arXiv:2211.10085v4 [cs.AI] 1 Aug 2025

Identifying Unique Spatial-Temporal Bayesian Network without Markov Equivalence

Mingyu Kang, Duxin Chen, Ning Meng, Gang Yan, and Wenwu Yu, *Senior Member, IEEE*

***Abstract*—Identifying vanilla Bayesian network to model spatial-temporal causality can be a critical yet challenging task. Different Markovian-equivalent directed acyclic graphs would be identified if the identifiability is not satisfied. To address this issue, Directed Cyclic Graph [1] is proposed to drop the directed acyclic constraint. But it does not always hold, and cannot model dynamical time-series process. Then, Full Time Graph [2] is proposed with introducing high-order time delay. Full Time Graph has no Markov equivalence class by assuming no instantaneous effects. But, it also assumes that the causality is invariant with varying time, that is not always satisfied in the spatio-temporal scenarios. Thus, in this work, a Spatial-Temporal Bayesian Network (STBN) is proposed to theoretically model the spatial-temporal causality from the perspective of information transfer. STBN explains the disappearance of network structure $X \to Z \to Y$ and $X \leftarrow Z \leftarrow Y$ by the principle of information path blocking. And finally, the uniqueness of STBN is proved. Based on this, a High-order Causal Entropy (HCE) algorithm is also proposed to uniquely identify STBN under time complexity $\mathcal{O}(n^3\tau_{max})$, where n is the number of variables and τ_{max} is the maximum time delay. Numerical experiments are conducted with comparison to other baseline algorithms. The results show that HCE algorithm obtains state-of-the-art identification accuracy. The code is available at https://github.com/KMY-SEU/HCE.**

***Impact Statement*—This work presents work whose goal is to advance the field of Artificial Intelligence. This work provides a new way to think in causality though the Spatial-Temporal Bayesian Network, and proposes a High-order Causal Entropy algorithm to identify the network structure uniquely with polynomial time complexity. The model and algorithm are efficient and explainable to be applied to discover the causal interactions with high-order time delays in collective behaviors (e.g., animal flocks and crowds), industrial system (e.g., smart grid, intelligent transportation system and wireless internet), etc.**

***Index Terms*—Spatial-temporal Bayesian network, Markov equivalence, causality, high-order causal entropy, time series.**

This work is supported by the National Key R&D Program of China under Grant No. 2022ZD0120003, the Zhishan Youth Scholar Program, the National Natural Science Foundation of China under Grant Nos. 62233004, 62273090, 62073076, and the Jiangsu Provincial Scientific Research Center of Applied Mathematics under Grant No. BK20233002. (corresponding authors: Duxin Chen, Wenwu Yu)

Mingyu Kang is with the School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China. (e-mail: kangmingyu@seu.edu.cn)

Duxin Chen is with the School of Mathematics, Southeast University, Nanjing 210096, China. (e-mail: chendx@seu.edu.cn)

Ning Meng is with the School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China, and also with the R&D department of Volkswagen China Technology Company (VCTC), Hefei 230601, China. (e-mail: ning.meng@volkswagen-anhui.com)

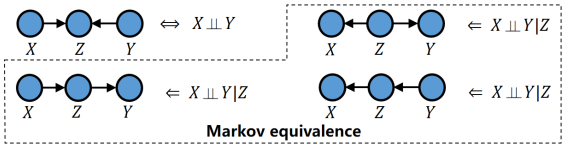
Gang Yan is with the School of Physics Science and Engineering, Tongji University, Shanghai 200092, China. (e-mail: gyan@tongji.edu.cn).

Wenwu Yu is with the Frontiers Science Center for Mobile Information Communication and Security, School of Mathematics, Southeast University, Nanjing 210096, China, and also with the Purple Mountain Laboratories, Nanjing 211102, China. (e-mail: wwyu@seu.edu.cn).

I. Introduction

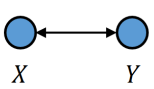
Identifying spatial-temporal causality can be a critical yet challenging task in many data-intensive scenarios, including collective behaviors [3], water systems [4], brain networks [5], transportation systems [6] and electricity networks [7], etc. A common point is that they can be modeled as a multivariate nonlinear system on the spatio-temporal interactions between variables, and the interactions usually have time delays.

Many algorithms [8]–[17] have been proposed to identify the spatio-temporal causality, that is modeled as a (causal) Bayesian network with a constraint of directed acyclic graph (DAG). However, the DAG can be hard or impossible to be identified if the "structure identifiability" [2] is not satisfied. On the one hand, the independence test cannot distinguish the fork structure $X \leftarrow Z \rightarrow Y$ from the chain structures $X \rightarrow Z \rightarrow Y$ and $X \leftarrow Z \leftarrow Y$, if any three variables X, Y, Z are given [18]. As shown in Fig. 1(a), there are only four fundamental structures in a DAG. But $X \leftarrow Z \rightarrow Y$, $X \rightarrow Z \rightarrow Y$ and $X \leftarrow Z \leftarrow Y$ cannot be distinguished, because they conform to the same conditional independence $X \perp\!\!\!\perp Y | Z$. Then, the forward and reverse directions are both detected, but only one is true. Thus, a class of Markov-equivalent DAGs would be detected finally. On the other hand, the DAG representation does not match the functional mechanism in data generation, especially in time-series process. E.g., as shown in Fig. 1(b), the two variables are interactive from the view of functional mechanism, but the DAG only allows one of the causal links. In this case, the causality would also be non-identifiable, and different DAGs would be identified as Markov equivalence class.



(a)

$$X_t = Y_{t-1} + \xi_{x,t}$$
$$Y_t = X_{t-1} + \xi_{y,t}$$
$$\Rightarrow$$



(b)

Fig. 1. Markov equivalence. Here, X_t and Y_t represent the variables X and Y at time t. ξ represents noise variable.

To address this issue, some special networks are designed, such as Directed Cyclic Graph [1]. Like the DAG-based Bayesian network, Directed Cyclic Graph can also be used

to build a Structural Causal Model without further assumptions [19]. However, Directed Cyclic Graph does not always hold because the common cause principle [20] is not always satisfied. Moreover, that network does not consider time delay, so that cannot model dynamical time-series process. Thus, Full Time Graph [2], a temporal Bayesian network, is proposed with introducing high-order time delay, and it has no Markov equivalence class if without instantaneous effects (see Theorem 10.1 in [2]). However, Full Time Graph assumes that the causality is invariant with varying time, that is not always satisfied in the spatio-temporal scenarios.

Thus, in this work, a Spatial-Temporal Bayesian Network (STBN) is proposed to theoretically model the spatial-temporal causality from the perspective of information transfer. Transfer entropy [21] is introduced here to quantify the information transfer. Based on this, the equivalence is proved between zero transfer entropy, time-series unpredictability and d-separation (equal to conditional independence). Then, information path blocking explains the disappearance of $X \to Z \to Y$ and $X \leftarrow Z \leftarrow Y$, and this further induces the uniqueness of STBN. Thus, the corresponding algorithm can be designed to identify the unique network structure of STBN by the transfer entropy with information path blocking. We call it as high-order causal entropy (HCE) algorithm here.

Thus, the main contributions are summarized as follows:

1. STBN is proposed to address the issue of Markov equivalence.
2. HCE algorithm is proposed to identify the unique network structure of STBN.
3. Sufficient experiments are conducted to support the conclusion of state-of-the-art accuracy of HCE algorithm on nonlinear data with time-varying distribution shift, compared to the other widely-used baseline algorithms.

II. Background and Notation

Throughout the paper, the following syntactic conventions are used. An upper case letter (e.g., X_i, Y) is denoted as a variable, and the same letter in lower case (e.g., x_i, y) is denoted as the state value of that variable. A bold-face capitalized letter (e.g., $\mathbf{X}_i, \mathbf{Y}$) is denoted as a set of variables, and a corresponding bold-face lower-case letter (e.g., $\mathbf{x}_i, \mathbf{y}$) is denoted as the assigned state values to each variable in that set.

Then, a general nonlinear system is considered. It is represented as a coupling of multiple stochastic variables $\mathbf{V} = \{X_1, \ldots, X_n\}$ as follow:

$$X_{i,t} = f_i(X_{j,t-\tau}, j = 1, \ldots, n, \tau = 1, \ldots, \tau_{max}) + \xi_{i,t}, \quad (1)$$

for each $X_i \in \mathbf{V}$ at time t. Here, τ_{max} is the maximum time delay, and $\xi_{i,t}$ is additive noise. Note that, all our works are built on this system.

