Data Driven I/O Structure Learning with Contemporaneous Causality

John A.W.B. Costanzo, Osman Yağan

***Abstract*—In the era of big data, industry and public policy are able to make use of large amounts of data for policy decisions. The proliferation of cheap sensors and fast communication enables policy makers to consider complex networks as a whole, using time series data from many sources to model the system. The Input/Output structures of such systems are helpful in understanding how they work and designing new control laws. This paper introduces the Causal Dynamic Graph (CDG) model which defines this structure explicitly. We provide a data-driven method for recovering the Input/Output structure of a CDG when every process is measured. We then discuss some of the implications of incomplete measurements on the graphical modeling and structural identification problem; we show that many relevant cases are equivalent to the simpler case where sensors are either perfect or completely missing. This will make the problem of graphically modeling such systems more tractable.**

***Index Terms*—Cause effect analysis, directed acyclic graph, graphical models, inference algorithms, network theory (graphs), signal processing algorithms, systems modeling**

NOMENCLATURE

$[p]$	$= \{0, 1, \cdots, p-1\}$.
$x[t_1 : t_2]$	$= \big(x[t_1] \quad x[t_1+1] \quad \cdots \quad x[t_2]\big)$.
$E(\mathcal{G})$	$= \mathcal{E}$. Edge set of graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.
$V(\mathcal{G})$	$= \mathcal{V}$. Vertices (nodes) of graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

I. INTRODUCTION

KNOWING a complex system's structure is invaluable in understanding how that system works. It is essential for determining what factors should be considered in predicting future trends, where to add new control laws, where to intervene to affect a desired outcome, or determining the cause of a mishap. Structural modeling has widespread applications, including in neurology [1], the spread of rumors [2], and the effects of default on financial systems [3]; for cyber physical systems, it has implications in observability [4] and cybersecurity [5], [6]. This paper aims to develop methods for determining the structure of a system by only using passively obtained measurements of the variables within the system.

There are many notions of "structure" such as the purely probabilistic Bayesian network [7] or the related functional dependency structure of a causal model [8]. We are interested in what we refer to as the *input/output structure*; a structure which relates entire time series to one another as inputs and outputs of causal processes. This allows us to make sense of a complex physical system and identify what the effects of failures or new control laws will be throughout the rest of the system. Probabilistic graphical models cannot make such guarantees because the factorization of a joint probability into conditional probabilities is not unique in general, and hence they do not characterize the effects of interventions [8], [9].

In large, complex systems, the input/output structure may not be known a priori, and it is not always possible to alter or inject signals into the system to observe how they propagate. Hence, the ability to extract information about the input/output structure from passively obtained data is of great interest. The most common formulation of such a structure is that used by [10] and [11] which assigns one node to each time series. However, for systems with instantaneous causality we will show that this is too simplified to retain a unique interpretation as an independence model and thus cannot be *directly* recovered from data. We solve this issue by giving each time series two nodes instead. This paper considers two scenarios.

In the first scenario, we assume we are able to measure the output of every process at every time instance. *Inductive Causation* [8] is capable of learning the full causal structure of such systems, up to some ambiguity in the direction of some of the instantaneously causal associations; however, it does so at high computational cost. Our algorithm, Extended Directed Information, learns the input/output structure of a system directly, without having to learn the full causal structure (which is a bigger graph). It achieves a better cost by ignoring many of the unnecessary tests Inductive Causation performs.

Our method improves upon [12] in that it can handle systems with nonlinear dynamics. It improves upon [10] in that it can handle systems with instantaneous causality. Finally, it improves upon methods based on arbitrarily sparsifying predictive models such as [13] in that it can provide guarantees about when the causal associations implied algebraically by the trained model hold true in the real system as well.

In the second scenario, we begin work toward modeling and learning the structure of systems when not all of the phenomena within a system is measured and available to the modeler. There are may ways in which we might only have partial information about a system. There may be dynamic variables that evolve too quickly for our sensors to keep up with, or there may be entire variables unmeasured or even unknown to the system supervisor.

Given a system with p variables taking on T discrete values over the window of study, our set of measurements may correspond to any subset of these pT values, of which there are very many. On the other hand, it is relatively simple to graphically represent systems where each variable is either

Manuscript received MONTH XX, YEAR; revised MONTH XX, YEAR; accepted MONTH XX, YEAR; Date of publication MONTH XX, YEAR; date of current version June 20, 2020; This work was supported in part by the TODO and in part by the TODO Recommended by TODO.

J. Costanzo and O. Yağan are with the Dept. of Elec. & Comput. Eng., Carnegie Mellon University, 1000 Forbes Ave, Pittsburgh, Pennsylvania, United States (e-mail: costanzo@cmu.edu; oyagan@andrew.cmu.edu).

measured perfectly or entirely missing. This can be done by taking the input/output structure of the full system and label some of its nodes "hidden". We will show that systems with subsampled measurements can be recast to fit this case as well, which justifies devoting particular attention to this graphical model.

Section II reviews the background on causality and reviews some existing models for networked dynamical systems in the literature. Section III discusses the limitations of those existing models and methods of structure learning. Section IV presents the Causal Dynamic Graph (CDG) model which we will be working with in the remainder of the paper. Section V presents the Extended Directed Information (EDI) algorithm, which recovers the structure of a CDG given a few of its conditional independence statements. Section VII addresses the problem of I/O structure learning in the presence of partial information sets, presents the *Latent* Causal Dynamic Graph model that models partial observation, and provides a preliminary result by which the analysis of these latent CDGs can be simplified, admitting a compact graphical representation. Finally, Section VIII concludes the paper.

II. Background on Graphical Models and Causality

An input/output relationship between two time series indicates causal relationships between the elements of these time series. The study of causality involves answering three broad classes of questions about a collection of random variables $\mathbf{X} = \{X_j : j \in [p]\}$ [8]. Different graphical models differ in their ability to answer three types of questions: *inference/prediction*, *intervention*, and *counterfactual*. Probablistic or predictive graphical models answer questions of the first type; they do not provide any guarantees as to the effect of interventions.

In this section we present the background theory of causality in more detail and explain the connection to causal dynamical systems; we refer the reader to [8] for more detail. We will use these ideas to build the Causal Dynamic Graph model we will use in the remainder of the paper.

A. Graph Theory Preliminary Definitions

Definition 1. *A graph $\mathcal{G}$ is said to be* **directed** *if $E(\mathcal{G}) \subset V(\mathcal{G}) \times V(\mathcal{G})$, and* **undirected** *if $E(\mathcal{G}) \subset \{\{i,j\} : i,j \in V(\mathcal{G})\}$. An edge (u,v) in a directed graph can also be called an* **arrow** *and denoted $u \to v$. If $u \to v$ then v is a* **child** *of u and u is a* **parent** *of v. The set of all parents of v in $\mathcal{G}$ is denoted $\rho_{\mathcal{G}}(v)$.*

If $u = u_1 \to u_2 \to \cdots \to u_n = v$ then v is called a **descendant** *of u. A graph is called a* **directed acyclic graph** *(DAG) if no node is its own descendant.*

Definition 2. *For $v \in V(\mathcal{G})$, the* **neighbors** *of v are* $\mathrm{ne}_{\mathcal{G}}(v) = \begin{cases} \{x \in V(\mathcal{G}) : \{x,v\} \in E(\mathcal{G})\} & \mathcal{G} \text{ undirected} \\ \{x \in V(\mathcal{G}) : (x,v) \in E(\mathcal{G}) \text{ or } (v,x) \in E(\mathcal{G})\} & \mathcal{G} \text{ directed} \end{cases}$.
Two nodes are **adjacent** *if they are each others' neighbors.*

Definition 3. *[7] Given a collection of p random variables $\mathbf{X} = \{X_i\}_{i \in [p]}$ with probability distribution $\Pr(\mathbf{X})$ and a directed acyclic graph with vertex set $[p]$ corresponding to the variables in $\mathbf{X}$, we say this probability distribution is* **Markov relative to** *$\mathcal{G}$ and $\mathcal{G}$ is a* **Bayesian network** *for $\Pr(\mathbf{X})$ if $\Pr(\mathbf{X})$ factors according to $\mathcal{G}$ as:*

$$\Pr(\mathbf{X}) = \prod_{i \in [p]} \Pr\left(X_i \mid X_{\rho_{\mathcal{G}}(i)}\right), \tag{1}$$

where $\rho_{\mathcal{G}}(i)$ are the parents of i in $\mathcal{G}$.

