Towards Practical Federated Causal Structure Learning

Zhaoyu Wang[0009−0009−6892−1264], Pingchuan Ma†[0000−0001−7680−2817], and Shuai Wang[0000−0002−0866−0308]

The Hong Kong University of Science and Technology, Hong Kong SAR
1950574@tongji.edu.cn
{pmaab,shuaiw}@cse.ust.hk
†corresponding author

Abstract. Understanding causal relations is vital in scientific discovery. The process of causal structure learning involves identifying causal graphs from observational data to understand such relations. Usually, a central server performs this task, but sharing data with the server poses privacy risks. Federated learning can solve this problem, but existing solutions for federated causal structure learning make unrealistic assumptions about data and lack convergence guarantees. FedC^{2}SL is a federated constraint-based causal structure learning scheme that learns causal graphs using a federated conditional independence test, which examines conditional independence between two variables under a condition set without collecting raw data from clients. FedC^{2}SL requires weaker and more realistic assumptions about data and offers stronger resistance to data variability among clients. FedPC and FedFCI are the two variants of FedC^{2}SL for causal structure learning in causal sufficiency and causal insufficiency, respectively. The study evaluates FedC^{2}SL using both synthetic datasets and real-world data against existing solutions and finds it demonstrates encouraging performance and strong resilience to data heterogeneity among clients.

Keywords: federated learning · Bayesian network · probabilistic graphical model · causal discovery.

1 Introduction

Learning causal relations from data is a fundamental problem in causal inference. Causal structure learning is a popular approach to identifying causal relationships in multivariate datasets, represented as a causal graph. This technique has been successfully applied in various fields such as medicine [3, 33, 37], economics [1], earth science [34] and data analytics [24].

Causal structure learning is performed on a central server with plaintext datasets. However, in applications like clinical data analysis, data may be distributed across different parties and may not be shared with a central server. To address this problem, federated learning is an emerging paradigm that allows data owners to collaboratively learn a model without sharing their data in plaintext [6,10]. However, current federated learning solutions are designed primarily for machine learning tasks that aggregate models trained on local datasets.

Several solutions have been proposed for federated causal structure learning [12, 28, 30, 36]. However, these solutions have prerequisites that may hinder their general applicability. For instance, NOTEARS-ADMM [30], which is the state-of-the-art solution for federated causal structure learning, collects parameterized causal graphs from clients and uses an ADMM procedure to find the consensus causal graph in each iteration. Since local graphs jointly participate in the ADMM procedure, it is non-trivial to employ secure aggregation to protect individual causal graphs, resulting in a considerable sensitive information leak to the central server. Additionally, the assumption that data is generated in a known functional form is deemed unrealistic in many real-life applications.

In general, many solutions attempt to locally learn a causal graph and aggregate them together, but this practice is not optimal for federated causal structure learning. Causal structure learning is known to be error-prone in small datasets, and local datasets may suffer selection bias with respect to the global dataset due to the potential heterogeneity of different clients. The causal graph independently learned from each local dataset may manifest certain biases with respect to the true causal graph of the whole dataset.

To address this issue, we propose a novel federated causal structure learning with constraint-based methods. This paradigm interacts data only with a set of statistical tests on conditional independence and deduces graphical structure from the test results. The key idea of our solution is to provide a federated conditional independence test protocol. Each client holds a local dataset and computes their local statistics, which are then securely aggregated to derive an unbiased estimation of the global statistics. With the global statistics, we can check the global conditional independence relations and conduct constraint-based causal structure learning accordingly.

We evaluate our solution with synthetic data and a real-world dataset and observe better results than baseline federated causal structure learning algorithms, including state-of-the-art methods NOTEARS-ADMM [30], RFCD [28], and four voting-based algorithms. Our solution also shows strong resiliency to client heterogeneity while other baseline algorithms encounter notable performance downgrades in this setting. Furthermore, our solution facilitates causal feature selection (CFS) and processes real-world data effectively.

In summary, we make the following contributions: (1) we advocate for federated causal structure learning with constraint-based paradigms; (2) we design a federated conditional independence test protocol to minimize privacy leakages and address client heterogeneity; and, (3) we conduct extensive experiments to assess the performance of our solution on both synthetic and real-world datasets. We release our implementation, FedC^{2}SL, on `https://github.com/wangzhaoyu07/FedC2SL` for further research and comparison.

2 Preliminary

In this section, we review preliminary knowledge of causal structure learning.
Notations. Let X and $\boldsymbol{X}$ represent a variable and a set of variables, respectively. In a graph, a node and a variable share the same notation. The sets of nodes and edges in a causal graph G are denoted as V_G and E_G, respectively. The notation $X \to Y \in E_G$ indicates that X is a parent of Y, while $X \leftrightarrow Y \in E_G$ indicates that X and Y are connected by a bidirected edge. The sets of neighbors and parents of X in G are denoted as $N_G(X)$ and $Pa_G(X)$, respectively. The notation $[K] \coloneqq \{1, \cdots, K\}$ is defined as the set of integers from 1 to K.

2.1 Causal Structure Learning

In causal inference, the relationship between data is often presented as a causal graph, which can take the form of a directed acyclic graph (DAG) or maximal ancestral graph (MAG). These two representations are used to depict causal relationships under different assumptions. In the following paragraphs, we introduce the corresponding causal graphs and formalize these canonical assumptions.

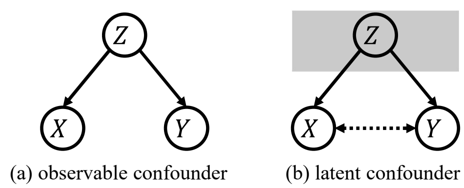


Fig. 1: Examples of observable and latent confounders.

Graphical Representation. Causal relations among variables in a multivariate dataset can be depicted using a causal graph. The causal graph can either be a directed acyclic graph (DAG), where adjacent variables are connected by a directed edge, or a mixed acyclic graph (MAG), which allows for bidirected edges to indicate shared latent confounders between two variables. In the DAG format, if a latent confounder is not observed, statistical associations between variables can exist without their true causal relations being well-represented. MAG overcomes this shortcoming and can be constructed from a true DAG and a set of latent variables, using a set of construction criteria [50]. See Fig. 1 (a) for an example of a DAG depicting a directed edge ($Z \to X$).

latent confounder can be e.g. genetic predisposition to smoke and also to get lung cancer. this is not observer. whereas observable confounder Z could be Age. Age may affect smoking propoensity and also lung cancer risk

Causal Sufficiency. Learning a DAG assumes a causally sufficient set of variables [38]. $\boldsymbol{X}$ is causally sufficient if there is no hidden cause $Z \notin \boldsymbol{X}$ causing more than one variable in $\boldsymbol{X}$. However, real-world data may not satisfy this assumption. MAG addresses this issue by introducing a bidirected edge $\leftrightarrow$. See Fig. 1 (b) for an example where a bidirected edge between X, Y due to the absence of Z.

Global Markov Property (GMP). The Global Markov Property (GMP) [21] connects graphical structures and statistical properties. It can be stated as: $X \perp\!\!\!\perp_G Y \mid \boldsymbol{Z} \implies X \perp\!\!\!\perp Y \mid \boldsymbol{Z}$. Here, $\perp\!\!\!\perp_G$ represents graphical separation and $\perp\!\!\!\perp$ represents statistical conditional independence in the joint probability distribution $P_{\boldsymbol{X}}$. D-separation is a structural constraint for directed acyclic graphs (DAGs), while m-separation is a constraint for mixed graphs (MAGs). We present their definitions in Supplementary Material.