A. Bayesian Network and Markov Equivalence

Bayesian network can be built on $\mathbf{V} = \{X_1, \ldots, X_n\}$ in Eq. (1). It is a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ that represents probabilistic knowledge [22], where $\mathbf{E}$ is the set of directed edges. Note that, if the Structural Causal Model is mentioned, the Bayesian network is causal, because every Structural Causal Model is associated with a graphical causal model, and the Bayesian network is one realization of the graphical causal models [22]. In Bayesian network, DAG is a type of graphical language to describe all of conditional independence in joint distribution. It factorizes the joint distribution into a product of multiple conditional probability distributions as follow:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \mathbf{Pa}_i), \quad (2)$$

where $\mathbf{Pa}_i$ is the direct predecessors of X_i in $\mathcal{G}$, that is also called as causal parent. Thus, with the Bayesian network, conditional independence is satisfied, that is, $X_i \perp\!\!\!\perp \{X_1, X_2, \ldots, X_{i-1}\} \backslash \mathbf{Pa}_i | \mathbf{Pa}_i$.

To identify the conditional independence from observation, d-separation provided a graphical criterion, that is defined as follow:

Definition 1 (d-separation [23])**.** *Let* $\mathbf{X}$*,* $\mathbf{Y}$ *and* $\mathbf{Z}$ *be three disjoint subsets of* $\mathbf{V}$*, and let* $\mathbf{p}$ *be any path from a node in* $\mathbf{X}$ *to a node in* $\mathbf{Y}$ *regardless of direction.* $\mathbf{Z}$ *is said to block* $\mathbf{p}$ *if and only if there is a node* $v \in \mathbf{p}$ *satisfying one of the following items.*

1) v *has v-structure (two nodes* $a, b \in \mathbf{p}$ *pointing to* v*, namely* $a \to v \leftarrow b$*), and neither* v *nor its any descendants are in* $\mathbf{Z}$*;*
2) v *in* $\mathbf{Z}$ *and* v *does not have v-structure.*

Then, $\mathbf{Z}$ *d-separate* $\mathbf{X}$ *and* $\mathbf{Y}$*, denoted as* $\mathbf{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{Y} | \mathbf{Z}$*, if* $\mathbf{Z}$ *blocks any* $\mathbf{p}$*.*

Based on d-separation, the assumptions of causal Markov and faithfulness can be defined as follows:

Definition 2 (Causal Markov [23])**.** *The joint distribution* P *is Markovian to the DAG* $\mathcal{G}$ *if*

$$\mathbf{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{Y} | \mathbf{Z} \Rightarrow \mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}, \quad (3)$$

where $\mathbf{X}$*,* $\mathbf{Y}$ *and* $\mathbf{Z}$ *are three disjoint subsets of* $\mathbf{V}$*.*

Definition 3 (Faithfulness [23])**.** *Probability distribution* P *is faithful to the DAG* $\mathcal{G}$ *if*

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z} \Rightarrow \mathbf{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{Y} | \mathbf{Z} \quad (4)$$

for all disjoint subsets $\mathbf{X}$*,* $\mathbf{Y}$ *and* $\mathbf{Z}$*.*

The two assumptions are also used to identify DAG-based network structure from observation directly by independence test. However, it is often not easy because of Markov equivalence. As shown in Fig. 1(a), there are only four structures in a Bayesian network if given any three variables X, Y, Z. If $X \perp\!\!\!\perp Y | Z$ is detected, the three equivalent structures are all corresponding to it. Thus, the identified Bayesian network are often a group of DAGs conforming to the Markov equivalence.

The issue of Markov equivalence can induce many problems that restrict the usage. Firstly, Markov blanket is not equal to the $\mathbf{Pa}_i$. The Markov blanket is defined as

Definition 4 (Markov blanket [23])**.** *For a DAG* $\mathcal{G}$ *and a target variable* X*, if there is a variable set* $\mathbf{M} \subset \mathbf{V}$ *satisfied*

$$X \perp\!\!\!\perp_{\mathcal{G}} \mathbf{V} \backslash (\mathbf{M} \cup \{X\}) | \mathbf{M}, \quad (5)$$

then the minimal $\mathbf{M}$ *is a Markov blanket, which contains all knowledge for predicting* X.

If Markov equivalence class exists, the causal direction is hard to identified truly, because forward and reverse directions may be both detected, but only one is true in DAG. Thus, the children may be used to predict parents. That is, $\mathbf{Pa}_i \subseteq \mathbf{M}$ as defined, which is usually not what we want.

Secondly, it is hard to infer counterfactuals. Based on Pearl's causal ladder [22], inferring counterfactual has three steps: identify a (causal) Bayesian network, intervene the network, and infer counterfactuals. However, if the network structure is not unique, the inferred counterfactual would not be unique. This is usually unacceptable.

B. Predictability and Information Transfer

Predictability was first proposed in Granger Causality (GC) test [8] to detect and quantify causality in time-series linear system by vector autoregressive model. The predictability to variable X provided by variable Y, $\mathcal{F}_{Y\to X}$, is equal to the logarithmic variance ratio between with and without Y in Gaussian vector autoregression process [24], [25]. That is,

$$\begin{aligned}\mathcal{F}_{Y\to X} &= \log\left[\frac{Var(\xi_t)}{Var(\xi'_t)}\right],\\ \xi_t &= X_t - \sum_{\tau=1}^{\tau_{max}} X_{t-\tau}w_{t-\tau},\\ \xi'_t &= X_t - \sum_{\tau=1}^{\tau_{max}} X_{t-\tau}w_{t-\tau} - \sum_{\tau=1}^{\tau_{max}} Y_{t-\tau}u_{t-\tau},\end{aligned} \tag{6}$$

where $Var(\cdot)$ denotes the variance for samples as $t = \tau_{max}+1,\dots,k$ if k samples is given. w and u are the numerical weights of X and Y, respectively. Thus, $\mathcal{F}_{Y\to X}$ in GC is the reduction of uncertainty from Y to X actually.

The nonlinear GC can be sophisticated, because of the nonlinearity in multivariate system Eq. (1). On the one hand, nonlinear autoregressive model is hard to build. On the other hand, it is time consuming to pair-wisely detect causality between each pair of variables. To address this issue, Barnett, et al. [26] proposed an approach to quantify bi-variate nonlinear GC based on τ_{max}-lag transfer entropy $\mathcal{T}_{Y\to X}$. It is defined as

$$\begin{aligned}&\mathcal{T}_{Y\to X}\\ =&H(X_t|X_{t-1}^{(\tau_{max})}) - H(X_t|X_{t-1}^{(\tau_{max})}, Y_{t-1}^{(\tau_{max})})\\ =&-\frac{1}{k-\tau_{max}}\sum_{t=\tau_{max}+1}^{k}\log\frac{f(x_{t-1}^{(\tau_{max})};\boldsymbol{\theta}_0)}{f(x_{t-1}^{(\tau_{max})}, y_{t-1}^{(\tau_{max})};\boldsymbol{\theta})}\\ =&-\frac{1}{k-\tau_{max}}\log\frac{\mathcal{L}(\hat{\boldsymbol{\theta}}_0|X,Y)}{\mathcal{L}(\hat{\boldsymbol{\theta}}|X,Y)}\\ =&\mathcal{F}_{Y\to X}.\end{aligned} \tag{7}$$

Here, $X_{t-1}^{(\tau_{max})} = \{X_{t-1},\dots,X_{t-\tau_{max}}\}$, and $x_{t-1}^{(\tau_{max})}$ is the sampled state value. $Y_{t-1}^{(\tau_{max})}$ and $y_{t-1}^{(\tau_{max})}$ is the same as this. Moreover, $H(Y|X) = -\sum_{x,y}P(x,y)\log P(y|x)$, and $f(\cdot|\boldsymbol{\theta})$ is nonlinear function with parameter $\boldsymbol{\theta}$. $\mathcal{L}(\cdot|X,Y)$ is the probabilistic likelihood of X and Y. $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and $\boldsymbol{\Theta}$ is full parameter set. $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}\in\boldsymbol{\Theta} | f(x_{t-1}^{(\tau_{max})}, y_{t-1}^{(\tau_{max})};\boldsymbol{\theta}) \text{ does not depend on } y_{t-1}^{(\tau_{max})}\}$. $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}$ are the corresponding estimation, respectively. $\mathcal{T}_{Y\to X}$ quantify the reduction of uncertainty and the information transfer from Y to X by the log-likelihood ratio.

III. Spatial-Temporal Bayesian Network

In this work, STBN is proposed as a type of Bayesian network, that represents spatial-temporal causality in interactions between variables with high-order time delays. The proofs for the uniqueness of STBN are also provided. Moreover, HCE algorithm is proposed to uniquely identify the network structure of STBN by measuring high-order causal entropy.

