2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

A Minimal Approach to Causal Inference on Topologies with Bounded Indegree

Christopher Quinn and Negar Kiyavash and Todd Coleman

Abstract— **The structure of the causal interdependencies between processes in a causal, stochastic dynamical system can be succinctly characterized by a generative model. Inferring the structure of the generative model, however, requires calculating divergences using the full joint statistics. For the case when an upperbound on the indegree of each process is known, we describe a computationally efficient method using directed information which does not require the full statistics and recovers the parents of each process independently from finding the parents of other processes.**

I. INTRODUCTION

There are many research problems in economics, biology, physics, and other disciplines involving stochastic dynamical systems with processes that interact causally. For some problems, it is important to characterize the *structure* of these interactions. Graphical models, where nodes represent processes and edges represent relationships between processes, can be simple, accessible, and complete depictions of the structure. In the related context of interdependent random variables, graphical models such as Bayesian networks, Markov random fields, and factor graphs have been widely used to succinctly represent statistical relationships between variables [1].

For dynamical systems of causally interdependent random processes, there is a similar graphical representation known as generative model graphs [2]. For these graphs, each process is represented as a node, and there are directed edges between processes such that a process is causally independent of all processes it does not have an incoming edge from, causally conditioned on those it does have an incoming edge from [2].

While generative model graphs succinctly represent the structure of the causal interdependencies in the system, computing them requires knowledge of the full joint statistics and a large number of divergence calculations. In [2], an alternative graphical representation, the directed information graph, was introduced. These graphs are based on directed information, an information theoretic quantity formally introduced by Massey [3].

Directed information has been shown to measure statistical causation between processes [4]–[6]. Directed information graphs were shown to be equivalent to generative model graphs [2]. Inferring directed information graphs also require knowledge of the full joint statistics but requires fewer divergence calculations.

For an arbitrary causal dependence structure, the requirement of full knowledge of the joint statistics is reasonable. However, often some information about the structures of the system is known apriori. For example, some gene networks, ecosystems, and the spread of some computer viruses have random graph structures, while some metabolic networks, neuronal networks, and social networks have small-world network structures [7]. Optimal transportation systems and the blood vessel system are known to have tree structures [8], [9].

When the structure is known to have certain properties such as having upper bounds on the indegree, where indegree corresponds to the number of parents for each process, one might expect that more efficient algorithms exist that can recover the causal dependencies. In fact, in this work we will show that such algorithms exists.

Our contribution can be summarized as follows:

- For the case when the underlying topology is known to be a tree, we propose an algorithm that identifies the structure only using pairwise statistics (directed information) and does not use any coupled optimization step.
- We generalize this algorithm to one which recovers the structure if an upper bound to the indegree for each process is known. This algorithm recovers the parents of each process independently from finding the parents of other processes, and it only uses statistics for groups the size of the bound.

In short, these algorithms have the advantage that they do not require the knowledge of the full joint and only utilize local statistics (such as between pairs, triplets, or K-sized groups where K is the upper bound on indegree). Moreover, they identify the topology in a distributed manner, i.e, finding the parents for one process is done independently from finding the parents of other processes. The latter property is specially advantageous for parallelizing the computations.

This material is based upon work supported in part by the U.S. Air Force Office of Scientific Research (AFOSR) under grant numbers MURI FA9550-10-1-0573 and FA9550-11-1-0016, and in part by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370. C. Quinn was supported by the Department of Energy Computational Science Graduate Fellowship, which is provided under grant number DE-FG02-97ER25308.

C. Quinn is with the Department of Electrical and Computer Engineering, University of Illinois, Urbana, Illinois 61801 quinn7@illinois.edu

N. Kiyavash is with the Department of Industrial and Enterprise Systems Engineering, University of Illinois, Urbana, Illinois 61801 kiyavash@illinois.edu

T. Coleman was with the Department of Electrical and Computer Engineering, University of Illinois, Urbana, Illinois 61801. He is now with the Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093 tpcoleman@ucsd.edu

II. RELATED WORK

There are some procedures, related to Chow and Liu's dependence tree approximation procedure for Bayesian networks [10], that find the best tree approximation of a stochastic dynamical system with an arbitrary structure. They compute pairwise statistical relationships, such as directed information and a spectral coherence based distance function, and use a global optimization step (maximum weight spanning tree algorithm) to find the best tree approximation [11], [12]. When the structure is a tree, these algorithms will recover the correct tree. However, they would still use a global optimization step which is not necessary with our proposed method. There are also algorithms related to Chow and Liu which can recover tree structures even with missing variables [13]. They use pairwise distances and joint optimization steps to find the *minimal* latent tree among the class of consistent latent trees.

There is an algorithm for Bayesian networks known as the SGS algorithm which, when there is a known upperbound of the indegree, uses a series of conditional independence tests using local statistics (from pairwise up to the size of the upperbound) to recover the structure [14]. While it only uses local statistics, it requires a large number of independence tests. An alternative approach to identifying sparse structures is Group Lasso, based on the model selection technique Lasso [15].

