Sample Quality Heterogeneity-aware Federated Causal Discovery through Adaptive Variable Space Selection

Xianjie Guo[1], **Kui Yu**[1*], **Hao Wang**[1], **Lizhen Cui**[2], **Han Yu**[3] and **Xiaoxiao Li**[4]
[1]School of Computer Science and Information Engineering, Hefei University of Technology, China
[2]School of Software, Shandong University, China
[3]College of Computing and Data Science, Nanyang Technological University, Singapore
[4]Department of Electrical and Computer Engineering, The University of British Columbia, Canada
xianjieguo@mail.hfut.edu.cn, {yukui, jsjxwangh}@hfut.edu.cn, clz@sdu.edu.cn, han.yu@ntu.edu.sg, xiaoxiao.li@ece.ubc.ca

Abstract

Federated causal discovery (FCD) aims to uncover causal relationships among variables from decentralized data across multiple clients, while preserving data privacy. In practice, the sample quality of each client's local data may vary across different variable spaces, referred to as *sample quality heterogeneity*. Thus, data from different clients might be suitable for learning different causal relationships among variables. Model aggregated under existing FCD methods requires the entire model parameters from each client, thereby being unable to handle the *sample quality heterogeneity* issue. In this paper, we propose the Federated Adaptive Causal Discovery (`FedACD`) method to bridge this gap. During federated model aggregation, it adaptively selects the causal relationships learned under the "good" variable space (i.e., one with high-quality samples) from each client, while masking those learned under the "bad" variable space (i.e., one with low-quality samples). This way, each client only needs to send the optimal learning results to the server, achieving accurate FCD. Extensive experiments on various types of datasets demonstrate significant advantages of `FedACD` over existing methods. The source code is available at https://github.com/Xianjie-Guo/FedACD.

1 Introduction

Causal discovery (CD) aims to identify causal relationships among variables using observational data [He *et al.*, 2021; Guo *et al.*, 2022], with applications across diverse fields, including medicine [Anderson *et al.*, 2023], trustworthy artificial intelligence [Huo *et al.*, 2023a; Yu *et al.*, 2011; Huo *et al.*, 2023b] and computer science [Pearl, 2018; Zhang *et al.*, 2023b]. For example, studying the causal relationships between different lifestyles (e.g., diet and exercise) and chronic diseases (e.g., cardiovascular diseases and diabetes) can help guide public health policies and recommendations. CD can be divided into two main categories: 1) combinatorial optimization-based methods, and 2) continuous optimization-based methods [Yu *et al.*, 2020; Vowels *et al.*, 2022]. The former heuristically evaluates the goodness-of-fit between structural combinations of variables and the dataset used to learn optimal causal relationships. The latter employs gradient descent to optimize a weight adjacency matrix to fit causal relationships among variables.

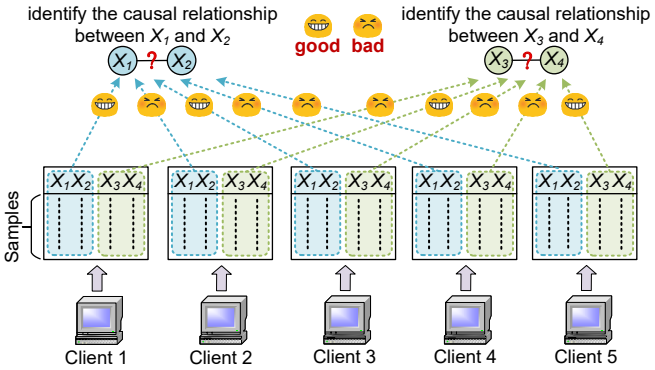

Figure 1: Examples of FL client sample quality heterogeneity.

Existing CD methods are always performed at a centralized site where all data are stored [Sheng *et al.*, 2024; Xiang *et al.*, 2023]. In practice, data are often distributed across multiple parties (e.g., patient data across hospitals [McMahan *et al.*, 2017; Liu *et al.*, 2024a; Zhang *et al.*, 2023a]), and cannot be moved to a central location due to data privacy concerns [Yang *et al.*, 2019; Liu *et al.*, 2022; Zhong *et al.*, 2023; Miao *et al.*, 2024; Liu *et al.*, 2024b]. As a result, each data owner holds a relatively small amount of samples, which severely limits the performance of existing CD methods. To address this issue, federated causal discovery (FCD) [Ng and Zhang, 2022] has emerged to uncover the underlying causal relationships among variables from decentralized data in a privacy-preserving manner. Existing FCD methods train models on the entire variable space of each client's local data. The entire model parameters from each client are involved in federated learning (FL) model aggregation at the FL server. However, for each FL client, the sample quality varies across different variable spaces. The *sample quality heterogeneity* issue results in data from different clients being suitable for

*Corresponding Author

learning different causal relationships among variables.

For example (Figure 1), when identifying the causal relationship between X_1 and X_2, only samples from clients 1 and 3 in the variable space of X_1 and X_2 exhibit "good" quality (i.e., can correctly determine the causal relationship between X_1 and X_2). In contrast, samples from other clients in the variable space of X_1 and X_2 are of "bad" quality (i.e., cannot correctly learn the causal relationship between X_1 and X_2). Similarly, when determining the causal relationship between variables X_3 and X_4, only samples from clients 2 and 5 in the variable space of X_3 and X_4 are of "good" quality, while samples from other clients in the variable space of X_3 and X_4 are of "bad" quality.

Existing FCD methods cannot deal with the *sample quality heterogeneity* issue since they train models on the entire variable space of each local data. To bridge this gap, we propose a first-of-its-kind Federated Adaptive Causal Discovery (`FedACD`) method. At the core of `FedACD` is a novel strategy to adaptively select the causal relationships learned under the "good" variable space (i.e., the space with high-quality samples) at each FL client, and only send them to the server. The causal relationships learned under the "bad" variable space (i.e., the space with low-quality samples) are masked. This way, each FL client only communicates the optimal learning results with the FL server during each round of training, thereby achieving accurate FCD. Under reasonable assumptions, theoretical analysis proves the superiority of `FedACD` over existing methods. Extensive experiments on benchmark Bayesian network data, synthetic Non-IID data, and real-world data demonstrate significant advantages of `FedACD` against five state-of-the-art methods, improving the F1 score by 19.65% and reducing the Structural Hamming Distance by 23.57% compared to the best baseline on average.

2 Related Work

The Fusion Problem in Causal Discovery. Although FCD and the fusion problem in causal discovery [Mooij *et al.*, 2020; Perry *et al.*, 2022; Huang *et al.*, 2020] share similarities in integrating causal relationships from multiple datasets, they have distinct characteristics and challenges. FCD focuses on learning causal relationships from decentralized data while preserving privacy and reducing communication costs, which are not primary concerns in the fusion problem. Additionally, in the fusion problem, the causal relationships between variables in the ground truth causal graph are typically required to satisfy specific assumptions, and not any arbitrary directed acyclic graph structure is applicable.