A. Spatial-Temporal Causality

Spatial-temporal causality is defined here as the probabilistic knowledge describing the causal interactions between variables with high-order time delays during a dynamical process, e.g., in Eq. (1). Consider an STBN $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ built on Eq. (1). $\mathbf{V} = \{X_1,\dots,X_n\}$ is the variable set, and $\mathbf{E}$ is the edge set. Note that, due to the introducing of high-order time delay, thus, the nodes in STBN actually denote variables at some times. For example, $X_{i,t}$ and $X_{i,t-1}$ are both variable X_i in traditional Bayesian network, but they denote X_i at different times here, that is, they are slightly different. To define spatial-temporal causality, two assumptions are firstly defined as:

1. **Temporal assumption:** the cause precedes the effect.
2. **Causal assumption:** the state of effect is predictable by the state of its direct cause.

The two assumptions are fundamental enough, and widely accepted in [2], [9], [10], [12], [27]–[29].

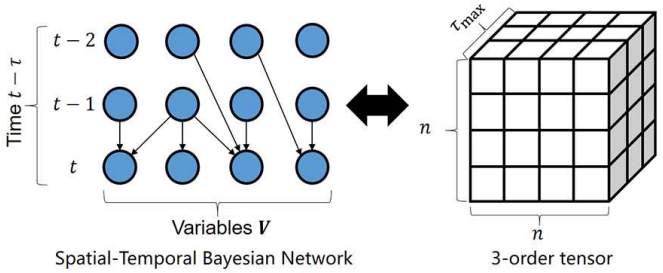


Fig. 2. Diagram of Spatial-Temporal Bayesian Network.

Thus, the adjacency of $\mathcal{G}$ can be denoted as a real 3-order tensor $A \in \mathbb{R}^{n\times n\times\tau_{max}}$, as shown in Fig. 2. The weight $A_{ji\tau}$ denotes the causal strength of $X_{j,t-\tau} \to X_{i,t}$ with time lag τ. Intuitively, STBN describes the stable multivariate causal relationship during a period of time, or the momentary causality at time t in time series. Then,

Proposition 1 (Acyclic)**.** *STBN* $\mathcal{G}$ *is a DAG.*

Proof: If $\mathcal{G}$ is not a DAG, there is at least one path $\mathbf{p}$ which is a cycle. Then, for any two different nodes $X, Y \in \mathbf{p}$, there must be two different path $X \to \cdots \to Y$ and $Y \to \cdots \to X$ without intersection except for them two. Then, X is the direct or indirect cause of Y, and the direct or

indirect effect of Y at the same time. This is contradictory to the **temporal assumption**. ■

Thus, due to the **Proposition 1**, the assumptions of causal Markov and faithfulness (see **Definition 2** and **Definition 3**) can also be applied to STBN. This enables to identify network structure from observation.

B. Proofs for Uniqueness

Here, transfer entropy [21] is introduced to quantify information transfer. Firstly, the causal equivalence between zero transfer entropy, unpredictability and d-separation is proved. Based on this, it is proved that parent set $\mathbf{Pa}_X$ is equal to Markov blanket $\mathbf{M}_X$ for arbitrary variable X, due to information path blocking. Finally, the uniqueness of STBN is obtained. Thus, first of all, two lemmas are introduced to prove the causal equivalence as follows:

Lemma 1. *For a variable X and another variable Y preceding X, $X \perp\!\!\!\perp_{\mathcal{G}} Y|\mathbf{Z}$ if and only if the transfer entropy $\mathcal{T}_{Y\to X|\mathbf{Z}} = 0$, where $\mathbf{Z}$ is the condition set of X.*

Proof: Given the variables X and Y, then $X \perp\!\!\!\perp_{\mathcal{G}} Y|\mathbf{Z} \Rightarrow X \perp\!\!\!\perp Y|\mathbf{Z}$ because of causal Markov (see **Definition 2**). Thus, $\mathcal{T}_{Y\to X|\mathbf{Z}} = H(X|\mathbf{Z}) - H(X|Y,\mathbf{Z}) = 0$. On the other hand, if $\mathcal{T}_{Y\to X|\mathbf{Z}} = 0$, then $X \perp\!\!\!\perp Y|\mathbf{Z}$. Thus, $X \perp\!\!\!\perp Y|\mathbf{Z} \Rightarrow X \perp\!\!\!\perp_{\mathcal{G}} Y|\mathbf{Z}$ because of faithfulness (see **Definiton 3**). ■

Lemma 2. *For a variable X and another variable Y preceding X, the transfer entropy $\mathcal{T}_{Y\to X|\mathbf{Z}} = 0$ if and only if $\mathcal{F}_{Y\to X|\mathbf{Z}} = 0$, where $\mathbf{Z}$ is a condition set of X.*

Proof: If given variables X, Y and $\mathbf{Z}$, their k joint samples $(x^{(k)}, y^{(k)}, \mathbf{z}^{(k)})$ are collected from system Eq. (1). That is, $x^{(k)} = \{x_1, \ldots, x_k\}$, and the other two are the same with one-to-one temporal correspondence. They are all assumed to be ergodic, that indicates the dynamical system have visited all parts of the state space with varying time [30]. This implies the expectation of functions (see Eq. (1)) on the observed samples would converge to the expectation on the state space.

Thus, transfer entropy can be defined as

$$\mathcal{T}_{Y\to X|\mathbf{Z}} = H(X|\mathbf{Z}) - H(X|Y,\mathbf{Z}). \tag{8}$$

Then, the prediction model f with parameters $\boldsymbol{\theta}$ can be defined as

$$f(x|y,\mathbf{z};\boldsymbol{\theta}) = P(X|Y,\mathbf{Z};\boldsymbol{\theta}). \tag{9}$$

Here, $P(\cdot|\cdot;\boldsymbol{\theta})$ is parameterized conditional probability distribution function. The model f is assumed to be identifiable and the parameters are assumed to be unique such that $f(\cdot|\cdot;\boldsymbol{\theta}_1) \neq f(\cdot|\cdot;\boldsymbol{\theta}_2)$ for $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$. Thus, the unique true parameter $\boldsymbol{\theta}^*$ satisfy

$$f(x|y,\mathbf{z};\boldsymbol{\theta}^*) = P(X|Y,\mathbf{Z};\boldsymbol{\theta}^*) = P(X|Y,\mathbf{Z}). \tag{10}$$

Thus, according to ergodic assumption, x_t is only up to y_t and $\mathbf{z}_t$ for any $t = 1, \ldots, k$ in f, the likelihood for parameters $\boldsymbol{\theta}$ can be written as

$$\begin{aligned}
&\mathcal{L}_k(\theta|x^{(k)}, y^{(k)}, \mathbf{z}^{(k)}) \\
=&P(x^{(k)}, y^{(k)}, \mathbf{z}^{(k)};\boldsymbol{\theta}) \\
=&P(x_k|x^{(k-1)}, y^{(k)}, \mathbf{z}^{(k)};\boldsymbol{\theta}) \\
&\times P(y_k, \mathbf{z}_k|x^{(k-1)}, y^{(k-1)}, \mathbf{z}^{(k-1)};\boldsymbol{\theta}) \\
&\times P(x^{(k-1)}, y^{(k-1)}, \mathbf{z}^{(k-1)};\boldsymbol{\theta}) \\
=&f(x_k|y_k,\mathbf{z}_k;\boldsymbol{\theta})q(y_k,\mathbf{z}_k)\mathcal{L}_{k-1} \\
=&\prod_{t=1}^{k} f(x_t|y_t,\mathbf{z}_t;\boldsymbol{\theta})q(y_t,\mathbf{z}_t),
\end{aligned} \tag{11}$$

where $P(x_t|x^{(t-1)}, y^{(t)}, \mathbf{z}^{(t)};\boldsymbol{\theta}) = f(x_t|x^{(t-1)}, y^{(t)}, \mathbf{z}^{(t)};\boldsymbol{\theta}) = f(x_t|y_t,\mathbf{z}_t;\boldsymbol{\theta})$. Moreover, the parameter set $\boldsymbol{\theta}$ only affects X, thus $P(y_t,\mathbf{z}_t|x^{(t-1)}, y^{(t-1)}, \mathbf{z}^{(t-1)};\boldsymbol{\theta})$ can be simplified as a distribution $q(y_t,\mathbf{z}_t)$. $q(y_t,\mathbf{z}_t)$ does not equal to zero almost everywhere, and not refer to $\boldsymbol{\theta}$. Then, to maximize the likelihood $\mathcal{L}_k$, the average log-likelihood is obtained as

$$\begin{aligned}
\ell(\boldsymbol{\theta}|x^{(k)}, y^{(k)}, \mathbf{z}^{(k)}) &= \frac{1}{k}\log\mathcal{L}_k(\boldsymbol{\theta}|x^{(k)}, y^{(k)}, \mathbf{z}^{(k)}) \\
&\sim \frac{1}{k}\sum_{t=1}^{k}\log f(x_t|y_t,\mathbf{z}_t;\boldsymbol{\theta}).
\end{aligned} \tag{12}$$

Thus, the unique parameter $\boldsymbol{\theta}^*$ for the prediction model f can always be obtained as $k \to \infty$. Thus, according to Birkhoff-Khinchin ergodic theorem [30] and Eq. (10), it is obtained that

$$\ell \xrightarrow{a.s.} \mathbb{E}[\log f(x|y,\mathbf{z};\boldsymbol{\theta})], \tag{13}$$

and then,

$$\mathbb{E}[\log f(x|y,\mathbf{z};\boldsymbol{\theta})] \xrightarrow{a.s.} -H(X|Y,\mathbf{Z}) \tag{14}$$

as $k \to \infty$. Here, $\mathbb{E}[\cdot]$ is expectation. In practical experiments, this is to converge with small k.

