FedCSL: A Scalable and Accurate Approach to Federated Causal Structure Learning

Xianjie Guo[1], **Kui Yu**[1*], **Lin Liu**[2], **Jiuyong Li**[2]

[1]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
[2]UniSA STEM, University of South Australia, Adelaide, Australia
xianjieguo@mail.hfut.edu.cn, yukui@hfut.edu.cn, {lin.liu, jiuyong.li}@unisa.edu.au

Abstract

As an emerging research direction, federated causal structure learning (CSL) aims at learning causal relationships from decentralized data across multiple clients while preserving data privacy. Existing federated CSL algorithms suffer from scalability and accuracy issues, since they require computationally expensive CSL algorithms to be executed at each client. Furthermore, in real-world scenarios, the number of samples held by each client varies significantly, and existing methods still assign equal weights to the learned structural information from each client, which severely harms the learning accuracy of those methods. To address these two limitations, we propose FedCSL, a scalable and accurate method for federated CSL. Specifically, FedCSL consists of two novel strategies: (1) a federated local-to-global learning strategy that enables FedCSL to scale to high-dimensional data for tackling the scalability issue, and (2) a novel weighted aggregation strategy that does not rely on any complex encryption techniques while preserving data privacy for tackling the accuracy issue. Extensive experiments on benchmark datasets, high-dimensional synthetic datasets and a real-world dataset verify the efficacy of the proposed FedCSL method. The source code is available at https://github.com/Xianjie-Guo/FedCSL.

Introduction

Causal structure learning (CSL) is a primary approach to uncover causal relationships among variables (Lampinen et al. 2022), and it has been widely applied in various fields including social science (Hedström and Ylikoski 2010), artificial intelligence (Glymour, Scheines, and Spirtes 2014; Wu et al. 2023a), and systems biology (Lagani et al. 2016).

Related Work During the last decades, many CSL methods have been proposed (Vowels, Camgoz, and Bowden 2022), which can be broadly categorized into two classes: combinatorial optimization-based and continuous optimization-based methods. Methods in the former, such as PC (Spirtes et al. 2000), GES (Chickering 2002) and A* (Yuan, Malone, and Wu 2011), heuristically evaluate the goodness-of-fit between the structural combinations over variables and the dataset used to learn the optimal causal structure. In contrast, methods in the latter class, such as NOTEARS (Zheng et al. 2018) and DAG-GNN (Yu et al. 2019), utilize gradient descent to optimize a weight adjacency matrix to fit causal relationships among variables, providing a new research approach for causal structure learning.

In practice, the performance of a learning algorithm is largely attributed to the number of available samples. To improve the learning performance, users often try to collect data from multiple decentralized sources and aggregate them into a large-scale dataset. However, due to concerns about data privacy, data owners are increasingly reluctant to share their local data with others (Yang et al. 2019). To address this issue, federated learning (FL) has emerged as a novel learning paradigm that trains a model from locally stored data while preserving privacy (McMahan et al. 2017).

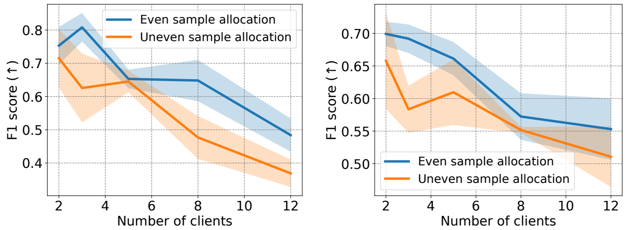
Although there has been significant research in the field of FL (Zhang et al. 2023; Yu et al. 2023), the problem of causal structure learning in the FL setting has received limited attentions so far. Some notable work in this area includes NOTEARS-ADMM (Ng and Zhang 2022), FedDAG (Gao et al. 2023) and FedPC (Huang et al. 2023a). Specifically, NOTEARS-ADMM directly applies the distributed optimization algorithm ADMM (Boyd et al. 2011) to optimize the NOTEARS method. FedDAG adopts a two-level structure for each local model, where the first level learns causal structure by communicating with the server, and the second level approximates variable relationships on each local data for handling data heterogeneity. FedPC proposes a layer-wise aggregation strategy for seamless integration of the PC algorithm into the federated learning paradigm, which has achieved promising performance on multiple types of data.

Challenges Existing federated CSL algorithms, however, face the following two main challenges.

Challenge 1: Scalability issue. Learning a global causal structure from a single data source is a computationally expensive task and is proven to be an NP-hard problem (Chickering, Heckerman, and Meek 2004). Often complex optimization methods or sophisticated neural network models are required for achieving satisfactory performance (Yu et al. 2019). Existing federated CSL algorithms primarily rely on global learning strategies, leading to significant computational challenges for these algorithms to deal with high-dimensional data across multiple clients (data sources).

Challenge 2: Accuracy issue. In real-world scenarios, the

[*]Corresponding Author



(a) Child Bayesian network (b) Insurance Bayesian network

Figure 1: Results of even and uneven sample allocations on the different numbers of clients.

data volume of each client is often different (called uneven sample allocation). In such cases, the quality of the learned structures varies across different clients and is typically positively correlated with sample size at clients. However, existing methods assign an equal weight to each client when aggregating the structures learned from different clients, leading to unsatisfactory performance. To illustrate the issue, we conduct experiments on two benchmark Bayesian network (BN) datasets, Child and Insurance (Tsamardinos, Brown, and Aliferis 2006), using a state-of-the-art federated CSL algorithm FedPC (Huang et al. 2023a). By keeping the total sample size unchanged, we allocate samples to each client in even and uneven ways, respectively. The experimental results in Fig. 1 clearly indicate that by assigning an equal weight to each client, the performance in the uneven sample allocation setting is significantly worse than that in the even sample allocation setting. To alleviate this issue, a feasible strategy is to assign different weights to clients based on the sample sizes held by them (Ma et al. 2022). However, the sample size information of each client is often private and not known to the server (Kairouz et al. 2021). For example, in a FL system for financial institutions, preserving sample size privacy ensures that the exact number of transactions or customer accounts held by each institution remains confidential. Thus, without using complex encryption techniques, the task of designing a weighted aggregation strategy becomes a significant challenge when the sample sizes held by different clients are unknown.