A random variable X_a is *conditionally independent of* X_b *given* the variables X_C (denoted $X_a \perp\!\!\!\perp X_b \mid X_C$) if [7]:

$$\Pr(X_a \mid X_b, X_C) = \Pr(X_a \mid X_C). \tag{2}$$

Bayesian networks are useful because they graphically encode all of the conditional independence relations among all subsets of variables in the network via d-separation:

Definition 4. *[7] Given $i,j \in [p]$ and $Z \subset [p] \setminus \{i,j\}$, we say that a trail q between i and j is* **d-separated in** *$\mathcal{G}$* **given** *Z if q either*

- *contains a* chain $q_{i-1} \to q_i \to q_{i+1}$, *satisfying $q_i \in Z$;*
- *contains a* fork $q_{i-1} \leftarrow q_i \to q_{i+1}$, *satisfying $q_i \in Z$;*
- *contains a* collider $q_{i-1} \to q_i \leftarrow q_{i+1}$, *satisfying that neither q_i nor any of its descendants in $\mathcal{G}$ are in Z.*

If all trails between i and j are d-separated given Z, then we say i and j are d-separated given Z, denoted $I_{\mathcal{G}}(i,Z,j)$. For sets A,B,C we say $I_{\mathcal{G}}(A,C,B)$ if and only if $I_{\mathcal{G}}(i,C,j)$ for every $i \in A$ and $j \in B$.

By Verma and Pearl [14], if $\mathcal{G}$ is a Bayesian network for $\mathbf{X}$, then if $I_{\mathcal{G}}(A,C,B)$, then $X_A \perp\!\!\!\perp X_B \mid X_C$. Given two nonadjacent nodes a and b, to find a set that separates them, it suffices to consider only the parents of a or the parents of b:

Fact 1 (Parental Markov Condition)**.** *[8] A collection of random variables $\mathbf{X}$ is Markov relative to a directed acyclic graph $\mathcal{G}$ if and only if each X_i is conditionally independent of its nondescendants in $\mathcal{G}$, given its parents in $\mathcal{G}$.*

B. Causal Graphical Models

Probablistic graphical models such as Bayesian networks do not model causality because interventions are a violation of the joint probability distribution. There are often many Bayesian networks that can factor a given probability distribution, and not all of them agree in the directions of their arrows—and therefore, on the effects of interventions. Since a causal model must specify more than just the joint probability, Pearl [8] developed the following model:

Definition 5. *[15] A* **Functional Causal Model** *is a pair $(\mathcal{G}, \Theta)$ where $\mathcal{G}$ is a directed acyclic graph with vertex set $[p]$ called the* **causal structure** *of the model, and Θ is a set of parameters assigning a distribution to pairwise independent random variables $\{W_i\}$ and a set of measurable functions f_i, where:*

$$X_i = f_i(X_{\rho_{\mathcal{G}}(i)}, W_i), \quad i = 1, \cdots, p \tag{3}$$

where $\rho_{\mathcal{G}}(i)$ denotes the parents *of i in $\mathcal{G}$.*

Every FCM admits a probability distribution $\Pr_\Theta(\mathbf{X})$ on the random variables $\mathbf{X}$; moreover, $\mathcal{G}$ is a Bayesian network for $\Pr_\Theta(\mathbf{X})$. Hence, functional causal models inherit all of the algorithms designed for Bayesian networks; see, e.g., [16], [17], [18].

C. Latent Models and Causal Sufficiency

For situations where only certain variables $\mathbf{X}_\mathcal{O}$ are measured, Pearl introduces the following model:

Definition 6. *[8] A* **latent structure** *is a pair $L = (\mathcal{G}, \mathcal{O})$ where $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a causal structure and $\mathcal{O} \subset \mathcal{V}$ represents the observable nodes. If $u \in \mathcal{V} \setminus \mathcal{O}$, then u is called a* **latent node** *of L.*

The drawback of the latent structure as a graphical model is that it depends on the latent nodes, which the modeler may not even be aware of. The following graphical model remedies this.

Definition 7 (Projection)**.** *[8] A latent structure $L_\mathcal{O} = (\mathcal{G}_\mathcal{O}, O)$ is a* **projection** *of another latent structure L if and only if*

- *every latent node in $\mathcal{G}_\mathcal{O}$ has no parents and two nonadjacent children in $\mathcal{O}$; and*
- *for every model with latent structure L, there exists a model with latent structure $L_\mathcal{O}$ that produces the same output $\mathbf{X}_\mathcal{O}$.*

We represent latent projections as mixed graphs with bidirected edges representing its latent nodes; that is, $a \leftarrow L \rightarrow b$ is represented by $a \leftrightarrow b$ and L is omitted entirely.

By Verma [19], every latent structure has at least one projection.

Definition 8. *[8] A latent structure is said to be* **causally sufficient** *if every node $j \in \mathcal{V}$ that has more than one child satisfies $j \in \mathcal{O}$.*

The projection of a causally sufficient system is always a DAG, and hence "causally sufficient" is often used interchangeably with "not latent".

D. Equivalence and Dependency Equivalence

Two structures are *observationally equivalent* if no data-driven experiment can tell them apart; i.e., if any joint distribution generated by one structure could have been generated by the other. *Dependency equivalence* is a coarser partition:

Definition 9. *[8] Two models $\mathcal{M}_1$ and $\mathcal{M}_2$ are said to be* **dependency equivalent** *if $I_{\mathcal{M}_1}(X, Z, Y) \iff I_{\mathcal{M}_2}(X, Z, Y)$.*

For causally sufficient systems, observational equivalence is equivalent to dependency equivalence [20]. Thus, the set of conditional independence statements of a system tell us everything we can know about the causal structure without making additional assumptions about f_i or $\{W_i\}$.

Definition 10. *[8] An* **open collider** *in a DAG is a triple of vertices x_i, x_j, x_k such that $x_i \rightarrow x_j \leftarrow x_k$ in the DAG but x_i and x_j are non-adjacent.*

Definition 11. *[8] The* **pattern** *of a DAG $\mathcal{G}$ is the mixed graph consisting of the skeleton of $\mathcal{G}$ and its open colliders.*

By the Pearl-Verma Theorem [8], two DAGs are equivalent if and only if they have the same pattern. Therefore, the pattern is the most that any data-driven, independence-based method can determine about a system's causal structure.

III. Limitations of Existing Dynamic Causal Inference Methods

Work in recovering the "structure" of collections of coupled processes has been a subject of study at least since Granger proposed [21] the statistic his name has become synonymous with. One branch of research in this field attempts to recover the structure through sparse estimation; see [13] for an example. But the sparsity of the resulting predictor is not guaranteed to relate to the structure of the system. Past values of process X may be very relevant in predicting the future values of process Y, but not actually have a causal effect on such values.

Early work in the field of structural identification includes the study of the structures of linear tree structures [22], ultimately leading to the development of the Linear Dynamic Graph model. Research in the field of recovering linear dynamic graphs continued in [23] and [24] leading up to the Extended Granger Filter [12], which can recover the structure of any "causally well-posed" linear dynamic graph.

In the area of nonlinear systems, [10] shows that for any system with a probability distribution that can be written as a product of factors of the form $\prod_j P(x_j[t]|\mathbf{x}_{S_j}[0:t-1])$, where the S_j are sets of indices, then it is true that $i \in S_j$ if and only if the conditional directed information from i to j is positive.

All of these methods have limitations: heuristic methods such as [13] do not provide correctness guarantees; [23] recovers only the undirected version of the system, and finds spurious links; [24] requires the system have no feedback loops; [12] requires that the system be linear and "causally well-posed"; and [10] requires that the system be strictly causal.

A. Strict Causality vs Causal Well-Posedness

Methods like Directed Information (DI) Graphs [10] require the assumption that the system is strictly causal and fail for systems with contemporaneous causality. Consider the system:

$$x_j[t] = \begin{cases} f_j(x_j[t-1], x_4[t-1], w_j[t]) & j = 0 \\ f_j(x_j[t-1], x_{j-1}[t], w_j[t]) & j \in \{1,2,3,4\} \end{cases} \tag{4}$$

While we might wish the DI graph of this system were the cycle $0 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 0$, reflecting the direct dependence of process j on process $j-1 (\mod 5)$, the directed information $I(x_4 \rightarrow x_j || x_{\overline{j3}})$ is positive for every $j \in \{0,1,2,3\}$, and therefore in the DI graph every node has an erroneous arrow into 4.

Granger [21] posits that contemporaneous causality is an indication of having too large of a sampling period. More generally, others [25] suggest that contemporaneous causality between two variables should always be interpreted as the effect of a latent node.

This claim seems rather strong. The Extended Granger Filter [12], for instance, requires only that the system is "causally well-posed", that is, instantaneous causality is allowed provided merely that there are no instantaneous causality *loops*. Hence, whether or not instantaneous causality is actually physically possible, from a practical perspective, there is no need to rule it out entirely.

One might then ask if the assumptions could be weakened even further. One can, at least mathematically, describe stable systems which have loops without delays in them. A simple example is seen in Figure 1. But while this system makes sense algebraically, it cannot exist as depicted, because information cannot make a round trip faster than light. In such cases, we must conclude, as Granger does, that incomplete data is causing the loop.