Then, a nested null model can be defined as $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} | X \text{ are indenpendent of } Y \text{ given } \mathbf{Z}\}$, and $\boldsymbol{\Theta}_0 \subseteq \boldsymbol{\Theta}$. $\boldsymbol{\Theta}$ is full parameter set. Then, the likelihood ratio can be defined as

$$\Lambda(x^{(k)}, y^{(k)}, \mathbf{z}^{(k)}) = \frac{\mathcal{L}_k(\hat{\boldsymbol{\theta}}_0|x^{(k)}, y^{(k)}, \mathbf{z}^{(k)})}{\mathcal{L}_k(\hat{\boldsymbol{\theta}}|x^{(k)}, y^{(k)}, \mathbf{z}^{(k)})}, \tag{15}$$

where $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}$ are maximum likelihood estimators for $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}$, respectively. It is intuitive that $\Lambda(x^{(k)}, y^{(k)}, \mathbf{z}^{(k)})$ measures the dependence degree of predicting X given Z to Y. Thus, the degree of predictability can be defined as

$$\mathcal{F}_{Y\to X|\mathbf{Z}} = -\frac{1}{k}\log\Lambda(x^{(k)}, y^{(k)}, \mathbf{z}^{(k)}), \tag{16}$$

Where $\mathcal{F}_{Y\to X|\mathbf{Z}} \in [0, +\infty)$. Then, $\mathcal{F}_{Y\to X|\mathbf{Z}} \xrightarrow{a.s.} \mathcal{T}_{Y\to X|\mathbf{Z}}$ as $k \to \infty$ obviously according to Eq. (8), Eq. (13) and Eq. (14). Thus, $\mathcal{T}_{Y\to X|\mathbf{Z}} = 0$ if and only if $\mathcal{F}_{Y\to X|\mathbf{Z}} = 0$. ■

With the two lemmas, the causal equivalence can be stated as follows:

Proposition 2 (Causal equivalence)**.** *In STBN $\mathcal{G}$ built on System Eq. (1), zero transfer entropy, time-series unpredictability and d-separation in $\mathcal{G}$ are equivalent, that is,*

$$\mathcal{F}_{Y\to X|\mathbf{Z}} = 0 \iff \mathcal{T}_{Y\to X|\mathbf{Z}} = 0 \iff X \perp\!\!\!\perp_{\mathcal{G}} Y|\mathbf{Z}, \tag{17}$$

for any variable Y preceding the target variable X, and $\mathbf{Z}$ is the set of some variables preceding X as condition set.

In **Proposition 2**, the transfer entropy bridges the predictability and the d-separation. This enables to understand the disappearance of $X \to Z \to Y$ and $X \leftarrow Z \leftarrow Y$ for arbitrary variables X, Y, Z. If $X \to Z \to Y$ is detected in STBN, then $X \perp\!\!\!\perp_{\mathcal{G}} Y|Z \Rightarrow \mathcal{T}_{X\to Y|Z} = 0 \Rightarrow \mathcal{F}_{X\to Y|Z} = 0$. Thus, the information path is blocked, and X provides no more predictability to Y if Z is detected at the same time. Thus, in $X \to Z \to Y$, the edge $X \to Z$ can be overlooked, and this is the same to do in $X \leftarrow Z \leftarrow Y$, because the minimal variable set to predict target variable is really what we want.

Thus, with the bridge of causal equivalence, parent set $\mathbf{Pa}_X = \mathbf{M}_X$ for any target variable X can be obtained in STBN. This is because all variables in $\mathbf{Pa}_X$ precede X, due to the **temporal assumption**, and as shown in Fig. 2. Note that, $\mathbf{M}_X$ is the Markov blanket of X (see **Definition 4**), that is the minimal variable set for prediction. $\mathbf{Pa}_X = \mathbf{M}_X$ can be proved by the following proofs from **Proposition 3** to **Corollary 2**.

Proposition 3. *$\forall Y \in \mathbf{Pa}_X$, the variable Y satisfies $\mathcal{F}_{Y\to X|\mathbf{Pa}_X\backslash\{Y\}} > 0$.*

Proof: For any variable $Y \in \mathbf{Pa}_X$, no variable set can d-separate X and Y. Thus, $\mathcal{F}_{Y\to X|\mathbf{Pa}_X\backslash\{Y\}} > 0$ is obtained from $Y \not\perp\!\!\!\perp_{\mathcal{G}} X|\mathbf{Pa}_X\backslash\{Y\}$ according to **Proposition 2** (i.e., $\mathcal{F}_{Y\to X|\mathbf{Pa}_X\backslash\{Y\}} = 0 \iff Y \perp\!\!\!\perp_{\mathcal{G}} X|\mathbf{Pa}_X\backslash\{Y\}$). ■

Independence is used to search Markov blanket (see **Definition 4**) in traditional Bayesian network [31]. But due to Markov equivalence, the identified Markov blanket $\mathbf{M}_X$ are larger than or equal to the true parent set $\mathbf{Pa}_X$. That is, spurious edges are to be identified inevitably. However, this can be addressed in STBN, because **temporal assumption** prevents causal information flow from effect to cause. Thus,

Proposition 4. *$\forall Y$ preceding X and $Y \notin \mathbf{Pa}_X$, the variable Y satisfies $\mathcal{F}_{Y\to X|\mathbf{Pa}_X} = 0$.*

Proof: In STBN $\mathcal{G}$, for any variable Y preceding X and $Y \notin \mathbf{Pa}_X$, $Y \perp\!\!\!\perp_{\mathcal{G}} X|\mathbf{Pa}_X$ because all information paths from Y to X are blocked by $\mathbf{Pa}_X$. Thus, $Y \perp\!\!\!\perp_{\mathcal{G}} X|\mathbf{Pa}_X\backslash\{Y\}$ because $\mathbf{Pa}_X = \mathbf{Pa}_X\backslash\{Y\}$. Thus, $\mathcal{F}_{Y\to X|\mathbf{Pa}_X\backslash\{Y\}} = 0$ according to **Proposition 2**. ■

Then, it is obtained that

Corollary 1. *$\forall Y$ preceding X, $Y \notin \mathbf{Pa}_X$ if and only if $\mathcal{F}_{Y\to X|\mathbf{Pa}_X\backslash\{Y\}} = 0$.*

Thus, if $\mathbf{V}$ is the full variable set, then

Proposition 5. *For any nested variable set $\mathbf{M} \supseteq \mathbf{Pa}_X$, and all variables in $\mathbf{M}$ precede X, then $\forall Y \in \mathbf{V}\backslash\mathbf{M} \cup \{X\}$, $Y \perp\!\!\!\perp_{\mathcal{G}} X|\mathbf{M}$, that is, $\mathcal{F}_{Y\to X|\mathbf{M}} = 0$.*

Proof: Let $\mathbf{M} \supseteq \mathbf{Pa}_X$ and all variables in $\mathbf{M}$ precede X. For any selected $Y \in \mathbf{V}\backslash\mathbf{M}\cup\{X\}$, then $Y \notin \mathbf{Pa}_X$. According to **Proposition 4**, $\mathcal{F}_{Y\to X|\mathbf{Pa}_X} = 0$ is obtained. And then, $Y \perp\!\!\!\perp_{\mathcal{G}} X|\mathbf{Pa}_X$ due to the causal equivalence. That means $\mathbf{Pa}_X$ blocks any information paths between X and Y, as shown in Fig. 3(a). Thus, $\mathbf{M}$ also blocks any information paths between X and Y because $\mathbf{M} \supseteq \mathbf{Pa}_X$. Thus, $Y \perp\!\!\!\perp_{\mathcal{G}} X|\mathbf{M}$. ■