Contributions In this paper, we propose a novel federated CSL method, called FedCSL, to tackle the two challenges above and make the following main contributions.

- To address Challenge 1, we design a federated local-to-global learning strategy (instead of global learning strategy adopted by exiting methods), which consists of three steps: *federated causal neighbor learning*, *federated global skeleton construction*, and *federated skeleton orientation*. This strategy enables the FedCSL algorithm to scale to high-dimensional data.
- To address Challenge 2, we design a novel strategy to calculate the weights of different clients in scenarios with uneven sample allocation. This strategy preserves data privacy without relying on encryption techniques, and can infer the relative sample sizes held by each client for computing the weights of different clients. Based on this strategy, at each step of the local-to-global learning process, FedCSL performs weighted aggregation of learned structures by clients using different weights.
- We conduct comprehensive experiments on benchmark BN datasets, high-dimensional synthetic datasets, and real dataset to evaluate the performance of FedCSL.

Notations and Assumptions

Let $X = \{X_1, X_2, ..., X_d\}$ be a set of d variables under consideration, $\mathcal{C} = \{c_1, c_2, ..., c_m\}$ be a set of m different clients, and $\mathcal{D}^{c_k} \in \mathbb{R}^{n_{c_k} \times d}$ represent the local dataset owned by client c_k, and n_{c_k} is the number of samples in $\mathcal{D}^{c_k}$, where $k \in \{1, 2, ..., m\}$.

A causal structure over X is often represented using a causal directed acyclic graph (DAG) (Huang et al. 2023b). In a causal DAG, if there is a direct edge $X_i \rightarrow X_j$ ($i, j \in \{1, 2, ..., d\}$), X_i is a direct cause (parent) of X_j, and X_j is a direct effect (child) of X_i (Wu et al. 2022). In this paper, if there is $X_i \rightarrow X_j$ or $X_j \rightarrow X_i$, we say X_i and X_j are causal neighbors to each other. Moreover, when X_i and X_j are causal neighbors, $X_i \not\perp\!\!\!\perp X_j | X_z$ always holds, where $X_z \subseteq X \setminus \{X_i, X_j\}$ and we use $\not\perp\!\!\!\perp$ (or $\perp\!\!\!\perp$) to represent the dependence (or independence) relation. In this paper, we use the G^2 test (Spirtes et al. 2000), which is an alternative to the χ^2 test, to conduct conditional independence (CI) tests between variables. Assume that ρ is the p-value returned by the G^2 test and α is a given significance level. Under the null hypothesis of "$H_0 : X_i \perp\!\!\!\perp X_j | X_z$", for a CI test of X_i and X_j given X_z, $X_i \perp\!\!\!\perp X_j | X_z$ holds if and only if $\rho > \alpha$.

In this work, we consider a horizontal federated learning setting, and different clients share the same feature space but have different sample space. Furthermore, we assume that all local datasets $\mathcal{D}^{\mathcal{C}} = \{\mathcal{D}^{c_1}, \mathcal{D}^{c_2}, ..., \mathcal{D}^{c_m}\}$ are uniformly sampled from the same causal DAG $\mathcal{G}$.

Federated causal structure learning aims to identify a causal DAG $\mathcal{G}$ from $\mathcal{D}^{\mathcal{C}}$ in a privacy-preserving way. Considering the practical application scenarios, throughout this paper, $n_{c_{k_1}} \neq n_{c_{k_2}}$ if $k_1 \neq k_2$ for $\forall k_1, k_2 \in \{1, 2, ..., m\}$ (i.e., clients have different data sample sizes), and $\mathcal{D}^{c_{k_1}}$ (or $\mathcal{D}^{c_{k_2}}$) can be high-dimensional. Given a local dataset $\mathcal{D}^{c_k}$ held by client c_k, based on the law of large numbers, the results of CI tests conducted on $\mathcal{D}^{c_k}$ will approximate the ground truth infinitely as the sample size in $\mathcal{D}^{c_k}$ approaches towards infinity. Thus, we make the following assumption.

Assumption 1. *Given any two local datasets $\mathcal{D}^{c_{k_1}}$ and $\mathcal{D}^{c_{k_2}}$ for $\forall k_1, k_2 \in \{1, 2, ..., m\}$, if $n_{c_{k_2}} > n_{c_{k_1}}$, the reliability of CI tests performed on $\mathcal{D}^{c_{k_2}}$ is higher than that of CI tests performed on $\mathcal{D}^{c_{k_1}}$.*

Method

As shown in Fig. 2, the proposed FedCSL consists of three steps: *federated causal neighbor learning*, *federated global skeleton construction* and *federated skeleton orientation*.

Federated Causal Neighbor Learning

Step 1-1: Learning the Potential Causal Neighbors of Each Variable Independently. At each client, we first

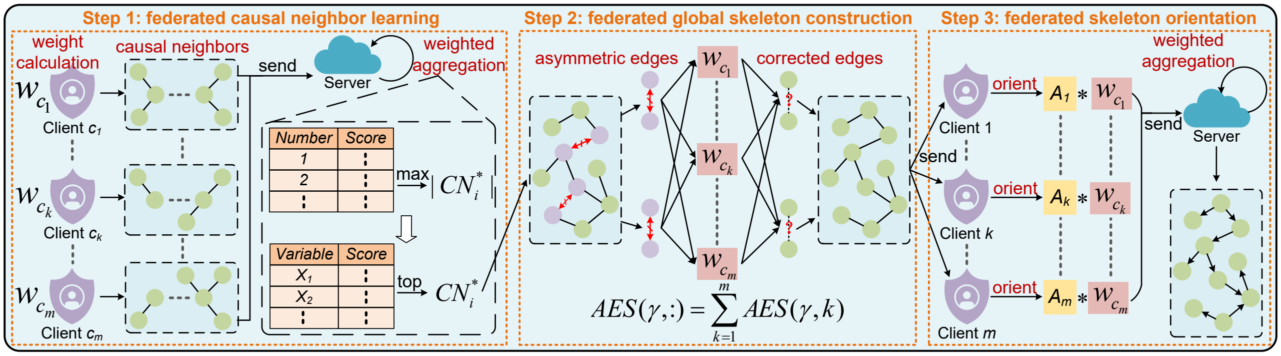


Figure 2: The framework of FedCSL, which consists of three steps.

employ a well-established local causal structure learning algorithm, HITON-PC (Aliferis et al. 2010), which utilizes CI tests, to learn the potential causal neighbors of each variable. That is, for client c_k, in this step, we obtain the causal neighbor sets of all variables, i.e. $CN^{c_k} = \{CN_i^{c_k}\}_{i\in\{1,2,...,d\}} = \{CN_1^{c_k}, CN_2^{c_k}, ..., CN_d^{c_k}\}$. Learning a set of causal neighbors for each variable rather than a global structure over all variables makes FedCSL scalable to high-dimensional data, since the search space for global structure learning grows exponentially with d; whereas the search space of local structure learning grows exponentially with the number of causal neighbors of the target variable but linearly with d.