Federated Causal Discovery. Notable FCD methods include NOTEARS-ADMM [Ng and Zhang, 2022], FED-CD [Abyaneh *et al.*, 2022], DARLS [Ye *et al.*, 2022], PERI [Mian *et al.*, 2023], FedDAG [Gao *et al.*, 2023], FedPC [Huang *et al.*, 2023a] and FedCSL [Guo *et al.*, 2024]. Specifically, NOTEARS-ADMM directly applies the distributed optimization algorithm ADMM [Boyd *et al.*, 2011] to optimize the NOTEARS method [Zheng *et al.*, 2018]. FED-CD is designed for a federated learning scenario that contains both observational and interventional data. As for DARLS and PERI, the former utilizes the distributed annealing strategy [Arshad and Silaghi, 2004] to search for the optimal causal graph, while the latter aggregates the results of the local greedy equivalent search [Chickering, 2002] and chooses the worst-case regret for each iteration. FedDAG adopts a two-level structure for each local model, where the first level learns causal relationships by communicating with the server, and the second level approximates variable relationships at each client for handling data heterogeneity. FedPC proposed a layer-wise aggregation strategy to adapt PC [Spirtes *et al.*, 2000] into FL settings. To address the scalability and accuracy limitations of existing methods, FedCSL designs a federated local-to-global learning strategy and a highly privacy-preserving weighted aggregation scheme, respectively.

However, these existing studies are not designed to deal with the FL client *sample quality heterogeneity* issue, which limits their applicability in practice. To the best of our knowledge, `FedACD` is the first FCD method to bridge this gap.

3 Preliminaries

In this work, we consider a horizontal FL setting, where different clients have large overlaps in the variable space but little overlap in the sample space. Let $\mathcal{C} = \{c_k\}_{k\in\{1,2,...,m\}} = \{c_1, c_2, ..., c_m\}$ be a set of m clients, $\mathcal{X} = \{X_i\}_{i\in\{1,2,...,d\}} = \{X_1, X_2, ..., X_d\}$ be a set of d variables at each client, and $\mathcal{D}^{c_k} \in \mathbb{R}^{n_{c_k}\times d}$ represent the local dataset owned by client c_k. Here, n_{c_k} is the number of samples in $\mathcal{D}^{c_k}$. Causal relationships over $\mathcal{X}$ are often represented by a causal directed acyclic graph (DAG) [Guo *et al.*, 2023]. 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}.

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. The causal relationship between variables at different clients satisfies the following Assumption 1.

Assumption 1 (Invariant Causal DAG [Gao *et al.*, 2023])**.** *All local datasets are uniformly sampled from the same causal DAG $\mathcal{G}$, and the probability distribution of samples for the same variable space can differ across different clients.*

In this paper, if there is $X_{i_1} \to X_{i_2}$ or $X_{i_2} \to X_{i_1}$, we say that X_{i_1} and X_{i_2} are causal neighbors to each other. We use $CN_i^{c_k}$ to represent the causal neighbor of X_i learned at client c_k. Under the faithfulness and causal sufficiency assumptions [Pearl, 1988], if $X_{i_1} \in CN_{i_2}^{c_k}$ or $X_{i_2} \in CN_{i_1}^{c_k}$ hold at client c_k, $X_{i_1} \not\perp\!\!\!\perp X_{i_2}|\boldsymbol{Z}$ always holds, where $\boldsymbol{Z} \subseteq \mathcal{X} \setminus \{X_{i_1}, X_{i_2}\}$. We use $\not\perp\!\!\!\perp$ (or $\perp\!\!\!\perp$) to represent the dependence (or independence) relation. We employ the G^2 test [Spirtes *et al.*, 2000] for discrete data and Fisher's Z test [Pena, 2008] for continuous data to conduct conditional independence (CI) tests for identifying causal relationships between variables. Assume that ρ is the p-value returned by CI tests, and α is a given significance level. Under the null hypothesis of "$H_0 : X_{i_1} \perp\!\!\!\perp X_{i_2}|\boldsymbol{Z}$", for a CI test of X_{i_1} and X_{i_2} given $\boldsymbol{Z}$, $X_{i_1} \perp\!\!\!\perp X_{i_2}|\boldsymbol{Z}$ holds if and only if $\rho > \alpha$.

When conducting CI tests for two variables under all subsets consisting of their respective causal neighbors, if there exists a CI test that makes $\rho \in (\alpha, 1]$ hold, then the CI test with a larger p-value is more reliable [Ramsey, 2016]. Thus, we make the following assumption.

Assumption 2. *Given a local dataset $\mathcal{D}^{c_k}$, if for $\forall \mathbf{Z}_1$ ($\subseteq CN_{i_1}^{c_k}$ or $\subseteq CN_{i_2}^{c_k}$), there exists CI tests that makes $X_{i_1} \perp\!\!\!\perp X_{i_2} | \mathbf{Z}$ hold, with the maximum p-value across all CI tests denoted as $\hat{p}(X_{i_1}, X_{i_2})$, and similarly if for $\forall \mathbf{Z}_2$ ($\subseteq CN_{i_3}^{c_k}$ or $\subseteq CN_{i_4}^{c_k}$), there exists CI tests that makes $X_{i_3} \perp\!\!\!\perp X_{i_4} | \mathbf{Z}$ hold, with the maximum p-value denoted as $\hat{p}(X_{i_3}, X_{i_4})$, then the test result between X_{i_1} and X_{i_2} is more reliable than that between X_{i_3} and X_{i_4} if $\hat{p}(X_{i_1}, X_{i_2}) > \hat{p}(X_{i_3}, X_{i_4})$.*

4 The Proposed FedACD Method

As shown in Figure 2, the proposed FedACD method consists of two phases: 1) *federated causal skeleton learning* (FCSL, details in Section 4.1) and 2) *federated skeleton orientation* (FSO, details in Section 4.2). Due to space limit, the detailed pseudo-code of FedACD is provided in Appendix B, and its time complexity is analyzed in Appendix C.

4.1 Federated Causal Skeleton Learning (FCSL)

As shown in Figure 2, Phase 1 is an iterative process. Specifically, FedACD first constructs a complete undirected graph over $\mathcal{X} = \{X_i\}_{i \in \{1,2,...,d\}}$ and sends it to each client. Then, at each client, it utilizes the conditional independence (CI) tests with the size of the conditioning set as l (start with 0) to remove the false causal edges. Next, it adaptively masks causal edges that may be deleted by mistake and sends all masked causal skeletons to the server for aggregation. Finally, it determines whether the aggregated causal skeleton meets the convergence condition: if so, Phase 1 ends; otherwise, it increments the size of l for the next iteration.

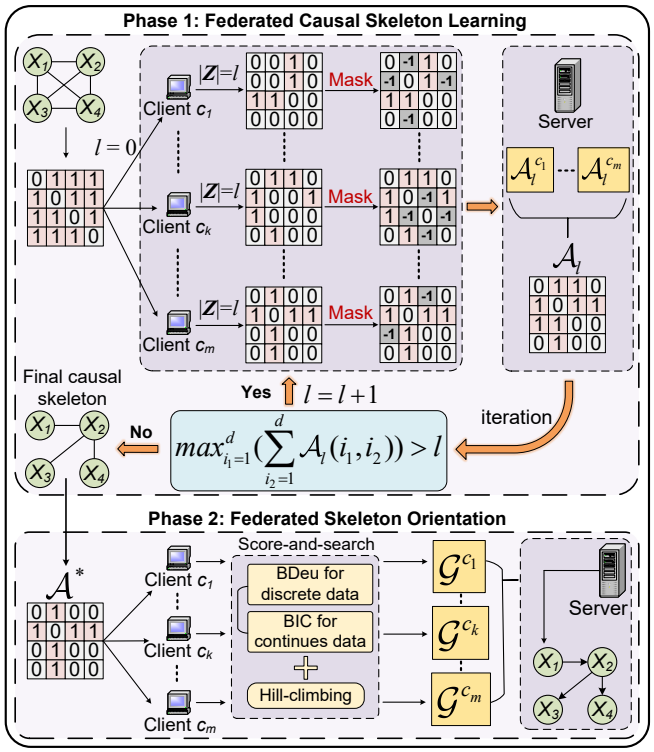

Figure 2: An overview of the proposed FedACD method.