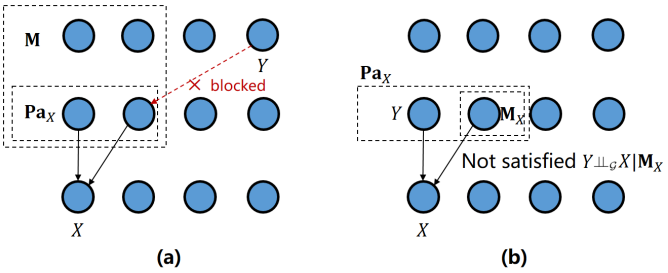


Fig. 3. Diagrams for **Proposition 5** and **Corollary 2**. (a) In **Proposition 5**, for any $\mathbf{M} \supseteq \mathbf{Pa}_X$, $\forall Y \in \mathbf{V}\backslash\mathbf{M}\cup\{X\}$, $\mathbf{M}$ blocks any information paths from Y to X. (b) In **Corollary 2**, if $\mathbf{Pa}_X \supset \mathbf{M}_X$, $\exists Y \in \mathbf{Pa}_X$, but $Y \notin \mathbf{M}_X$, then $Y \not\perp\!\!\!\perp_{\mathcal{G}} X|\mathbf{M}_X$ can be obtained.

Thus, $\mathbf{Pa}_X = \mathbf{M}_X$ can be obtained with few proofs.

Corollary 2. *$\mathbf{Pa}_X = \mathbf{M}_X$ for variable X in STBN $\mathcal{G}$,*

Proof: Let's discuss it in categories. Firstly, we suppose $\mathbf{M}_X$ is the Markov blanket of the variable X according to **Definition 4**. Then, on the one hand, if $\mathbf{Pa}_X \subseteq \mathbf{M}_X$, $\mathbf{Pa}_X = \mathbf{M}_X$ obviously due to **Proposition 5**, as shown in Fig. 3(a). Thus, $\mathbf{M}_X$ is the minimal realization of $\mathbf{M}$. On the other hand, if $\mathbf{Pa}_X \supset \mathbf{M}_X$, then $\exists Y$ precedes X, $Y \in \mathbf{Pa}_X$, but $Y \notin \mathbf{M}_X$. Then, we obtain $Y \not\perp\!\!\!\perp_{\mathcal{G}} X|\mathbf{M}_X$ because there is a direct connection between Y and X, and $\mathbf{M}_X$ cannot d-separates them, as shown in Fig. 3(b). But due to **Definition 4**, and $\mathbf{M}_X$ is the Markov blanket of X, thus we obtain $Y \perp\!\!\!\perp_{\mathcal{G}} X|\mathbf{M}_X$. Thus, this is contradictory. Thus, $\mathbf{Pa}_X = \mathbf{M}_X$ only.

■

Thus, the variables providing prediction knowledge to X is only up to causal parents $\mathbf{Pa}_X$ in STBN $\mathcal{G}$. Meanwhile, that means $\mathbf{Pa}_X$ can be identified in parallel, because the neighbor variables at the same time provide no any prediction knowledge to X. This induces decomposability property of STBN $\mathcal{G}$, as follows:

Proposition 6 (Decomposability)**.** *STBN $\mathcal{G}$ at time t can be decomposed into n subgraphs $\mathcal{G}_{i,t} = (\mathbf{V}, \mathbf{E}_{i,t})$ for simultaneous variables $X_{i,t}, i = 1, \ldots, n$. Then, d-separation in $\mathcal{G}$ is the same as that in $\mathcal{G}_{i,t}$, that is, $\perp\!\!\!\perp_{\mathcal{G}} \iff \perp\!\!\!\perp_{\mathcal{G}_{i,t}}$.*

Proof: According to **Corollary 2**, the predictability for each target variable $X_{i,t}, i = 1, \ldots, n$ is only up to their parents $\mathbf{Pa}_i$. And, there must be no connection between these variables at the same time t, and no connection backtracking from time t to the past, because of the **temporal assumption**. Thus, (i) For each variable at time t, e.g., $X_{1,t}$, the variables $X_{2,t}, \ldots, X_{n,t}$ would not be in the set $\mathbf{Pa}_1$, where $\mathbf{Pa}_1$ is the causal parents of $X_{1,t}$. (ii) For each variable at time t, e.g., any causal pathways pointing to $X_{1,t}$ through $X_{2,t}, \ldots, X_{n,t}$ would be blocked. Thus, the causal links regarding $X_{2,t}, \ldots, X_{n,t}$ and the other pathways through them can be deleted. Then, the reduced subgraph $\mathcal{G}_{1,t}$ is obtained. It is the same to $X_{2,t}, \ldots, X_{n,t}$, and $\mathcal{G}_{2,t}, \ldots, \mathcal{G}_{n,t}$ can also be obtained, respectively. Thus, $\mathcal{G}$ can be decomposed into n subgraphs $\mathcal{G}_{i,t} = (\mathbf{V}, \mathbf{E}_{i,t})$. Here, $\mathbf{E}_{i,t}$ is the directed connections from each variable in $\mathbf{Pa}_i$ to X_i. That is, $\mathcal{G}$ is the linear combination of all $\mathcal{G}_{i,t}$. Moreover, $\perp\!\!\!\perp_{\mathcal{G}} \iff \perp\!\!\!\perp_{\mathcal{G}_{i,t}}$

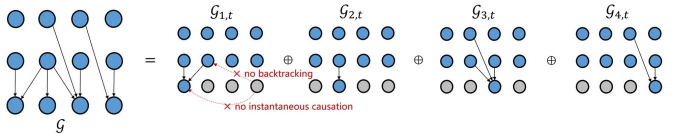


Fig. 4. Diagram of decomposing STBN and combining subgraphs. Here, $\bigoplus$ represents linear combination.

is trivial to see in Fig. 4. ■

Then, the uniqueness property of STBN $\mathcal{G}$ can be finally obtained as follows:

Proposition 7 (Uniqueness)**.** *The network structure of STBN $\mathcal{G}$ is unique.*

Proof: According to the principle of decomposability, we have n subgraphs $\mathcal{G}_{i,t}$ of $\mathcal{G}$. In $\mathcal{G}_{i,t}$, the parents of the target variable $X_{i,t}$ is unique. we can first prove it.

Let X be any target variable $X_{i,t}$. Then, $\mathbf{Pa}_X$ and $\widetilde{\mathbf{Pa}}_X$ are both the parents of X in $\mathcal{G}_{i,t}$ and $\widetilde{\mathcal{G}}_{i,t}$, respectively. $\forall Y \in \widetilde{\mathbf{Pa}}_X$, no variable set $\mathbf{Z}$ excluding X and Y can satisfy $Y \perp\!\!\!\perp_{\widetilde{\mathcal{G}}_{i,t}} X|\mathbf{Z}$. Thus, $\mathbf{Pa}_X\backslash\{Y\}$, one possible $\mathbf{Z}$, is also not able to do so. Thus, $Y \not\perp\!\!\!\perp_{\widetilde{\mathcal{G}}_{i,t}} X|\mathbf{Pa}_X\backslash\{Y\}$, and then $Y \not\perp\!\!\!\perp_{\mathcal{G}_{i,t}} X|\mathbf{Pa}_X\backslash\{Y\}$ because $\perp\!\!\!\perp_{\widetilde{\mathcal{G}}_{i,t}} \Rightarrow \perp\!\!\!\perp_{\mathcal{G}} \Rightarrow \perp\!\!\!\perp_{\mathcal{G}_{i,t}}$ (see decomposability in **Proposition 6**). Then, $Y \in \mathbf{Pa}_X$ is obtained by **Corollary 1** and the causal equivalence in **Proposition 2**. Thus, $\widetilde{\mathbf{Pa}}_X \subseteq \mathbf{Pa}_X$. In a similar way, $\widetilde{\mathbf{Pa}}_X \supseteq \mathbf{Pa}_X$ can be also obtained. Thus, $\mathbf{Pa}_X = \widetilde{\mathbf{Pa}}_X$. Thus, $\mathcal{G}_{i,t} = \widetilde{\mathcal{G}}_{i,t}$.