Faithfulness Assumption. Faithfulness assumption states that conditional independence on the joint distribution implies d-separation (or m-separation) on the causal graph. Formally, $X \perp\!\!\!\perp Y \mid \boldsymbol{Z} \implies X \perp\!\!\!\perp_G Y \mid \boldsymbol{Z}$

In the remainder of the paper, we assume GMP and faithfulness assumption always hold. Moreover, we assume causal sufficiency in FedPC and propose FedFCI that is also tolerant to causally insufficient data.

Markov Equivalence Class (MEC). Given the Markov condition and faithfulness assumption, statistical tests can be performed on data to deduce graph structures through graphical separation constraints. However, inferring the full structure of a causal graph from data is difficult and can lead to multiple causal graphs being compatible with the constraints deduced from conditional independence. To address this, causal structure learning algorithms aim to recover a MEC, which summarizes a set of causal graphs sharing the same set of d-separations (or m-separations) [32]. The MEC is represented as a CPDAG for DAG learning and as a PAG for MAG learning [50]. FedC^{2}SL follows standard conventions [23, 32, 50] in recovering the MEC of a given dataset.

Constraint-based Causal Structure Learning. Constraint-based methods are commonly used for causal structure learning, identifying the MEC from observational datasets. The PC algorithm (see details in Supplementary Material) is a representative constraint-based causal structure learning algorithm. This algorithm involves two phases: learning the causal skeleton and orienting the edges. During the former phase, the adjacency relations between variables are learned and an undirected graph is created. In this graph, the edges represent the underlying causal graph's skeleton. In the latter phase, a set of orientation rules is applied to assign a causal direction to the undirected edges of the skeleton. In comparison to the PC algorithm, which performs DAG learning, the FCI algorithm [50] is designed for MAG learning, incorporating another set of orientation rules while using a similar skeleton learning procedure of the PC algorithm.

3 Research Overview

This section presents the research overview. We begin by providing an overview of FedC^{2}SL in Sec. 3.1, covering the problem setup and threat model. Sec. 3.2 provides a comparison between our solution and existing approaches.

3.1 Problem Setup

In this paper, we consider two causal discovery problems: FedPC and FedFCI.

FedPC. Assuming causal sufficiency, FedPC involves a DAG $G = (V, E)$ that encodes the causal relationships among a variable vector $\boldsymbol{X} = \{X_1, \cdots, X_d\}$

with a joint probability distribution $P_{\boldsymbol{X}}$ satisfying the Global Markov Property (GMP) with respect to G, and G is faithful to $P_{\boldsymbol{X}}$.

FedFCI. In causal insufficient data, FedFCI involves a MAG $M = (V, E)$ that encodes the causal relationships among a variable vector $\boldsymbol{X} = \{X_1, \cdots, X_d\}$ with a joint probability distribution $P_{\boldsymbol{X} \cup L}$, where L is a set of unknown latent variables. Here, $P_{\boldsymbol{X}}$ is the observable distribution with $P_{\boldsymbol{X} \cup L}$ being marginalized on L. If L is empty, then the setting is equivalent to FedPC. We assume that $P_{\boldsymbol{X}}$ satisfies GMP with respect to G and M is faithful to $P_{\boldsymbol{X}}$.

We now describe the setting of clients that are identical for either FedPC or FedFCI. Suppose that there are K local datasets $\mathcal{D} \coloneqq \{D^1, \cdots, D^K\}$ and $D^i = \{\boldsymbol{x}_1^i, \cdots, \boldsymbol{x}_{n_i}^i\}$. We denote $\boldsymbol{x}_{j,k}^i$ be the k-th element of the j-th record in the i-th local dataset. Each record in the global dataset $\mathcal{D}$ is sampled *i.i.d.* (independent and identically distributed) from $P_{\boldsymbol{X}}$. We allow for selection bias on client local datasets as long as the global dataset is unbiased with respect to $P_{\boldsymbol{X}}$, which is one of the main challenges in federated learning [16]. For example, local datasets from different hospitals may be biased on different patient subpopulations. However, with a sufficient number of clients, the global dataset (by pooling all local datasets) becomes unbiased. This assumption is weaker than the Invariant DAG Assumption in DS-FCD [12]. We borrow the concept from general causal structure learning [42] and formally define this property as follows.

Definition 1 (Client Heterogeneity). *Let $\{X_1, \cdots, X_d\}$ be the visible variables in the dataset and G be the ground-truth causal graph. To represent client-wise heterogeneity, we assume that there is an implicit surrogate variable $C : [K]$ be the child variable of $S \subseteq \{X_1, \cdots, X_d\}$ in an augmented causal graph and the i-th client holds the records with $C = i$. When $S = \emptyset$, the local datasets are homogeneous.*

The overall goal is to recover MEC of G or M from distributed local datasets in the presence of client heterogeneity while minimizing data leakages.

Threat Model. Our threat model aligns with the standard setting [47]. We assume that all parties including the central server and clients are honest but curious, meaning that they will follow the protocol but are interested in learning as much private information as possible from others. We are concerned with the leakage of private client data, and we do not consider any coalitions between participants. We will show later that FedC^{2}SL is resilient to client dropouts, although we do not explicitly consider this during algorithm design.

Security Objective. The federated learning paradigms aim to prevent raw data sharing, and only aggregated results are released [45]. We aim to achieve MPC-style security to ensure that the semi-honest server only knows the aggregated results and not individual updates. To establish this property formally, we define client indistinguishability in federated causal structure learning.

Definition 2 (Client Indistinguishability). *Let $\boldsymbol{x} \in D^i$ be a record that only exists in the i-th client (i.e., $\forall j \neq i, \forall \boldsymbol{x}' \in D^j, \boldsymbol{x} \neq \boldsymbol{x}'$). Let $\mathbb{P}(A)$ be the public* *knowledge (e.g., intermediate data and final results) revealed in the protocol A. A is said to be client indistinguishable for an adversary if* $\forall i, j \in [K], P(\boldsymbol{x} \in D^i \mid \mathbb{P}(A)) = P(\boldsymbol{x} \in D^j \mid \mathbb{P}(A))$.

3.2 Comparison with Existing Solutions

In this section, we review existing solutions and compare them with FEDC^{2}SL. We summarize existing works and FEDC^{2}SL, in terms of assumptions, application scope, and leakage, in Table 1.

Table 1: Comparing existing works and FEDC^{2}SL.

Solution	Input	Assumption	Heterogeneity	Application Scope	Client Leakage	Global Leakage
NOTEARS-ADMM [30]	Data	Additive Noise	✗	DAG	Individual Graph + Parameter	Graph + Parameter
DS-FCD [12]	Data	Additive Noise	✗	DAG	Individual Graph + Parameter	Graph + Parameter
RFCD [28]	Data	Additive Noise	✗	DAG	Individual Graph Fitness	Graph
K2 [36]	Data + Order	Faithfulness	✓	DAG	Aggregated Fitness	Graph
FEDC^{2}SL	Data	Faithfulness	✓	DAG & MAG	Aggregated Low-dim. Distribution	Graph

Comparison of Prerequisites. Most federated causal structure learning solutions assume additive noise in the data generating process, which is considered stronger than the faithfulness assumption in K2 [36] and FEDC^{2}SL. However, K2 requires prior knowledge of the topological order of nodes in a ground-truth DAG, which is impractical. Additionally, solutions that learn local graphs independently are intolerant of client heterogeneity, as their performance would degrade arbitrarily in theory. FEDC^{2}SL is the only solution that supports MAG learning, allowing for causal structure learning on causally insufficient data, making it a more practical option. (See Sec. 2.1 for more details.)