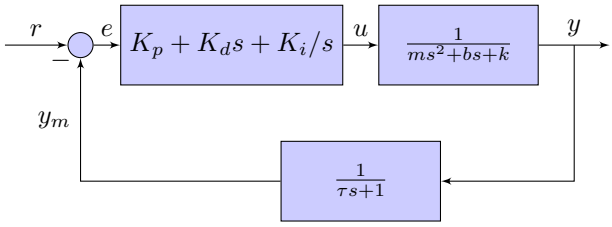


Fig. 1. A block diagram of a feedback control system common in elementary control theory.

B. Causal Aliasing

Unfortunately, the association between contemporaneous causality and incomplete data is not strong. Not only is contemporaneous causality not always spurious, it is also neither a necessary effect nor the only potential effect of unobserved variables. They are separate issues.

Allowing for the possibility of unmeasured variables has more profound consequences than just contemporaneous causality or causal ill-posedness. For instance, if we have three signals, $x[t]$, $y[t]$, and $z[t]$, given by:

$$\begin{aligned} x[t] = e_1[t], \quad y[t] = a_1 x[t-1] + e_2[t], \\ z[t] = b_1 y[t-1] + e_3[t], \end{aligned} \tag{5}$$

where $e_i[t]$ are all mutually independent, and consider what happens if we only have measurements of $x[2t]$, $y[2t]$, and $z[2t]$. We obtain:

$$\begin{aligned} x[2t] = e_1[2t], \quad y[2t] = a_1 e_1[2t-1] + e_2[2t], \\ z[2t] = b_1 a_1 x[2(t-1)] + b_1 e_2[2t-1] + e_3[2t]. \end{aligned} \tag{6}$$

This shows that this dataset is consistent with an alternative model in which the z time series is directly a function of the x time series, with y depending on neither. That alternative model, while being the simplest explanation of the data, is not true; using such a model to make policy decisions could have unintended consequences. We call this *causal aliasing*: the condition in which classes of structures become indistinguishable from one another due to sampling. In a later section, we propose a latent node graphical model whose representation does not explicitly depend on missing nodes. First, we propose a graphical model for fully observed systems which the latent node model will be based on.

IV. The Causal Dynamic Graph Model

In this section, we will present our model for large networked systems. The model we present is *stochastic* in that it allows each of its outputs a degree of autonomy, and *dynamic* in that each output influences each other over time.

In general, a stochastic process is defined as follows:

Definition 12. *[26] Let $(\Omega, \mathcal{F}, P)$ be a probability space, (S, Σ) be a measurable space, and $\mathcal{T}$ be any set. An S-valued, $\mathcal{T}$-indexed stochastic process is a function $X : \mathcal{T} \times \Omega \to S$ such that for every $t \in \mathcal{T}$, $X(t, \cdot)$ is $(\Sigma, \mathcal{F})$-measurable.*

For any $\omega \in \Omega$, the mapping $X(\cdot, \omega) : \mathcal{T} \to S$ is called a **realization** *or* **sample path** *of the stochastic process.*

We denote $X(t) : \Omega \to S$ defined $X(t)(\omega) = X(t, \omega)$ as the random variable representing the value of the process X at time t. When $\mathcal{T} = [T]$ for some $T \in \mathbb{N}$, we instead write $X[t]$ to make clear that it is a discrete-time process. Physical systems often have an explicit dependence of the present on the past that is not captured by Definition 12.

In large scale systems, we have many co-evolving phenomena and many sensors producing data; we receive a collection of measurements at each time. We shall denote a realization of the jth process variable at time t as $x_j[t]$, and say $\mathbf{x}[t]$ is a realization of all variables at time t in vector form.

For each j there is some subset $\rho^C(j) \subset [p]$ such that $x_j[t]$ depends on $x_k[\tau]$ for some $\tau < t$, if and only if $k \in \rho^C(j)$. A dynamic variable may depend on its own history, so we do allow $j \in \rho^C(j)$. A causal system, if it is not "strictly" causal, may also satisfy that $x_j[t]$ depends on $x_i[t]$ if and only if $i \in \rho^I(j) \subset [p] \setminus \{j\}$, but this must not introduce a directed cycle as this would be paradoxical.

The above considerations motivate the following model:

Definition 13 (Causal Dynamic Graph)**.** *Let $p, T \in \mathbb{Z}$, $\{f_i : i \in [p]\}$ be a set of measurable functions, and $\mathbf{W} = \{W_i[\tau] : i \in [p], \tau \in [T]\}$ be a stochastic process with index set $[p] \times [T]$ satisfying $W_i[\tau] \perp W_j[\tau']$ for $(i, \tau) \neq (j, \tau')$ and that for each i, $\{W_i[\tau]\}_{\tau \in [T]}$ is an i.i.d. process.*

Let $\mathcal{G}^{\mathcal{C}}$ be a directed graph with vertex set $[p]$ and $\mathcal{G}^{\mathcal{I}}$ be a directed acyclic *graph with vertex set $[p]$. For $j \in [p]$, let $\rho^{\mathcal{C}}(j)$ be the parents of j in $\mathcal{G}^{\mathcal{C}}$ and $\rho^{\mathcal{I}}(j)$ be the parents of j in $\mathcal{G}^{\mathcal{I}}$.*

A **Causal Dynamic Graph** *(CDG) is the tuple $\mathcal{M} = (\{f_j\}_{j \in [p]}, \{W_j[t]\}_{j \in [p], t \in [T]}, \mathcal{G}^{\mathcal{C}}, \mathcal{G}^{\mathcal{I}})$ producing output random variables $\mathbf{X} = \{X_j[t] : j \in [p], t \in [T]\}$ via the equations:*

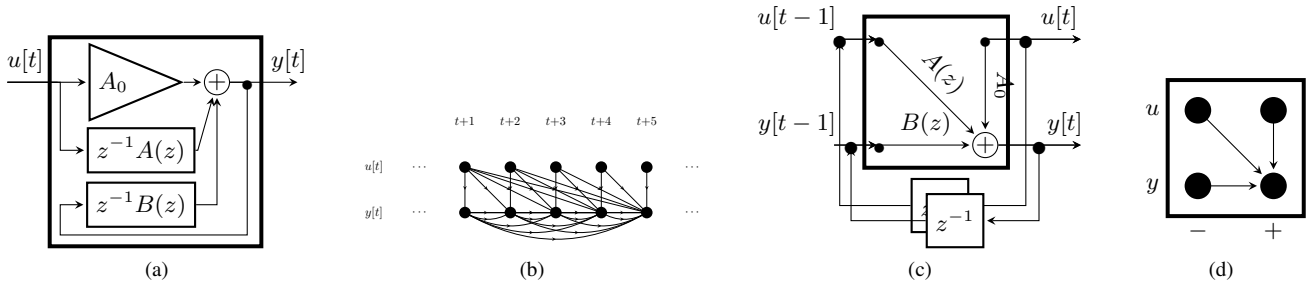
$$X_j[t] = f_j\left(X_{\rho^{\mathcal{C}}(j)}[0 : t-1], X_{\rho^{\mathcal{I}}(j)}[t], W_j[t]\right). \tag{7}$$

Every realization $\mathbf{W} = \mathbf{w}$ produces a realization $\mathbf{X} = \mathbf{x}$:

$$x_j[t] = f_j\left(x_{\rho^{\mathcal{C}}(j)}[0 : t-1], x_{\rho^{\mathcal{I}}(j)}[t], w_j[t]\right). \tag{8}$$

$(\mathcal{G}^{\mathcal{C}}, \mathcal{G}^{\mathcal{I}})$ are called the **component graphs** *of $\mathcal{M}$; $\mathcal{G}^{\mathcal{C}}$ is called the* **causal part** *and $\mathcal{G}^{\mathcal{I}}$ is called the* **instantaneous part***. We refer to the different x_j as* **variables** *and to the particular $x_j[t]$ as* **measurements***.*

Every CDG is a functional causal model [8], and as such, has a *causal structure*, given below:



(a) (b) (c) (d)

Fig. 2. (a) IIR filter. $A(z)$ and $B(z)$ are polynomial in z^{-1}. (b) Part of causal structure if A and B have degree ≥ 4. (c) Strictly causal elements represented as feeedback. (d) I/O structure.

Definition 14. *The* **causal structure** *of a CDG is the DAG $\mathcal{G}$ with vertex set $[p] \times [T]$ and an arrow $(i,t) \to (j, t+t')$ if $x_j[t+t']$ is a non-constant function of $x_i[t]$.*

Not all DAGs are the causal structures of CDGs; therefore we distinguish them with the following terminology.

Definition 15. *A DAG $\mathcal{G}_{DAG}$ with vertex set $[p] \times [T]$ is called a* **Dynamic DAG** *if there exists a Causal Dynamic Graph with causal structure $\mathcal{G}_{DAG}$.*

Algorithms for finding the structure of a time invariant CDG by searching among all possible DAGs can be improved in their efficiency and precision by limiting their search space to the much smaller class of dynamic DAGs.