For this step, any state-of-the-art local causal structure learning algorithm (Wu et al. 2019, 2020, 2023b,c; Guo et al. 2022c,b; Yang et al. 2023; Yu et al. 2020) can be utilized.

Step 1-2: Calculating the Weights of Different Clients. According to Assumption 1, we can infer that the larger the sample size held by a client, the more reliable the results of the CI tests performed at that client. Therefore, we need to assign higher weights to the causal structures learned from the clients with larger sample sizes. However, due to high data privacy protection, clients often are unwilling to disclose the sizes of the samples they hold. In this case, to indirectly infer the relative sample sizes of each client, we first propose Theorems 1 and 2 to establish the relationship between p-values of the CI tests and sample sizes. The proofs of Theorems 1 and 2 are given in Appendix A.

Theorem 1. *Let ρ be the p-value obtained by conducting G^2 test to determine whether two variables X_i and X_j are conditionally independent given an empty set, using a dataset $\mathcal{D}^{c_k}$. If $X_i \perp\!\!\!\perp X_j|\emptyset$ holds true in the underlying causal structure behind $\mathcal{D}^{c_k}$, $\rho \to 1^-$ when $n_{c_k} \to \infty$.*

Theorem 2. *If $X_i \not\perp\!\!\!\perp X_j|\emptyset$ holds true in the underlying causal structure behind $\mathcal{D}^{c_k}$, $\rho \to 0^+$ when $n_{c_k} \to \infty$.*

Given a dataset $\mathcal{D}^{c_k}$, for a CI test of X_i and X_j given an empty set, Theorem 1 implies that if the null hypothesis of "$H_0 : X_i \perp\!\!\!\perp X_j|\emptyset$" is accepted, i.e., $\alpha < \rho \leq 1$, then a larger ρ could indicate a larger n_{c_k}. In contrast, Theorem 2 states that if the null hypothesis of "$H_0 : X_i \perp\!\!\!\perp X_j|\emptyset$" is rejected, i.e., $0 \leq \rho \leq \alpha$, then a larger ρ could indicate a smaller n_{c_k}. Thus, based on Theorems 1 and 2, we have the following hypothesis. We have experimentally validated Hypothesis 1 in the experimental section.

Hypothesis 1. *Let $\rho_{ij}^{c_{k_1}}$ and $\rho_{ij}^{c_{k_2}}$ be the p-values obtained by conducting G^2 tests to determine whether $X_i \perp\!\!\!\perp X_j|\emptyset$ holds, using two local datasets $\mathcal{D}^{c_{k_1}}$ and $\mathcal{D}^{c_{k_2}}$, respectively. If $X_i \perp\!\!\!\perp X_j|\emptyset$ holds true on both $\mathcal{D}^{c_{k_1}}$ and $\mathcal{D}^{c_{k_2}}$ (i.e. $\rho_{ij}^{c_{k_1}}, \rho_{ij}^{c_{k_2}} \in (\alpha, 1]$), $n_{c_{k_2}} \geq n_{c_{k_1}}$ if $\rho_{ij}^{c_{k_2}} \geq \rho_{ij}^{c_{k_1}}$; if $X_i \not\perp\!\!\!\perp X_j|\emptyset$ holds true on both $\mathcal{D}^{c_{k_1}}$ and $\mathcal{D}^{c_{k_2}}$ (i.e. $\rho_{ij}^{c_{k_1}}, \rho_{ij}^{c_{k_2}} \in [0, \alpha]$), $n_{c_{k_2}} \leq n_{c_{k_1}}$ if $\rho_{ij}^{c_{k_2}} \geq \rho_{ij}^{c_{k_1}}$.*

According to Hypothesis 1, without the use of complex encryption techniques, we design a novel strategy to calculate the weights of different clients without knowing the actual sample sizes of each client, thus highly preserving data privacy. In Step 1-1, we have obtained the p-values returned by conducting CI tests between every pair of variables given an empty set. To facilitate the calculation of the weight values of different clients, we need to uniformly scale the p-values in [0, α] and (α, 1] to the interval [0, 1] using the following two rules. At client c_k, for each p-value $\rho_{ij}^{c_k}$ ($i, j \in \{1, 2, ..., d\}$ and $i < j$) obtained by conducting CI tests between X_i and X_j conditioning on an empty set, if $\rho_{ij}^{c_k} \in (\alpha, 1]$, we proportionally scale up $\rho_{ij}^{c_k}$ to the [0, 1] interval, since the p-values in (α, 1] are directly proportional to the sample size, as indicated by Hypothesis 1. If $\rho_{ij}^{c_k} \in [0, \alpha]$, we first proportionally scale up $\rho_{ij}^{c_k}$ to the [0, 1] interval, and then apply a central symmetric flip to the values within this interval. This is because, as indicated by Hypothesis 1, p-values in [0, α] are inversely proportional to the sample size. Here, we denote the transformed $\rho_{ij}^{c_k}$ in the two rules mentioned above as $\hat{\rho}_{ij}^{c_k}$. Thus, we have:

$$\hat{\rho}_{ij}^{c_k} = \begin{cases} \frac{\alpha-\rho_{ij}^{c_k}}{\alpha} & if\ \rho_{ij}^{c_k} \in [0, \alpha] \\ \frac{\rho_{ij}^{c_k}-\alpha}{1-\alpha} & if\ \rho_{ij}^{c_k} \in (\alpha, 1] \end{cases} \tag{1}$$

Finally, we calculate the average of all the values in $\{\hat{\rho}_{ij}^{c_k}\}_{i,j\in\{1,2,...,d\},i<j}$, denoted as w_{c_k}, and use it as the weight for client c_k. That is,

$$w_{c_k} = \frac{1}{\frac{d(d-1)}{2}} \sum_{i=1}^{d} \sum_{j=i+1}^{d} \hat{\rho}_{ij}^{c_k}. \tag{2}$$

By following this procedure, we can compute the weights for all clients, $w_{\mathcal{C}} = \{w_{c_k}\}_{k\in\{1,2,...,m\}} = \{w_{c_1}, w_{c_2}, ..., w_{c_m}\}$, and the higher the weight of a client, the larger the size of the samples the client has. Note that the total number of CI tests performed on different clients may vary, but the total number of CI tests between any two variables under the empty set condition is always the same (i.e., $d(d-1)/2$ CI tests). Therefore, to ensure fairness, we only utilize the p-values returned from the CI tests given an empty set to estimate the weights of different clients.