Phase 1-1: Learning Causal Skeleton.
Since the sample space at different clients almost does not overlap in the horizontal FL scenario, the causal skeletons learned at each client may differ greatly, resulting in poor aggregation results. Therefore, we aggregate and update the causal skeletons learned at each client by limiting the size of the condition set for CI tests.

At client c_k ($k \in \{1, 2, ..., m\}$), for two variables X_{i_1} and X_{i_2} ($X_{i_1}, X_{i_2} \in \mathcal{X}$) that are causal neighbors, FedACD utilizes CI tests to determine whether they are conditionally independent given the conditioning set $\mathbf{Z}$ ($\mathbf{Z} \subseteq CN_{i_1}^{c_k}$ or $\mathbf{Z} \subseteq CN_{i_2}^{c_k}$) with $|\mathbf{Z}| = l$. If $X_{i_1} \perp\!\!\!\perp X_{i_2} | \mathbf{Z}$ holds, the causal edge between X_{i_1} and X_{i_2} is removed. By performing the above operations at each client, we obtain m causal skeletons and their corresponding adjacency matrices $\{A_l^{c_k}\}_{k \in \{1,2,...,m\}}$. $A_l^{c_k}(i_1, i_2) = 1$ indicates that on the local dataset $\mathcal{D}^{c_k}$, X_{i_1} and X_{i_2} are conditionally dependent given any $\mathbf{Z}$ ($\mathbf{Z} \subseteq CN_{i_1}^{c_k}$ or $\mathbf{Z} \subseteq CN_{i_2}^{c_k}$) with $|\mathbf{Z}| = l$. On the contrary, $A_l^{c_k}(i_1, i_2) = 0$ means that there exists a $\mathbf{Z}$ ($\mathbf{Z} \subseteq CN_{i_1}^{c_k}$ or $\mathbf{Z} \subseteq CN_{i_2}^{c_k}$) with $|\mathbf{Z}| = l$ that makes X_{i_1} and X_{i_2} conditionally independent. Since the causal skeleton is an undirected graph, $A_l^{c_k}(i_1, i_2) = A_l^{c_k}(i_2, i_1)$ holds for $\forall i_1, i_2, k, l$.

Phase 1-2: Adaptively Masking the Causal Skeleton.
As illustrated in Figure 1, the *sample quality heterogeneity* issue implies that the local datasets $\{\mathcal{D}^{c_k}\}_{k \in \{1,2,...,m\}}$ from different clients might be suitable for learning different causal relationships among variables. Therefore, we design an adaptive strategy to mask the causal edges that may be erroneously removed from $\{A_l^{c_k}\}_{k \in \{1,2,...,m\}}$, which has been learned in Phase 1-1 with "-1", rendering them ineffective in subsequent aggregation phase. The detailed process is as follows.

Let $r \in [0, 1]$ denote the masking rate for each adjacency matrix $A_l^{c_k}$, t^{c_k} represent the number of causal edges removed on the adjacency matrix $A_l^{c_k}$ in Phase 1-1, and $\mathcal{E}_j^{c_k}$ ($j \in \{1, 2, ..., t^{c_k}\}$) denote the j-th causal edge removed on $A_l^{c_k}$. Further, $\mathcal{E}_j^{c_k}(a)$ and $\mathcal{E}_j^{c_k}(b)$ represent the two variables linked by this edge. In Phase 1-1, there might be multiple conditioning sets that can make two variables $\mathcal{E}_j^{c_k}(a)$ and $\mathcal{E}_j^{c_k}(b)$ conditionally independent. However, the CI test with a larger p-value is considered more reliable [Ramsey, 2016]. Therefore, we only record the test results with the maximum p-value among all CI tests that indicate the conditional independence of two variables, and denote the maximum p-value as $\hat{p}(\mathcal{E}_j^{c_k}(a), \mathcal{E}_j^{c_k}(b))$. According to Assumption 2, the smaller $\hat{p}(\mathcal{E}_j^{c_k}(a), \mathcal{E}_j^{c_k}(b))$, the more likely outcome is that $\mathcal{E}_j^{c_k}$ is a mistakenly deleted causal edge. Conversely, a larger value of $\hat{p}(\mathcal{E}_j^{c_k}(a), \mathcal{E}_j^{c_k}(b))$ implies that there is no causal connection between $\mathcal{E}_j^{c_k}(a)$ and $\mathcal{E}_j^{c_k}(b)$ in the ground truth. Thus, we need to mask the causal edges that have been removed from $A_l^{c_k}$ learned in Phase 1-1 with low $\hat{p}(\cdot, \cdot)$. Let the indices of the edges that need to be masked in $A_l^{c_k}$ be stored in the vector Φ^{c_k}, we have:

$$\underset{k=1,2,...,m}{\Phi^{c_k}} = \mathring{Bottom}_{\lceil r * t^{c_k} \rceil}(\{\hat{p}(\mathcal{E}_j^{c_k}(a), \mathcal{E}_j^{c_k}(b))\}), \quad (1)$$

where $\mathring{Bottom}_{\lceil r * t^{c_k} \rceil}$ is used to obtain the indices corresponding to the bottom $\lceil r * t^{c_k} \rceil$ elements in a vector sorted in

descending order. We use $\mathcal{A}_l^{c_k}$ to denote the adjacency matrix after masking $A_l^{c_k}$:

$$\underset{\substack{k=1,2,\ldots,m\\ i_1,i_2=1,2,\ldots,d}}{\mathcal{A}_l^{c_k}(i_1,i_2)} = \begin{cases} \text{-1} & \begin{array}{l} if\ \exists \psi \in \Phi^{c_k}\ such\ that \\ \mathcal{E}_\psi^{c_k}(a) = i_1 \wedge \mathcal{E}_\psi^{c_k}(b) = i_2 \\ or\ \mathcal{E}_\psi^{c_k}(a) = i_2 \wedge \mathcal{E}_\psi^{c_k}(b) = i_1 \end{array} \\ A_l^{c_k}(i_1,i_2) & otherwise. \end{cases} \tag{2}$$

