}(\mathcal{X}), \mathbb{P}^{c_2}(\mathcal{X}), ..., \mathbb{P}^{c_m}(\mathcal{X})\}$ as the decentralized probability distribution set. Causal relationships over $\mathcal{X}$ are often represented by a causal directed acyclic graph (DAG). In a causal DAG, if there is a direct edge $X_{i_1} \to X_{i_2}$ ($i_1, i_2 \in \{1, 2, ..., d\}$), X_{i_1} is a direct cause of X_{i_2}, and X_{i_2} is a direct effect of X_{i_1} [33].

A causal DAG serves as the fundamental graphical structure for Bayesian networks (BNs), where a BN extends the DAG by associating each node with a conditional probability distribution that quantifies the relationship between a node and its parents, thereby transforming a purely structural representation into a complete probabilistic model capable of encoding both qualitative relationships through its graph structure and quantitative dependencies through its probability distributions.

Definition 1 (Bayesian Network [28]). *Given a causal DAG $\mathcal{G}$ and a joint probability distribution $\mathbb{P}$ over a set of random variables $\mathcal{X}$, the triplet $< \mathcal{X}, \mathcal{G}, \mathbb{P} >$ is called a Bayesian network if it satisfies the Markov condition: each variable in $\mathcal{G}$ is conditionally independent of any subset of its non-descendants given its parent variables.*

Definition 2 (D-Separation [28]). *In a causal DAG $\mathcal{G}$, a path γ is d-separated (or blocked) by a variable set $\boldsymbol{S} \subset \mathcal{X}$ if and only if: 1) γ contains a chain $X_{i_1} \to X_{i_3} \to X_{i_2}$ or a fork $X_{i_1} \leftarrow X_{i_3} \to X_{i_2}$ with the middle variable $X_{i_3} \in \boldsymbol{S}$, or 2) γ contains an inverted fork (or collider) $X_{i_1} \to X_{i_3} \leftarrow X_{i_2}$ with $X_{i_3} \notin \boldsymbol{S}$ and X_{i_3}'s descendants $\notin \boldsymbol{S}$.*

Two variables X_{i_1} and X_{i_2} are d-separated by a variable set $\boldsymbol{S}$ if and only if $\boldsymbol{S}$ blocks every path from X_{i_1} to X_{i_2}. We call such a set $\boldsymbol{S}$ a separation set of X_{i_1} from X_{i_2}.

Definition 3 (Faithfulness [28]). *Given a BN $< \mathcal{X}, \mathcal{G}, \mathbb{P} >$, the probability distribution $\mathbb{P}$ is faithful to the DAG $\mathcal{G}$ if and only if for any variables $X_i, X_j \in \mathcal{X}$ and any subset $\boldsymbol{S} \subseteq \mathcal{X} \setminus \{X_i, X_j\}$, the following equivalence holds: $X_i \perp\!\!\!\perp X_j | \boldsymbol{S}$ in $\mathbb{P} \iff X_i$ and X_j are d-separated by $\boldsymbol{S}$ in $\mathcal{G}$.*

The Faithfulness assumption establishes a relation between a probability distribution $\mathbb{P}$ and its underlying DAG $\mathcal{G}$. In a BN, this assumption implies that two variables $X_{i_1}, X_{i_2} \in \mathcal{X}$ that are conditionally independent given a subset $\boldsymbol{S} \subseteq \mathcal{X} \setminus \{X_{i_1}, X_{i_2}\}$ in $\mathbb{P}$ are d-separated by $\boldsymbol{S}$ in $\mathcal{G}$.

Definition 4 (Causal Sufficiency [33, 28]). *A set of variables $\mathcal{X}$ satisfies causal sufficiency if and only if for any variables $X_{i_1}, X_{i_2} \in \mathcal{X}$, every common cause of X_{i_1} and X_{i_2} is also in $\mathcal{X}$.*

Under the assumptions of Faithfulness and Causal Sufficiency, we can use conditional independence (CI) tests to find all dependencies or independencies entailed in a BN.