However, a time invariant system can also be described by the much more compact *input/output structure*:

Definition 16. *Let $\mathcal{M}$ be a $[p]$-process CDG with component graphs $(\mathcal{G}^{\mathcal{C}}, \mathcal{G}^{\mathcal{I}})$. The* **input/output (I/O) structure** *(also called the* **two-factor** *I/O structure to distinguish it from similar structures) of $\mathcal{M}$ is the graph $\mathcal{S} = ([p] \times \{-,+\}, \mathcal{E}_S)$ where $(i,-) \to (j,+)$ if $i \to j$ in $\mathcal{G}^{\mathcal{C}}$ and $(i,+) \to (j,+)$ if $i \to j$ in $\mathcal{G}^{\mathcal{I}}$.*

$\mathrm{comp}(\mathcal{S}) = (\mathcal{G}^{\mathcal{C}}, \mathcal{G}^{\mathcal{I}})$ are called the **component graphs** *of $\mathcal{S}$.*

The nodes of the form $(j,-)$ represent the entire past of process j, $x_j[0:t-1]$; the nodes of the form $(j,\cdot)$ represent the current value of process j, $x_j[t]$. (Since a CDG is time invariant, t is arbitrary.) Figure 2 demonstrates how this structure is obtained from the block diagram of a system by isolating feedback.

While a more typical structural model is the union of $\mathcal{G}^{\mathcal{C}}$ and $\mathcal{G}^{\mathcal{I}}$ (sometimes called the "generative graph"), such as in [10], [27], [28], and [23], for systems with instantaneous causality it is necessary (at least during the learning phase) to consider these components separately.

One way to demonstrate this is to consider what a node "represents". In [10], every arrow $i \to j$ corresponds to a positive directed information $DI(\mathbf{x}_i \to \mathbf{x}_j \mid \mathbf{x}_{\overline{ij}})$, which implies a relationship between $x_i[0:t-1]$ and $x_j[t]$. In this case, node i represents $x_i[0:t-1]$. But if there was another arrow *into* i, then that same node i represents $x_i[t]$. Therefore, what the node represents depends on which arrow one is looking at and which of its endpoints that node is. With the two-factor I/O structure proposed here, the present $x_i[t]$ and past $x_i[0:t-1]$ each get their own node; this is necessary when a system has contemporaneous causality because an arrow can connect two present nodes in this case.

Moreover, while the graph used in [23] to describe the system is allowed to be cyclic, our two-factor I/O structure cannot be cyclic. Therefore, the causal well-posedness of a system is explicit, allowing us to guarantee that the resulting system satisfies the Causal Markov Condition without having to assume that there is no instantaneous causality, as [28] does. This will allow us to obtain the exact topology up to the orientation of some contemporaneous links. We contend that this is also (implicitly) the graph that the method provided in [12] uses for linear systems; making this graph explicit allows for extension to nonlinear systems.

Finally, for a given $\mathcal{S}$, there exists a unique dynamic DAG $\mathcal{G}^*$ such that, for any CDG with I/O structure $\mathcal{S}$ and causal structure $\mathcal{G}$, we have $\mathcal{G}$ is a subgraph of $\mathcal{G}^*$. The same is not true if we replace $\mathcal{S}$ with the "generative graph" $\mathcal{G}^{\mathcal{C}} \cup \mathcal{G}^{\mathcal{I}}$.

A. Connection Between I/O Structure and Causal Structure

Definition 17. *Let $\mathcal{G} = ([p] \times [T], E)$ be a dynamic DAG. The* input/output structure *of $\mathcal{G}$, denoted $\mathrm{IO}(\mathcal{G})$, is the graph with vertex set $([p] \times \{-,+\}, \mathcal{E}_S)$ with an edge from $(i,-) \to (j,+)$ if there exists a $\tau > 0$ such that $(i, T-\tau) \to (j,T)$ in $\mathcal{G}$ and $(i,+) \to (j,+)$ if $(i,T) \to (j,T)$ in $\mathcal{G}$.*

Remark 1. *If $\mathcal{M}$ is a time invariant CDG with causal structure $\mathcal{G}$ and I/O structure $\mathcal{S}$, then $\mathcal{S} = \mathrm{IO}(\mathcal{G})$.*

V. Recovering the I/O Structure of Fully Observed Time Invariant Systems

In this section we present an algorithm that, given measurements of every variable in a time invariant system at every time step, recovers its I/O structure up to some ambiguity in the direction of the contemporaneous causal links; see Fig. 4 for an illustration. Strict causality is not assumed.

A first approach at the problem might be to use inductive causation [8] to find the pattern (Def. 11) of the full causal structure of the CDG, additionally using temporal information to orient any adjacencies $(i,t_1) \to (j,t_2)$ if $t_1 < t_2$. The I/O structure is derived from this. However, if the I/O structure (or even a simpler structure) is all we wish to obtain, inductive causation performs a lot of unnecessary work, resulting in a

naïve time complexity of up to 2^{pT}, pT being the number of nodes in the causal structure. Efficient implementations such as the PC algorithm [29] do better, with a time complexity $\mathcal{O}\left((pT)^k\right)$ where k is the maximum degree of $\mathcal{G}$. However, given that the I/O structure does not grow with T, one might ask if we can do even better by not reconstructing the causal structure at all—it does, after all, contain far more information than we need. By removing many of the tests that are not relevant to the I/O structure, we propose a faster algorithm for finding the I/O structure directly.

A. The Extended Directed Information (EDI) Algorithm

Inductive causation [8] is not a statistical algorithm but rather a method of interpreting statistics. It requires, as its input, all statements of the form $x_i[t_1] \perp\!\!\!\perp x_j[t_2] \mid Z$ with $Z \subset \mathbf{x}[0:T] \setminus \{x_i[t_1], x_j[t_2]\}$. On the other hand, we require only the following conditional independence statements:

$$x_j[t] \perp\!\!\!\perp x_i[t] \mid \mathbf{x}[0:t-1], \mathbf{x}_C[t] \tag{9}$$

for all $C \subset [p] \setminus \{i, j\}$ and

$$x_j[t] \perp\!\!\!\perp x_i[0:t-1] \mid \mathbf{x}_{\bar{i}}[0:t-1], \mathbf{x}_S[t] \tag{10}$$

for all $S \subset [p] \setminus \{j\}$, where $\bar{i} = \mathcal{V} \setminus \{i\}$, and $\mathbf{x}_S[\tau] = \{x_s[\tau] : s \in S\}$.

Since $\mathbf{x}$ is the output of a stationary CDG, the variable t in (9) and (10) is just a placeholder. In Section VI, we discuss a few classical methods that could be used to obtain (9) and (10) for certain systems, though we emphasize that our intended contribution is not a method for determining conditional independence but rather interpreting conditional independence. In other words, no matter how these statements are provided, the structure identification can be performed by Algorithm 1 proposed by us.

The $\mathcal{G}^I$ returned by EDI is partially directed. An arrow in $\mathcal{G}^C$ or $\mathcal{G}^I$ corresponds to an arrow in the corresponding I/O structure element, whereas an undirected edge in $\mathcal{G}^I$ means that some orientation of that edge is in the instantaneous part of the I/O structure. We call $(\mathcal{G}^C, \mathcal{G}^I)$ the I/O *pattern* of the system; rather than being a single I/O structure, it represents a class of I/O structures whose members may disagree on the orientation of certain links in the instantaneous part.

B. Proof of Correctness

Our EDI algorithm is derived from Inductive Causation (IC) [8], but whereas IC would check every pair $x_i[t], x_j[t']$ for conditional independence given every possible subset of $\{x_k[\tau] : k \in [p], \tau \in [T]\}$, we incorporate temporal directionality and symmetry to render most of these tests irrelevant. The correctness of EDI follows from its compatibility with IC.