Comparison on Privacy Protection. We review the privacy protection mechanisms in proposed solutions. On the client side, NOTEARS-ADMM [30] and DS-FCD [12] require clients to update the local causal graph and corresponding parameters in plaintext to the global server. These graph parameters consist of multiple regression models trained on local datasets, which are vulnerable to privacy attacks on ML models. RFCD [28] requires clients to send the fitness score of global causal graphs on the local dataset, which poses a privacy risk for adversaries to infer the source of particular data samples and violate the client indistinguishability. In contrast, K2 [36] and FEDC^{2}SL use secure aggregation or secure multi-party computation protocols such that only aggregated results are revealed. K2 employs a score function to measure the fitness of a local structure and the global structure is established by selecting the best local structure in a greedy manner. The score function is computed over the distributed clients with MPC schemes such that individual updates are protected. FEDC^{2}SL uses a constraint-based strategy to learn the causal graph and securely aggregates the data distribution marginalized over multiple low-dimensional subspaces to the global server. The marginalized low-dimensional distributions are strictly less informative than NOTEARS-ADMM and DS-FCD. The global server asserts conditional independence on the aggregated distributions and deduces graphical separations by faithfulness accordingly.

Asymptotic Convergence. FedC^{2}SL is inherited from constraint-based methods that offers asymptotic convergence to the MEC of the ground-truth causal graph under certain assumptions. In contrast, NOTEARS-ADMM and DS-FCD use continuous optimization to the non-convex function and only converge on stationary solutions. RFCD and K2 use greedy search over the combinatorial graph space, which does not provide global convergence guarantees.

4 FedC^{2}SL

In this section, we present FedC^{2}SL, a novel federated causal structure learning algorithm with minimized privacy leakage compared to its counterparts.

4.1 Causal Structure Learning

As discussed in Sec. 2.1, the causal graph is learned by testing conditional independence in the dataset. A MEC of the causal graph contains all conditional independence, as per GMP and faithfulness assumption. Moreover, the MEC can be recovered from the set of all conditional independence relations. Therefore, the set of all conditional independence relations in a dataset is both necessary and sufficient to represent the MEC of the underlying causal graph.

Remark 1. Under GMP and faithfulness assumption, a Markov equivalence class of causal graph encodes all conditional independence relations among data. Once the Markov equivalence class is revealed, all conditional independence relations are revealed simultaneously. Therefore, the conditional independence stands for the minimal information leak of federated causal structure learning.

Instead of creating specific federated causal structure learning methods, we propose using a federated conditional independence test procedure. This procedure is fundamental to all constraint-based causal structure learning algorithms, such as the PC algorithm. By implementing our federated version, we can replace the centralized conditional independence tester in any existing constraint-based causal structure learning algorithm and make it federated. In this paper, we apply our federated conditional independence test procedure to two well-known causal structure learning algorithms, namely, FedPC and FedFCI, which are based on the PC algorithm [38] and FCI algorithm [38, 50], respectively.

4.2 Federated Conditional Independence Test

To enhance privacy protection, a multiparty secure conditional independence test that releases only the conditional independence relations would be ideal. However, implementing such a solution using MPC would result in impractical computational overheads for producing real-world datasets. Therefore, we propose a practical trade-off that boosts computation efficiency while causing negligible privacy leakage on relatively insensitive information.

To introduce our federated conditional independence test protocol, we first explain how to test conditional independence in a centralized dataset. Consider three random variables X, Y, and Z from a multivariate discrete distributions. The conditional independence of X and Y given Z is defined as follows:

Definition 3 (Conditional Independence). *X and Y are conditionally independent given Z if and only if, for all possible $(x, y, z) \in (X, Y, Z)$, $P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z)$.*

While Def. 3 is straightforward to verify, it is non-trivial to statistically test this property with finite samples. The most popular way is to use χ^2-test [4], whose null hypothesis and alternative hypothesis are defined as follows:

$$H_0 : X \perp\!\!\!\perp Y|Z, H_1 : X \not\perp\!\!\!\perp Y|Z \tag{1}$$

The statistic $\hat{Q}$ is computed as $\hat{Q} = \sum_{x,y,z} \frac{(v_{xyz} - \frac{v_{xz}v_{yz}}{v_z})^2}{\frac{v_{xz}v_{yz}}{v_z}}$ where v_{xyz} is the number of samples with $X = x$, $Y = y$ and $Z = z$; and so on. Under null hypothesis H_0, $\hat{Q}$ follows a $\chi^2_{\texttt{dof}}$ distribution where $\texttt{dof} = \sum_{z \in Z}(|X_{\boldsymbol{Z}=z}| - 1)(|Y_{\boldsymbol{Z}=z}| - 1)$ is the degree of freedom and $|X|, |Y|$ denote the cardinality of the multivariate discrete random variable. Let $1 - \alpha$ be the significance level. The null hypothesis is rejected if $\hat{Q} > \chi^2_{\texttt{dof}}(1 - \alpha)$.

Why a Voting Scheme Is Not Suitable? One potential approach to test conditional independence in a federated setting is to perform standard χ^2-tests on each client independently and use the voted local conditional independence as the conditional independence on the global dataset. However, this approach is not feasible for two reasons. Firstly, the χ^2-test requires that all v_{xyz} are larger than 5 to ensure its validity [40]. This requirement is often unattainable on small local datasets, leading to inaccurate test results. Secondly, even if the requirement is met, the voting result may not reflect the global conditional independence in the presence of selection bias on the client dataset. As will be shown in Sec. 5, simple voting strategies often yield inaccurate results.

To preserve privacy while computing $\hat{Q}$ on the global dataset, we can perform secure aggregation over the four counts $(v_z, v_{xz}, v_{yz}, v_{xyz})$ instead of using the voting scheme. Securely summing up $v^1_{xyz}, \cdots, v^K_{xyz}$ from all clients can obtain v_{xyz}. However, if Z contains multiple variables, releasing v_{xyz} could raise privacy concerns due to its encoding of the joint distribution of multiple variables. We discuss the privacy implications of releasing such high-dimensional distribution in the following paragraph.

High-Dim. Distribution vs. Low-Dim. Distribution. We note that high-dimensional distribution is more sensitive than low-dimensional distribution, which allows adversaries to localize a particular instance (e.g., patient of a minority disease). Hence, we anticipate to avoid such leakages. In contrast, the joint distribution marginalized over low-dimensional subspace is generally less sensitive. It can be deemed as a high-level summary of data distribution and individual privacy is retained on a reasonable degree.