Thus, the network structure of STBN $\mathcal{G}$ is unique because all subgraphs $\mathcal{G}_{i,t}$ are unique, and $\mathcal{G}$ is just the linear combination of multiple unique subgraphs. Thus, there is no Markov equivalence issue in STBN. ■

C. High-order Causal Entropy Algorithm

To identify unique network structure of STBN, HCE algorithm is proposed by the transfer entropy under the principle of information path blocking. It is a special case of transfer entropy, thus we call it high-order causal entropy here. The pseudocode of HCE algorithm is shown in **Algorithm 1**.

Due to the decomposability property (see **Proposition 6**), the network can be decomposed into n subgraphs. These subgraphs can be identified independently and combined to be a complete network finally. Thus, HCE algorithm can be implemented by n child processes. Each child process searches causal parents $\mathbf{Pa}_i$ for each variable $X_i \in \{X_1, \ldots, X_n\}$. This is in a paralleled way instead of n-times sequential iterations.

Then, each child process can be decomposed into two modules by parameters τ_{max}, α and β. In the first module, for target variable X_i, HCE algorithm calculates the causal entropy for each historical variable and each time delay $\tau = 1, \ldots, \tau_{max}$. This module saves part of variables from the given full condition set, and finally obtains possible $\mathbf{Pa}_i$. In the second module, HCE algorithm removes spurious causation from $\mathbf{Pa}_i$, and the subgraph is obtained at the end.

The parameters α and β are both used to control the bound of zero causal entropy according to **Corollary 1**. If the value of causal entropy is less than α or β, the value is to approximate to be zero. Thus, the obtained $\mathbf{Pa}_i$ becomes narrower with bigger α and β. Finally, the network can be obtained by combining all subgraphs.

Moreover, the time complexity of HCE algorithm is $\mathcal{O}(n^3\tau_{max})$. The main computation is up to searching causal parents for each variable at time t and estimating causal entropy for each possible edge. The time to search causal parents is $\mathcal{O}(n\tau_{max})$ because all historical variables are to be traversed. Then, note that HCE algorithm can be in paralleled, thus this time is the same for each variable, and the time to combine all subgraphs is $\mathcal{O}(1)$, that can be ignored. Meanwhile, KSG estimator [32] is adopted to measure the causal entropy with time complexity $\mathcal{O}(n^2)$ in worst case. But it can be accelerated by using KD-tree neighbor search.

Algorithm 1: High-order Causal Entropy Algorithm

Input: Observation variables $\mathbf{V} = \{X_1, \ldots, X_n\}$ and maximum time lag τ_{max}.
Output: An STBN $\mathcal{G}$.

for $i = 1, \ldots, n$ *in parallel* **do**
  Initialize causal parent set $\mathbf{Pa}_i = \varnothing$;
  for $\tau = 1, \ldots, \tau_{max}$ **do**
    Let condition set $\mathbf{Z} = \{X_{1,t-\tau}, \ldots, X_{n,t-\tau}\}$;
    for $j = 1, \ldots, n$ **do**
      $\mathcal{T}_{X_{j,t-\tau} \to X_{i,t}|\mathbf{Z}\backslash\{X_{j,t-\tau}\}} = H(X_{i,t}|\mathbf{Z}\backslash\{X_{j,t-\tau}\}) - H(X_{i,t}|\mathbf{Z})$;
      if $\mathcal{T}_{X_{j,t-\tau} \to X_{i,t}|\mathbf{Z}\backslash\{X_{j,t-\tau}\}} > \alpha$ **then**
        append $X_{j,t-\tau}$ into $\mathbf{Pa}_i$;
      else
        remove $X_{j,t-\tau}$ from $\mathbf{Z}$;
  for each $Y \in \mathbf{Pa}_i$ **do**
    $\mathcal{T}_{Y \to X_{i,t}|\mathbf{Pa}_i\backslash\{Y\}} = H(X_{i,t}|\mathbf{Pa}_i\backslash\{Y\}) - H(X_{i,t}|\mathbf{Pa}_i)$;
    if $\mathcal{T}_{Y \to X_{i,t}|\mathbf{Pa}_i\backslash\{Y\}} < \beta$ **then**
      remove Y from $\mathbf{Pa}_i$;
Construct the network $\mathcal{G}$ with $\mathbf{Pa}_1, \ldots, \mathbf{Pa}_n$;

IV. Experiments

A. Synthetic Data

In the experiments, the performance of HCE algorithm is tested on synthetic data generated from STBNs. For example, as shown in Fig. 5, f_1, f_2 are defined as $f_1(x) = x$ and $f_2(x) = x + 5x^2e^{-\frac{x^2}{20}}$. The state of variable at time t is certainly calculated by the historical states of all variables, such that $x_{1,t} = 0.2f_2(x_{1,t-1}) + 0.3f_1(x_{env,t-1}) + \mathcal{N}(0,1)$, and $\mathcal{N}(0,1)$ is noise conforming to standard normal distribution. Moreover, the environment variable X_{env} is the common parent for variable X_1 and variable X_4, and X_{env} has no causal parents.

Then, to generate trending time-series data, the state of environment variable X_{env} is set to be a uniformly increasing sequence as shown in Fig. 6. Clearly, the trajectory of X_{env} is in the range of $1.3 \sim 8.0$ within 2000 time steps. Meanwhile, the other observation variables are also presented in Fig. 6.

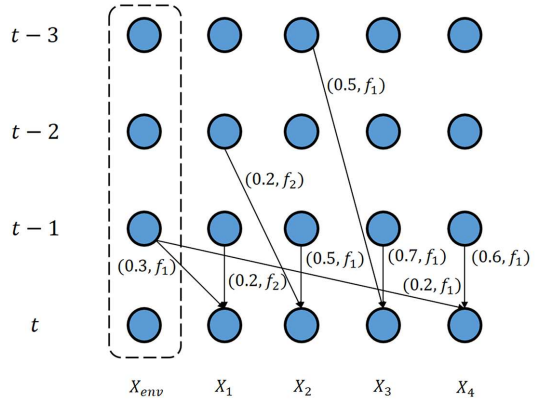


Fig. 5. An example of STBN.

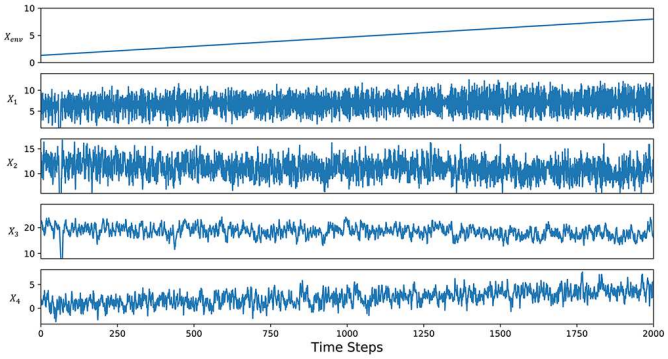


Fig. 6. Trending time series of five variables.

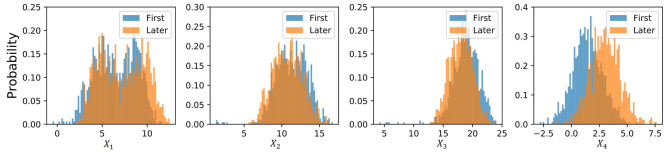


Fig. 7. Dataset shift of the four observation variables.

They all fluctuate in different and limited ranges, and have no property of periodicity. As shown in Fig. 7, it is presented the distributions of four observation variables within first 1000 time steps and within later 1000 time steps respectively. Clearly, dataset shift occurs, so that regression-based models are not to work.

Moreover, all data are randomly generated from STBNs like Fig. 5, but we do not list them all here. In practice, their network size n and the number of edges are kept in a certain proportion, and the number of edges is at least twice the network size. All the edges are set at random to guarantee the randomization of the experiments. Moreover, the sample size k is always fixed to be 2000.

B. Evaluation Metrics

In the experiments, two indices are adopted to measure the performance of model, including true positive rate (TPR) and false positive rate (FPR). They are defined as

$$\begin{aligned} TPR &= \frac{true\ positive}{true\ positive + false\ negative}, \\ FPR &= \frac{false\ positive}{false\ positive + true\ negative}. \end{aligned} \tag{18}$$

C. Results

To conduct comparative experiments, HCE algorithm is compared with GC test [8], convergent cross mapping (CCM) algorithm [9], PCMCI algorithm [12], and DYNOTEARS algorithm [13]. Among them, GC test is a classic algorithm to identify time-series causality by vector regression model. CCM algorithm is proposed to detect causality with high-order time delay in complex ecosystem based on the tool of nonlinear state space reconstruction. PCMCI algorithm is proposed to identify Full Time Graph from large time-series dataset. DYNOTEARS algorithm is a score-based algorithm that minimizes a penalized loss subject to an acyclicity constraint. But note that, it is based on linear modeling, and it allows instantaneous causation.