Definition 5 (Conditional Independence). *Two variables X_{i_1} and X_{i_2} ($i_1, i_2 \in \{1, 2, ..., d\}$) are conditionally independent given a variable set $\boldsymbol{S} \subseteq \mathcal{X} \setminus \{X_{i_1}, X_{i_2}\}$ if $P(X_{i_1}, X_{i_2}|\boldsymbol{S}) = P(X_{i_1}|\boldsymbol{S})P(X_{i_2}|\boldsymbol{S})$; otherwise, they are conditionally dependent given $\boldsymbol{S}$.*

We denote conditional independence between variables X_{i_1} and X_{i_2} given a set of variables $\boldsymbol{S}$ as $X_{i_1} \perp\!\!\!\perp X_{i_2}|\boldsymbol{S}$. To evaluate these conditional independence relationships in our experiments, we employ the G^2 test [28], a likelihood-ratio alternative to the conventional χ^2 test. The G^2 statistic is defined as:

$$G^2 = 2 \sum_{a,b,\boldsymbol{v}} H^{a,b,\boldsymbol{v}}_{i_1,i_2,\boldsymbol{S}} \ln\left(\frac{H^{a,b,\boldsymbol{v}}_{i_1,i_2,\boldsymbol{S}} H^{\boldsymbol{v}}_{\boldsymbol{S}}}{H^{a,\boldsymbol{v}}_{i_1,\boldsymbol{S}} H^{b,\boldsymbol{v}}_{i_2,\boldsymbol{S}}}\right), \tag{1}$$

where $H^{a,b,\boldsymbol{v}}_{i_1,i_2,\boldsymbol{S}}$ denotes the number of counts satisfying $X_{i_1} = a$, $X_{i_2} = b$ and $\boldsymbol{S} = \boldsymbol{v}$, and $H^{\boldsymbol{v}}_{\boldsymbol{S}}$, $H^{a,\boldsymbol{v}}_{i_1,\boldsymbol{S}}$ and $H^{b,\boldsymbol{v}}_{i_2,\boldsymbol{S}}$ are defined similarly. The G^2 statistic is asymptotically distributed as χ^2 with degrees of freedom (df) calculated as:

$$df = (r_{i_1} - 1)(r_{i_2} - 1) \prod_{X_{i_3} \in \boldsymbol{S}} r_{i_3}, \tag{2}$$

where r_{i_1}, r_{i_2} and r_{i_3} are the domains (number of distinct values) of X_{i_1}, X_{i_2} and X_{i_3}, respectively. Given a significance level α and the p-value ρ returned by the G^2 test, under the null hypothesis $H_0 : X_{i_1} \perp\!\!\!\perp X_{i_2}|\boldsymbol{S}$, we conclude that $X_{i_1} \perp\!\!\!\perp X_{i_2}|\boldsymbol{S}$ holds if and only if $\rho > \alpha$.

Federated causal discovery (FCD) aims to identify a causal DAG $\mathcal{G}$ from all local datasets $\mathcal{D}^{c_k}{}_{k \in \{1,2,...,m\}}$ in a privacy-preserving manner. We make the following assumption:

Assumption 1 (Invariant Causal DAG [16]). *All local datasets are uniformly sampled from the same causal DAG $\mathcal{G}$, although the probability distribution of samples for the same variable space can differ across different clients.*

By performing CI tests and identifying d-separation relationships among random variables, we can infer the entire graph structure in a federated setting while preserving data privacy.

3.2 Bootstrapping Technique

This paper proposes a novel Bootstrapping-based approach to enhance causal discovery in federated settings with limited local samples. To understand the rationale behind this choice, it's important first to discuss the concept of ensemble learning and its connection to Bootstrapping.

Ensemble learning is a widely adopted machine learning approach that combines multiple models to solve a problem, often achieving superior generalization performance compared to single models [34]. Bootstrapping [24], a resampling technique, plays a crucial role in creating diverse datasets for ensemble methods. Given an original dataset D_{orig}, the process of resampling through Bootstrapping for generating a sub-dataset D_1 is as follows:

- Randomly selecting an instance from D_{orig} and adding it to D_1. The selected instance is then returned to D_{orig}, allowing for potential resampling.
- Repeating this procedure n times to create D_1 containing n instances.