Algorithm 1: $\texttt{Fed-CI}(X \perp\!\!\!\perp Y \mid \boldsymbol{Z})$

Input: Data in K clients: $\mathcal{D} := \{D^1, \cdots, D^K\}$; Statistical Significance: $1-\sigma$.
Output: Whether reject $X \perp\!\!\!\perp Y \mid \boldsymbol{Z}$.

if $\boldsymbol{Z} = \emptyset$ **then** $\boldsymbol{Z} \leftarrow \{\mathbb{1}\}$;
foreach $z \in \boldsymbol{Z}$ **do**
  // i) compute marginal distribution
  // client side:
  let v_z^i be the count of $\boldsymbol{Z} = z$ on D^i;
  let v_{xz}^i, v_{yz}^i be the count of $X = x$ (or $Y = y$) with $\boldsymbol{Z} = z$ on D^i;
  // server side:
  $v_z \leftarrow \texttt{SecureAgg}(\{v^i\}_{i\in[K]})$;
  foreach $x \in X$ **do** $v_x \leftarrow \texttt{SecureAgg}(\{v_{xz}^i\}_{i\in[K]})$;
  foreach $y \in Y$ **do** $v_y \leftarrow \texttt{SecureAgg}(\{v_{yz}^i\}_{i\in[K]})$;
  foreach $(x, y) \in X, Y$ **do** broadcast $\bar{v}_{xyz} = \frac{v_{xz}v_{yz}}{v_z}$;
  sample $\boldsymbol{P}$ from $\mathcal{Q}_{2,0}^{l\times m}$ and broadcast $\boldsymbol{P}$;
  // ii) compute χ^2 statistics
  // client side:
  $\boldsymbol{u}_z^i[\mathbb{I}(x, y)] \leftarrow \frac{v_{xyz}^i - \frac{\bar{v}_{xyz}}{K}}{\sqrt{\bar{v}_{xyz}}}$;
  $\boldsymbol{e}^i \leftarrow \boldsymbol{P} \times \boldsymbol{u}_{\boldsymbol{z}}^{\boldsymbol{i}}$;
  // server side:
  $\boldsymbol{e} \leftarrow \texttt{SecureAgg}(\{\boldsymbol{e}^i\}_{i\in[K]})$;
  $\hat{Q}_z \leftarrow \frac{\sum_{k=1}^{l} |e_k|^{2/l}}{(\frac{2}{\pi}\Gamma(\frac{2}{l})\Gamma(1-\frac{1}{l})\sin(\frac{\pi}{l}))^l}$;
  $\texttt{dof}_z \leftarrow (|X_{\boldsymbol{Z}=z}| - 1)(|Y_{\boldsymbol{Z}=z}| - 1)$;
end
// iii) aggregate χ^2 statistics
$\hat{Q} \leftarrow \sum \hat{Q}_z$;
$\texttt{dof} \leftarrow \sum \texttt{dof}_z$;
if $\hat{Q} > \chi^2_{\texttt{dof},1-\sigma}$ **then return** *reject null hypothesis*;
else return *fail to reject null hypothesis*;

To alleviate the direct release of high-dimensional distributions, we leverage the idea in [45] to recast $\hat{Q}$ statistic into a second frequency moment estimation problem and employ random projection to hide the distribution. Specifically, let $\bar{v}_{xyz} = \frac{v_{xz}v_{yz}}{v_z}$. For each client, we compute $\boldsymbol{u}_z^i[\mathbb{I}(x, y)] = \frac{v_{xyz}^i - \frac{\bar{v}_{xyz}}{K}}{\sqrt{\bar{v}_{xyz}}}$ where $\mathbb{I} : [|X|] \times [|Y|] \to [|X||Y|]$ is an index function. The $\hat{Q}$ can be rewritten as

$$
\begin{aligned}
\hat{Q} &= \sum_{x,y,z} \frac{(v_{xyz} - \frac{v_{xz}v_{yz}}{v_z})^2}{\frac{v_{xz}v_{yz}}{v_z}} = \sum_{z}\sum_{x,y} \left(\frac{v_{xyz} - \bar{v}_{xyz}}{\sqrt{\bar{v}_{xyz}}}\right)^2 \\
&= \sum_{z} \| \sum_{i\in[K]} \boldsymbol{u}_z^i \|_2^2 = \sum_{z} \|\boldsymbol{u}_z\|_2^2
\end{aligned}
\tag{2}
$$

It is worth noting that the above recasting does not obviously conceal v_{xyz} because it can still be derived from $\boldsymbol{u}_z$. To protect $\boldsymbol{u}_z^i$, a random projection is employed to encode $\boldsymbol{u}_z^i$ into $\boldsymbol{e}^i$ and a geometric mean estimation is performed over the encoding. Then, the main result of [22,45] implies the following theorem.

Theorem 1. *Let $\boldsymbol{P}$ be a projection matrix whose values are independently sampled from a α-stable distribution [15] $\mathcal{Q}_{2,0}^{l\times m}$ ($m = |X||Y|$), $\boldsymbol{e}^i = P \times \boldsymbol{u}_z^i$ be the encoding on the i-th client and $\boldsymbol{e} = \sum_{i\in[K]} \boldsymbol{e}^i$ be the aggregated encoding. $\hat{d}_{(2),gm} = \frac{\sum_{k=1}^{l} |e_k|^{2/l}}{(\frac{2}{\pi}\Gamma(\frac{2}{l})\Gamma(1-\frac{1}{l})\sin(\frac{\pi}{l}))^l}$ is the unbiased estimation on $\|\boldsymbol{u}_z\|_2^2$.*

Accordingly, we can compute $\hat{Q}_z$ for each $z \in Z$ and sum them up to obtain $\hat{Q}$. Using secure aggregation, the joint distribution of X, Y, Z on local datasets is already perfectly invisible to the central server. The encoding scheme in the above theorem provides additional privacy protection to the distribution on the global dataset. Specifically, under appropriate parameters, after receiving the aggregated encoding $\boldsymbol{e}$, the server cannot revert back to the original $\boldsymbol{u}_z$. Indeed, given $\boldsymbol{e}$, $\boldsymbol{u}_z$ is concealed into a subspace with exponential feasible solutions according to Theorem 2 in [45]. We now outline the workflow of our federated conditional independence test protocol in Alg. 1. To make Alg. 1 compatible to empty condition set (i.e., $\boldsymbol{Z} = \emptyset$), we add a dummy variable $\mathbb{1}$ to $\boldsymbol{Z}$ (line 1) and the subsequent loop (lines 2–21) only contains one iteration applied on the entire (local) datasets (e.g., v_z^i is the count of total samples in D^i, and so on). In each iteration where a possible value of $\boldsymbol{Z}$ is picked, each client counts $v_z^i, v_{xz}^i, v_{yz}^i$ privately (lines 5–6) and securely aggregates to the server (lines 8–10). The server then broadcasts $\bar{v}_{xyz} = \frac{v_{xz}v_{yz}}{v_z}$ for each $(x, y) \in X, Y$ and the projection matrix $\boldsymbol{P}$ to all clients (lines 11-12). Then, the client computes $\boldsymbol{u}_z^i$ and generates $\boldsymbol{e}^i$ (lines 15–16). The server aggregates encodings (line 18), perform geometric mean estimation to derive $\hat{Q}_z$ (line 19) and compute degree of freedom $\texttt{dof}_z$ (line 20). After enumerating all $z \in Z$, the total χ^2 statistics and the total degree of freedom is computed by summing $\hat{Q}_z$ and $\texttt{dof}_z$ up, respectively (lines 23–24). Finally, $\hat{Q}$ is compared against $\chi^2_{\texttt{dof},1-\sigma}$ and Alg. 1 decides whether to reject null hypothesis (lines 25–26).