Step 1-3: Calculating the Optimal Number of Causal Neighbors for Each Variable. In Step 1-1, the learned causal neighbor set of each variable may vary across different clients, so it is necessary to first compute the optimal number of causal neighbors for each variable. Specifically, for X_i, we first send the client weights $w_{\mathcal{C}}$ and the causal neighbor sets of X_i obtained at all clients $\{CN_i^{c_k}\}_{k\in\{1,2,...,m\}}$ to the server. Let $|CN_i^{c_k}|$ represent the number of causal neighbors for X_i learned at client c_k. To facilitate weighted aggregation, we utilize a mask matrix $\mathit{\Psi}_i \in \mathbb{R}^{m\times \max_{k=1}^{m}(|CN_i^{c_k}|)}$ to record the number of causal neighbors for X_i obtained across m clients as follows.

$$\underset{\substack{k=1,2,...,m \\ \xi=1,2,...,\max_{k=1}^{m}(|CN_i^{c_k}|)}}{\mathit{\Psi}_i(k,\xi)} = \begin{cases} 1 & if\ \xi = |CN_i^{c_k}| \\ 0 & otherwise \end{cases}, \quad (3)$$

where $\max_{k=1}^{m}(|CN_i^{c_k}|)$ denotes the maximum number of causal neighbors for X_i learned across all clients, and ξ represents an indicator variable for the number of causal neighbors ($\xi \in \{1, 2, ..., \max_{k=1}^{m}(|CN_i^{c_k}|)\}$). Eq. (3) denotes that if the number of causal neighbors for X_i learned at client c_k is ξ, $\mathit{\Psi}_i(k,\xi) = 1$; otherwise, $\mathit{\Psi}_i(k,\xi) = 0$.

At the server side, for each ξ, we add up the weights of the clients on which the number of causal neighbors for X_i learned is equal to ξ (as shown in Fig. 2). Then we choose the number of causal neighbors that corresponds to the maximum sum of weights as the best estimate for the optimal number of causal neighbors for X_i, denoted as $|CN_i^*|$. Eq. (4) below summarize the above described operation.

$$|CN_i^*| = \begin{cases} \overset{\circ}{M}ax([w_{c_1}, w_{c_2}, ..., w_{c_m}]\mathit{\Psi}_i) & if\ \parallel \mathit{\Psi}_i \parallel_1 \neq 0 \\ 0 & otherwise \end{cases}, \quad (4)$$

where $\overset{\circ}{M}ax(\cdot)$ is utilized to select the index value corresponding to the maximum value in a vector, and $\parallel \mathit{\Psi}_i \parallel_1$ denotes the calculation of the L1 norm of matrix $\mathit{\Psi}_i$ to determine if $\mathit{\Psi}_i$ is a zero matrix. Finally, we obtain the optimal number of federally learned causal neighbors of each variable over all clients, i.e., $\{|CN_i^*|\}_{i\in\{1,2,...,d\}}$.

Step 1-4: Determining the Optimal Causal Neighbors of Each Variable. According to the optimal number of causal neighbors for each variable obtained in Step 1-3, this step aims to federally learn an optimal set of causal neighbors for each variable across all clients. Specifically, for X_i, its potential causal neighbor sets learned at all clients are recorded in a mask matrix $B_i \in \mathbb{R}^{m\times d}$ as follows.

$$\underset{k=1,2,...,m;j=1,2,...,d}{B_i(k,j)} = \begin{cases} 1 & if\ X_j \in CN_i^{c_k} \\ 0 & otherwise \end{cases}, \quad (5)$$

where if X_j is a causal neighbor of X_i at client c_k, then $B_i(k,j) = 1$; otherwise, $B_i(k,j) = 0$.

After obtaining all potential causal neighbors of X_i in Step 1-1, for each potential causal neighbor X_j of X_i, we check if X_j is learned at client c_k as a potential causal neighbor of X_i. If it is, the score of X_j is increased by w_{c_k}; otherwise, its score remains unchanged (as shown in Fig. 2). This process is repeated for all clients to obtain the final score of X_j (as a potential neighbor of X_i). Finally, the scores of all potential causal neighbors of X_i are sorted, and the top $|CN_i^*|$ potential causal neighbors of X_i are selected as the federally learned optimal set of causal neighbors for X_i across all clients, denoted as CN_i^*. Therefore, we have:

$$CN_i^* = \overset{\circ}{T}op_{|CN_i^*|}([w_{c_1}, w_{c_2}, ..., w_{c_m}]B_i), \quad (6)$$

where $\overset{\circ}{T}op_{|CN_i^*|}$ is used to obtain the variable index corresponding to the top $|CN_i^*|$ elements in a vector based on their score order.

Federated Global Skeleton Construction

In this step, we utilize the learned optimal causal neighbor sets for all variables in Step 1-4, i.e., $\{CN_i^*\}_{i\in\{1,2,...,d\}}$, to construct a global skeleton. Given any two variables X_i and X_j, if there is an edge connecting X_i and X_j in the true causal structure, X_i and X_j must be each other's causal neighbors; otherwise, they are not causal neighbors of each other. Thus, if $X_i \in CN_j^*$ and $X_j \in CN_i^*$, we connect X_i and X_j with an undirected edge; if $X_i \notin CN_j^*$ and $X_j \notin CN_i^*$, we consider that there is no edge between X_i and X_j. However, as shown in Fig. 2, we may also encounter the case that $X_i \in CN_j^*$ but $X_j \notin CN_i^*$ (or $X_i \notin CN_j^*$ but $X_j \in CN_i^*$). In this case, we say that there is an asymmetric edge between X_i and X_j, denoted as $X_i \nleftrightarrow X_j$. How to tackle these asymmetric edges in a federated setting?