A key property of Bootstrapping, which makes it particularly useful for our purposes, is captured in the following proposition:

Proposition 1 ([24])**.** *When generating a sub-dataset D_1 from an original dataset D_{orig} using Bootstrapping, as $n \to \infty$, approximately 36.8% of the samples in D_{orig} will not appear in D_1.*

Proof. For each sample in D_{orig} (with n samples), the probability that it will never be picked up in n times sampling is $(1 - \frac{1}{n})^n$. Since $\lim_{n\to\infty}(1 - \frac{1}{n})^n = \frac{1}{e} \approx 0.368$, when the sample size tends to infinity, approximately 36.8% of the samples in D_{orig} will not be added to D_1 (with n samples). □

This property ensures diversity in the resampled datasets, which is crucial for the effectiveness of ensemble learning approaches. In the context of federated causal discovery with limited local samples, Bootstrapping offers a powerful tool to artificially increase the diversity of data available to each client, potentially improving the robustness and statistical reliability of the causal discovery process by generating multiple resampled datasets from the original data distribution.

To address the inherent challenges of limited local sample sizes in federated settings, our proposed method incorporates Bootstrapping to generate multiple statistical replicates of the local datasets. This statistical augmentation enables more reliable causal discovery even when individual clients possess datasets that may be insufficient for traditional causal discovery methods.

4 The Proposed `FedECD` Method

The `FedECD` method is designed to address the challenge of causal discovery in federated settings with limited local samples. Our approach leverages Bootstrapping techniques [24] to enhance the robustness and accuracy of causal discovery across distributed clients. The `FedECD` method consists of two main phases (Fig. 2): 1) *Federated Causal Skeleton Optimization Using Bootstrapping*; and 2) *Federated Causal Structure Refinement Leveraging Bootstrapping*. In both phases, we employ a two-layer (***client-layer*** and ***server-layer***) aggregation strategy: the ***client-layer*** aggregates results from Bootstrapped sub-datasets within each client, while the ***server-layer*** aggregates results across all clients at the server. This approach allows us to effectively utilize limited local samples while preserving data privacy in the federated setting.

4.1 Federated Causal Skeleton Optimization Using Bootstrapping

The first phase of `FedECD` focuses on learning an optimized causal skeleton through a federated process that incorporates Bootstrapping. We use $\mathcal{S}$ to denote the current global skeleton, and $\mathcal{S}(i_1, i_2) = \mathcal{S}(i_2, i_1) = 1$ $(i_1, i_2 \in \{1, 2, \ldots, d\})$ to represent that there is an undirected edge between X_{i_1} and X_{i_2} in $\mathcal{S}$. We start with a fully connected undirected graph (skeleton) and iteratively remove false edges based on CI tests [28] performed on the local dataset of each client and the resampled sub-datasets generated by using Bootstrapping. The skeleton optimization process proceeds as follows.

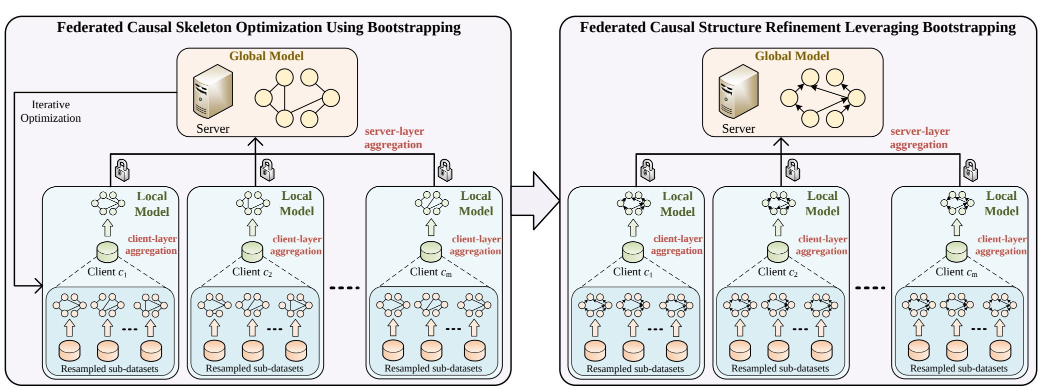

Figure 2: The framework of `FedECD`, which consists of two phases.