Theorem 1. *Let $\mathcal{G}$ be the causal structure of a time invariant CDG, let $\mathcal{G}_{IC}$ be its pattern as recovered by Inductive Causation, let $\operatorname{comp}(\mathcal{G}_{IC}) = \left(\mathcal{G}^C_{IC}, \mathcal{G}^I_{IC}\right)$ (see Def. 17,) and let $\left(\mathcal{G}^C_{EDI}, \mathcal{G}^I_{EDI}\right)$ be the I/O pattern as recovered by EDI. Then:*

$$\mathcal{G}^C_{EDI} = \mathcal{G}^C_{IC}; \quad \textit{and} \quad \operatorname{sk}\left(\mathcal{G}^I_{EDI}\right) = \operatorname{sk}\left(\mathcal{G}^I_{IC}\right)$$

Algorithm 1 Extended Directed Information (EDI)

for $i, j \in [p]$ **do**
  Find a set $C_{ij} \subset [p] \setminus \{i, j\}$ satisfying:
    $x_i[t] \perp\!\!\!\perp x_j[t] \mid \mathbf{x}[0:t-1], x_{C_{ij}}[t]$
  If no such set exists, then i and j are neighbors in $\mathcal{G}^I$.
end for
for $i, j \in [p]$ (including $i = j$) **do**
  Find a set $S_{ij} \subset \operatorname{ne}_{\mathcal{G}^I}(j)$ satisfying:
    $x_j[t] \perp\!\!\!\perp x_i[0:t-1] \mid x_{\bar{i}}[0:t-1], x_{S_{ij}}[t]$
  If no such set exists, add an arrow $i \to j$ in $\mathcal{G}^C$.
end for
for every i, j that are adjacent in $\mathcal{G}^I$ **do**
  if If there exists a k (including if $k = i$ or $k = j$) such that: $k \to i$ in $\mathcal{G}^C$ **and** $k \not\to j$ in $\mathcal{G}^C$ **then**
    if $i \notin S_{kj}$, orient $j \to i$ in $\mathcal{G}^I$ **else** orient $i \to j$ in $\mathcal{G}^I$.
  end if
  if there exists a k other than i or j such that: k and i are adjacent in $\mathcal{G}^I$ **and** k and j are *not* adjacent in $\mathcal{G}^I$, **then**
    if $i \notin C_{kj}$ **then** orient $k \to i$ and $j \to i$ in $\mathcal{G}^I$.
  end if
end for
Of any remaining unoriented adjacencies in $\mathcal{G}^I$, if one orientation would result in a new open collider or a directed cycle in $\mathcal{G}^I$, reject that orientation and conclude the opposite.
return $(\mathcal{G}^C, \mathcal{G}^I)$.

where $\operatorname{sk}(D)$ (short for "skeleton") denotes the graph formed by unorienting the edges in the partially oriented graph D.

Moreover, if $i \to j$ in $\mathcal{G}^I_{EDI}$, then $i \to j$ in $\mathcal{G}^I_{IC}$ as well.

Proof. Fix $t \in [T]$ arbitrarily.

1) ($\mathcal{G}^C_{EDI} \subseteq \mathcal{G}^C_{IC}$.) Suppose IC determines $i \not\to j$ in $\mathcal{G}^C$. Then, for all $\tau > 0$, we must have $(i, t-\tau) \not\to (j, t)$ in $\mathcal{G}$. Let $S = \rho((j, t))$ be the parents of (j, t) in $\mathcal{G}$. Since no $(i, t-\tau)$ can be a descendant of (j, t) in a DDAG, this means that, by Fact 1, S is a separating set between (j, t) and $A = \{(i, t-\tau) : \tau > 0\}$. The set $B = \{(k, t-\tau) : k \neq i, \tau > 0\}$ is also a set of nondescendants of (j, t), so S d-separates (j, t) from $B \setminus S$ as well.
   By the weak union axiom [30], it follows that $S \cup (B \setminus S)$ d-separates (j, t) from A.
   Now that we have established that $x_j[t]$ is conditionally independent of $x_i[0:t-1]$ given the random variables represented by the nodes in $S \cup B$, we need only finally note that all of the elements of S that are not in B have time index t. Since, S is the set of parents of (j, t), all of the nodes in $S \setminus B$ therefore have spatial index corresponding to a neighbor of j in $\mathcal{G}^I$. Therefore, step 2(a) in EDI will find a set ruling out the adjacency of $i \to j$ in $\mathcal{G}^C$.
2) ($\mathcal{G}^C_{IC} \subseteq \mathcal{G}^C_{EDI}$.) $i \to j$ in $\mathcal{G}^C_{IC}$ implies $(i, t-\tau)$ and (j, t) are adjacent in $\mathcal{G}$ and therefore no set, including the sets step 2 considers, can possibly separate them.
3) ($\operatorname{skeleton}\left(\mathcal{G}^I_{EDI}\right) \subseteq \operatorname{skeleton}\left(\mathcal{G}^I_{IC}\right)$.) Suppose IC determines i and j are not adjacent in $\mathcal{G}^I$. Then there exists an S such that for all $t > 0$ $x_i[t] \perp\!\!\!\perp x_j[t] \mid S$. Without loss of generality suppose (i, t) is not a descendant of (j, t). By

Fact 1, the parents of (j,t) in $\mathcal{G}$ ($S = \rho((j,t))$) separate (j,t) from (i,t). Let $B = \{(k, t-\tau) : k \in [p], \tau > 0\}$. Then again by the weak union axiom [30], $B \cup S$ separates (j,t) from (i,t). Moreover $B \cup S$ is the type of set that step 1(a) of EDI would look for to separate $x_j[t]$ from $x_i[t]$, and therefore EDI would conclude i and j are not adjacent in $\mathcal{G}^I$ either.

4) (skeleton $\left(\mathcal{G}^I_{IC}\right) \subseteq$ skeleton $\left(\mathcal{G}^I_{EDI}\right)$.) Similar to 2).
5) ($i \to j$ in $\mathcal{G}^I_{EDI}$ implies $i \to j$ in $\mathcal{G}^I_{IC}$.) Suppose step 3 of EDI determines $i \to j$ in $\mathcal{G}^I$. This implies either:
   - There is a $k \in [p]$ such that $k \to i$ in $\mathcal{G}^C$, $k \not\to j$ in $\mathcal{G}^C$, and $i \in S_{kj}$. This implies that there exists a τ such that $(k, t-\tau) \to (i,t)$ in $\mathcal{G}$. A previous result already implies (i,t) and (j,t) are adjacent in $\mathcal{G}$. Orienting $(j,t) \to (i,t)$ would create an open collider $(k, t-\tau) \to (i,t) \leftarrow (j,t)$ in $\mathcal{G}$, but since $i \in S_{kj}$ this open collider cannot exist. Therefore $(i,t) \to (j,t)$ in $\mathcal{G}$.
   - There is a $k \in [p]$ such that $k \to j$ in $\mathcal{G}^C$, $k \not\to i$ in $\mathcal{G}^C$, and $j \notin S_{ki}$. This implies that there exists a $\tau > 0$ such that $(k, t-\tau) \to (j,t)$ in $\mathcal{G}$. Since (j,t) and (i,t) are at least adjacent in $\mathcal{G}$, the fact that (j,t) does not separate $(k, t-\tau)$ from (i,t) implies that $(k, t-\tau) \to (j,t) \leftarrow (i,t)$ must be an open collider in $\mathcal{G}$.
   - There is a $k \in [p] \setminus \{i,j\}$ such that k and j are adjacent, k and i are nonadjacent, and $j \notin C_{ki}$. This implies $(k,t) \to (j,t) \leftarrow (i,t)$ is an open collider in $\mathcal{G}$.

   As to the correctness of step 4, it is obvious that $\mathcal{G}^I$ cannot have any directed cycles, so rejecting any orientation leading to one is correct. On the other hand, every open collider in $\mathcal{G}^I_{IC}$ corresponds to a sequence of open colliders in $\mathcal{G}$, since all adjacencies between triplets (i,t), (j,t) and (k,t) in a dynamic DAG are represented in $\mathcal{G}^I$. Since Inductive Causation detects all open colliders, this implies that $\mathcal{G}^I_{IC}$ has all of its open colliders oriented, and that any orientation of any edge that would result in a new open collider would have been rejected by the last step of IC.

   Hence we need only show that all open colliders in $\mathcal{G}^I_{IC}$ will be added to $\mathcal{G}^I_{EDI}$ in step 3. For $i \to k \leftarrow j$ to be open collider in $\mathcal{G}^I_{IC}$ would mean $(i,t) \to (k,t) \leftarrow (j,t)$ in $\mathcal{G}$. Since (i,t) and (j,t) are not adjacent, this means there is a set that separates them; moreover, this set cannot include (k,t). Since (k,t) cannot represent any of the variables in $\mathbf{x}[0:t-1]$, it suffices to ensure that k is not an element of C_{ij} in step 1 of EDI. This is precisely what step 3(b) does, and hence any ambiguities leading to open colliders after step 3 terminates are correctly rejected in step 4.

□

VI. Simulation Results

A. Simulation Set-up

Extended Directed Information is a method of interpreting statistics, not computing them. As such, it is necessary to first decide on a method appropriate to the data for determining the conditional independence statements required. If the system is linear and Gaussian, for example, whether $x_j[t] \perp\!\!\!\perp x_i[t] \mid \mathbf{x}[0:t], \mathbf{x}_C[t]$ is true can be determined through the *partial correlation* [31] of $x_j[t]$ and $x_i[t]$ given $\mathbf{x}[0:t], \mathbf{x}_C[t]$, which can be found by solving the least squares problem:

$$\min_{\mathbf{w}} \sum_t \left(x_j[t] - \sum_{k \in Z} w_k[0] x_k[t] - \sum_{\substack{\tau > 0 \\ k \in [p]}} w_k[\tau] x_k[t-\tau] \right)^2 \quad (11)$$

with $Z = C \cup \{i\}$ and determining whether $w_i[0] \neq 0$ is statistically significant (see e.g., [32] for details). To be clear, no one optimization problem suffices to find the *structure* of the system. They only provide the statistics by which EDI infers the structure.