5 Evaluation

In this section, we evaluate FedC^{2}SL to answer the following three research questions (RQs): **RQ1: Effectiveness.** Does FedC^{2}SL effectively recover causal relations from data with different variable sizes and client numbers? **RQ2: Resiliency.** Does FedC^{2}SL manifest resiliency in terms of client dropouts or client heterogeneity? **RQ3: Real-world Data.** Does FedC^{2}SL identify reasonable causal relations on real-world data? We answer them in the following sections.

5.1 Experimental Setup

Baselines. We compare the performance of FedC^{2}SL with seven baselines, including two SOTA methods: NOTEARS-ADMM [30] and RFCD [28]. We also implement two baseline algorithms, PC-Voting and PC-CIT-Voting, which aggregate and vote on local causal graphs to form a global causal graph. Additionally, we compare with the centralized PC algorithm [38], as well as FCI algorithm and two voting-based baselines (FCI-Voting and FCI-CIT-Voting). We report the hyperparameters in Supplementary Material.

Dataset. We evaluate FEDC^{2}SL on synthetic and real-world datasets. We describe the generation of synthetic datasets in Supplementary Material. We use the discrete version of the Sachs dataset [35], a real-world dataset on protein expressions involved in human immune system cells.

Metrics. We use Structural Hamming Distance (SHD) between the Markov equivalence classes of learned causal graph and the ground truth (lower is better). We also record the processing time. For each experiment, we repeat ten times and report the averaged results.

5.2 Effectiveness

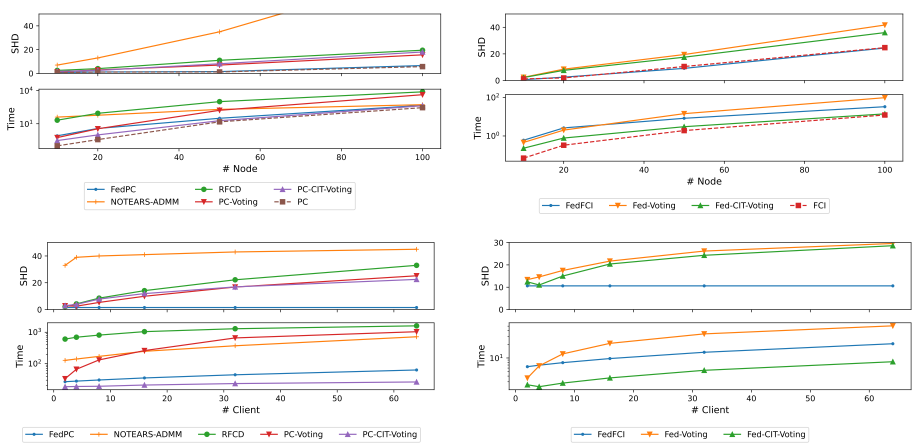


Fig. 2: Performance on different variable sizes and client numbers.

High-dimensional datasets pose challenges for causal structure learning. We evaluate the performance of FEDC^{2}SL on datasets with varying variable sizes and fixed client size ($K = 10$) in Fig. 2. We report the results for federated causal structure learning for DAG and MAG. We observe that FEDC^{2}SL consistently outperforms all other methods (excluding its centralized version) in terms of SHD on all scales. Its accuracy is closely aligned with PC and FCI (i.e., its centralized version), indicating negligible utility loss in the federated procedure.

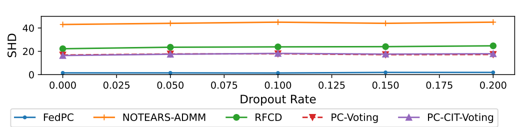
Furthermore, the processing time of FEDC^{2}SL is slightly higher but acceptable and often lower than other counterparts. On datasets with 100 variables, FedC^{2}SL shows a much lower SHD than other federated algorithms, indicating its scalability to high-dimensional data. In contrast, NOTEARS-ADMM performs poorly on datasets with 100 variables due to its assumption on additive noise being violated in discrete datasets, which is further amplified by high-dimensional settings.

We also studied the effectiveness of FedC^{2}SL with different client sizes ($K \in \{2, 4, 8, 16, 32, 64\}$) under a fixed variable size ($d = 50$) in Fig. 2. With the growth of client size, most algorithms show an increasing trend in terms of SHD. However, FedC^{2}SL consistently has the lowest SHD with a mild increase of processing time. In contrast, local causal graph learning-based methods have notable difficulty in handling large client sizes due to the low stability of local datasets and reaching a high-quality consensus on the global causal graph.

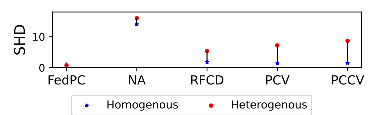
Answer to RQ1: *FedC^{2}SL effectively recovers causal graphs from federated datasets with high accuracy for varying variable sizes and client sizes, outperforming existing methods and having negligible utility loss.*

5.3 Resiliency

We evaluate the performance of different algorithms in federated learning with respect to client dropouts and heterogeneous datasets.



(a) Resiliency to client dropouts.



(b) Resiliency to client heterogeneity.

Fig. 3: Resiliency evaluation.

In terms of resiliency to client dropouts, most algorithms, including our FedC^{2}SL, do not explicitly consider it in their design. However, our experiments show that the SHD of FedC^{2}SL and other algorithms does not notably downgrade even if 20% clients drop out. This emphasizes the robustness of de facto causal structure learning algorithms to client dropouts.

Regarding resiliency to client heterogeneity, FedPC performs consistently well in both homogeneous and heterogeneous datasets ($d = 20, K = 4$). Notably, FedPC demonstrates negligible performance degradation in the presence of client heterogeneity, while other solutions, such as NOTEARS-ADMM and RFCD, suffer notable increases in SHD (on average, 4.7 increase on SHD). This limitation results from their assumption that local datasets accurately represent the joint probability distribution, which is invalid under heterogeneity. Actually, the local causal graph could arbitrarily diverge from the true causal graph.

Answer to RQ2: *FedC^{2}SL shows resilience to both client dropouts and client heterogeneity. Compared to other solutions, FedC^{2}SL consistently performs well in homogeneous and heterogeneous datasets.*

5.4 Real-world Data

We evaluate FedC^{2}SL's performance on the protein expression dataset from the real-world dataset, Scahs [35], which contains 853 samples and 11 variables with a ground-truth causal graph having 17 edges. We split the dataset into $K \in \{2, 4, 8, 16, 32, 64\}$ clients and perform federated causal structure learning. Each algorithm runs ten times for each setting and we report the average results in Fig. 4. The results show that FedC^{2}SL demonstrates the best and highly stable performance on this dataset compared to other algorithms.