Phase 1-3: Aggregating All Masked Causal Skeletons. After obtaining all the masked matrices $\{\mathcal{A}_l^{c_k}\}_{k\in\{1,2,\ldots,m\}}$ in Phase 1-2, the FL server aggregates and updates these matrices. Since a value of "-1" in a cell of the masked matrices indicates that the corresponding learning result is unreliable, such masked causal relationships are invalid during aggregation. In other words, each element in $\mathcal{A}_l^{c_k}$ only participates in voting if its value is not "-1". Let the aggregated adjacency matrix be $\mathcal{A}_l$, we have:

$$\underset{i_1,i_2=1,2,\ldots,d}{\mathcal{A}_l(i_1,i_2)} = \begin{cases} 1 & if\ y \geq \frac{m-|x|}{2} \\ 0 & otherwise. \end{cases}, \tag{3}$$

where x represents the number of adjacency matrices in $\{\mathcal{A}_l^{c_k}\}_{k\in\{1,2,\ldots,m\}}$ where the causal relationship between X_{i_1} and X_{i_2} is masked, i.e.,

$$x = \sum_{k=1}^{m} \mathcal{A}_l^{c_k}(i_1,i_2)\ subject\ to\ \mathcal{A}_l^{c_k}(i_1,i_2) = -1, \tag{4}$$

and y denotes the number of adjacency matrices in $\{\mathcal{A}_l^{c_k}\}_{k\in\{1,2,\ldots,m\}}$ where there exists a causal edge between X_{i_1} and X_{i_2}, i.e.,

$$y = \sum_{k=1}^{m} \mathcal{A}_l^{c_k}(i_1,i_2)\ subject\ to\ \mathcal{A}_l^{c_k}(i_1,i_2) = 1. \tag{5}$$

Phase 1-4: Iterating Phases 1-1 to 1-3 until Convergence. For the newly aggregated causal skeleton $\mathcal{A}_l$, the server first determines whether it has converged (i.e., whether any erroneous causal edges have not been removed). Specifically, the server checks whether there are variables in $\mathcal{A}_l$ whose number of causal neighbors is greater than l. If such variables exist, it means that new CI tests, where the size of the conditioning set is greater than l, can be performed to remove possible erroneous causal edges further; otherwise, it means that $\mathcal{A}_l$ has converged and no more CI test is needed. The convergence condition is formalized as:

$$\max_{i_1=1}^{d} \left(\sum_{i_2=1}^{d} \mathcal{A}_l(i_1,i_2) \right) > l. \tag{6}$$

If Eq. (6) holds, l is incremented by 1, and the server sends $\mathcal{A}_l$ as the new initial causal skeleton to each client to repeat Phases 1-1 to 1-3; otherwise, the optimal causal skeleton is returned.

Analytical Evaluation of FCSL

Here, under reasonable assumptions, we theoretically show the superiority of `FedACD` over existing methods in Phase 1. Let $\acute{A}$ represent the adjacency matrix corresponding to the causal skeleton in the ground truth. According to the adaptive masking strategy in Eqs. (1) and (2), at client c_k, the probability of not being masked in $\{\mathcal{E}_j^{c_k}\}_{j\in\{1,2,\ldots,t^{c_k}\}}$ is $(1-r)$, i.e., $P(\mathcal{A}_l^{c_k}(\mathcal{E}_j^{c_k}(a),\mathcal{E}_j^{c_k}(b)) \neq -1)$. Thus, when $P(\acute{A}(\mathcal{E}_j^{c_k}(a),\mathcal{E}_j^{c_k}(b)) = 0) \geq \frac{(1-r)}{2}$ holds, in Phase 1-2, the proportion of correctly removed $\mathcal{E}_j^{c_k}$ in all unmasked $\mathcal{E}_j^{c_k}$ is 50% or more in $\mathcal{A}_l^{c_k}$. Therefore, we have Theorem 1.

Theorem 1. *If* $P(\acute{A}(\mathcal{E}_j^{c_k}(a),\mathcal{E}_j^{c_k}(b)) = 0) \geq \frac{(1-r)}{2}$, *then* $P(\acute{A}(\mathcal{E}_j^{c_k}(a),\mathcal{E}_j^{c_k}(b)) = 0 | \mathcal{A}_l^{c_k}(\mathcal{E}_j^{c_k}(a),\mathcal{E}_j^{c_k}(b)) \neq -1) \geq \frac{1}{2}$.

For existing voting-based FCD methods (e.g., FedPC), to ensure that each client has correctly removed causal edges in more than half of all removed causal edges in each round of iteration, $P(\acute{A}(\mathcal{E}_j^{c_k}(a),\mathcal{E}_j^{c_k}(b)) = 0) \geq \frac{1}{2}$ must be satisfied, whereas our method only needs to satisfy $P(\acute{A}(\mathcal{E}_j^{c_k}(a),\mathcal{E}_j^{c_k}(b)) = 0) \geq \frac{(1-r)}{2}$. According to the aggregation rule in Eq. (3), we have the following Theorem 2.

Theorem 2. *Given* $\forall X_{i_1}, X_{i_2} \in \mathcal{X}$, *in* $\{\mathcal{A}_l^{c_k}\}_{k\in\{1,2,\ldots,m\}}$ *containing* m *matrices, if* $P(\mathcal{A}_l^{c_k}(i_1,i_2) = 0 | \acute{A}(i_1,i_2) = 0) > \frac{1-P(\mathcal{A}_l^{c_k}(i_1,i_2)=-1|\acute{A}(i_1,i_2)=0)}{2}$ *holds, the false causal edge between* X_{i_1} *and* X_{i_2} *can be correctly removed in Phase 1-3; if* $P(\mathcal{A}_l^{c_k}(i_1,i_2) = 1 | \acute{A}(i_1,i_2) = 1) \geq \frac{1-P(\mathcal{A}_l^{c_k}(i_1,i_2)=-1|\acute{A}(i_1,i_2)=1)}{2}$ *holds, the true causal edge between* X_{i_1} *and* X_{i_2} *does not be discarded in Phase 1-3.*

Existing voting-based FCD methods (e.g., FedPC) require $P(\mathcal{A}_l^{c_k}(i_1,i_2) = 1 | \acute{A}(i_1,i_2) = 1) \geq \frac{1}{2}$ and $P(\mathcal{A}_l^{c_k}(i_1,i_2) = 0 | \acute{A}(i_1,i_2) = 0) > \frac{1}{2}$ to ensure that the true causal edge between X_{i_1} and X_{i_2} does not be discarded while the false causal edge between X_{i_1} and X_{i_2} can be correctly removed in each round of aggregation. In contrast, based on Theorem 2, `FedACD` only requires $P(\mathcal{A}_l^{c_k}(i_1,i_2) = 0 | \acute{A}(i_1,i_2) = 0) > \frac{1-P(\mathcal{A}_l^{c_k}(i_1,i_2)=-1|\acute{A}(i_1,i_2)=0)}{2}$ and $P(\mathcal{A}_l^{c_k}(i_1,i_2) = 1 | \acute{A}(i_1,i_2) = 1) \geq \frac{1-P(\mathcal{A}_l^{c_k}(i_1,i_2)=-1|\acute{A}(i_1,i_2)=1)}{2}$ to achieve optimal performance, which greatly improves the accuracy of the aggregated causal skeleton in practice. The proofs of Theorems 1 and 2 are given in Appendix A.

4.2 Federated Skeleton Orientation (FSO)

As shown in Figure 2, after obtaining the optimal causal skeleton, Phase 2 orients the undirected edges at each client for learning m causal DAGs, and then aggregates all causal DAGs at the server side to produce the final causal DAG.