Structure identification has two components: the gathering of the statistics, and the interpretation of those statistics into statements about the structure. We have proven that EDI performs the second component correctly. We will now show that these components do work together to yield a practical algorithm. To demonstrate the benefit of EDI over other methods, we consider the following two systems with instantaneous causality. The first scenario is:

$$\mathbf{x}[t] = \mathbf{e}[t] + 0.4 \begin{bmatrix} 0 & & & 0 \\ 1 & & & 0 \\ 0 & 1 & & 0 \\ 0 & & 1 & 0 \end{bmatrix} \mathbf{x}[t] + \begin{bmatrix} 0.5 & & & 0.4 \\ 0 & 0.5 & & 0 \\ 0 & & 0.5 & 0 \\ 0 & & & 0.5 \end{bmatrix} \mathbf{x}[t-1], \quad (12)$$

and the second scenario is:

$$\mathbf{x}[t] = \mathbf{e}[t] + 0.1 \begin{bmatrix} 0 & & & 0 \\ 1 & 0 & 1 & 0 \\ 0 & & & 0 \\ 0 & 1 & 0 & 1 \\ 0 & & & 0 \end{bmatrix} \mathbf{x}[t] + \begin{bmatrix} 0.5 & & & 0.1 \\ 0 & 0.5 & & 0 \\ 0 & & 0.5 & 0 \\ 0 & & & 0.5 \end{bmatrix} \mathbf{x}[t-1], \quad (13)$$

with I/O structures and generative graphs shown in Figure 3.

In order to test our proposed approach, we generated a realization of these systems and used EDI to recover their structure. The conditional independence statements were obtained as follows. For each $j \in [p]$ and each $Z \subset [p] \setminus \{j\}$, we solved the problem in (11) using LassoCV from the `scikit-learn` package in Python to obtain a sparse $\mathbf{w}_j^{(Z)}$. The set Z corresponds to which *contemporaneous* coefficients are active in that model; that is, $x_i[0]$ is not included as a regressor if $i \notin Z$. We define

$$d_{ji}^{(Z)} = \sum_{\tau > 0} \left(w_{ji}^{(Z)}[\tau] \right)^2. \quad (14)$$

We concluded that (9) is true if either $w_{ji}^{(C \cup \{i\})}[0] \neq 0$ or $w_{ij}^{(C \cup \{j\})}[0] \neq 0$, and that (10) is true if $d_{ji}^{(S)} \neq 0$. With these criteria, we used EDI to construct the two-factor I/O structure of the system. The original and reconstructed I/O structures of these systems are shown in Figure 3. As that figure demonstrates, the reconstruction is correct.

B. Comparison with Prior Methods

For linear, Gaussian systems, the Extended Granger Filter [12] recovers the system as well, and does so in a very similar manner to EDI. This is analogous to the equivalence of Granger Causality to Directed Information under the same conditions [33].

EDI is analogous to Directed Information graphs [10] for strictly causal systems, by the following lemma:

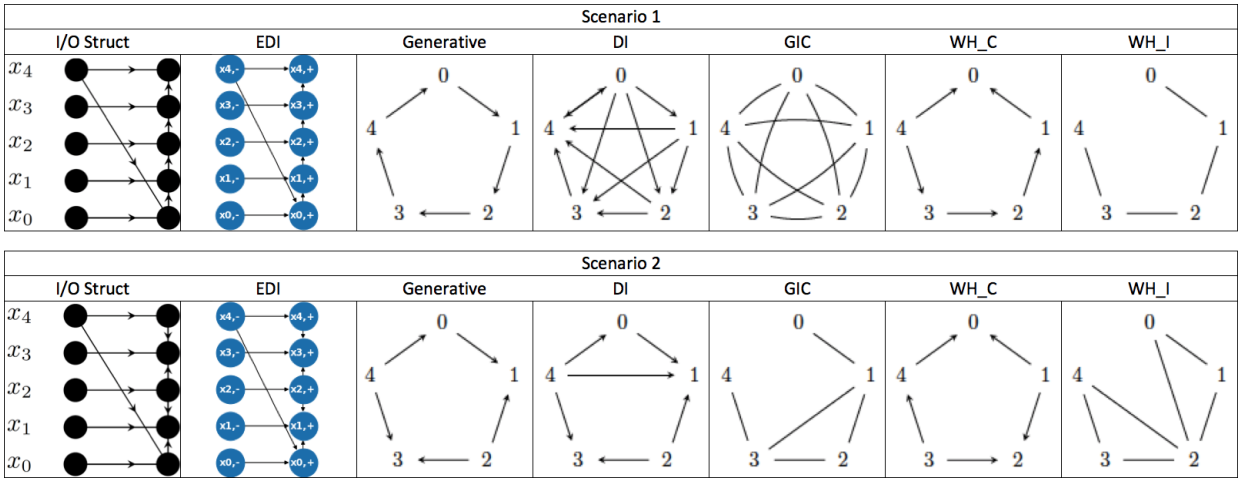


Fig. 3. I/O Struct: The true I/O structures of the example systems in (12) and (13). EDI: Our results from simulated data. Generative: A more common notion of structure that does not separate instantaneous causality. DI, GIC, WH_C, WH_I are all existing methods that attempt to find the generative graph directly, but find spurious links and reverse orientations of arrows where ours does not.

Corollary 1.1. *If all of the system's dynamics are strictly causal, then Extended Directed Information method is equivalent to the Directed Information Graph [10] of the system.*

Proof. If all of the system's dynamics are strictly causal, then $\mathcal{G}^I$ is empty, rendering steps 1, 3, and 4 irrelevant. Moreover, for each i, j, $C_{ij} = \varnothing$. Hence, the conditional independence condition for adjacency of $i \to j$ simplifies to the single test

$$x_j[t] \perp\!\!\!\perp x_i[0:t-1] \mid x_{\bar{i}}[0:t-1]$$

for every t, or, equivalently:

$$I\left(x_j[t]; x_i[0:t-1] \mid x_{\bar{i}}[0:t-1]\right) = 0$$

where I is the mutual information [34]. By time invariance, this is equivalent to:

$$\sum_{t=0}^{T-1} I\left(x_j[t]; x_i[0:t-1] \mid x_{\bar{i}}[0:t-1]\right) = 0,$$

which is the definition of directed information [10]:

$$I(x_i \to x_j \mid x_{\bar{i}}) = 0.$$

□

When the system is not strictly causal, we will show that Directed Information fails. We also compare EDI to other methods, such as Wiener-Hopf filter sparsity [23] or Granger's test for instantaneous causality [21]. All of these algorithms can be implemented using statistics already collected for use by EDI; for instance, the directed information graph [10] has $i \to j$ if and only if $d_{ji}^{(\varnothing)} \neq 0$. When using LASSO to obtain this parameter, one obtains the method proposed in [13].

However, we can also provide a best-case scenario for these algorithms by deriving *exactly* the conditional independence statements they are implicitly estimating. This shows that their failure is not a statistical anomaly or matter of numerical precision but rather a fundamental limitation in the information they use.

The competing graphs we constructed are:

- The Directed Information (DI) graph [10], with $i \to j$ if $(i, -)$ and $(j, +)$ are not d-separated given $\{(k, -) : kin[p], k \neq i\}$ in $\mathcal{S}$;
- The Granger Instantaneous Causality (GIC) graph [21], with $i - j$ if $(i, +)$ and $(j, +)$ are not d-separated given $\{(k, -) : k \in [p]\}$ in $\mathcal{S}$;
- The Wiener-Hopf (WH) graph [23], which is the union of the two component graphs:
  - WH_C, with $i \to j$ if $(i, -)$ and $(j, +)$ are not d-separated given $V(S) \setminus \{(i, -), (j, +)\}$, and
  - WH_I, with $i - j$ if $(i, +)$ and $(j, +)$ are not d-separated given $V(S) \setminus \{(i, +), (j, +)\}$.

We constructed these graphs for the systems in (12) and (13). These graphs can be compared to the generative graph in Figure 3.

Granger [21] states that "instantaneous causality $x_i \to x_j$ is occurring" if the condition represented by the edge $i - j$ in the GIC graph is met. Actually, this edge only indicates a collider-free path between i and j in $\mathcal{G}^{\mathcal{I}}$, and does not otherwise relate to the structure of the system; Figure 3 illustrates this.

Directed Information [10] purports to recover $\mathcal{G}^{\mathcal{C}}$ or the generative graph (which are the same under the assumptions of DI) but does neither when $\mathcal{G}^{\mathcal{I}}$ is not empty. It was known that Wiener-Hopf sparsity [23] gives an undirected graph and spuriously joins spouses. We attempted to remedy this by considering the instantaneous and strictly causal components separately, but we see neither WH_I nor WH_C nor their union give the generative graph.