We will first introduce notations and definitions. We will then review some properties of generative model graphs and directed information graphs. Then we will introduce this simpler procedure for tree structures. Then we introduce the generalized version.

III. DEFINITIONS

A. Notation and Information Theoretic Definitions

- For a sequence $a_1, a_2, \ldots$, denote a_i^j as $(a_i, \ldots, a_j)$ and $a^k \triangleq a_1^k$.
- Denote $[m] \triangleq \{1, \ldots, m\}$ and the power set $2^{[m]}$ on $[m]$ to be the set of all subsets of $[m]$. Let $[m]_i \triangleq [m] \backslash \{i\}$.
- For any Borel space Z, denote its Borel sets by $\mathcal{B}(\mathsf{Z})$ and the space of probability measures on $(\mathsf{Z}, \mathcal{B}(\mathsf{Z}))$ as $\mathcal{P}(\mathsf{Z})$.
- Consider two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{P}(\mathsf{Z})$. $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$ ($\mathbb{P} \ll \mathbb{Q}$) if $\mathbb{Q}(A) = 0$ implies that $\mathbb{P}(A) = 0$ for all $A \in \mathcal{B}(\mathsf{Z})$. If $\mathbb{P} \ll \mathbb{Q}$, denote the Radon-Nikodym derivative as the random variable $\frac{d\mathbb{P}}{d\mathbb{Q}} : \mathsf{Z} \to \mathbb{R}$ that satisfies
$$\mathbb{P}(A) = \int_{z \in A} \frac{d\mathbb{P}}{d\mathbb{Q}}(z) \mathbb{Q}(dz), \quad A \in \mathcal{B}(\mathsf{Z}).$$
- The *Kullback-Leibler divergence* between $\mathbb{P} \in \mathcal{P}(\mathsf{Z})$ and $\mathbb{Q} \in \mathcal{P}(\mathsf{Z})$ is defined as
$$D(\mathbb{P}\|\mathbb{Q}) \triangleq \mathbb{E}_{\mathbb{P}}\left[\log \frac{d\mathbb{P}}{d\mathbb{Q}}\right] = \int_{z \in \mathsf{Z}} \log \frac{d\mathbb{P}}{d\mathbb{Q}}(z) \mathbb{P}(dz) \quad (1)$$
if $\mathbb{P} \ll \mathbb{Q}$ and ∞ otherwise.
- Throughout this paper, we will consider m random processes where the ith (with $i \in \{1, \ldots, m\}$) random process at time j (with $j \in \{1, \ldots, n\}$), takes values in a Borel space X.
- For a sample space Ω, sigma-algebra $\mathcal{F}$, and probability measure $\mathbb{P}$, denote the probability space as $(\Omega, \mathcal{F}, \mathbb{P})$.
- Denote the ith random variable at time j by $X_{i,j} : \Omega \to \mathsf{X}$, the ith random process as $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,n}) : \Omega \to \mathsf{X}^n$, the subset of random processes $\underline{\mathbf{X}}_{\mathcal{I}} = (\mathbf{X}_i : i \in \mathcal{I})^T : \Omega \to \mathsf{X}^{|\mathcal{I}|n}$, and the whole collection of all m random processes as $\underline{\mathbf{X}} \triangleq \underline{\mathbf{X}}_{[m]} : \Omega \to \mathsf{X}^{mn}$. Denote the whole collection of all m random processes from time $j = 1$ to n' as $\underline{\mathbf{X}}_{(1:n')} \triangleq (X_{1,1}, \ldots, X_{1,n'}, \ldots X_{m,1}, \ldots, X_{m,n'}) : \Omega \to \mathsf{X}^{m(n')}$.
- The probability measure $\mathbb{P}$ thus induces a probability distribution on $X_{i,j}$ given by $P_{X_{i,j}}(\cdot) \in \mathcal{P}(\mathsf{X})$, a joint distribution on $\mathbf{X}_i$ given by $P_{\mathbf{X}_i}(\cdot) \in \mathcal{P}(\mathsf{X}^n)$, and a joint distribution on $\underline{\mathbf{X}}_{\mathcal{I}}$ given by $P_{\underline{\mathbf{X}}_{\mathcal{I}}}(\cdot) \in \mathcal{P}\left(\mathsf{X}^{|\mathcal{I}|n}\right)$.
- A distribution $P_{\underline{\mathbf{X}}}$ is called *positive* if there exists a measure ϕ such that $P_{\underline{\mathbf{X}}} \ll \phi$ and $\frac{dP_{\underline{\mathbf{X}}}}{d\phi}(\underline{\mathbf{x}}) > 0$ for all $\underline{\mathbf{x}}$ in the support of $P_{\underline{\mathbf{X}}}$.
- With slight abuse of notation, denote $\mathbf{Y} \equiv \mathbf{X}_i$ for some i and $\mathbf{X} \equiv \mathbf{X}_k$ for some $i \neq k$ and denote the conditional distribution and *causally conditioned* distribution of $\mathbf{Y}$ given $\mathbf{X}$ as
$$\begin{aligned} P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(d\mathbf{y}) &\triangleq P_{\mathbf{Y}|\mathbf{X}}(d\mathbf{y}|\mathbf{x}) \\ &= \prod_{j=1}^{n} P_{Y_j|Y^{j-1},X^n}\left(dy_j|y^{j-1}, x^n\right) \quad (2) \\ P_{\mathbf{Y}\|\mathbf{X}=\mathbf{x}}(d\mathbf{y}) &\triangleq P_{\mathbf{Y}\|\mathbf{X}}(d\mathbf{y}\|\mathbf{x}) \\ &\triangleq \prod_{j=1}^{n} P_{Y_j|Y^{j-1},X^{j-1}}\left(dy_j|y^{j-1}, x^{j-1}\right). \quad (3) \end{aligned}$$
Note the similarity with regular conditioning in (3), except in causal conditioning the future (x_j^n) is not conditioned on [16] [1]. The notation for $P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$ and $P_{\mathbf{Y}\|\mathbf{X}=\mathbf{x}}$ is used to emphasize that $P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}} \in \mathcal{P}(\mathsf{X}^n)$ and $P_{\mathbf{Y}\|\mathbf{X}=\mathbf{x}} \in \mathcal{P}(\mathsf{X}^n)$.
- With slight abuse of notation, denote $\mathbf{Y} \equiv \mathbf{X}_i$ for some i with $\mathsf{Y} = \mathsf{X}^n$ and $\underline{\mathbf{W}} \equiv \underline{\mathbf{X}}_{\mathcal{I}}$ for some $\mathcal{I} \subseteq [m]_i$ with $\underline{\mathcal{W}} = \mathsf{X}^{|\mathcal{I}|n}$. Consider two sets of causally conditioned distributions $\{P_{\mathbf{Y}\|\underline{\mathbf{W}}=\underline{\mathbf{w}}} \in \mathcal{P}(\mathsf{Y}) : \underline{\mathbf{w}} \in \underline{\mathcal{W}}\}$ and $\{Q_{\mathbf{Y}\|\underline{\mathbf{W}}=\underline{\mathbf{w}}} \in \mathcal{P}(\mathsf{Y}) : \underline{\mathbf{w}} \in \underline{\mathcal{W}}\}$ along with a marginal distribution $P_{\underline{\mathbf{W}}} \in \mathcal{P}(\underline{\mathcal{W}})$. Then the conditional KL divergence is given by
$$\begin{aligned} &D\left(P_{\mathbf{Y}\|\underline{\mathbf{W}}} \| Q_{\mathbf{Y}\|\underline{\mathbf{W}}} | P_{\underline{\mathbf{W}}}\right) \\ &\quad = \int_{\underline{\mathcal{W}}} D\left(P_{\mathbf{Y}\|\underline{\mathbf{W}}=\underline{\mathbf{w}}} \| Q_{\mathbf{Y}\|\underline{\mathbf{W}}=\underline{\mathbf{w}}}\right) P_{\underline{\mathbf{W}}}(d\underline{\mathbf{w}}) \quad (4) \end{aligned}$$
The following Lemma will be useful throughout:
Lemma 3.1: $D\left(P_{\mathbf{Y}\|\underline{\mathbf{W}}} \| Q_{\mathbf{Y}\|\underline{\mathbf{W}}} | P_{\underline{\mathbf{W}}}\right) = 0$ if and only if $P_{\mathbf{Y}\|\underline{\mathbf{W}}=\underline{\mathbf{w}}}(d\mathbf{y}) = Q_{\mathbf{Y}\|\underline{\mathbf{W}}=\underline{\mathbf{w}}}(d\mathbf{y})$ with $P_{\underline{\mathbf{W}}}$ probability one.