To deal with this issue, we first utilize $\{CN_i^*\}_{i\in\{1,2,...,d\}}$ obtained in Step 1-4 to construct an initial global skeleton that may contain some asymmetric edges. Then we design a weighted scoring strategy to determine whether each asymmetric edge should be preserved as an undirected edge in the initial global skeleton or removed from it. We assign a score to each asymmetric edge as follows. For an asymmetric edge $X_i \nleftrightarrow X_j$, if the learned causal neighbor results at client c_k indicate that X_i is a causal neighbor of X_j and X_j is also a causal neighbor of X_i, we let the score of this edge be $[(1+1) * w_{c_k})]$. If it is determined that X_i is not a causal neighbor of X_j and X_j is also not a causal neighbor of X_i, we set the score of this edge to $[(-1-1) * w_{c_k}]$. In all other cases, i.e., $X_i \in CN_j^{c_k}$ but $X_j \notin CN_i^{c_k}$ (or $X_i \notin CN_j^{c_k}$ but $X_j \in CN_i^{c_k}$), we set the score of this edge to $[(-1+1) * w_{c_k}] = 0$. The score of the γ-th asymmetric edge at client c_k is denoted as $AES(\gamma, k)$, and we have:

$$AES(\gamma,k) = \begin{cases} (1+1) * w_{c_k} & if\ X_i \in CN_j^{c_k} \wedge X_j \in CN_i^{c_k} \\ (-1-1) * w_{c_k} & if\ X_i \notin CN_j^{c_k} \wedge X_j \notin CN_i^{c_k} \\ (-1+1) * w_{c_k} & if\ X_i \in CN_j^{c_k} \wedge X_j \notin CN_i^{c_k} \\ (1-1) * w_{c_k} & if\ X_i \notin CN_j^{c_k} \wedge X_j \in CN_i^{c_k}. \end{cases} \quad (7)$$

Then the overall score of the γ-th asymmetric edge considering all clients is

$$AES(\gamma,:) = \sum_{k=1}^{m} AES(\gamma,k). \tag{8}$$

Finally, if the overall score of the γ-th asymmetric edge is greater than 0, then it is retained in the initial global skeleton; otherwise, it is removed from the initial global skeleton. We denote the final global skeleton as S^*.

Federated Skeleton Orientation

As shown in Fig. 2, after having obtained the federally learned global skeleton S^*, the next step aims to orient the edges in S^* by a weighted aggregation strategy.

Specifically, the server first sends S^* to each client, then we use a Bayesian score criteria, BDeu (Scutari 2016), and a search procedure, hill-climbing (Gámez, Mateo, and Puerta 2011) to greedily orient the undirected edges in S^* at each client to obtain a global causal structure with the highest score at each client. Let $\mathcal{G}_k$ denote the global causal structure learned at client c_k and A_k denote the adjacency matrix corresponding to $\mathcal{G}_k$, and "$A_k(i,j) = 1$" denotes that there is an edge from X_i to X_j in $\mathcal{G}_k$.

Subsequently, we sends all adjacency matrices (i.e., A_1,A_2,...,A_m) back to the server to compute the aggregated adjacency matrix A^* as follows.

$$A^* = (A_1 * w_{c_1}) \oplus (A_2 * w_{c_2}) \oplus \cdots \oplus (A_m * w_{c_m}), \tag{9}$$

where $\oplus$ represents the element-wise addition of matrices. Finally, we compare the elements at corresponding positions on the diagonal of matrix A^* for obtaining the final causal structure (marked as $\mathcal{G}^*$). Specifically, if $A^*(i,j) > A^*(j,i)$, then there exists a directed edge from X_i to X_j; if $A^*(i,j) \leq A^*(j,i)$ and $A^*(j,i) \neq 0$, then there exists a directed edge from X_j to X_i; otherwise, there is no edge between X_i and X_j. To summarize, we have

$$\mathcal{G}^* \Leftarrow \begin{cases} X_i \rightarrow X_j & if\ A^*(i,j) > A^*(j,i) \\ X_i \leftarrow X_j & if\ A^*(i,j) \leq A^*(j,i) \wedge A^*(j,i) \neq 0 \\ X_i \nLeftrightarrow X_j & if\ A^*(i,j) = 0 \wedge A^*(j,i) = 0, \end{cases} \tag{10}$$

where "$X_i \nLeftrightarrow X_j$" denotes that there is no edge connected between X_i and X_j.

Experiments

Experiment Setting

Datasets. We utilize the following three types of datasets.

- **Benchmark BN datasets.** We use five benchmark BN datasets: Child with 20 variables, Insurance with 27 variables, Alarm with 37 variables, Pigs with 441 variables and Gene with 801 variables, and each dataset contains 5,000 samples (Tsamardinos, Brown, and Aliferis 2006).
- **High-dimensional synthetic datasets.** We first construct a causal DAG with 5,000 variables, where the maximum number of parents for each variable is 3, and the average degree of each variable is 2. Then, based on this causal DAG, we use an open-source software package[1] to generate three datasets, each containing 5,000 samples.
- **Real-world datasets.** We also compare the proposed method with the baselines on the Sachs (Sachs et al. 2005) dataset. Sachs is a benchmark graphical model representing protein signaling networks in human cells. It consists of 11 nodes (cell types) and 17 edges. Our experiments use 7,466 commonly used observational samples.

In our experiments, the local datasets at different clients have different sizes. Let $n = \sum_{k=1}^{m} n_{c_k}$ be the sum of sample sizes owned by the m clients, the sample size of each local dataset is set as follows.

$$\begin{aligned} & n_{c_1} = \lfloor \frac{n}{2m} \rfloor, n_{c_k} - n_{c_{k-1}} = \lfloor \frac{2(n - mn_{c_1})}{m(m-1)} \rfloor, \\ & n_{c_k} = n_{c_1} + \lfloor \frac{2(n - mn_{c_1})}{m(m-1)} \rfloor (k-1), k \in 2, 3, ..., m. \end{aligned} \tag{11}$$

Evaluation metrics. We use the Structural Hamming Distance (SHD) and F1 score (Guo et al. 2022a) to evaluate the discovered causal structures in a federated setting.