In the experiments, the truth of STBN is hidden firstly, and then the network structure is identified by these algorithms. The TPR and FPR are both calculated after running these algorithms on the synthetic data. The parameters of HCE are set as that $\tau_{max} = 5, \alpha = 0.01$ and $\beta = 0.02$.

As shown in Fig. 8, the receiver operating characteristic (ROC) curve is drawn with TPRs and FPRs on time series generated by STBN in Fig. 5. The ROC curve of HCE covers the curves generated from the other algorithms. That is, the area under curve (AUC) of HCE is higher than that of other algorithms. Thus, HCE algorithm has better generalization.

Moreover, the performance of HCE algorithm is tested on dataset with bigger network size. As shown in Fig. 9, it is clear that HCE algorithm performs generally better than the other baseline algorithms with increasing network size. Note that all results for boxplot are tested on dataset generated from STBNs with different network sizes and different network structures. To eliminate test errors, the data generation processes are implemented at least 20 times. For all algorithms, the FPRs are fixed to be below 0.1 as possible, so that the TPRs are easy to compare.

As shown in Fig. 9, the TPRs of HCE algorithm gradually decrease with the increase of network size, but it is still higher than that of PCMCI algorithm with the same FPRs. By contrast, GC and CCM algorithms have lower TPRs and larger variances at each running time. Moreover, DYNOTEARS algorithm poorly performs on the TPRs when the FPRs are fixed, with increasing network size. Also, it poorly performs on the ROC curve, as shown in Fig. 8. The reasons are as follows: (i) The modeling of DYNOTEARS is linear, but the data generation is nonlinear in the experiments of this work. (ii) DYNOTEARS considers the instantaneous effects between variables, that induces the huge error to detect many spurious instantaneous causal links, as shown in Fig. 10. Clearly, there are many non-zero causal links when time lag $\tau = 0$, but actually, these links are spurious. Thus, if given some observation variables, and the variable number is less

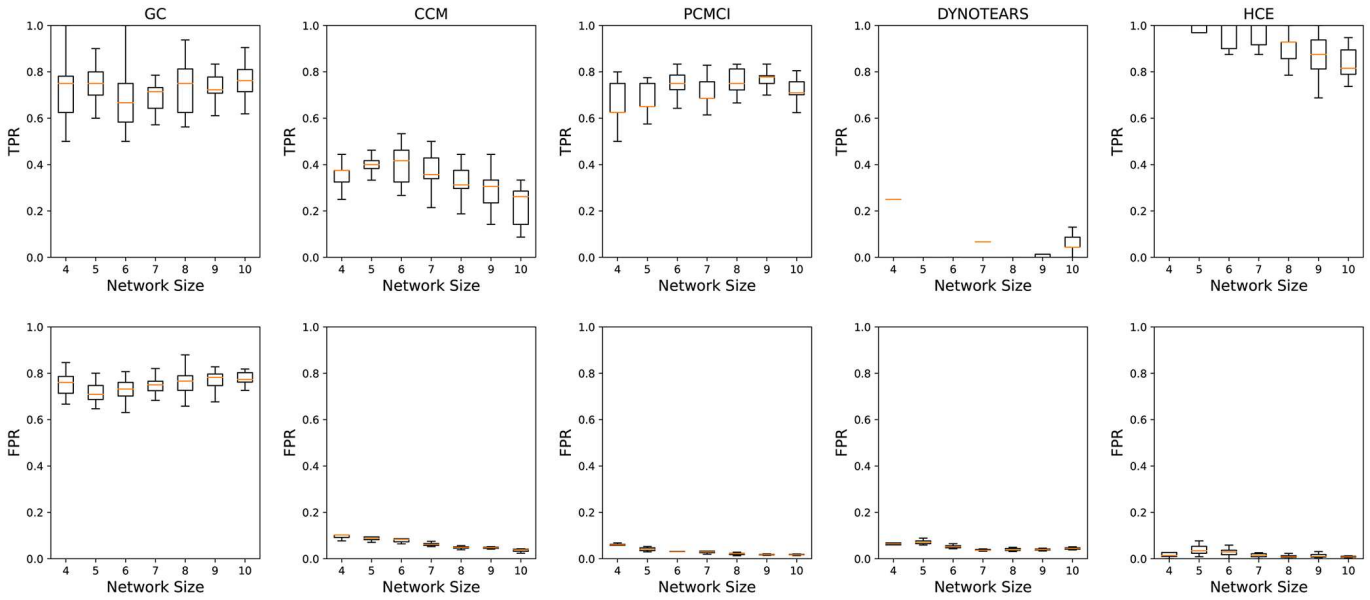


Fig. 9. The comparison to identification accuracy of HCE algorithm and other baseline algorithms.

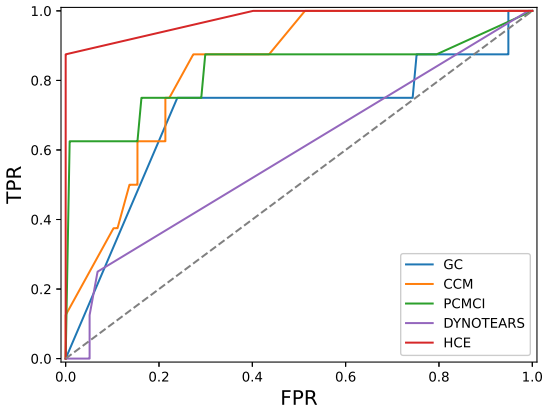


Fig. 8. The ROC curve of HCE and other algorithms.

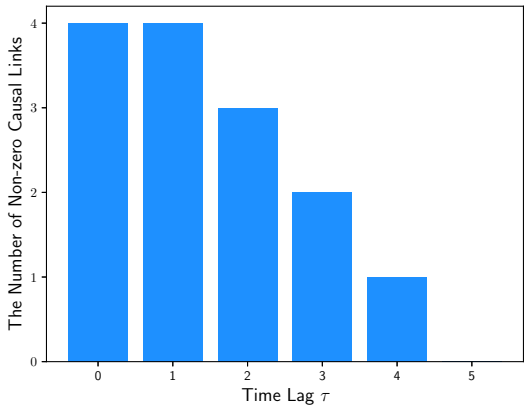


Fig. 10. The distribution of non-zero causal links on an example of DYNOTEARS.

than 10, HCE algorithm can identify the true structure of STBN, even though the data distribution of time series is shifted with time.

V. Conclusion and Future Work

In this work, STBN is proposed to model the spatial-temporal causality by introducing the principle of information transfer. From this perspective, information path blocking explains the disappearance of $X \to Z \to Y$ and $X \leftarrow Z \leftarrow Y$ in STBN, and further induces the uniqueness of network structure. Based on this, HCE algorithm is designed to identify the unique network structure by measuring transfer entropy with information path blocking. Sufficient experiments are conducted to support the main conclusions regarding the performance of HCE algorithm, compared with GC test, CCM algorithm, PCMCI algorithm, and DYNOTEARS algorithm. The results show the state-of-the-art identification accuracy of HCE algorithm on dataset with time-varying shift and increasing network size.

In future works, we will focus on the goal of causal inference based on the proposed STBN model, that is, to further build a framework for inferring spatial-temporal counterfactuals. One possible approach is to use STBN to initialize the network in interactive dynamics [33]. The interactive dynamics can be sparsely identified with the network. And then, the network structure can be intervened, and the counterfactuals can be solved by resetting the initialization values. Another solution to realize this can be found in paper [34], but we find it only generalized on one-order Markov process, which is not universal. A non-temporal but important case can be found in paper [35]. If use the deep generative model to model the network, the supported network size would be larger, but the explainability (or interpretability) would be reduced.

Moreover, we find it different in the intervening operations

between STBNs and the non-temporal Bayesian networks. As stated in [22], if a non-temporal Bayesian network is intervened, the network structure would be changed, and the intervened probability distributions are to be recalculated though the backdoor criterion and front-door criterion. But it is different in STBNs, an STBN models the momentary causality at time t, thus an intervention on STBN would only be defined at that time, but not during a period of times. Thus, to describe the momentary interventions accurately, a group of interventions at different timestamps are needed. This makes the intervening operations to STBNs more complex, and more flexible meanwhile. This is also one focus of the future works.

REFERENCES

[1] P. Spirtes, "Directed cyclic graphical representations of feedback models," in *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 1995, p. 491–498.

[2] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[3] W. M. Lord, J. Sun, N. T. Ouellette, and E. M. Bollt, "Inference of causal information flow in collective animal behavior," *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, vol. 2, no. 1, pp. 107–116, 2016.