[1] Note the slight difference in conditioning upon x^{j-1} in this definition as compared to x^j in the original causal conditioning definition. The purpose for doing this will be clear later in the manuscript.

- Let $\mathbf{X} \equiv \mathbf{X}_i$ for some i, $\mathbf{Y} \equiv \mathbf{X}_k$ for some k and $\underline{\mathbf{W}} \equiv \underline{\mathbf{X}}_{\mathcal{I}}$ for some $\mathcal{I} \subseteq [m]_{i,k}$. The mutual information, *directed information* [3], and causally conditioned directed information [16] are given by

$$\mathrm{I}(\mathbf{X};\mathbf{Y}) \triangleq D\big(P_{\mathbf{Y}|\mathbf{X}} \| P_{\mathbf{Y}} | P_{\mathbf{X}}\big) \quad (5)$$

$$\mathrm{I}(\mathbf{X} \to \mathbf{Y}) \triangleq D\big(P_{\mathbf{Y}\|\mathbf{X}} \| P_{\mathbf{Y}} | P_{\mathbf{X}}\big) \quad (6)$$

$$\mathrm{I}(\mathbf{X} \to \mathbf{Y} \| \underline{\mathbf{W}}) \triangleq D\big(P_{\mathbf{Y}\|\mathbf{X},\underline{\mathbf{W}}} \| P_{\mathbf{Y}\|\underline{\mathbf{W}}} | P_{\mathbf{X},\underline{\mathbf{W}}}\big) \quad (7)$$