Phase 2-1: Orienting Undirected Edges at Each Client. Let $\mathcal{A}^*$ represent the adjacency matrix corresponding to the optimal causal skeleton obtained in Phase 1. The server first sends $\mathcal{A}^*$ to each client, then the score-and-search strategy is adopted to greedily orient the undirected edges in $\mathcal{A}^*$ for obtaining a causal DAG with the highest score at each client. Let $\mathcal{G}^{c_k}$ denote the causal DAG learned at client c_k, and "$\mathcal{G}^{c_k}(i_1,i_2) = 1$" denotes that there is an edge from X_{i_1} to X_{i_2} in $\mathcal{G}^{c_k}$. For discrete datasets, we utilize a Bayesian score, BDeu [Scutari, 2016], and a search procedure, hill-climbing [Gámez *et al.*, 2011], to implement the above score-and-search process. Here, the BDeu score for the causal DAG

$\mathcal{G}^{c_k}$ learned on client dataset $\mathcal{D}^{c_k}$ is defined as:

$$\text{BDeu}(\mathcal{G}^{c_k}, \mathcal{D}^{c_k}) = \ln P(\mathcal{D}^{c_k}|\beta, \mathcal{G}^{c_k}) + \ln P(\beta, \mathcal{G}^{c_k}), \quad (7)$$

where $P(\mathcal{D}^{c_k}|\beta, \mathcal{G}^{c_k})$ denotes the probability of the local dataset $\mathcal{D}^{c_k}$ given the equivalent sample size parameter β and the causal DAG $\mathcal{G}^{c_k}$. It measures how well the causal DAG predicts the observed discrete data. $P(\beta, \mathcal{G}^{c_k})$ represents the prior probability of β. It serves as a regularization term and influences the strength of prior beliefs about the density of the causal DAG.

For continuous datasets, we use an information-theoretic score, BIC [Watanabe, 2013], to calculate the fitting score between $\mathcal{G}^{c_k}$ and $\mathcal{D}^{c_k}$. The BIC score for $\mathcal{G}^{c_k}$ learned on $\mathcal{D}^{c_k}$ is defined as:

$$\text{BIC}(\mathcal{G}^{c_k}, \mathcal{D}^{c_k}) = -2 \cdot \ln(\hat{L}(\mathcal{G}^{c_k}, \mathcal{D}^{c_k})) + \mu \cdot \ln(n_{c_k}), \quad (8)$$

where $\hat{L}(\mathcal{G}^{c_k}, \mathcal{D}^{c_k})$ denotes the ability of the causal DAG to explain the observed continuous dataset $\mathcal{D}^{c_k}$. A higher likelihood indicates a better fit. μ represents the average density of $\mathcal{G}^{c_k}$. The penalty term "$\mu \cdot \ln(n_{c_k})$" discourages overly dense causal DAGs, favoring sparser causal DAGs to reduce the risk of overfitting.

Phase 2-2: Aggregating All Causal DAGs at the Server. `FedACD` sends all causal DAGs $\{\mathcal{G}^{c_1}, \mathcal{G}^{c_2}, \cdots, \mathcal{G}^{c_m}\}$ learned in Phase 2-1 back to the server to compute the aggregated causal DAG $\mathcal{G}^*$ as:

$$\mathcal{G}^* = \mathcal{G}^{c_1} \oplus \mathcal{G}^{c_2} \oplus \cdots \oplus \mathcal{G}^{c_m}, \quad (9)$$

where $\oplus$ represents the element-wise addition of matrices. Finally, we compare the elements at corresponding positions on the diagonal of matrix $\mathcal{G}^*$ for obtaining the final causal DAG. Specifically, if $\mathcal{G}^*(i_1, i_2) > \mathcal{G}^*(i_2, i_1)$, then there exists a directed edge from X_{i_1} to X_{i_2}. If $\mathcal{G}^*(i_1, i_2) \leq \mathcal{G}^*(i_2, i_1)$ and $\mathcal{G}^*(i_2, i_1) \neq 0$, there exists a directed edge from X_{i_2} to X_{i_1}; Otherwise, there is no edge between X_{i_1} and X_{i_2}. To summarize, we have:

$$\begin{cases} \mathcal{G}^*(i_1, i_2) = 1 \wedge \mathcal{G}^*(i_2, i_1) = 0 \ \ if\ \mathcal{G}^*(i_1, i_2) > \mathcal{G}^*(i_2, i_1) \\ \mathcal{G}^*(i_1, i_2) = \mathcal{G}^*(i_2, i_1) = 0 \qquad if\ \mathcal{G}^*(i_1, i_2) = \mathcal{G}^*(i_2, i_1) = 0 \\ \mathcal{G}^*(i_1, i_2) = 0 \wedge \mathcal{G}^*(i_2, i_1) = 1 \ \ \text{otherwise}, \end{cases} \quad (10)$$

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

4.3 Privacy and Cost Analysis

Privacy Preservation Capability of `FedACD`

In Phases 1-1 and 1-2, all CI tests requiring raw data are performed by the respective FL clients, and the adaptive masking strategy is locally executed by each client based on the p-values obtained from the CI tests. In Phase 2-1, the score-and-search process, which requires raw data, is also performed locally by each client. As a result, `FedACD` only exchanges structural information represented by the adjacency matrices throughout the entire FL process, without exposing clients' raw data or any statistical information from the CI tests.

To further mitigate the risk of inferring local data from the learned structural information, `FedACD` can be combined with the following techniques. (1) Additive homomorphic encryption (Paillier's scheme [Paillier, 1999]) can be applied in the aggregation process of Phase 1-3 and Phase 2-2 to prevent leakage of the *sample quality heterogeneity* reflected in the masking information on each client. (2) To prevent the leakage of semantic information about each variable during communication between clients and the server, we incorporate an easily implementable privacy protection strategy [Huang *et al.*, 2023a] into `FedACD`. Specifically, the remote server instructs each client to assign unique identifiers (e.g., "1", "2", "3", etc.) to the variable semantics following their alphabetical order. If multiple variables share the same initial letter, they are further sorted by subsequent letters. Each client then sends only these assigned identifiers to the remote server for aggregation, protecting the variable semantics.

Communication Cost of `FedACD`

Communication is a critical bottleneck in FL. Therefore, developing communication-efficient methods for FL model training is essential. Here, we argue that `FedACD` introduces relatively low communication overhead. In Phase 1, during each iteration, all clients need to send a masked adjacency matrix $\mathcal{A}_l^{c_k}$ of size $d * d$ to the server. The server, in turn, sends back the aggregated adjacency matrix $\mathcal{A}_l$ to each client, requiring $O(m|\mathcal{A}_l^{c_k}| + m|\mathcal{A}_l|) = O(md^2 + md^2) = O(md^2)$ information cost. Assume that Phase 1 requires L (usually $L \leq 3$) iterations in total. Therefore, Phase 1 incurs a total communication cost of $O(md^2L)$. In Phase 2, firstly, the adjacency matrix $\mathcal{A}^*$ corresponding to the final causal skeleton needs to be sent from the server to each client. Then, the adjacency matrices $\mathcal{G}^{c_k}$ learned by each client are sent back to the server to obtain the final causal DAG through an aggregation strategy, resulting in a total information cost of $O(m|\mathcal{A}^*| + m|\mathcal{G}^{c_k}|) = O(md^2 + md^2) = O(md^2)$.