1. Initialize the conditioning set size $|\boldsymbol{S}| = 0$ for any CI tests between X_{i_1} and X_{i_2} conditioning on a subset $\boldsymbol{S} \subseteq \mathcal{X} \setminus \{X_{i_1}, X_{i_2}\}$ ($i_1, i_2 \in \{1, 2, \ldots, d\}$).
2. For each client $c_k \in \mathcal{C}$, perform CI tests on its local dataset $\mathcal{D}^{c_k}$:

$$\mathcal{S}^{c_k}(i_1, i_2) = \mathcal{S}^{c_k}(i_2, i_1) = \begin{cases} 0, & \text{if } X_{i_1} \perp\!\!\!\perp X_{i_2} | \boldsymbol{S} \\ 1, & \text{otherwise.} \end{cases} \tag{3}$$

The server aggregates skeletons from all clients:

$$\mathcal{S}(i_1, i_2) = \frac{1}{m} \sum_{k=1}^{m} \mathcal{S}^{c_k}(i_1, i_2), \quad \mathcal{S}(i_1, i_2) = \begin{cases} 1, & \text{if } \mathcal{S}(i_1, i_2) \geq 0.5 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Determine the set of edges that may be removed: $\mathcal{E}_{\text{del}} = \{(i_1, i_2) | \mathcal{S}(i_1, i_2) = 0\}$.
3. For each client $c_k \in \mathcal{C}$: a) Generate B sub-datasets $\{\mathcal{D}_1^{c_k}, \mathcal{D}_2^{c_k}, ..., \mathcal{D}_B^{c_k}\}$ using Bootstrapping. b) For each sub-dataset $\mathcal{D}_j^{c_k}$, compute the weighted skeleton $\hat{\mathcal{S}}_j^{c_k}$:

$$\hat{\mathcal{S}}_j^{c_k}(i_1, i_2) = \begin{cases} w(\rho), & \text{if } (i_1, i_2) \in \mathcal{E}_{\text{del}} \wedge X_{i_1} \perp\!\!\!\perp X_{i_2} | \boldsymbol{S} \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $w(\rho)$ is a function mapping p-values ρ returned by CI tests to the $[0, 1]$ interval. c) Aggregate weighted skeletons within the client (***client-layer aggregation***):

$$\hat{\mathcal{S}}^{c_k}(i_1, i_2) = \frac{1}{B} \sum_{j=1}^{B} \hat{\mathcal{S}}_j^{c_k}(i_1, i_2). \tag{6}$$

4. The server aggregates weighted skeletons from all clients (***server-layer aggregation***) and updates the global skeleton:

$$\hat{\mathcal{S}}(i_1, i_2) = \frac{1}{m} \sum_{k=1}^{m} \hat{\mathcal{S}}^{c_k}(i_1, i_2), \quad \mathcal{S}(i_1, i_2) = \begin{cases} 0, & \text{if } (i_1, i_2) \in \mathcal{E}_{\text{del}} \wedge \mathcal{S}(i_1, i_2) > \tau \\ 1, & \text{if } (i_1, i_2) \in \mathcal{E}_{\text{del}} \wedge \mathcal{S}(i_1, i_2) \leq \tau, \end{cases} \tag{7}$$

where $\tau = 0.05$ is a predefined threshold.
5. If any edges are removed, the conditioning set size $|\boldsymbol{S}| = |\boldsymbol{S}| + 1$ and return to step 2. Otherwise, terminate the algorithm, yielding the optimal global causal skeleton $\mathcal{S}^*$.

This process optimizes the global causal skeleton by incrementally increasing the conditioning set size and leveraging Bootstrapping techniques, all while preserving data privacy in the FL setting. The use of weighted skeletons and aggregation at both ***client-layer*** and ***server-layer*** allows for a more robust estimation of the true causal skeleton, particularly in scenarios with limited local samples.