Baselines. We compare FedCSL with five state-of-the-art federated CSL methods, including NOTEARS-ADMM, NOTEARS-MLP-ADMM (Ng and Zhang 2022), GS-FedDAG, AS-FedDAG (Gao et al. 2023) and FedPC (Huang et al. 2023a), on the benchmark BN datasets and the real-world dataset. In addition, as existing federated CSL methods do not scale up to high-dimensional data, we develop the following four baselines using an efficient and effective CSL method, ADL (Guo et al. 2023), and compare them with our method on the high-dimensional synthetic datasets.

- ADL-AllData. We centralize all clients' data to a single dataset and run the ADL algorithm on it.
- ADL-Avg. We first run ADL at each client for obtaining m causal structures, and then calculate the average value of the metrics corresponding to m causal structures.
- ADL-Best. We first run ADL at each client independently to get m causal structures, and then select the causal structure with the highest F1 score as the final output.
- ADL-Voting. We apply a voting method (Na and Yang 2010) to the ADL algorithm.

Results on Benchmark BN Data

In this section, we report the experimental results of FedCSL and the baselines on benchmark BN datasets in terms of SHD, F1 score and Time (i.e., running time) metrics.

From Fig. 3, we can observe that regardless of the number of clients, FedCSL consistently achieves the lowest SHD value and the best F1 score on all datasets, which validates the superiority of our method. Especially, FedCSL achieves significant performance improvement when the number of clients reaches 12. This highlights the effectiveness of our designed weighted aggregation strategy in addressing the challenge of uneven sample allocation in the horizontal federated learning scenario, as the number of

[1] The code is available at https://www.cs.ubc.ca/ murphyk/Software/DAGlearn/index.html.

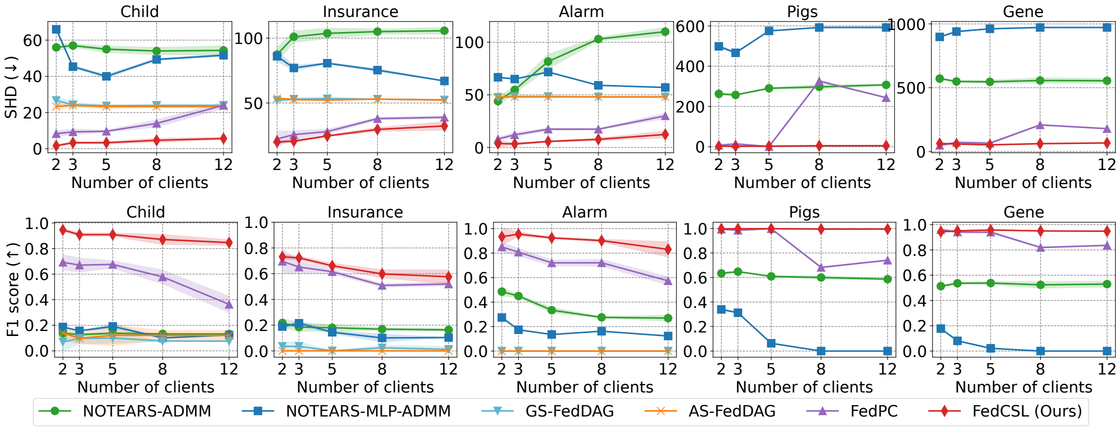


Figure 3: Structure learning results on the benchmark datasets. There are 5,000 samples in total, allocated unevenly across {2, 3, 5, 8, 12} clients. We show the performance of all methods in two metrics (SHD and F1 score from top to bottom). Note that due to insufficient memory, GS-FedDAG and AS-FedDAG are unable to produce results on the Pigs and Gene datasets.

Method	Time (↓)				
	Child	Insurance	Alarm	Pigs	Gene
NT-ADMM	31.9	42.2	57.4	8270.6	14149.4
NT-M-ADMM	57.2	80.6	121.5	12611.2	44131.0
GS-FedDAG	3274.6	4322.7	7956.8	OOM	OOM
AS-FedDAG	3052.8	4674.5	7238.4	OOM	OOM
FedPC	2.6	3.7	4.3	3198.5	1069.2
FedCSL (Ours)	**0.7**	**0.9**	**1.2**	**322.6**	**220.9**

"NT-ADMM" refers to NOTEARS-ADMM.
"NT-M-ADMM" refers to NOTEARS-MLP-ADMM.
OOM: OUT OF MEMORY.

Table 1: Running time (in seconds) of each algorithm on the benchmark datasets when the number of clients is set to 12.

clients increases. Compared with the best baseline FedPC, the F1 score of FedCSL is 20%˜42% higher on the Child BN dataset. The SHD values achieved by our method are significantly smaller than those achieved by NOTEARS-ADMM, NOTEARS-MLP-ADMM, GS-FedDAG and AS-FedDAG, since it is hard for these four baselines to select a suitable threshold to prune false directed edges, the causal structures learned by these four baselines often contain a larger number of false edges, leading to inaccurate causal structures.

Table 1 shows the execution time of all methods, and we can see that FedCSL is significantly faster than all the baselines. Especially, on large-scale datasets (e.g., Pigs and Gene), our method is more than 10 times faster than all baselines, indicating that the proposed local-to-global strategy indeed enhances the efficiency of FedCSL. NOTEARS-ADMM, NOTEARS-MLP-ADMM, GS-FedDAG and AS-FedDAG incur significant computational costs on large-scale BN datasets due to the adoption of complex neural network models or sophisticated optimization methods.

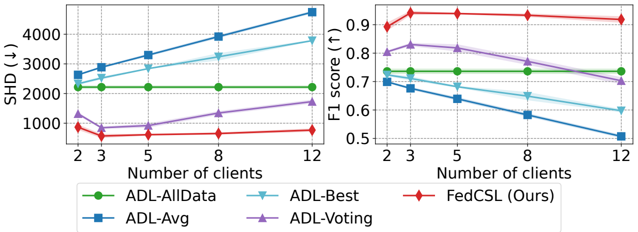


Figure 4: Structure learning results on high-dimensional synthetic datasets with 5,000 variables.