5 Experimental Evaluation

5.1 Experiment Setting

Datasets. We utilize the following three types of datasets.

- **Benchmark BN datasets.** We use three benchmark BN datasets: *Child* with 20 variables, *Insurance* with 27 variables and *Alarm* with 37 variables, and each dataset contains 5,000 samples [Tsamardinos *et al.*, 2006].
- **Synthetic Non-IID datasets.** We employ the publicly available code from [Gao *et al.*, 2023] to generate two batches of Non-IID datasets, distributed among a total of 5,000 samples across {3, 5, 10, 15, 20} clients, with each client's data originating from a different distribution. The ground truth for each batch of data across different clients is consistent, comprising 20 variables with 40 edges and 30 variables with 60 edges, respectively.
- **Real-world datasets.** We also compare the proposed method with the baselines on two non-overlapping networks of sizes {8, 20} from the lung cancer gene-expression dataset, REGED [Statnikov *et al.*, 2015]. For each network, we generate {3, 5, 10, 15, 20} distinct environments (i.e., FL clients) to create a horizontal FL scenario, with a total of 1,000 samples.

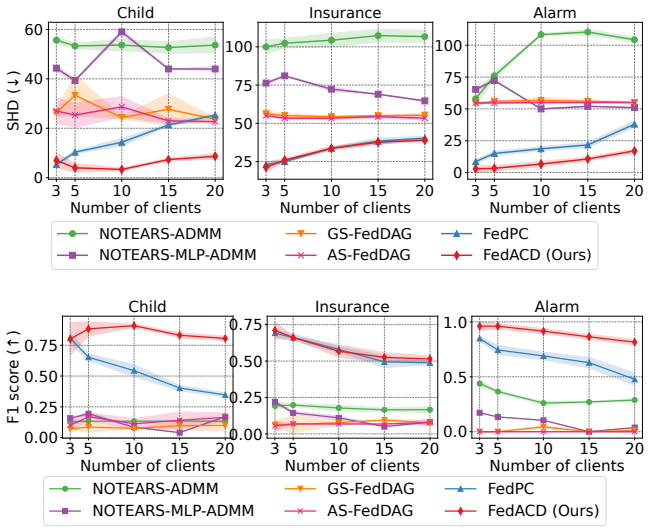

Figure 3: Experimental results on benchmark BN datasets. There are 5,000 samples in total, allocated evenly across {3, 5, 10, 15, 20} clients. We show the performance of all methods in two metrics (SHD and F1 score from top to bottom).

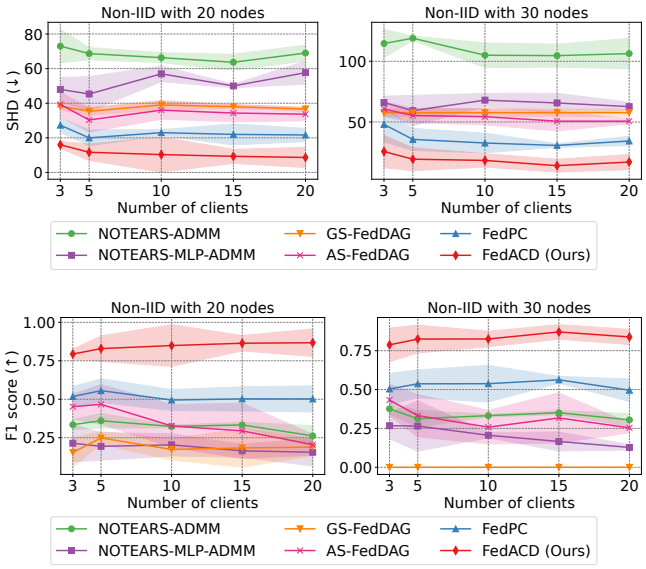

Figure 4: Experimental results on synthetic Non-IID datasets. There are 5,000 samples in total, allocated evenly across {3, 5, 10, 15, 20} clients. We show the performance of all methods in two metrics (SHD and F1 score from top to bottom).

Evaluation metrics. We adopt the Structural Hamming Distance (SHD) (the lower the better) and F1 score (the higher the better) [Huang *et al.*, 2023b] to evaluate the learned causal DAGs in FL settings. We have also adopted other metrics (details can be found in Appendix D).

Comparison methods. `FedACD` is compared with five state-of-the-art FCD methods: 1) NOTEARS-ADMM [Ng and Zhang, 2022], 2) NOTEARS-MLP-ADMM [Ng and Zhang, 2022], 3) GS-FedDAG [Gao *et al.*, 2023], 4) AS-FedDAG [Gao *et al.*, 2023] and 5) FedPC [Huang *et al.*, 2023a]. We adopt these baselines partly because of their publicly available implementations.

Implementation details of the `FedACD` algorithm and the baselines are provided in Appendix F.

5.2 Results on Benchmark Bayesian Network Data

In this section, we report the comparison results between `FedACD` and the baselines on benchmark Bayesian network datasets in terms of SHD and F1 score.

From Figure 3, it can be observed that `FedACD` achieves the lowest SHD value and the highest F1 score in most scenarios, which validates its superiority. In particular, on the Child and Alarm datasets, when the number of clients reaches 20, `FedACD` shows significantly superior performance compared to other baselines. This is attributed to the fact that, with a larger number of clients, the allocated sample size per client becomes smaller. In such scenarios, the sample quality varies significantly across different variable spaces at each client, yet existing FCD methods cannot discern which variable spaces exhibit better sample quality. In contrast, `FedACD`'s adaptive masking strategy renders the learning results from samples in low-quality variable spaces at each client ineffective, while preserving the results from samples in high-quality variable spaces.

Compared with the best baseline, FedPC, the F1 score of `FedACD` is 42% higher on the Child dataset when the number of clients is 20. The F1 scores achieved by `FedACD` are significantly higher than those achieved by NOTEARS-ADMM, NOTEARS-MLP-ADMM, GS-FedDAG and AS-FedDAG, since it is hard for them to select a suitable threshold to prune false causal edges. As a result, the causal DAGs learned by these four baselines often contain a larger number of false causal edges. In addition, as the number of clients increases, all algorithms experience significant performance degradation, whereas `FedACD` maintains good stability.

5.3 Results on Synthetic Non-IID Data

In real-world scenarios, clients' local datasets are often Non-IID. Therefore, in this section, we also evaluate the performance of all algorithms on synthesized Non-IID data. The results are shown in Figure 4. It can be observed that regardless of the number of clients, `FedACD` consistently achieves the lowest SHD values and the highest F1 scores. This indicates that the adaptive masking strategy designed in `FedACD` is well suited to the horizontal FL setting with Non-IID local data. This is because in such scenarios, the significant differences in sample quality across variable spaces at each client are obvious, and `FedACD` is designed to address this issue.