Conceptually, mutual information and directed information are related. However, while mutual information quantifies statistical correlation (in the colloquial sense of statistical interdependence), directed information quantifies statistical *causation*. For example, $\mathrm{I}(\mathbf{X};\mathbf{Y}) = \mathrm{I}(\mathbf{Y};\mathbf{X})$, but $\mathrm{I}(\mathbf{X} \to \mathbf{Y}) \neq \mathrm{I}(\mathbf{Y} \to \mathbf{X})$ in general. Note that as a consequence of Lemma 3.1 and (7), we have:

Corollary 3.2: $\mathrm{I}(\mathbf{X} \to \mathbf{Y} \| \underline{\mathbf{W}}) = 0$ if and only if $\mathbf{X}$ is causally conditionally independent of $\mathbf{Y}$ given $\underline{\mathbf{W}}$:

$$P_{\mathbf{Y}\|\mathbf{X}=\mathbf{x},\underline{\mathbf{W}}=\underline{\mathbf{w}}}(d\mathbf{y}) = P_{\mathbf{Y}\|\underline{\mathbf{W}}=\underline{\mathbf{w}}}(d\mathbf{y}), \quad P_{\mathbf{X},\underline{\mathbf{W}}} - a.s.$$

Equivalently, we denote that $\mathbf{X} \Rightarrow \underline{\mathbf{W}} \Rightarrow \mathbf{Y}$ form a *causal Markov chain*.

B. Generative Model Graphs and Directed Information Graphs

We will now consider succinct representations of causal interactions between processes in stochastic dynamical systems. For these representations, the causal interactions in the dynamical system need not have a tree structure. The first representation that will be discussed is generative models. The second will be directed information graphs, which were recently introduced in [2]. In that paper, it was shown that these two representations are equivalent and that the procedure to identify directed information graphs is more efficient than the procedure to identify generative models (fewer divergence calculations for general structures). The following definitions are from [2].

Let $\underline{\mathbf{X}}$ be a set of m random processes, each of time length n, in a causal, stochastic dynamical system with a joint distribution $P_{\underline{\mathbf{X}}}$. The distribution can be factorized over time as

$$P_{\underline{\mathbf{X}}}(d\underline{\mathbf{x}}) = \prod_{j=1}^{n} P_{\underline{\mathbf{X}}_{(j)}|\underline{\mathbf{X}}_{(1:j-1)}}(d\underline{\mathbf{x}}_{(j)}|\underline{\mathbf{x}}_{(1:j-1)}). \quad (8)$$

Since this system is causal, given the full past, the future (next step) of each of the processes is independent. Thus, factorizing over the processes

$$P_{\underline{\mathbf{X}}}(d\underline{\mathbf{x}}) = \prod_{j=1}^{n}\prod_{i=1}^{m} P_{X_{i,j}|\underline{\mathbf{X}}_{(1:j-1)}}(dx_{i,j}|\underline{\mathbf{x}}_{(1:j-1)}). \quad (9)$$

Assumption 1: For the remainder of this paper, we only consider joint distributions $P_{\underline{\mathbf{X}}}$ which are positive and satisfy (9).

Equation (9) can be rewritten as

$$P_{\underline{\mathbf{X}}}(d\underline{\mathbf{x}}) \quad = \quad \prod_{i=1}^{m} P_{\mathbf{X}_i \| \underline{\mathbf{X}}_{[m]_i}}(d\mathbf{x}_i \parallel \underline{\mathbf{x}}_{[m]_i}). \quad (10)$$

While (10) fully characterizes the dynamical system it might be the case that not every process $\mathbf{X}_i$ depends causally on *all* the other processes $\underline{\mathbf{X}}_{[m]_i}$. For notational simplity, consider a function $A : [m] \to 2^{[m]}$ which for each process $i \in [m]$, specifies the subset of other processes that $\mathbf{X}_i$ causally depends on. We can fully describe the dynamics with the following factorization

$$P_A(d\underline{\mathbf{x}}) = \prod_{i=1}^{m} P_{\mathbf{X}_i \| \underline{\mathbf{X}}_{A(i)}}(d\mathbf{x}_i \parallel \underline{\mathbf{x}}_{A(i)}). \quad (11)$$

We will call such simpler factorizations "generative models." In particular, we will focus on those where each for each process i, the subset of other processes $A(i)$ causally conditioned on is of minimal cardinality.

Definition 3.3 ([2]): Under Assumption 1, for a joint distribution $P_{\underline{\mathbf{X}}}$, a *minimal generative model* is a function $A : [m] \to 2^{[m]}$ such that for each process $i \in [m]$, $i \notin A(i)$ and $|A(i)|$ is minimal such that

$$D\big(P_{\underline{\mathbf{X}}} \| P_A\big) = 0, \quad (12)$$

where P_A is defined in (11).

Lemma 3.4 ([2]): Under Assumption 1, for any distribution $P_{\underline{\mathbf{X}}}$ there is a unique, minimal generative model.