Results on High-dimensional Synthetic Data

Since existing federated causal structure learning methods do no scale up to high-dimensional datasets (e.g. with 5,000 variables), we develop the four new algorithms as baselines to verify the effectiveness and efficiency of our method. From Fig. 4, we can observe that FedCSL outperforms its baselines significantly in terms of SHD and F1 score metrics, surpassing even ADL-AllData. Furthermore, as the number of clients increases (with a reduced allocation of samples per client), the performance of ADL-Avg, ADL-Best and ADL-Voting deteriorates noticeably. However, our method maintains its ideal performance. This is attributed to two factors: on the one hand, during federated global skeleton construction, our designed weighted scoring strategy enhances the accuracy of causal neighbor learning by correcting asymmetric edges. On the other hand, our proposed weighted aggregation strategy helps mitigate the interference caused by erroneous structure learning results at clients with a lower allocation of samples.

The running times of FedCSL and the baselines on high-dimensional datasets are shown Table 2. It is worth noting that, except for ADL-AllData, ADL-Avg and ADL-Best are also executed on a single dataset, since their execution processes at each client are separated. Even so, we find that FedCSL has a minimal difference in runtime compared to ADL-AllData, ADL-Avg and ADL-Best. In comparison to ADL-Voting, FedCSL exhibits a clear efficiency advantage, and as the number of clients increases, the efficiency gap

Method	Time (↓)				
	2 clients	3 clients	5 clients	8 clients	12 clients
ADL-AllData	2141.8	2103.7	2122.1	2167.6	2120.5
ADL-Avg	2176.6	2544.1	3350.6	4888.9	7084.3
ADL-Best	2014.0	2080.7	2126.0	3105.5	4421.2
ADL-Voting	5088.8	9272.0	22373.9	42635.8	86896.9
FedCSL (Ours)	1925.0	2397.8	3368.9	4773.5	6662.8

Table 2: Running time (in seconds) on the synthetic datasets.

Method	SHD (↓)				
	2 clients	3 clients	5 clients	8 clients	12 clients
NT-ADMM	32	32	34	32	29
NT-M-ADMM	21	25	23	32	27
GS-FedDAG	22	20	**19**	20	13
AS-FedDAG	23	20	**19**	19	**12**
FedPC	41	40	40	41	26
FedCSL (Ours)	**12**	**18**	22	**15**	17

"NT-ADMM" refers to NOTEARS-ADMM.
"NT-M-ADMM" refers to NOTEARS-MLP-ADMM.

Table 3: Results on the real Sachs dataset.

between FedCSL and ADL-Voting further widens.

Results on Real-World Data

The experimental results on the real dataset, Sachs (Sachs et al. 2005), are presented in Table 3. From the table, it can be observed that FedCSL achieves the lowest SHD value when the number of clients is 2, 3, and 8. Although our method does not outperform the baselines in terms of performance when the number of clients is 5 and 12, the difference in achieved SHD values is minimal. Since the Sachs dataset is relatively small in scale, the execution time of each algorithm is negligible, and thus it is not reported in Table 3.

Ablation Study

To validate the effectiveness of the proposed weighted aggregation strategy, we conduct ablation experiments in this section. Specifically, we first develop a variant of FedCSL, denoted as "FedCSL w/o weighting", which maintains equivalent weights for all clients (i.e., $w_{c_{k_1}} = w_{c_{k_2}}$ holds for $\forall k_1, k_2 \in \{1, 2, ..., m\}$) throughout the learning process. Then FedCSL is compared with "FedCSL w/o weighting" using five benchmark BN datasets. The results are presented in Fig 5. We observe that FedCSL achieves higher F1 scores and lower SHD values than "FedCSL w/o weighting" on all benchmark BN datasets, demonstrating the effectiveness of our designed weighted aggregation strategy in the horizontal federated learning setting, especially when dealing with the scenario of uneven sample sizes held by different clients.

Validation of Hypothesis 1

In this section, we validate Hypothesis 1 by conducting experiments on the benchmark datasets. Specifically, we first generate three datasets using Child BN, each containing 300, 400, and 500 samples, respectively. Then, we perform causal neighbor learning on these datasets and record the p-values returned by the CI tests when the condition set is empty. For ease of observation, we randomly select 20 CI test results with p-values indicating conditional independence and 20 CI test results with p-values indicating conditional dependence for each dataset. The results are presented in Fig. 6.

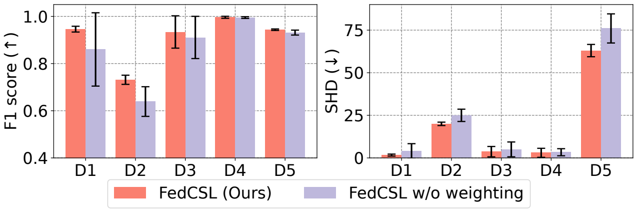


Figure 5: Results of ablation experiments. ("D1", "D2", "D3", "D4", and "D5" represent Child, Insurance, Alarm, Pigs, and Gene BN datasets, respectively.)

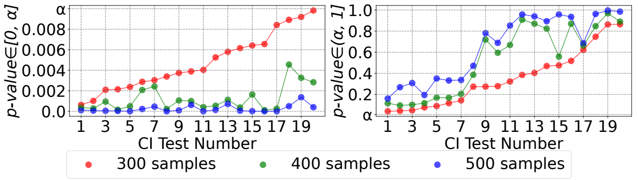


Figure 6: Experimental Verification of Hypothesis 1.

As shown in the left sub-figure of Fig. 6, when two variables are conditional dependent (i.e., the p-values is in $[0, \alpha]$), the obtained p-values on datasets with smaller sample sizes are consistently higher than those on datasets with larger sample sizes. Here, the red points are higher than the green points, and the green points are higher than the blue points. In contrast, as illustrated in the right sub-figure of Fig. 6, when two variables are conditionally independent (i.e., the p-values is in $(\alpha, 1])$, the obtained p-values on datasets with smaller sample sizes are consistently lower than those on datasets with larger sample sizes. Here, the red points are lower than the green points, and the green points are lower than the blue points. Overall, the results presented in Fig. 6 perfectly validate the correctness of Hypothesis 1.