5.4 Results on Real-World Data

The experimental results on the two real-world datasets, REGED with 8 nodes and 20 nodes [Statnikov *et al.*, 2015], are presented in Figure 5. It can be observed that under REGED with 8 nodes, `FedACD` achieves the lowest SHD value when the number of clients is 3, 10, 15 and 20. In addition, `FedACD` achieves the highest F1 score under REGED with 20 nodes, except when the number of clients is 20. Com-

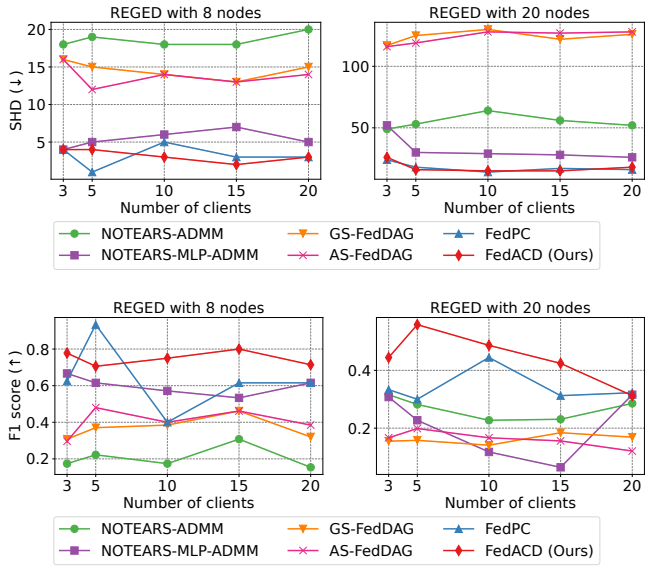

Figure 5: Experimental results on real-world datasets. There are 1,000 samples in total, allocated evenly across {3, 5, 10, 15, 20} clients. We show the performance of all methods in two metrics (SHD and F1 score from top to bottom).

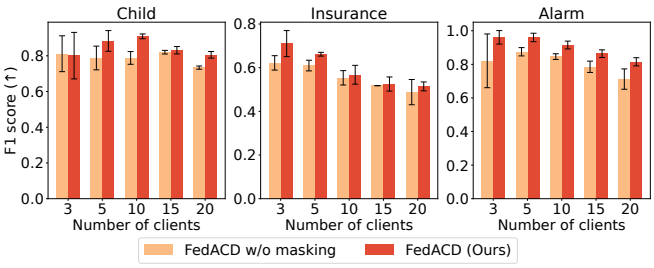

Figure 6: Ablation study results.

pared to other baselines, our method achieves more stable performance under REGED with 8 nodes with different numbers of clients.

5.5 Ablation Study

In this section, we conduct ablation experiments to validate the effectiveness of the proposed adaptive masking strategy. Specifically, we first develop a variant of FedACD, denoted as "FedACD w/o masking", which maintains a constant masking rate of 0 throughout the entire learning process. We then compare FedACD with "FedACD w/o masking" using three benchmark BN datasets, Child, Insurance and Alarm, across {3,5,10,15,20} clients. The experimental results are presented in Figure 6. It can be observed that FedACD consistently achieves higher F1 scores and lower SHD values than "FedACD w/o masking". This demonstrates the effectiveness of our proposed adaptive masking strategy under a horizontal FL setting, especially when dealing with potentially significant differences in sample quality across different variable spaces at each client.

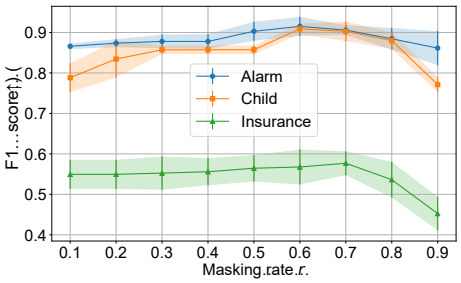

Figure 7: Sensitivity analysis of FedACD to the masking rate r.

5.6 Sensitivity Analysis

In this section, we conduct sensitivity analysis on the masking rate r of FedACD using the benchmark datasets. Specifically, we first generate three batches of data using Child, Insurance and Alarm BNs, respectively. For each dataset, there are 5,000 samples in total, allocated evenly across 10 clients. Then, we test FedACD on these datasets by varying the masking rate from 0.1 to 0.9, and record the F1 scores obtained. The experimental results are presented in Figure 7.

It can be observed that the F1 score achieved by FedACD increases with the increase of the masking rate r when $r \leq 0.6$. However, when $r \geq 0.7$, the F1 score achieved by FedACD drops significantly. The reason behind this phenomenon lies in Theorem 1, which indicates that when $P(\acute{A}(\mathcal{E}_j^{c_k}(a), \mathcal{E}_j^{c_k}(b)) = 0)$ is fixed, a larger masking rate r implies a higher probability of successfully removing a false causal edge in Phase 1-2. However, when the masking rate becomes excessively large (e.g., $r \geq 0.7$), $\exists i_1, i_2$ such that "$\sum_{k=1}^{m} \mathcal{A}_l^{c_k}(i_1, i_2) \neq -m$" holds, leading to the ineffectiveness of the aggregation strategy in Eq. (3). Consequently, the performance of FedACD experiences a noticeable decline. In our experiments, the masking rate r for FedACD is set to 0.6 on all types of datasets.

6 Conclusions and Future Work

In the horizontal FL scenarios, the sample quality under different variable spaces at each FL client can be different. As a result, the local data of each client might only be suitable for learning the causal relationships among certain variables, while the effectiveness of learning causal relationships among other variables may be limited. Existing FCD methods cannot effectively handle this *sample quality heterogeneity* issue. Our proposed FedACD bridges this crucial gap by adaptively selecting and sending only the optimal causal relationships learned under the "good" variable space (i.e., the space with high-quality samples) at each FL client to the server, while masking the causal relationships learned under the "bad" variable space (i.e., the space with low-quality samples) at each FL client, during FL model training. In this way, each client only communicates the optimal learning results to the server, achieving accurate FCD. Theoretical analysis and extensive experimental results demonstrate the efficacy of FedACD.

In the future, we plan to explore FCD in more challenging scenarios, such as data containing hidden variables and data combining observational and interventional sources.

Acknowledgments

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); the China Scholarship Council (under grant 202306690023); the National Natural Science Foundation of China (under grant 92367202); the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (No. AISG2-RP-2020-019); and the RIE 2020 Advanced Manufacturing and Engineering (AME) Programmatic Fund (No. A20G8b0102), Singapore.

References

[Abyaneh *et al.*, 2022] Amin Abyaneh, Nino Scherrer, Patrick Schwab, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. FED-CD: Federated causal discovery from interventional and observational data. *arXiv preprint arXiv:2211.03846*, 2022.

[Anderson *et al.*, 2023] Lisa M Anderson, Kelvin O Lim, Erich Kummerfeld, et al. Causal discovery analysis: A promising tool in advancing precision medicine for eating disorders. *International Journal of Eating Disorders*, 56(11):2012–2021, 2023.

[Arshad and Silaghi, 2004] Muhammad Arshad and Marius C Silaghi. Distributed simulated annealing. *Distributed Constraint Problem Solving and Reasoning in Multi-Agent Systems*, 112, 2004.

[Boyd *et al.*, 2011] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, 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.

[Chickering, 2002] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

[Gámez *et al.*, 2011] José A Gámez, Juan L Mateo, and José M Puerta. Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. *Data Mining and Knowledge Discovery*, 22(1):106–148, 2011.

[Gao *et al.*, 2023] Erdun Gao, Junjia Chen, Li Shen, Tongliang Liu, Mingming Gong, and Howard Bondell. FedDAG: Federated DAG structure learning. *Transactions on Machine Learning Research*, 2023, 2023.