4.2 Federated Causal Structure Refinement Leveraging Bootstrapping

In the second phase, `FedECD` refines the causal structure by determining the direction of the edges using a Bootstrapping-based approach in the federated setting. This phase builds upon the optimized skeleton $\mathcal{S}^*$ from the first phase and aims to orient the edges to form a causal DAG [28]. The refinement process proceeds as follows.

1. For each $c_k \in \mathcal{C}$: a) Generate B Bootstrapped sub-datasets $\{\mathcal{D}_1^{c_k}, \mathcal{D}_2^{c_k}, ..., \mathcal{D}_B^{c_k}\}$ from $\mathcal{D}^{c_k}$. b) For each sub-dataset $\mathcal{D}_j^{c_k}$, learn a DAG $\mathcal{G}_j^{c_k}$ using a score-based method [35] constrained by $\mathcal{S}^*$:

$$\mathcal{G}_j^{c_k} = \arg\max_{\mathcal{G} \in \mathcal{G}(\mathcal{S}^*)} Score(\mathcal{G}, \mathcal{D}_j^{c_k}), \tag{8}$$

where $\mathcal{G}(\mathcal{S}^*)$ is the set of all DAGs consistent with $\mathcal{S}^*$, and $Score(\cdot)$ is a scoring function (e.g., BDeu [36]). c) Aggregate the B DAGs into a ***client-layer*** DAG $\mathcal{G}^{c_k}$ using majority voting:

$$\mathcal{G}^{c_k} = \sum_{j=1}^{B} \mathcal{G}_j^{c_k}, \quad \begin{cases} \mathcal{G}^{c_k}(i_1, i_2) = 1 \wedge \mathcal{G}^{c_k}(i_2, i_1) = 0 & if\ \mathcal{G}^{c_k}(i_1, i_2) > \mathcal{G}^{c_k}(i_2, i_1) \\ \mathcal{G}^{c_k}(i_1, i_2) = \mathcal{G}^{c_k}(i_2, i_1) = 0 & if\ \mathcal{G}^{c_k}(i_1, i_2) = \mathcal{G}^{c_k}(i_2, i_1) = 0 \\ \mathcal{G}^{c_k}(i_1, i_2) = 0 \wedge \mathcal{G}^{c_k}(i_2, i_1) = 1 & \text{otherwise,} \end{cases} \tag{9}$$

where $i_1 = 1, 2, \ldots, d$ and $i_2 = 1, 2, \ldots, (i_1 - 1)$.

2. The server aggregates the ***client-layer*** DAGs into a ***server-layer*** DAG $\mathcal{G}^*$:

$$\mathcal{G}^* = \sum_{k=1}^{m} \mathcal{G}^{c_k}, \quad \begin{cases} \mathcal{G}^*(i_1, i_2) = 1 \wedge \mathcal{G}^*(i_2, i_1) = 0 & if\ \mathcal{G}^*(i_1, i_2) > \mathcal{G}^*(i_2, i_1) \\ \mathcal{G}^*(i_1, i_2) = \mathcal{G}^*(i_2, i_1) = 0 & if\ \mathcal{G}^*(i_1, i_2) = \mathcal{G}^*(i_2, i_1) = 0 \\ \mathcal{G}^*(i_1, i_2) = 0 \wedge \mathcal{G}^*(i_2, i_1) = 1 & \text{otherwise,} \end{cases} \tag{10}$$

where $i_1 = 1, 2, \ldots, d$ and $i_2 = 1, 2, \ldots, (i_1 - 1)$.

The resulting global DAG $\mathcal{G}^*$ represents the final causal structure learned by `FedECD`. This refined process leverages Bootstrapping to enhance the robustness of edge orientation in the federated setting, particularly when dealing with limited local samples. The use of majority voting at both ***client-layer*** and ***server-layer*** helps mitigate the impact of potential instabilities in individual DAG estimates.