When we say "generative models," we refer to minimal generative models. Note that Lemma 3.4 means that for each process $\mathbf{Y} \equiv \mathbf{X}_i$, there is a unique set of processes, indexed by $A(i)$, such that (12) holds. Not only is $A(i)$ of minimal cardinality, but there is only one $A(i)$ of that cardinality [2].

The structure of the generative model, which is the structure of causal dependencies in the stochastic, dynamical system, can be represented graphically.

Definition 3.5 ([2]): A *generative model graph* is a directed graph for a generative model where each process is represented by a node, and there is a directed edge from $\mathbf{X}_k$ to $\mathbf{X}_i$ for $i, k \in [m]$ iff $k \in A(i)$.

Generative model graphs are one representation of the causal dynamics of a system. They characterize causally conditioned independences between processes. An alternative, related representation characterizes the causally conditioned directed information between processes.

Definition 3.6 ([2]): A *directed information graph* is a directed graph over a set of random processes $\underline{\mathbf{X}}$ where each node represents a process and there is a directed edge from process $\mathbf{X} \equiv \mathbf{X}_k$ to process $\mathbf{Y} \equiv \mathbf{X}_i$ (for some $i, k \in [m]$) iff

$$\mathrm{I}(\mathbf{X} \to \mathbf{Y} \parallel \underline{\mathbf{X}}_{[m]_{i,k}}) > 0. \quad (13)$$

An edge is drawn from $\mathbf{X}$ to $\mathbf{Y}$ if, even with full knowledge of the past of all other processes, the past of $\mathbf{X}$ still influences the future of $\mathbf{Y}$. From the definition, directed information graphs are unique. We note that directed information graphs are analogous to Markov networks [1], but instead of conditional independence, for directed information graphs the criterion is causally conditional independence. Although they are different characterizations of the causal dependencies, minimal generative model graphs and directed information graphs are equivalent [2].

We will now consider the problem of inferring these structures from known statistics.

IV. DISCOVERING GENERATIVE MODEL GRAPHS WITH TREE TOPOLOGIES

For the case that the generative model graph is known to be a tree, we can recover the structure directly from the definition of generative model graphs or from the definition of directed information graphs. Both use the full statistics to recover each parent. However, since the structure is simple, using the full statistics might not be necessary. This is discussed in the following example.

Example 1: Consider a set of six processes $\underline{\mathbf{X}} = \{\mathbf{X}_1, \ldots, \mathbf{X}_5, \mathbf{Y}\}$ with a joint distribution $P_{\underline{\mathbf{X}}}$. Assume it has a minimal generative model with a tree structure. We want to efficiently determine which process is the parent of $\mathbf{Y}$. This is represented graphically in Figure 1, where there is a question mark next to each edge we will check.

We could identify the parent through the definition of generative models. To test if $\mathbf{X}_3$ is the parent, calculate

$$D\Big(P_{\mathbf{Y}\|\underline{\mathbf{X}}_{[5]}} \| P_{\mathbf{Y}\|\mathbf{X}_3} \big| P_{\underline{\mathbf{X}}_{[5]}}\Big). \tag{14}$$

If this value is 0, then $\mathbf{X}_3$ is the parent. Otherwise, check another. There are six subsets to check in total: $\{\emptyset, \mathbf{X}_1, \ldots, \mathbf{X}_5\}$.

Alternatively, the definition of directed information graphs could be used. To test if $\mathbf{X}_3$ is the parent, calculate

$$\mathrm{I}(\mathbf{X}_3 \to \mathbf{Y} \parallel \underline{\mathbf{X}}_{[5]_3}) = D\Big(P_{\mathbf{Y}\|\underline{\mathbf{X}}_{[5]}} \| P_{\mathbf{Y}\|\underline{\mathbf{X}}_{[5]_3}} \big| P_{\underline{\mathbf{X}}_{[5]}}\Big). \tag{15}$$

If this value is greater than 0, then $\mathbf{X}_3$ is the parent.

Note that to calculate (14) or (15) requires the full joint distribution to determine $P_{\mathbf{Y}\|\underline{\mathbf{X}}_{[5]}}$. This might seem necessary because there is no knowledge of the structure of $\{\mathbf{X}_1, \ldots, \mathbf{X}_5\}$ other than that each process has at most one parent.

However, there is a method, described below, that uses the same number of divergence calculations as the above two methods, but only uses *pairwise* statistics and can discover the parents separately. In particular, to check if $\mathbf{X}_3$ is the parent, it only needs to calculate

$$\mathrm{I}(\mathbf{X}_3 \to \mathbf{Y}) = D\big(P_{\mathbf{Y}\|\mathbf{X}_3} \| P_{\mathbf{Y}} \big| P_{\mathbf{X}_3}\big). \tag{16}$$

This calculation only requires knowledge of $P_{\mathbf{Y},\mathbf{X}_3}$, not $P_{\underline{\mathbf{X}}}$. Also, finding the parent of $\mathbf{Y}$ can be done *independently* of finding the structure of $\{\mathbf{X}_1, \ldots, \mathbf{X}_5\}$. No global optimization step is necessary. The motivation for this algorithm comes from the following data processing inequality for causal Markov chains.