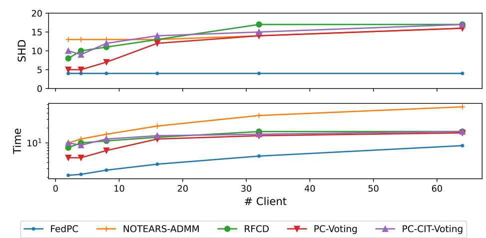
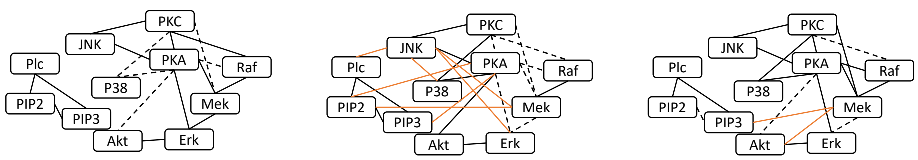


Fig. 4: Performance on the Sachs dataset.

With 64 clients, FedC^{2}SL obtains a minimal SHD of 5.6 while the minimal SHD of other algorithms is 15.7. This indicates that most edges are incorrect in causal graphs learned by previous algorithms. We interpret that FedC^{2}SL offers a unique advantage on learning from federated small datasets.



(a) FedPC (SHD=5) (b) NOTEARS-ADMM (SHD=14) (b) RFCD (SHD=6)

Fig. 5: Causal graphs learned by FedPC, NOTEARS-ADMM, and RFCD. Black solid lines denote correct edges learned by the algorithm; orange lines denote erroneous edges learned by the algorithm; dashed lines denote missing edges.

We present the Markov equivalence classes of causal graphs learned by FedPC, NOTEARS-ADMM, and RFCD under two clients in Fig. 5. In general, FedPC generates the most accurate causal graph with the lowest SHD without any erroneous edge. In contrast, both NOTEARS-ADMM and RFCD have incorrect edges, and NOTEARS-ADMM generates a considerable number of erroneous edges, which would significantly undermine human comprehension of the underlying causal mechanisms behind the data.

Answer to RQ3: *FedC^{2}SL outperforms other methods on the real-world dataset, Sachs, demonstrating the best and highly stable performance with a much lower SHD.*

6 Related Work

Private Causal Inference. Several studies have focused on privacy protections in the causal inference process. Xu et al. [46], Wang et al. [44], and Ma et al. [25] independently propose differentially private causal structure learning methods. Kusner et al. [20] present a differentially private additive noise model for inferring pairwise cause-effect relations, while Niu et al. [31] propose a differentially private cause-effect estimation algorithm. Murakonda et al. [29] study the privacy risks of learning causal graphs from data.

Federated Statistical Tests. The federated χ^2 test [45] is closely related to our work. It is a federated *correlation test*, whereas the χ^2-test in FedC^{2}SL is designed for *conditional independence test*. Our work applies federated statistical tests to enable practical federated causal structure learning, a crucial step in understanding the causal relations of data and enabling causal inference. Bogdanov et al. [5] design an MPC-based federated Student's t-test protocol, while Yue et al. [48] propose a federated hypothesis testing scheme for data generated from a linear model. Furthermore, Gaboardi et al. [11] use local differential privacy to secure the χ^2-test, and Vepakomma et al. [41] propose a differentially private independence testing across two parties.

Federated Machine Learning. Federated learning refers to the process of collaboratively training a machine learning model from distributed datasets across clients and has been studied extensively [16]. McMahan et al. [26] originally coined the term, and since then, there have been various proposals [14, 17, 18, 39, 47] to address practical challenges, such as communication costs and non-IID data across different clients. These proposals include update quantization [2, 19], fine-tuning homomorphic encryption precision [49], and optimizing non-IID data [9, 27, 43].

7 Conclusion

In this paper, we propose FedC^{2}SL, a federated constraint-based causal structure learning framework. FedC^{2}SL is the first work that applies federated conditional independence test protocol to enable federated causal structure learning and is tolerant to client heterogeneity. We instantiate two algorithms with FedC^{2}SL, namely FedPC and FedFCI, to handle different assumptions about data. Through extensive experiments, we find FedC^{2}SL manifests competitive performance on both synthetic data and real-world data.

Acknowledgement

We thank the anonymous reviewers for their insightful comments. We also thank Qi Pang for valuable discussions. This research is supported in part by the HKUST 30 for 30 research initiative scheme under the the contract Z1283 and the Academic Hardware Grant from NVIDIA.

References

1. Addo, P.M., Manibialoa, C., McIsaac, F.: Exploring nonlinearity on the co2 emissions, economic production and energy use nexus: A causal discovery approach. Energy Reports **7**, 6196–6204 (2021)
2. Amiri, M.M., Gunduz, D., Kulkarni, S.R., Poor, H.V.: Federated learning with quantized global model updates. arXiv preprint arXiv:2006.10672 (2020)
3. Belyaeva, A., Squires, C., Uhler, C.: Dci: learning causal differences between gene regulatory networks. Bioinformatics **37**(18), 3067–3069 (2021)
4. Bishop, Y.M., Fienberg, S.E., Fienberg, S.E., Holland, P.W.: Discrete multivariate analysis (1976)
5. Bogdanov, D., Kamm, L., Laur, S., Pruulmann-Vengerfeldt, P., Talviste, R., Willemson, J.: Privacy-preserving statistical data analysis on federated databases. In: Annual Privacy Forum. pp. 30–55. Springer (2014)
6. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K.: Practical secure aggregation for privacy-preserving machine learning. In: proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. pp. 1175–1191 (2017)
7. Buntine, W.: Theory refinement on bayesian networks. In: Uncertainty in Artificial Intelligence: Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence, UCLA, at Los Angeles, July 13-15, 1991. p. 52. Elsevier (2014)
8. Cawley, G.C.: Causal & non-causal feature selection for ridge regression. In: Causation and Prediction Challenge. pp. 107–128. PMLR (2008)
9. Chai, Z., Chen, Y., Zhao, L., Cheng, Y., Rangwala, H.: Fedat: A communication-efficient federated learning method with asynchronous tiers under non-iid data. arXiv preprint arXiv:2010.05958 (2020)
10. Fereidooni, H., Marchal, S., Miettinen, M., Mirhoseini, A., Möllering, H., Nguyen, T.D., Rieger, P., Sadeghi, A.R., Schneider, T., Yalame, H., et al.: Safelearn: secure aggregation for private federated learning. In: 2021 IEEE Security and Privacy Workshops (SPW). pp. 56–62. IEEE (2021)
11. Gaboardi, M., Rogers, R.: Local private hypothesis testing: Chi-square tests. In: International Conference on Machine Learning. pp. 1626–1635. PMLR (2018)
12. Gao, E., Chen, J., Shen, L., Liu, T., Gong, M., Bondell, H.: Federated causal discovery. arXiv preprint arXiv:2112.03555 (2021)
13. Gilbert, E.N.: Random graphs. The Annals of Mathematical Statistics **30**(4), 1141–1144 (1959), `http://www.jstor.org/stable/2237458`
14. He, C., Li, S., So, J., Zhang, M., Wang, H., Wang, X., Vepakomma, P., Singh, A., Qiu, H., Shen, L., et al.: Fedml: A research library and benchmark for federated machine learning. arXiv preprint (2020)
15. Indyk, P.: Stable distributions, pseudorandom generators, embeddings, and data stream computation. Journal of the ACM (JACM) **53**(3), 307–323 (2006)
16. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al.: Advances and open problems in federated learning. Foundations and Trends® in Machine Learning **14**(1–2), 1–210 (2021)
17. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T.: Scaffold: Stochastic controlled averaging for federated learning. In: International Conference on Machine Learning. pp. 5132–5143. PMLR (2020)
18. Khan, L.U., Pandey, S.R., Tran, N.H., Saad, W., Han, Z., Nguyen, M.N., Hong, C.S.: Federated learning for edge networks: Resource optimization and incentive mechanism. IEEE Communications Magazine **58**(10), 88–93 (2020)