5 Experimental Evaluation

5.1 Experiment Settings

Datasets. We use eight benchmark BN datasets [35], including *Child*, *Child3*, *Child5*, *Insurance*, *Insurance3*, *Insurance5*, *Alarm* and *Alarm3*[4], and each dataset contains 10,000 samples, allocated evenly across {5, 10, 15, 20, 25, 30} clients.

Metrics. We adopt the Ar_F1 (the higher the better) and TPR (the higher the better) metrics [37] to evaluate the learned causal DAGs in FL settings.

Baselines. `FedECD` is compared with seven state-of-the-art FCD methods, including F2SL-Best, F2SL-Voting, F2SL-Avg, PC-stable-Best, F2SL-stable-Voting, F2SL-stable-Avg and FedPC [17][5]. Among them, F2SL [38] and PC-stable [39] are traditional CD algorithms[6]. We have adapted them with specific strategies to function in an FL setting. The suffix "-Best" indicates that the algorithm is first run independently on each client to generate m DAGs, after which the DAG with the highest Ar_F1 score is selected as the final output. The "-Voting" variant applies a voting method [40] to the algorithm, while "-Avg" means that the algorithm is first run independently on each client to obtain m DAGs, and then the average Ar_F1 score across the m DAGs is calculated.

5.2 Results and Discussion

Figures 3-4 demonstrate `FedECD`'s superior performance across various scenarios, particularly when local sample sizes are limited:

- As the number of clients increases and local samples decrease, `FedECD` maintains robust performance while other algorithms decline sharply.
- For datasets like *Child*, *Child3*, and *Child5*, `FedECD`'s performance remains stable even with 20-30 clients, where each client has very limited samples.
- This consistent performance in challenging scenarios highlights `FedECD`'s effectiveness in addressing causal discovery with limited local samples in federated settings.

These observations highlight the efficacy of `FedECD`'s resampling technique and two-layer aggregation strategy in real-world FL scenarios where client-level data scarcity is common.

6 Conclusions and Future Work

In this paper, we proposed `FedECD`, a novel FCD approach that effectively addresses the challenge of limited local samples in FL scenarios. The core innovation of `FedECD` lies in its two-phase structure, each employing a two-layer (***client-layer*** and ***server-layer***) aggregation strategy: (1) *Federated Causal Skeleton Optimization* and (2) *Federated Causal Structure Refinement*. In both

[4]These benchmark BNs are publicly available at `http://www.bnlearn.com/bnrepository/`.
[5]The code is available at `https://github.com/Xianjie-Guo/FedPC`.
[6]The source codes of PC-stable and F2SL are available at `https://github.com/kuiy/CausalLearner`.

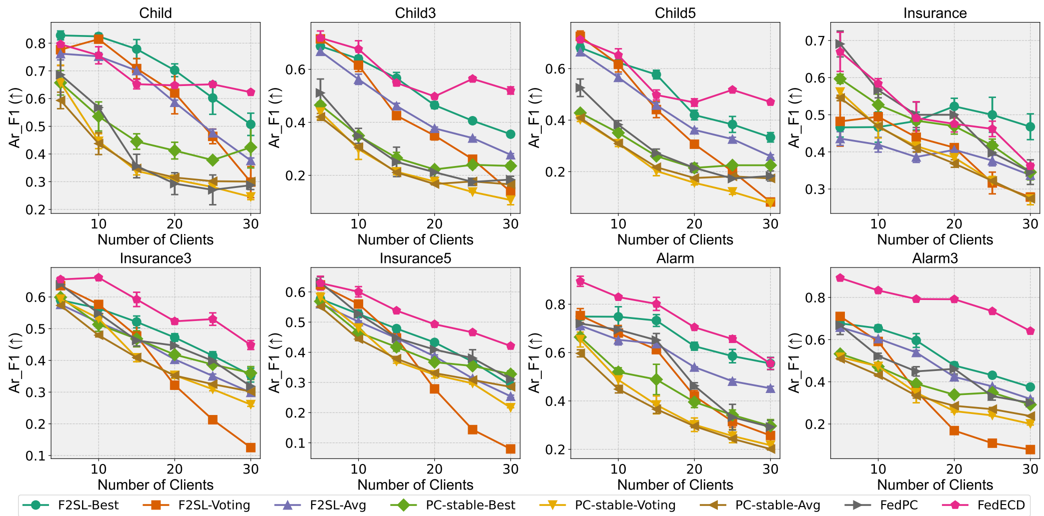

Figure 3: Experimental results on benchmark BN datasets. (Ar_F1 metric).