Lemma 4.1: Let $\{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$ be a set of processes with a causal markov chain generative model: $\mathbf{X} \Rightarrow \mathbf{Y} \Rightarrow \mathbf{Z}$. Then

$$\mathrm{I}(\mathbf{X} \to \mathbf{Z}) < \mathrm{I}(\mathbf{Y} \to \mathbf{Z}). \tag{17}$$

Proof: By applying the chain rule for directed information twice (Ch. 3 in [16]),

$$\begin{aligned} \mathrm{I}(\{\mathbf{X}, \mathbf{Y}\} \to \mathbf{Z}) &= \mathrm{I}(\mathbf{X} \to \mathbf{Z}) + \mathrm{I}(\mathbf{Y} \to \mathbf{Z} \parallel \mathbf{X}) && (18)\\ &= \mathrm{I}(\mathbf{Y} \to \mathbf{Z}) + \mathrm{I}(\mathbf{X} \to \mathbf{Z} \parallel \mathbf{Y}) && (19) \end{aligned}$$

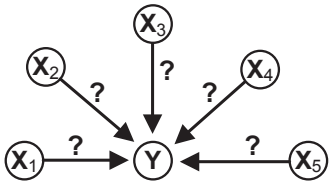

Fig. 1. A graph of candidate parents for process $\mathbf{Y}$ in Example 1. It is known that the minimal generative model has a tree structure, but it is unknown which tree that is. The sub-tree structure of $\{\mathbf{X}_1, \ldots, \mathbf{X}_5\}$ is unknown and not depicted.

By definition of edges in a generative model,

$$D\big(P_{\mathbf{Z}\|\{\mathbf{X},\mathbf{Y}\}} \| P_{\mathbf{Z}\|\mathbf{Y}} \big| P_{\{\mathbf{X},\mathbf{Y}\}}\big) = 0, \tag{20}$$

which by definition of causally conditional directed information implies that $\mathrm{I}(\mathbf{X} \to \mathbf{Z} \parallel \mathbf{Y}) = 0$. Thus,

$$\mathrm{I}(\mathbf{Y} \to \mathbf{Z}) = \mathrm{I}(\mathbf{X} \to \mathbf{Z}) + \mathrm{I}(\mathbf{Y} \to \mathbf{Z} \parallel \mathbf{X}) \tag{21}$$

Consider the case that $\mathrm{I}(\mathbf{Y} \to \mathbf{Z} \parallel \mathbf{X})$ was also 0. This would imply (by definition of causally conditioned directed information) that

$$D\big(P_{\mathbf{Z}\|\{\mathbf{X},\mathbf{Y}\}} \| P_{\mathbf{Z}\|\mathbf{X}} \big| P_{\{\mathbf{X},\mathbf{Y}\}}\big) = 0. \tag{22}$$

This means that either there is an edge from $\mathbf{X}$ to $\mathbf{Z}$ in the generative model (thus multiple minimal generative models), or $\mathbf{Z}$ is causally conditionally independent of $\{\mathbf{X}, \mathbf{Y}\}$ so there is no edge into $\mathbf{Z}$. Both of these possibilities contradict the uniqueness of minimal generative models Lemma 3.4. Thus,

$$\mathrm{I}(\mathbf{Y} \to \mathbf{Z} \parallel \mathbf{X}) > 0 \tag{23}$$

which implies from (21) that

$$\mathrm{I}(\mathbf{X} \to \mathbf{Z}) < \mathrm{I}(\mathbf{Y} \to \mathbf{Z}).$$

■

Since processes in directed trees have at most one parent, such as $\mathbf{Y}$ is $\mathbf{Z}$'s parent in the above lemma, the parent of a process can be found by calculating all pairwise directed informations to that process and picking the one with maximal value. This is formally described in the following algorithm.

Algorithm 1. TreeRecovery

Input: set of all second order distributions $\{P_{\mathbf{X}_i,\mathbf{X}_l}\}_{i,l\in[m]}$.

1. Initialize $parent(1 \ldots m) \leftarrow \emptyset$
2. **For** $i, l \in [m]$ with $i \neq l$
3. Compute $\mathrm{I}(\mathbf{X}_l \to \mathbf{X}_i)$
4. **For** $i \in [m]$
5. $k = \arg\max_{l \in [m]_i} \mathrm{I}(\mathbf{X}_l \to \mathbf{X}_i)$
6. **If** $\mathrm{I}(\mathbf{X}_k \to \mathbf{X}_i) > 0$
7. $parent(i) \leftarrow k$

Theorem 4.2: If the minimal generative model for $P_{\underline{\mathbf{X}}}$ has a directed tree structure, Algorithm 1 will recover it.

	Distributed Search	Pairwise Statistics
CL		X
GMG and DIG	X	
Alg. 1	X	X

Table 1. A comparison of properties of Chow and Liu based algorithms [11], [12], algorithms implicit in the definitions of generative model graphs (GMG) and directed information graphs (DIG) [2], and Algorithm 1. *Distributed search* means that the algorithm finds the parent of a process independently of finding the parents of other processes.