[Guo *et al.*, 2022] Xianjie Guo, Yujie Wang, Xiaoling Huang, Shuai Yang, and Kui Yu. Bootstrap-based causal structure learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 656–665, 2022.

[Guo *et al.*, 2023] Xianjie Guo, Kui Yu, Lin Liu, Peipei Li, and Jiuyong Li. Adaptive skeleton construction for accurate DAG learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10526–10539, 2023.

[Guo *et al.*, 2024] Xianjie Guo, Kui Yu, Lin Liu, and Jiuyong Li. FedCSL: A scalable and accurate approach to federated causal structure learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12235–12243, 2024.

[He *et al.*, 2021] Yue He, Peng Cui, Zheyan Shen, Renzhe Xu, Furui Liu, and Yong Jiang. Daring: Differentiable causal discovery with residual independence. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 596–605, 2021.

[Huang *et al.*, 2020] Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery from multiple data sets with non-identical variable sets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10153–10161, 2020.

[Huang *et al.*, 2023a] Jianli Huang, Xianjie Guo, Kui Yu, Fuyuan Cao, and Jiye Liang. Towards privacy-aware causal structure learning in federated setting. *IEEE Transactions on Big Data*, 9(6):1525–1535, 2023.

[Huang *et al.*, 2023b] Xiaoling Huang, Xianjie Guo, Yuling Li, and Kui Yu. A novel data enhancement approach to dag learning with small data samples. *Applied Intelligence*, 53(22):27589–27607, 2023.

[Huo *et al.*, 2023a] Cuiying Huo, Dongxiao He, Chundong Liang, Di Jin, Tie Qiu, and Lingfei Wu. Trustgnn: Graph neural network-based trust evaluation via learnable propagative and composable nature. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Huo *et al.*, 2023b] Cuiying Huo, Di Jin, Yawen Li, Dongxiao He, Yu-Bin Yang, and Lingfei Wu. T2-gnn: Graph neural networks for graphs with incomplete features and structure via teacher-student distillation. In *AAAI*, pages 4339–4346, 2023.

[Liu *et al.*, 2022] Jiaxin Liu, Liwei Deng, Hao Miao, Yan Zhao, and Kai Zheng. Task assignment with federated preference learning in spatial crowdsourcing. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1279–1288, 2022.

[Liu *et al.*, 2024a] Rui Liu, Pengwei Xing, Zichao Deng, Anran Li, Cuntai Guan, and Han Yu. Federated graph neural networks: Overview, techniques, and challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Liu *et al.*, 2024b] Ziqiao Liu, Hao Miao, Yan Zhao, Chenxi Liu, Kai Zheng, and Huan Li. LightTR: A lightweight framework for federated trajectory recovery. In *IEEE International Conference on Data Engineering*, 2024.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Mian *et al.*, 2023] Osman Mian, David Kaltenpoth, Michael Kamp, and Jilles Vreeken. Nothing but regrets—privacy-preserving federated causal discovery. In *International Conference on Artificial Intelligence and Statistics*, pages 8263–8278. PMLR, 2023.

[Miao *et al.*, 2024] Hao Miao, Xiaolong Zhong, Jiaxin Liu, Yan Zhao, Xiangyu Zhao, Weizhu Qian, Kai Zheng, and Christian S Jensen. Task assignment with efficient federated preference learning in spatial crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1800–1814, 2024.

[Mooij *et al.*, 2020] Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

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

[Paillier, 1999] Pascal Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *International Conference on The Theory and Applications of Cryptographic Techniques*, pages 223–238. Springer, 1999.

[Pearl, 1988] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[Pearl, 2018] Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 3–3, 2018.

[Pena, 2008] Jose M Pena. Learning gaussian graphical models of gene networks with false discovery rate control. In *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 165–176. Springer, 2008.

[Perry *et al.*, 2022] Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. *Advances in Neural Information Processing Systems*, 35:10904–10917, 2022.

[Ramsey, 2016] Joseph Ramsey. Improving accuracy and scalability of the PC algorithm by maximizing p-value. *arXiv preprint arXiv:1610.00378*, 2016.

[Scutari, 2016] Marco Scutari. An empirical-Bayes score for discrete Bayesian networks. In *Conference on Probabilistic Graphical Models*, pages 438–448. PMLR, 2016.

[Sheng *et al.*, 2024] Shaojing Sheng, Xianjie Guo, Kui Yu, and Xindong Wu. Local causal structure learning with missing data. *Expert Systems with Applications*, 238:121831, 2024.

[Spirtes *et al.*, 2000] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT Press, 2000.

[Statnikov *et al.*, 2015] Alexander Statnikov, Sisi Ma, Mikael Henaff, Nikita Lytkin, Efstratios Efstathiadis, Eric R Peskin, and Constantin F Aliferis. Ultra-scalable and efficient methods for hybrid observational and experimental local causal pathway discovery. *Journal of Machine Learning Research*, 16(1):3219–3267, 2015.

[Tsamardinos *et al.*, 2006] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 2006.

[Vowels *et al.*, 2022] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like DAGs? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

[Watanabe, 2013] Sumio Watanabe. A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14(1):867–897, 2013.

[Xiang *et al.*, 2023] Guodu Xiang, Hao Wang, Kui Yu, Xianjie Guo, Fuyuan Cao, and Yukun Song. Bootstrap-based layer-wise refining for causal structure learning. *IEEE Transactions on Artificial Intelligence*, 1(01):1–15, 2023.

[Yang *et al.*, 2019] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2):1–19, 2019.

[Ye *et al.*, 2022] Qiaoling Ye, Arash A Amini, and Qing Zhou. Distributed learning of generalized linear causal networks. *arXiv preprint arXiv:2201.09194*, 2022.

[Yu *et al.*, 2011] Han Yu, Siyuan Liu, Alex C Kot, Chunyan Miao, and Cyril Leung. Dynamic witness selection for trustworthy distributed cooperative sensing in cognitive radio networks. In *2011 IEEE 13th International Conference on Communication Technology*, pages 1–6. IEEE, 2011.

[Yu *et al.*, 2020] Kui Yu, Xianjie Guo, Lin Liu, Jiuyong Li, Hao Wang, Zhaolong Ling, and Xindong Wu. Causality-based feature selection: Methods and evaluations. *ACM Computing Surveys (CSUR)*, 53(5):1–36, 2020.

[Zhang *et al.*, 2023a] Honglei Zhang, Fangyuan Luo, Jun Wu, Xiangnan He, and Yidong Li. Lightfr: Lightweight federated recommendation with privacy-preserving matrix factorization. *ACM Transactions on Information Systems*, 41(4):1–28, 2023.

[Zhang *et al.*, 2023b] Yuxin Zhang, Jindong Wang, Yiqiang Chen, Han Yu, and Tao Qin. Adaptive memory networks with self-supervised learning for unsupervised anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12068–12080, 2023.

[Zheng *et al.*, 2018] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[Zhong *et al.*, 2023] Xiaolong Zhong, Hao Miao, Dazhuo Qiu, Yan Zhao, and Kai Zheng. Personalized location-preference learning for federated task assignment in spatial crowdsourcing. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3534–3543, 2023.