C. Discussion

The clear drawback of EDI over the aforementioned methods is the potential need to learn 2^{p-1} predictive models for each process index j. However, the failure of these comparison methods demonstrates that structure identification cannot be done by one predictive model alone. On the other hand, EDI recovers these structures correctly Moreover, a "smart" implementation of EDI would compute these models on an

as-needed basis and, for sparse systems, would not need most of them.

At the same time, EDI provides a significant computational improvement over Inductive Causation [8], which (though it, too, would identify the structure exactly) requires learning a staggering 2^{pT-1} models for each of its pT nodes. That said, there are certain "lucky cases" where Inductive Causation obtains orientations of arrows that EDI would miss. For instance, the system

$$\begin{aligned} x[t] &= e_1[t] + ax[t-1]; \\ y[t] &= e_2[t] + b_1 x[t] + b_2 x[t-2] \end{aligned} \tag{15}$$

(with, as usual, $e_1[t], e_2[t]$ i.i.d.,) has a fully identifiable and orientable structure (Figure 4) by Inductive Causation; the arrow $(x,t) \to (y,t)$ is orientable because it forms an open collider with $(x,t-2) \to (y,t)$. Unfortunately, (x,t) and $(x,t-1)$ *are* adjacent, so since EDI does not distinguish between nodes $(x,t-2)$ and $(x,t-1)$, it fails to orient this collider in step 3 of EDI. This potential loss of precision in fully identifying $\mathcal{G}^I$ is what is given up by the major improvement in computational tractability offered by EDI over Inductive Causation.

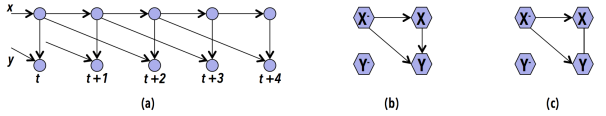


Fig. 4. (a): The causal structure of the system in (15); (b): the I/O structure of that system; (c): the I/O structure as recovered by EDI

VII. Dynamic Networks with Hidden Nodes

In previous sections, we discussed how to model causally coupled collections of stochastic processes and how to infer the I/O structure of these collections from statistics involving passively obtained data. We also discussed the sensitivity of this method to the effects of unmeasured influences within a network. In this section, we will present an augmented model for Causal Graph Processes with unmeasured influences, which we call "Latent CDGs". We will discuss the equivalence of latent CDGs and the significance of being able to represent these equivalence classes in a clear and concise manner. We then show that every latent CDG that is latent due to uniform sampling is equivalent to some latent CDG that is latent due to a completely missing node in the I/O structure through a simple change of basis, which significantly narrows the scope of future work into latent structure learning.

A. Latent Causal Dynamic Graphs

A *latent* system is just a normal system together with a set of outputs whose values are known to the observer.

Definition 18. *A* **latent** *CDG is a pair $(\mathcal{M}, \mathcal{O})$ where $\mathcal{M}$ is a CDG with p subprocesses and $\mathcal{O} \subset [p] \times [T]$ is a set of* observed measurements.

We call

$$\{k \in [p] : (\{k\} \times [T]) \cap \mathcal{O} \neq \varnothing\} \tag{16}$$

the *observed variables* of a latent system, and

$$\{k \in [p] : (\{k\} \times [T]) \subseteq \mathcal{O}\} \tag{17}$$

its *fully observed variables*.

The word *latent* is used in sources such as [8] to denote nodes that are in the underlying causal model but not in the set of observable nodes.

Failing to measure all of the outputs of a complex system may mean that we have entire variables being unmeasured (which we call "spatial" latency), but it may also mean that we do not have a measurement of every value a particular variable takes over time (which we call "temporal" latency).

Definition 19. *A latent CDG is called*

- **spatially latent** *if there exists a $j \in [p]$ such that $\{j\} \times [T] \cap \mathcal{O} = \varnothing$;*
- **temporally latent** *if there exists a j, t_1, t_2 such that $(j,t_1) \in \mathcal{O}$ but $(j,t_2) \notin \mathcal{O}$; and*
- **strictly spatially latent** *if it is not temporally latent.*

A latent CDG is strictly spatially latent if and only if all of its observed variables are fully observed variables.

Certain temporally latent systems where each variable is measured at a regular interval are called "periodically latent".

Definition 20. *A latent CDG is called* **periodic** *if for every observed variable j, there exists a $\tau_j \in [T]$ and a $k_j \in [\tau_j]$ such that for all $(j,t) \in [p] \times [T]$:*

$$(j,t) \in \mathcal{O} \qquad \iff \qquad t \bmod \tau_j = k_j.$$

This could apply, for instance, to systems with digital sensors that only generate a measurement at a particular interval. Note that a periodically latent CDG can have variables that are both completely unobserved and completely observed (the latter by selecting $\tau_j = 1$,) and can support a different sampling period for each other variable. Hence, this model is very general.

B. Equivalence of Periodic and Spatial Latency

The strictly spatially latent CDG is a versatile model in its own right. This is most obvious when a system has more variables than sensors. It also applies to cases like the one discussed in [27] where an entire time series is corrupted; here, the hidden true variable $x_j[t]$ is measured by a noisy sensor which sends $\widehat{x}_{j'}[t]$ to the supervisor. Less obvious, but quite consequential to the significance of the model, is that even periodically latent systems can be recast as strictly spatially latent systems. Therefore, any algorithm that can identify structures consistent with spatially latent data can identify structures consistent with periodically latent data as well.

Claim 2. *Every periodically latent CDG can be recast as a strictly spatially latent CDG by a change of coordinates.*

Proof. Let Q be the least common multiple of $\{\tau_j : j \in [p]\}$ and let $\mathbf{z}$ be the output of the Qp-process CDG satisfying:

$$\begin{aligned} x_j[t] &= z_{Qj+(t \bmod Q)}[t \operatorname{div} Q]; \\ \mathcal{O} &= \{(Qj + \tau_j s + k_j, t) : s \in [Q/\tau_j], t \in [T], j \in [p]\}. \end{aligned} \tag{18}$$

Then for every k, $z_k[t] = x_j[Qt+\tau]$ where $\tau = k \bmod Q$ and $j = k \operatorname{div} Q$. $z_k[t]$ is observable if and only if τ is a multiple of τ_j, and this is either true or false independent of t; □

This justifies restricting our analysis to the strictly spatially latent case—because this actually covers *all* latent cases except those in which the sampling may be nonuniform. This construction has Qp nodes, and hence grows as the sampling period of any one sensor increases. Specific situations may admit more efficient constructions.

Of course, cases of nonuniform sampling may arise in practice. For instance, nonuniform sampling occurs when a system has energy harvesting sensors [35], or sensors communicating over an unreliable channel [36]. The above result does not cover such cases.

That said, even some nonuniform sampling cases can be recast as a strictly spatially latent system. For instance, in the latter case, the supervisor is actually aware of missing measurements, so we can say that the augmented system has dynamic variables $\{x_j[t], y_j[t], c_j[t]\}$ and observable set $\mathcal{O} = \{y_j[t]\}$, where $x_j[t]$ for $j \in [p]$ are the physical process variables and $y_j[t]$ is either $x_j[t]$ or 0 depending on whether the hidden channel state variable $c_j[t]$ is zero or one. Here, $\{y_j\}$ represent the sensor measurements as received by the central controller, and are entirely observed; whereas the $\{x_j, c_j\}$ are entirely unobserved.

However, even when a system with nonuniform sampling can be recast as a strictly spatially latent system, the appropriate transformation likely depends on the particular scenario. Moreover, such recasting may allow one to make assumptions about the dynamics of the system that do not necessarily apply to a general strictly spatially latent system. Therefore, we set this scenario aside as a topic for future research and return to spatially and periodically latent systems.

C. Example

If we return to the example in (5) and (6), if we let $x_1[t] = x[2t+1]$ and $x_2[t] = x[2t]$ and define z_1, z_2, y_1, y_2 similarly, we find that the equivalent system is:

$$\begin{aligned} x_1[t] = e_1[2t+1]; \qquad x_2[t] = e_1[2t] \\ y_1[t] = a_1 x_2[t] + e_2[2t+1] \\ y_2[t] = a_1 x_1[t-1] + e_2[2t] \\ z_1[t] = b_1 y_2[t] + e_3[2t+1] \\ z_2[t] = b_1 y_1[t-1] + e_3[2t] \end{aligned} \tag{19}$$

where $\mathcal{O} = \{x_2, y_2, z_2\}$. Notice that this equivalent system is not strictly causal, even though the original system was.

D. Structure and Graphical Representations of Latent CDGs

For an arbitrary $\mathcal{O} \subset \mathcal{V}$, the only clear way to represent the structure of a latent CDG is by its latent causal structure (Definition 6.) This presents the same deficiency for dynamical systems as does the causal structure: namely, that, because new data are not statistically independent of previous data, a dynamical system's latent structure is large and increases in size linearly as more data are collected. Alternatively, one could start with the I/O structure and label certain nodes as "fully measured," "partially unmeasured," or "fully unmeasured;" the exact labels one uses depending on the level of precision desired. This graphical representation has the benefit of being bounded, having only as many nodes as there are process variables in the system. However, how some of these vague labels correspond to what guarantees can be made about the recoverability of the system is unclear.