Proof: Suppose that a process $\mathbf{Y} \equiv \mathbf{X}_i$ does not have a parent in the generative model. By definition of a generative model,

$$D\Big(P_{\mathbf{Y}\|\underline{\mathbf{X}}_{[m]_i}}\|P_{\mathbf{Y}}\big|P_{\underline{\mathbf{X}}_{[m]_i}}\Big) = 0. \tag{24}$$

By definition of directed information, $\mathrm{I}(\underline{\mathbf{X}}_{[m]_i} \to \mathbf{Y}) = 0$. For all $l \in [m]_i$, by the chain rule for directed information (Ch. 3 in [16]),

$$\mathrm{I}(\underline{\mathbf{X}}_{[m]_i} \to \mathbf{Y}) = \mathrm{I}(\mathbf{X}_l \to \mathbf{Y}) + \mathrm{I}(\underline{\mathbf{X}}_{[m]_{il}} \to \mathbf{Y} \parallel \mathbf{X}_l) = 0. \tag{25}$$

By nonnegativity of directed information, $\mathrm{I}(\mathbf{X}_l \to \mathbf{Y}) = 0$. Thus, the algorithm will correctly return $parent(i) = \emptyset$.

Otherwise, $\mathbf{Y}$ has exactly one parent, denoted as $\mathbf{X}_k$, with $\mathrm{I}(\mathbf{X}_k \to \mathbf{Y}) > 0$. Since this minimal generative model is unique by Lemma 3.4, for any $l \in [m]_{i,k}$, either $\mathbf{Y}$ is causally independent of $\mathbf{X}_l$, or $(\mathbf{X}_l, \mathbf{X}_k, \mathbf{Y})$ form a causal Markov chain. Thus, by the data processing inequality Lemma 4.1,

$$\mathrm{I}(\mathbf{X}_k \to \mathbf{Y}) > \mathrm{I}(\mathbf{X}_l \to \mathbf{Y}) \quad \forall\, l \in [m]_{i,k}. \tag{26}$$

■

There are two significant properties of Algorithm 1.

1) only pairwise statistics are needed, and
2) the parent of each process can be found *independently* of the parents of other processes.

A comparison with other methods is in Table 1. Methods like Chow and Liu [10] which approximate arbitrary structures with causal dependence trees [11], [12], only use pairwise statistics, but require a global optimization step. Both generative model graphs and directed information graphs find the parent for each process independently from finding the parents of other processes, but need the full joint statistics.

Thus, this method provides a simple procedure to recover the structure of a causal, stochastic dynamical system when the structure is a tree. A natural next question is whether there is an analogous procedure for more complicated structures.

V. DISCOVERING GENERATIVE MODEL GRAPHS WITH GENERAL TOPOLOGIES

As discussed in [2], if a process has more than one parent, pairwise directed information values can be misleading. This can be illustrated through an example.

Example 2: Let $\mathbf{W}$, $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ be four processes, with $\mathbf{W}$ and $\mathbf{X}$ independent Bernoulli($\frac{1}{2}$) processes and

$$\begin{aligned} Y_i &= W_{i-1} \oplus X_{i-1} + \epsilon_i \\ Z_i &= W_{i-2} \oplus X_{i-2} + \epsilon_i' \end{aligned}$$

for some i.i.d. Gaussian noises $\{\epsilon_i, \epsilon_i'\}_{i=1}^n$. Because of the properties of the XOR function $\oplus$, $\mathrm{I}(\mathbf{X} \to \mathbf{Z}) = 0$ while $\mathrm{I}(\mathbf{Y} \to \mathbf{Z}) > 0$. On the other hand, $\mathrm{I}(\mathbf{X} \to \mathbf{Z} \parallel \mathbf{W}, \mathbf{Y}) > 0$ and $\mathrm{I}(\mathbf{Y} \to \mathbf{Z} \parallel \mathbf{W}, \mathbf{X}) = 0$. Thus, Algorithm 1, which only uses pairwise tests, would fail at recovering this topology.

A natural question is whether there exists a generalization of the procedure in Algorithm 1 for generative model graphs with more general topologies. We will now develop a procedure similar to Algorithm 1 which can efficiently recover the true structure when an upper bound on the number of parents for each process is known.

Where in Algorithm 1, directed informations of the form $\mathrm{I}(\mathbf{X}_l \to \mathbf{Y})$ were calculated for individual processes $\mathbf{X}_l$, here directed informations of the form $\mathrm{I}(\underline{\mathbf{X}}_{\mathcal{I}_l} \to \mathbf{Y})$ will be calculated for subsets of processes $\mathcal{I}_l \subseteq [m]_i$. The subsets $\underline{\mathbf{X}}_{\mathcal{I}_l}$ that maximize the directed information to $\mathbf{Y}$ all will contain the true parents of $\mathbf{Y}$, and there will be no other process common to all of them.