19. Konečnỳ, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T., Bacon, D.: Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492 (2016)
20. Kusner, M.J., Sun, Y., Sridharan, K., Weinberger, K.Q.: Private causal inference. In: Artificial Intelligence and Statistics. pp. 1308–1317. PMLR (2016)
21. Lauritzen, S.L.: Graphical models. Clarendon Press (1996)
22. Li, P.: Estimators and tail bounds for dimension reduction in $l_\alpha (0 < \alpha \leq 2)$ using stable random projections. In: Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 10–19 (2008)
23. Ma, P., Ding, R., Dai, H., Jiang, Y., Wang, S., Han, S., Zhang, D.: Ml4s: Learning causal skeleton from vicinal graphs. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1213–1223 (2022)
24. Ma, P., Ding, R., Wang, S., Han, S., Zhang, D.: Xinsight: explainable data analysis through the lens of causality. arXiv preprint arXiv:2207.12718 (2022)
25. Ma, P., Ji, Z., Pang, Q., Wang, S.: Noleaks: Differentially private causal discovery under functional causal model. IEEE Transactions on Information Forensics and Security **17**, 2324–2338 (2022)
26. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
27. Mhaisen, N., Awad, A., Mohamed, A., Erbad, A., Guizani, M.: Analysis and optimal edge assignment for hierarchical federated learning on non-iid data. arXiv preprint arXiv:2012.05622 (2020)
28. Mian, O., Kaltenpoth, D., Kamp, M.: Regret-based federated causal discovery. In: The KDD'22 Workshop on Causal Discovery. pp. 61–69. PMLR (2022)
29. Murakonda, S.K., Shokri, R., Theodorakopoulos, G.: Quantifying the privacy risks of learning high-dimensional graphical models. In: International Conference on Artificial Intelligence and Statistics. pp. 2287–2295. PMLR (2021)
30. Ng, I., Zhang, K.: Towards federated bayesian network structure learning with continuous optimization. In: International Conference on Artificial Intelligence and Statistics. pp. 8095–8111. PMLR (2022)
31. Niu, F., Nori, H., Quistorff, B., Caruana, R., Ngwe, D., Kannan, A.: Differentially private estimation of heterogeneous causal effects. arXiv preprint arXiv:2202.11043 (2022)
32. Peters, J., Janzing, D., Schölkopf, B.: Elements of causal inference: foundations and learning algorithms. The MIT Press (2017)
33. Pinna, A., Soranzo, N., De La Fuente, A.: From knockouts to networks: establishing direct cause-effect relationships through graph analysis. PloS one **5**(10), e12912 (2010)
34. Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M.D., Muñoz-Marí, J., et al.: Inferring causation from time series in earth system sciences. Nature communications **10**(1), 1–13 (2019)
35. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P.: Causal protein-signaling networks derived from multiparameter single-cell data. Science **308**(5721), 523–529 (2005)
36. Samet, S., Miri, A.: Privacy-preserving bayesian network for horizontally partitioned data. In: 2009 International Conference on Computational Science and Engineering. vol. 3, pp. 9–16. IEEE (2009)

37. Shen, X., Ma, S., Vemuri, P., Simon, G.: Challenges and opportunities with causal discovery algorithms: application to alzheimer's pathophysiology. Scientific reports **10**(1), 1–12 (2020)
38. Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D.: Causation, prediction, and search. MIT press (2000)
39. T Dinh, C., Tran, N., Nguyen, T.D.: Personalized federated learning with moreau envelopes. NeurIPS (2020)
40. Triola, M.: Essentials of Statistics. Pearson Education (2014), `https://books.google.com.hk/books?id=QZN-AgAAQBAJ`
41. Vepakomma, P., Amiri, M.M., Canonne, C.L., Raskar, R., Pentland, A.: Private independence testing across two parties. arXiv preprint arXiv:2207.03652 (2022)
42. Versteeg, P., Mooij, J., Zhang, C.: Local constraint-based causal discovery under selection bias. In: Conference on Causal Learning and Reasoning. pp. 840–860. PMLR (2022)
43. Wang, H., Kaplan, Z., Niu, D., Li, B.: Optimizing federated learning on non-iid data with reinforcement learning. In: IEEE INFOCOM 2020-IEEE Conference on Computer Communications. pp. 1698–1707. IEEE (2020)
44. Wang, L., Pang, Q., Song, D.: Towards practical differentially private causal graph discovery. Advances in Neural Information Processing Systems **33**, 5516–5526 (2020)
45. Wang, L., Pang, Q., Wang, S., Song, D.: Fed-χ^2: Privacy preserving federated correlation test. arXiv preprint arXiv:2105.14618 (2021)
46. Xu, D., Yuan, S., Wu, X.: Differential privacy preserving causal graph discovery. In: 2017 IEEE Symposium on Privacy-Aware Computing (PAC). pp. 60–71. IEEE (2017)
47. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST) **10**(2), 1–19 (2019)
48. Yue, X., Kontar, R.A., Gómez, A.M.E.: Federated data analytics: A study on linear models. arXiv preprint arXiv:2206.07786 (2022)
49. Zhang, C., Li, S., Xia, J., Wang, W., Yan, F., Liu, Y.: {BatchCrypt}: Efficient homomorphic encryption for {Cross-Silo} federated learning. In: 2020 USENIX annual technical conference (USENIX ATC 20). pp. 493–506 (2020)
50. Zhang, J.: On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artificial Intelligence **172**(16-17), 1873–1896 (2008)

Supplementary Material

7.1 Workflow

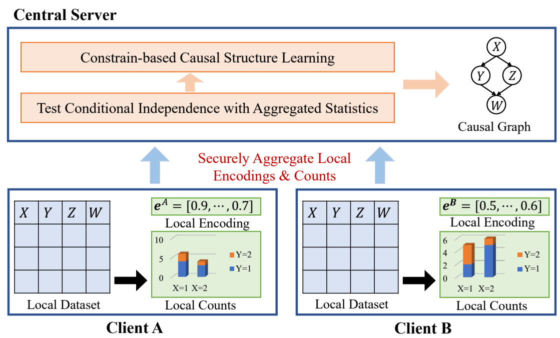


Fig. 6: FEDC^{2}SL workflow.