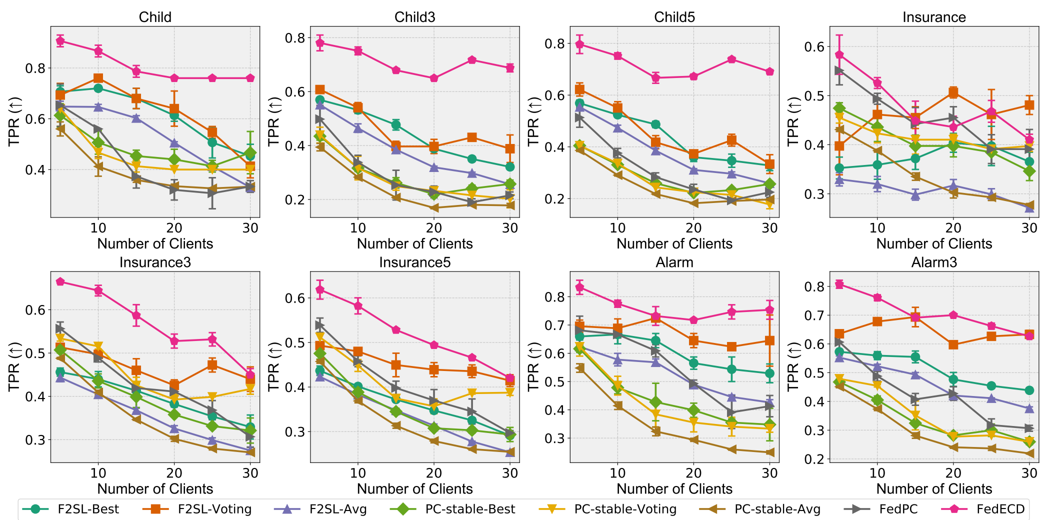

Figure 4: Experimental results on benchmark BN datasets. (TPR metric).

phases, `FedECD` leverages resampling techniques at the ***client-layer*** to mitigate the impact of limited samples, while the ***server-layer*** aggregates results across all clients. `FedECD` significantly enhances the robustness and accuracy of causal discovery in federated settings with limited local samples, while preserving privacy. Experiments on benchmark Bayesian network datasets demonstrate `FedECD`'s superior performance over existing FCD methods.

Future work will explore the application of `FedECD` to scenarios with hidden variables and mixed observational-interventional data.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Science and Technology Major Project (under grant 2020AAA0106100); the National Natural Science Foundation of China (under grant 62376087); and the China Scholarship Council (under grant 202306690023).

References

[1] Liu, C., K. Kuang. Causal structure learning for latent intervened non-stationary data. In *International Conference on Machine Learning*, pages 21756–21777. PMLR, 2023.

[2] Cai, R., Z. Huang, W. Chen, et al. Causal discovery with latent confounders based on higher-order cumulants. In *International Conference on Machine Learning*, pages 3380–3407. PMLR, 2023.

[3] Guo, X., K. Yu, L. Liu, et al. Causal feature selection with dual correction. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):938–951, 2022.

[4] —. Progressive skeleton learning for effective local-to-global causal structure learning. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[5] Guo, X., Y. Wang, X. Huang, et al. Bootstrap-based causal structure learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 656–665. 2022.

[6] Linardatos, P., V. Papastefanopoulos, S. Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[7] Moraffah, R., M. Karami, R. Guo, et al. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33, 2020.

[8] Yu, H., C. Miao, C. Leung, et al. Mitigating herding in hierarchical crowdsourcing networks. *Scientific Reports*, 6(1):4, 2016.

[9] Vuković, M., S. Thalmann. Causal discovery in manufacturing: A structured literature review. *Journal of Manufacturing and Materials Processing*, 6(1):10, 2022.