If a system has an equivalent strictly spatially latent form, on the other hand, the following graphical representation has as much precision has the I/O structure and requires only two node labels:

Definition 21. *A* **Latent I/O Structure** *is a triple* $\left(\mathcal{G}^{\mathcal{C}}, \mathcal{G}^{\mathcal{I}}, \mathcal{O}_1\right)$ *where* $(\mathcal{G}^{\mathcal{C}}, \mathcal{G}^{\mathcal{I}})$ *is an I/O structure and* $\mathcal{O}_1$ *represents the fully observed processes.*

The latent I/O structure is the dynamic analogue to the latent *causal* structure in Definition 6 and allows us to model systems with hidden dynamic variables. Before we can develop an algorithm similar to EDI to learn latent I/O structures from data, however, it will be necessary to characterize the independence structure of a latent I/O structure in terms of its observable variables, as the latent projection (Definition 7) does for latent causal structures. This will be a topic for future work.

VIII. Conclusion

We presented the Causal Dynamic Graph model to model complex systems as large networks of interdependent processes when every dependency is caused by an input/output relationship between a pair of processes.

We briefly visited the related ideas of contemporaneous causality and causal ill-posedness. Instantaneous causality is not necessarily spurious or due to hidden data, as Granger posited; however a *cycle* of instantaneous causality must be. We argued that it is in cases like this that we must conclude that hidden confounders are the source of such illusory causation. However, contemporaneous causality is by no means a reliable indicator of the presence of hidden information, and in general the assumption that all variables within a system are observed can never be justified by the data.

For cases in which we do assume to measure every variable, we provided the Extended Directed Information algorithm that can efficiently recover the input/output structure of a system, even one that may have true instantaneous causality. This algorithm produces a small class of structures that are all consistent with the data. While using Inductive Causation to recover the full causal structure of the system would result in a smaller class of structures, and hence a more precise result, Extended Directed Information achieves a smaller computational complexity owing to its disregard for the order of a process's dynamics. It recovers the strictly causal part of a system's I/O structure exactly, identifies all pairs of processes whose dependence has a contemporaneous component, and in some situations even determines the direction of such contemporaneous components. We showed with simulated data that EDI works in practice as well, and demonstrated with examples how other existing methods produce incorrect results.

We introduced the Latent CDG model for systems that are causal in nature but that we measure imperfectly. We showed that periodically latent causal graph processes could be recast as strictly spatially latent causal graph processes, therefore admitting a graphical representation based on the input/output structure of the system that the data were derived from. Therefore, any structure learning algorithm that works for strictly spatially latent causal graph processes will work for periodically latent ones as well, even though the latter is a superclass of the former in general. We leave the development of such an algorithm for future work; and while it will not completely solve the structural inference problem in the case of nonuniform sampling, it will be a powerful tool in the modeler's toolbox all the same.

REFERENCES

[1] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 17–44, 2011.

[2] J. Zhou, Z. Liu, and B. Li, "Influence of network structure on rumor propagation," *Physics Letters A*, vol. 368, no. 6, pp. 458–463, 2007.

[3] H. Amini and A. Minca, "Inhomogeneous financial networks and contagious links," *Operations Research*, vol. 64, no. 5, pp. 1109–1120, 2016.

[4] X. Liu, Y. Mo, S. Pequito, B. Sinopoli, S. Kar, and A. P. Aguiar, "Minimum robust sensor placement for large scale linear time-invariant systems: a structured systems approach," *IFAC Proceedings Volumes*, vol. 46, no. 27, pp. 417–424, 2013.

[5] S. Weerakkody, X. Liu, and B. Sinopoli, "Robust structural analysis and design of distributed control systems to prevent zero dynamics attacks," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pp. 1356–1361, IEEE, 2017.

[6] S. Weerakkody, X. Liu, S. H. Son, and B. Sinopoli, "A graph-theoretic characterization of perfect attackability for secure design of distributed control systems.," *IEEE Trans. Control of Network Systems*, vol. 4, no. 1, pp. 60–70, 2017.

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

[8] J. Pearl, *Causality*. Cambridge university press, 2009.

[9] J. Pearl, "Simpson's paradox: An anatomy," *Department of Statistics, UCLA*, 2011.

[10] C. J. Quinn, N. Kiyavash, and T. P. Coleman, "Directed information graphs," *IEEE Transactions on information theory*, vol. 61, no. 12, pp. 6887–6909, 2015.

[11] D. Materassi and M. V. Salapaka, "Graphoid-based methodologies in modeling, analysis, identification and control of networks of dynamic systems," in *American Control Conference (ACC), 2016*, pp. 4661–4675, IEEE, 2016.

[12] M. Dimovska and D. Materassi, "Granger-causality meets causal inference in graphical models: Learning networks via non-invasive observations," in *Decision and Control (CDC), 2017 IEEE 56th Annual Conference on*, pp. 5268–5273, IEEE, 2017.

[13] A. Chiuso and G. Pillonetto, "A bayesian approach to sparse dynamic network identification," *Automatica*, vol. 48, no. 8, pp. 1553–1565, 2012.

[14] T. Verma and J. Pearl, "Causal Networks: Semantics and Expressiveness," in *Proc. 4th Workshop Uncertainty in Artificial Intelligence*, pp. 352–359, 1988.

[15] J. Pearl and T. S. Verma, "A theory of inferred causation," *Studies in Logic and the Foundations of Mathematics*, vol. 134, pp. 789–811, 1995.

[16] N. L. Zhang and D. Poole, "A simple approach to bayesian network computations," in *Proceedings of the Biennial Conference-Canadian Society for Computational Studies of Intelligence*, pp. 171–178, CANADIAN INFORMATION PROCESSING SOCIETY, 1994.

[17] S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 50, no. 2, pp. 157–194, 1988.

[18] A. Braunstein, M. Mézard, and R. Zecchina, "Survey propagation: An algorithm for satisfiability," *Random Structures & Algorithms*, vol. 27, no. 2, pp. 201–226, 2005.

[19] T. Verma, "Graphical aspects of causal models," Technical Report R-191, UCLA, Computer Science Department, 1993.

[20] T. S. Verma and J. Pearl, "Equivalence and synthesis of causal models," Tech. Rep. R-150, UCLA, Computer Science Department, 1990.

[21] C. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, pp. 424–438, 1969.

[22] D. Materassi and G. Innocenti, "Topological identification in networks of dynamical systems," *IEEE Trans. Autom. Control*, vol. 55, no. 8, pp. 1860–1871, 2010.

[23] D. Materassi, M. V. Salapaka, and L. Giarrè, "Relations between structure and estimators in networks of dynamical systems," in *50th Annual Conf. Decision Control and European Control Conf.*, pp. 162–167, IEEE, 2011.

[24] D. Materassi and M. V. Salapaka, "Reconstruction of directed acyclic networks of dynamical systems," in *American Control Conference (ACC), 2013*, pp. 4687–4692, IEEE, 2013.

[25] M. Eichler and V. Didelez, "Causal reasoning in graphical time series models," *arXiv preprint arXiv:1206.5246*, 2012.

[26] J. Lamperti, *Stochastic processes: a survey of the mathematical theory*, vol. 23. Springer Science & Business Media, 2012.

[27] V. R. Subramanian, A. Lamperski, and M. V. Salapaka, "Inferring directed graphs for networks from corrupt data-streams," in *2018 IEEE Conference on Decision and Control (CDC)*, pp. 4493–4498, IEEE, 2018.

[28] J. Runge, "Causal network reconstruction from time series: From theoretical assumptions to practical estimation," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 7, p. 075310, 2018.

[29] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.

[30] J. Pearl and A. Paz, "Graphoids: a graph-based logic for reasoning about relevance relations. ucla computer science department technical report 850038," *Advances in Artificial Intelligence-II, North-Holland Publishing Co*, 1987.

[31] K. Baba, R. Shibata, and M. Sibuya, "Partial correlation and conditional correlation as measures of conditional independence," *Australian & New Zealand Journal of Statistics*, vol. 46, no. 4, pp. 657–664, 2004.

[32] F. Hayashi, *Econometrics*. Princeton University Press, 2000.

[33] L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for gaussian variables," *Physical review letters*, vol. 103, no. 23, p. 238701, 2009.

[34] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[35] O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, and A. Yener, "Transmission with energy harvesting nodes in fading wireless channels: Optimal policies," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 8, pp. 1732–1743, 2011.

[36] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. I. Jordan, and S. S. Sastry, "Kalman filtering with intermittent observations," *IEEE transactions on Automatic Control*, vol. 49, no. 9, pp. 1453–1464, 2004.