7.2 Graph Separation

Definition 4 (d-separation). *Two nodes X, Y are d-separated by a set of nodes $\boldsymbol{Z}$ in a causal graph G if and only if X, Y are not d-connected by $\boldsymbol{Z}$ in G. Two nodes X, Y are d-connected by $\boldsymbol{Z}$ in G if and only if an undirected path U connects them in such a way that for each collider W on U, either W or a descendant of W is in $\boldsymbol{Z}$, and no non-collider on U is in $\boldsymbol{Z}$. If $\rightarrow W \leftarrow$ exist in path U ($\rightarrow$ and $\leftarrow$ are directed edges), W is a collider.*

Definition 5 (m-separation). *Two nodes X, Y are m-separated by a set of nodes $\boldsymbol{Z}$ in a causal graph G if and only if there is no active path between X, Y relative to $\boldsymbol{Z}$ in G. A path U between X and Y is active relative to a (possible empty) set of nodes $\boldsymbol{Z}$ if (i) every non-collider on U is not a member of $\boldsymbol{Z}$ and (ii) every collider on U has a descendant in $\boldsymbol{Z}$.*

7.3 PC Algorithm

Algorithm 2: PC Algorithm

1 **Function** `Skeleton-Learning`($X_1, \cdots, X_d$):
2   let S_0 be a complete graph over $X_1, \cdots, X_d$;
3   let S be a copy of S_0;
4   **foreach** *edge* (X, Y) *in* S_0 **do**
5         $V \leftarrow \{X_1, \cdots, X_d\} \setminus \{X, Y\}$;
6   **if** $\exists \boldsymbol{Z} \subseteq V . X \perp\!\!\!\perp Y \mid \boldsymbol{Z}$ **then**
7   remove (X, Y) in S;
8   **end**
9   **return** S
10 **Function** `PC`($X_1, \cdots, X_d$):
11   // skeleton learning
12     $G \leftarrow$ `Skeleton-Learning`($X_1, \cdots, X_d$);
13   // orientation
14   **foreach** *unshielded triple* $(X - Z - Y)$ *in* G **do**
15   **if** $\forall X \perp\!\!\!\perp Y \mid G, Z \notin G$ **then**
16   orient $(X - Z - Y)$ as $(X \rightarrow Z \leftarrow Y)$;
17   **end**
18   **repeat**
19   **foreach** $(X \rightarrow Z - Y)$ **do** orient as $(X \rightarrow Z \leftarrow Y)$;
20   **foreach** $(X \rightarrow Z \rightarrow Y)$ *and* $(X - Y)$ **do** orient as $(X \rightarrow Y)$;
21   **foreach** $(X - Y)$, $(X - Z)$, $(X - W)$, $(Z \rightarrow Y)$, $(W \rightarrow Y)$ *and* Z, W *is nonadjacent* **do** orient as $(X \rightarrow Y)$;
22   **until** *no more edgs can be oriented*;
23   **return** G

We present the workflow of the PC algorithm [38] in Alg. 2. [1] In the first step (lines 1–9; line 12), edge adjacency is confirmed if there is no conditional independence between two variables (line 6). In the second step (lines 14–22), a set of orientation rules are applied based on conditional independence and graphical criteria.

7.4 Dataset and Hyperparameters

For the synthetic datasets, we use the Erdős–Rényi (ER) random graph model [13] to synthesize DAGs with $d \in \{10, 20, 50, 100\}$ nodes. We sample the graph parameters from a Dirichlet-multinomial distribution. To generate causally insufficient datasets, we randomly mask some variables in the DAG and generate its corresponding MAG. We use forward sampling to obtain 10,000 samples per dataset, which we split into $K \in \{2, 4, 8, 16, 32, 64\}$ partitions as local datasets. To generate client heterogeneous datasets, we add a surrogate variable to the DAG and split the dataset for K clients according to its value (Def. 1).

We set the hyperparameters of our evaluations as follows. The encoding size l in Theorem 1 is 50. The α of Alg. 1 is 0.05. NOTEARS-ADMM is concretized

[1] We use a simplified skeleton learning algorithm in Alg. 2 for the ease of presentation.

with Multi-Layer Perceptron (MLP) to handle the non-linearity in data. We use the default parameters in NOTEARS-ADMM and perform a grid search on the threshold τ with best performance. RFCD is concretized with the GES [38] algorithm with the BDeu score function [7] that is particularly designed for discrete data. We use standard χ^2-test with $\alpha = 0.05$ in PC, PC-Voting, PC-CIT-Voting, FCI, FCI-Voting, and FCI-CIT-Voting.

7.5 Evaluation on Downstream Application

Following [25], we also launch FedC^{2}SL in a downstream task of causal structure learning, namely, causal feature selection (CFS). Regression or classification models often suffer out-of-distribution (OOD) issues, which undermines their accuracy on test data. Here, OOD indicate that the distribution of test data is different from the distribution of training data. From the perspective of causality, OOD issues can be interpreted by causal mechanism shifts in the underlying causal graph. Such shifts, however, is prevalent in practice in the evolving environments. CSF is motivated by the observation that direct causal relations are often more reliable against OOD than indirect causal relations. Hence, CFS aims to pick a subset of variables with strong causal relations to improve the robustness of models on OOD data, instead of using full features.

We use the LUCAS dataset [8] that is particularly designed for assessing causal feature selection algorithms. This dataset contains a training set and two OOD test datasets—LUCAS1 and LUCAS2. LUCAS2 suffers from a higher degree of domain shifts (i.e., more "OOD") than that of LUCAS1. We use different federated causal structure learning algorithms to select the subset of features, perform model training on the selected features, and measure the regression errors in the form of Mean Square Error (MSE) on each test dataset (lower is better). The results are shown in Table 2. We observe that FedPC consistently finds the best features regardless of the client sizes. However, PC-CIT-Voting, while showing comparable performance with #Client=4/8, fails to identify the subset of features with strong causal relations when #Client reaches to eight. Other methods either over-aggressively rule out useful features or use too many features that undermine the performance on the OOD data.

Table 2: Performance in the LUCAS benchmark. Best test MSE is highlighted. "#Features" denotes the number of remaining features after feature selection.

#Client=2				
Dataset	**Method**	**#Features**	**Training MSE**	**Test MSE**
LUCAS 1	None	11	0.423	0.642
	FedPC	4	0.445	0.608
	NOTEARS-ADMM	3	0.457	0.631
	RFCD	4	0.435	0.656
	PC-Voting	4	0.445	0.608
	PC-CIT-Voting	3	0.457	0.631
LUCAS 2	None	11	0.423	0.631
	FedPC	4	0.445	0.574
	NOTEARS-ADMM	3	0.457	0.589
	RFCD	4	0.435	0.646
	PC-Voting	4	0.445	0.574
	PC-CIT-Voting	3	0.457	0.589
#Client=4				
LUCAS 1	None	11	0.423	0.642
	FedPC	4	0.445	0.608
	NOTEARS-ADMM	4	0.435	0.656
	RFCD	3	0.457	0.631
	PC-Voting	4	0.445	0.608
	PC-CIT-Voting	3	0.457	0.631
LUCAS 2	None	11	0.423	0.631
	FedPC	4	0.445	0.574
	NOTEARS-ADMM	4	0.435	0.646
	RFCD	3	0.457	0.589
	PC-Voting	4	0.445	0.574
	PC-CIT-Voting	3	0.457	0.589
#Client=8				
LUCAS 1	None	11	0.423	0.642
	FedPC	4	0.445	0.608
	NOTEARS-ADMM	6	0.443	0.628
	RFCD	2	0.480	0.616
	PC-Voting	3	0.457	0.631
	PC-CIT-Voting	3	0.457	0.631
LUCAS 2	None	11	0.423	0.631
	FedPC	4	0.445	0.574
	NOTEARS-ADMM	6	0.443	0.635
	RFCD	2	0.480	0.590
	PC-Voting	3	0.457	0.589
	PC-CIT-Voting	3	0.457	0.589