[10] Richens, J. G., C. M. Lee, S. Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, 11(1):3923, 2020.

[11] Li, X., H. Xiong, X. Li, et al. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64(12):3197–3234, 2022.

[12] Lecca, P. Machine learning for causal inference in biological networks: perspectives of this challenge. *Frontiers in Bioinformatics*, 1:746712, 2021.

[13] Kemp, K. Concealed data practices and competition law: why privacy matters. *European Competition Journal*, 16(2-3):628–672, 2020.

[14] Li, Q., Z. Wen, Z. Wu, et al. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3347–3366, 2021.

[15] Ng, I., K. Zhang. Towards federated Bayesian network structure learning with continuous optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 8095–8111. PMLR, 2022.

[16] Gao, E., J. Chen, L. Shen, et al. FedDAG: Federated DAG structure learning. *Transactions on Machine Learning Research*, 2023.

[17] Huang, J., X. Guo, K. Yu, et al. Towards privacy-aware causal structure learning in federated setting. *IEEE Transactions on Big Data*, 9(6):1525–1535, 2023.

[18] Li, L., I. Ng, G. Luo, et al. Federated causal discovery from heterogeneous data. In *International Conference on Learning Representations*. 2024.

[19] Guo, X., K. Yu, L. Liu, et al. FedCSL: A scalable and accurate approach to federated causal structure learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12235–12243. 2024.

[20] Yang, Q., Y. Liu, T. Chen, et al. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2):1–19, 2019.

[21] Kairouz, P., H. B. McMahan, B. Avent, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[22] Dwivedi, A. K., I. Mallawaarachchi, L. A. Alvarado. Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. *Statistics in medicine*, 36(14):2187–2205, 2017.

[23] Amalnerkar, E., T. H. Lee, W. Lim. Reliability analysis using bootstrap information criterion for small sample size response functions. *Structural and Multidisciplinary Optimization*, 62:2901–2913, 2020.

[24] Hesterberg, T. Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(6):497–526, 2011.

[25] Zheng, X., B. Aragam, P. K. Ravikumar, et al. DAGs with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[26] Boyd, S., N. Parikh, E. Chu, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[27] Yang, D., X. He, J. Wang, et al. Federated causality learning with explainable adaptive optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16308–16315. 2024.

[28] Spirtes, P., C. N. Glymour, R. Scheines, et al. *Causation, prediction, and search*. MIT Press, 2000.

[29] Wang, Z., P. Ma, S. Wang. Towards practical federated causal structure learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, vol. 14170, pages 351–367. Springer, 2023.

[30] Ye, Q., A. A. Amini, Q. Zhou. Federated learning of generalized linear causal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[31] Mian, O., D. Kaltenpoth, M. Kamp, et al. Nothing but regrets—privacy-preserving federated causal discovery. In *International Conference on Artificial Intelligence and Statistics*, pages 8263–8278. PMLR, 2023.

[32] Abyaneh, A., N. Scherrer, P. Schwab, et al. FED-CD: Federated causal discovery from interventional and observational data. *arXiv preprint arXiv:2211.03846*, 2022.

[33] Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[34] Jiang-She, Z. A survey of selective ensemble learning algorithms. *Chinese Journal of Computers*, 34(8):1399–1410, 2011.

[35] Tsamardinos, I., L. E. Brown, C. F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 2006.

[36] Suzuki, J. A theoretical analysis of the BDeu scores in Bayesian network structure learning. *Behaviormetrika*, 44:97–116, 2017.

[37] Guo, X., K. Yu, L. Liu, et al. Adaptive skeleton construction for accurate DAG learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10526–10539, 2023.

[38] Yu, K., Z. Ling, L. Liu, et al. Feature selection for efficient local-to-global Bayesian network structure learning. *ACM Transactions on Knowledge Discovery from Data*, 18(2):1–27, 2023.

[39] Colombo, D., M. H. Maathuis, et al. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

[40] Na, Y., J. Yang. Distributed Bayesian network structure learning. In *2010 IEEE International Symposium on Industrial Electronics*, pages 1607–1611. IEEE, 2010.