[4] H. Macian-Sorribes, J.-L. Molina, S. Zazo, and M. Pulido-Velázquez, "Analysis of spatio-temporal dependence of inflow time series through bayesian causal modelling," *Journal of Hydrology*, vol. 597, p. 125722, 2021.

[5] A. Bagheri, M. Pasande, K. Bello, B. N. Araabi, and A. Akhondi-Asl, "Discovering the effective connectome of the brain with dynamic bayesian dag learning," *NeuroImage*, vol. 297, p. 120684, 2024.

[6] W. Du, S. Chen, Z. Li, X. Cao, and Y. Lv, "A spatial-temporal approach for multi-airport traffic flow prediction through causality graphs," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 1, pp. 532–544, 2024.

[7] M. Kang, R. Zhu, D. Chen, C. Li, W. Gu, X. Qian, and W. Yu, "A cross-modal generative adversarial network for scenarios generation of renewable energy," *IEEE Transactions on Power Systems*, vol. 39, no. 2, pp. 2630–2640, 2024.

[8] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[9] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems," *Science*, vol. 338, no. 6106, pp. 496–500, 2012.

[10] J. Sun, D. Taylor, and E. M. Bollt, "Causal network inference by optimal causation entropy," *SIAM Journal on Applied Dynamical Systems*, vol. 14, no. 1, pp. 73–106, 2015.

[11] D. Malinsky and P. Spirtes, "Causal structure learning from multivariate time series in settings with unmeasured confounding," in *Proceedings of 2018 ACM SIGKDD workshop on causal discovery*, vol. 92. PMLR, 2018, pp. 23–47.

[12] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science Advances*, vol. 5, no. 11, p. eaau4996, 2019.

[13] R. Pamfil, N. Sriwattanaworachai, S. Desai, P. Pilgerstorfer, K. Georgatzis, P. Beaumont, and B. Aragam, "Dynotears: Structure learning from time-series data," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1595–1605.

[14] A. Bregoli, M. Scutari, and F. Stella, "A constraint-based algorithm for the structural learning of continuous-time bayesian networks," *International Journal of Approximate Reasoning*, vol. 138, pp. 105–122, 2021.

[15] W. Yao, Y. Sun, A. Ho, C. Sun, and K. Zhang, "Learning temporally causal latent processes from general temporal data," in *International Conference on Learning Representations*, 2022.

[16] W. Gong, J. Jennings, C. Zhang, and N. Pawlowski, "Rhino: Deep causal temporal relationship learning with history-dependent noise," in *International Conference on Learning Representations*, 2023.

[17] P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves, "Causal representation learning for instantaneous and temporal effects in interactive systems," in *International Conference on Learning Representations*, 2023.

[18] T. Verma and J. Pearl, "Equivalence and synthesis of causal models," in *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, USA, 1991, pp. 255–270.

[19] S. Bongers, P. Forré, J. Peters, and J. M. Mooij, "Foundations of structural causal models with cycles and latent variables," *The Annals of Statistics*, vol. 49, no. 5, pp. 2885–2915, 2021.

[20] J. Pearl, *Causality: Models, Reasoning, and Inference, 2nd edition*. Cambridge University Press, New York, USA., 2009.

[21] T. Schreiber, "Measuring information transfer," *Physical Review Letters*, vol. 85, no. 2, p. 461, 2000.

[22] M. Glymour, J. Pearl, and N. P. Jewell, *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[23] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[24] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *Journal of the American Statistical Association*, vol. 77, no. 378, pp. 304–313, 1982.

[25] J. Geweke, "Measures of conditional linear dependence and feedback between time series," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 907–915, 1984.

[26] L. Barnett and T. Bossomaier, "Transfer entropy as a log-likelihood ratio," *Physical Review Letters*, vol. 109, no. 13, p. 138105, 2012.

[27] P. Spirtes and K. Zhang, "Causal discovery and inference: concepts and recent methodological advances," *Applied Informatics*, vol. 3, no. 1, pp. 1–28, 2016.

[28] Y. Zhang, Y. Zhang, E. Swears, N. Larios, Z. Wang, and Q. Ji, "Modeling temporal interactions with interval temporal bayesian networks for complex activity recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2468–2483, 2013.

[29] X. Zhang, C. Song, J. Zhao, Z. Xu, and X. Deng, "Spatial-temporal causality modeling for industrial processes with a knowledge-data guided reinforcement learning," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 4, pp. 5634–5646, 2024.

[30] P. Walters, *An introduction to ergodic theory*. Springer Science & Business Media, 2000.

[31] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*. MIT press, 2000.

[32] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, 2004.

[33] T.-T. Gao and G. Yan, "Autonomous inference of complex network dynamics from incomplete and noisy data," *Nature Computational Science*, vol. 2, no. 3, pp. 160–168, 2022.

[34] R. Ness, K. Paneri, and O. Vitek, "Integrating markov processes with structural causal modeling enables counterfactual inference in complex systems," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[35] N. Pawlowski, D. Coelho de Castro, and B. Glocker, "Deep structural causal models for tractable counterfactual inference," *Advances in Neural Information Processing Systems*, vol. 33, 2020.



Mingyu Kang received M.S. degree of computer technology from Southeast University in China in 2021. He is currently a Ph.D candidate with the Jiangsu Key Laboratory of Networked Collective Intelligence, School of Cyber Science and Engineering, Southeast University, Nanjing, China. His research interests include complex system, complex network, deep learning and causal inference.



Duxin Chen received the B.S. degree in automatic control and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology,Wuhan,China, in 2013 and 2018, respectively. He is currently an associate professor with the School of Mathematics, and vice director of the Jiangsu Provincial Scientific Research Center of Applied Mathematics, Southeast University, Nanjing, China. He is also one of the direction leaders of the Huawei-SEUJoint Innovation Lab of Networked Collective Intelligence. His research interests include causal inference, prediction/generation, and system identification techniques for complex networks and systems science, artificial intelligence related theory and applications.



Ning Meng received M.S. degree of cyberspace security from Southeast University in China in 2023. He is currently a senior engineer in the R&D department of Volkswagen China Technology Company (VCTC).



Gang Yan received the B.S. and Ph.D. degrees in Electronic Engineering from University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He was with Temasek Laboratories at National University of Singapore, Singapore, and Network Science Institute at Northeastern University, Boston, USA as Research Scientist and Research Fellow in 2010-2013 and 2013-2016, respectively. He joined Tongji University, Shanghai, China in 2017 and is currently a Distinguished Professor.

His research lies at the interface between complex systems and artificial intelligence and his research results have been published in Nature, Nature Physics, Nature Computational Science, Nature Communications, NeurIPS, AAAI, etc. He was the recipient of Distinguished Young Scholar from National Science Foundation of China in 2022, and the Excellent Editor of the IEEE Transactions on Network Science and Engineering in 2021.



Wenwu Yu (S'07-M'12-SM'15) received the B.Sc. degree in information and computing science and M.Sc. degree in applied mathematics from the Department of Mathematics, Southeast University, Nanjing, China, in 2004 and 2007, respectively, and the Ph.D. degree from the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China, in 2010. Currently, he is the Dean in the School of Mathematics, the Deputy Associate Director of National Center of Applied Mathematics in Southeast University of Jiangsu, the Deputy Director of Jiangsu Provincial Scientific Research Center of Applied Mathematics, Deputy Associate Director of Jiangsu Provincial Key Laboratory of Networked Collective Intelligence, and a Full Professor with the Endowed Chair Honor in Southeast University, China.

Dr. Yu held several visiting positions in Australia, China, Germany, Italy, the Netherlands, and the USA. His research interests include multi-agent systems, complex networks and systems, disturbance control, distributed optimization, machine learning, game theory, cyberspace security, smart grids, intelligent transportation systems, big-data analysis, etc.

Dr. Yu serves as an Editorial Board Member of several flag journals, including IEEE Transactions on Circuits and Systems II, IEEE Trans. Industrial Cyber-Physical Systems, IEEE Transactions on Industrial Informatics, IEEE Transactions on Systems, Man, and Cybernetics: Systems, Science China Information Sciences, Science China Technological Sciences, etc.

He was listed by Clarivate Analytics/Thomson Reuters Highly Cited Researchers in Engineering in 2014-2023. He publishes about 100 IEEE Transactions journal papers with more than 20,000 citations. Moreover, Dr. Yu is also the recipient of the Second Prize of State Natural Science Award of China in 2016. He is also the Cheung Kong Scholars Programmer of Ministry of Education of China (Artificial Intelligence).