Conclusion

In this paper, we propose a novel method FedCSL, which overcomes the scalability and accuracy issues encountered by existing federated CSL algorithms. Specifically, we design a federated local-to-global learning strategy that enables FedCSL to scale to high-dimensional data. Based on theoretical analysis, we devise a highly privacy-preserving weighted aggregation strategy, which ensures that FedCSL achieves high learning accuracy even in scenarios with uneven sample allocations. Extensive experiments on various types of data demonstrate the accuracy and scalability of our method. In future, we plan to extend this work to more generalized scenarios, such as the scenarios with data heterogeneity and the presence of latent variables.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (under grant 2020AAA0106100), the National Natural Science Foundation of China (under grant 62376087), and China Scholarship Council (under grant 202306690023).

References

Aliferis, C. F.; Statnikov, A.; Tsamardinos, I.; Mani, S.; and Koutsoukos, X. D. 2010. Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(1).

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

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

Chickering, M.; Heckerman, D.; and Meek, C. 2004. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5: 1287–1330.

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

Gao, E.; Chen, J.; Shen, L.; Liu, T.; Gong, M.; and Bondell, H. 2023. FedDAG: Federated DAG Structure Learning. *Transactions on Machine Learning Research*.

Glymour, C.; Scheines, R.; and Spirtes, P. 2014. *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling*. Academic Press.

Guo, X.; Wang, Y.; Huang, X.; Yang, S.; and Yu, K. 2022a. Bootstrap-based Causal Structure Learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 656–665.

Guo, X.; Yu, K.; Cao, F.; Li, P.; and Wang, H. 2022b. Error-aware Markov blanket learning for causal feature selection. *Information Sciences*, 589: 849–877.

Guo, X.; Yu, K.; Liu, L.; Cao, F.; and Li, J. 2022c. Causal feature selection with dual correction. *IEEE Transactions on Neural Networks and Learning Systems*.

Guo, X.; Yu, K.; Liu, L.; Li, P.; and Li, J. 2023. Adaptive Skeleton Construction for Accurate DAG Learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(10): 10526–10539.

Hedström, P.; and Ylikoski, P. 2010. Causal mechanisms in the social sciences. *Annual Review of Sociology*, 36: 49–67.

Huang, J.; Guo, X.; Yu, K.; Cao, F.; and Liang, J. 2023a. Towards Privacy-Aware Causal Structure Learning in Federated Setting. *IEEE Transactions on Big Data*, 9(6): 1525–1535.

Huang, X.; Guo, X.; Li, Y.; and Yu, K. 2023b. A novel data enhancement approach to DAG learning with small data samples. *Applied Intelligence*, 53(22): 27589–27607.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2): 1–210.

Lagani, V.; Triantafillou, S.; Ball, G.; Tegnér, J.; and Tsamardinos, I. 2016. Probabilistic computational causal discovery for systems biology. *Uncertainty in Biology: A Computational Modeling Approach*, 33–73.

Lampinen, A. K.; Roy, N.; Dasgupta, I.; Chan, S. C.; Tam, A.; Mcclelland, J.; Yan, C.; Santoro, A.; Rabinowitz, N. C.; Wang, J.; et al. 2022. Tell me why! Explanations support learning relational and causal structure. In *International Conference on Machine Learning*, 11868–11890. PMLR.

Ma, X.; Zhang, J.; Guo, S.; and Xu, W. 2022. Layer-wised model aggregation for personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10092–10101.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Na, Y.; and Yang, J. 2010. Distributed Bayesian network structure learning. In *2010 IEEE International Symposium on Industrial Electronics*, 1607–1611. IEEE.

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

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

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

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, prediction, and search*. MIT Press.

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

Vowels, M. J.; Camgoz, N. C.; and Bowden, R. 2022. D'ya like DAGs? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4): 1–36.

Wu, L.; Zhao, H.; Li, Z.; Huang, Z.; Liu, Q.; and Chen, E. 2023a. Learning the Explainable Semantic Relations via Unified Graph Topic-Disentangled Neural Networks. *ACM Transactions on Knowledge Discovery from Data*, 17(8): 1–23.

Wu, X.; Jiang, B.; Wang, X.; Ban, T.; and Chen, H. 2023b. Feature Selection in the Data Stream Based on Incremental Markov Boundary Learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Wu, X.; Jiang, B.; Wu, T.; and Chen, H. 2023c. Practical Markov boundary learning without strong assumptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10388–10398.

Wu, X.; Jiang, B.; Yu, K.; Chen, H.; and Miao, C. 2020. Multi-label causal feature selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 6430–6437.

Wu, X.; Jiang, B.; Yu, K.; Chen, H.; et al. 2019. Accurate Markov boundary discovery for causal feature selection. *IEEE Transactions on Cybernetics*, 50(12): 4983–4996.

Wu, X.; Jiang, B.; Zhong, Y.; and Chen, H. 2022. Multi-target Markov boundary discovery: Theory, algorithm, and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4964–4980.

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

Yang, S.; Guo, X.; Yu, K.; Huang, X.; Jiang, T.; He, J.; and Gu, L. 2023. Causal Feature Selection in the Presence of Sample Selection Bias. *ACM Transactions on Intelligent Systems and Technology*, 14(5): 78:1–78:18.

Yu, K.; Guo, X.; Liu, L.; Li, J.; Wang, H.; Ling, Z.; and Wu, X. 2020. Causality-based feature selection: Methods and evaluations. *ACM Computing Surveys*, 53(5): 1–36.

Yu, Y.; Chen, J.; Gao, T.; and Yu, M. 2019. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*, 7154–7163. PMLR.

Yu, Y.; Liu, Q.; Wu, L.; Yu, R.; Yu, S. L.; and Zhang, Z. 2023. Untargeted attack against federated recommendation systems via poisonous item embeddings and the defense. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 4854–4863.

Yuan, C.; Malone, B.; and Wu, X. 2011. Learning optimal Bayesian networks using A* search. In *Proceedings of International Joint Conference on Artificial Intelligence*, 2186–2191.

Zhang, H.; Luo, F.; Wu, J.; He, X.; and Li, Y. 2023. LightFR: Lightweight federated recommendation with privacy-preserving matrix factorization. *ACM Transactions on Information Systems*, 41(4): 1–28.

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