The motivation for this algorithm comes from extending the intuition of the data processing inequality for causal Markov chains of Lemma 4.1. The parents of a process $\mathbf{Y}$ convey all of the relevant causal information that the entire set of processes $\underline{\mathbf{X}}_{[m]_i}$ does. Any set of processes $\underline{\mathbf{X}}_{\mathcal{I}_l}$ which includes all of the parents will have the maximal directed information. Any set of processes $\underline{\mathbf{X}}_{\mathcal{I}_l}$ which does not include all of the parents can be seen to form a causal Markov chain with the set of parent processes and $\mathbf{Y}$. In this case, $\underline{\mathbf{X}}_{\mathcal{I}_l}$ will have directed information strictly less than the parents. Before making this statement precise (in Lemma 5.2), first consider the following lemma which will be used to prove Lemma 5.2.

Lemma 5.1 ([2]): Let $\underline{\mathbf{X}}$ be a set of processes with a joint distribution $P_{\underline{\mathbf{X}}}$ satisfying Assumption 1. Let $\mathbf{Y} \equiv \mathbf{X}_i$ for some $i \in [m]$. Let $\mathcal{I}_A, \mathcal{I}_B \subseteq [m]_i$ be two subsets of indices that carry the same causal information about $\mathbf{Y}$:

$$D\Big(P_{\mathbf{Y}\|\underline{\mathbf{X}}_{\mathcal{I}_A}}\|P_{\mathbf{Y}\|\underline{\mathbf{X}}_{\mathcal{I}_B}}\big|P_{\underline{\mathbf{X}}_{\mathcal{I}_A\cup\mathcal{I}_B}}\Big) = 0. \tag{27}$$

Then their intersection contains all of the causal influence on $\mathbf{Y}$:

$$D\Big(P_{\mathbf{Y}\|\underline{\mathbf{X}}_{\mathcal{I}_A}}\|P_{\mathbf{Y}\|\underline{\mathbf{X}}_{\mathcal{I}_A\cap\mathcal{I}_B}}\big|P_{\underline{\mathbf{X}}_{\mathcal{I}_A}}\Big) = 0. \tag{28}$$

Lemma 5.2: Let $\mathbf{Y} \equiv \mathbf{X}_i$ be a process whose parent processes in the generative model are denoted by $\underline{\mathbf{Z}}$. Let $\underline{\mathbf{A}}$ be any subset of processes of $\underline{\mathbf{X}}_{[m]_i}$ containing all of the parents, $\underline{\mathbf{Z}} \subseteq \underline{\mathbf{A}}$, and $\underline{\mathbf{B}}$ any subset of processes of $\underline{\mathbf{X}}_{[m]_i}$. Then $\underline{\mathbf{B}} \Rightarrow \underline{\mathbf{A}} \Rightarrow \mathbf{Y}$ forms a causal Markov chain, and thus

$$\mathrm{I}(\underline{\mathbf{B}} \to \mathbf{Y}) \leq \mathrm{I}(\underline{\mathbf{A}} \to \mathbf{Y}). \tag{29}$$

Moreover, there is equality iff $\underline{\mathbf{B}}$ contains all of the parents of $\mathbf{Y}$: $\underline{\mathbf{Z}} \subseteq \underline{\mathbf{B}}$.

Proof: A graphical depiction of the subsets is in Figure 2. A sketch of the proof involves considering $\mathrm{I}(\underline{\mathbf{B}} \to \mathbf{Y} \parallel \underline{\mathbf{A}})$ and nonnegativity of directed information to show the inequality, $\mathrm{I}(\underline{\mathbf{A}} \to \mathbf{Y} \parallel \underline{\mathbf{B}})$ to show equality if $\underline{\mathbf{B}}$ contains all the parents, and Lemma 5.1 for strict inequality otherwise. ■

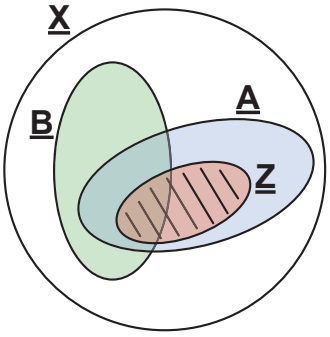

Fig. 2. A graph depicting the overlap of sets in Lemma 5.2. $\mathbf{Y} \equiv \mathbf{X}_i$ is a process with parents $\underline{\mathbf{Z}} \subseteq \underline{\mathbf{A}} \subseteq \underline{\mathbf{X}}_{[m]_i}$. $\underline{\mathbf{B}} \subseteq \underline{\mathbf{X}}_{[m]_i}$ is another, arbitrary subset of processes.

If we want to identify the structure of a generative model, Lemma 5.2 suggests some important properties that can be used to develop an efficient procedure. The first is that finding the parents of a process $\mathbf{Y}$ can be done independently of finding the rest of the structure. The second is that any subset of processes with all of the parents of $\mathbf{Y}$ conveys the maximal directed information to $\mathbf{Y}$ of any subset of processes (excluding $\mathbf{Y}$ itself). Any subset not including all of the parents conveys strictly less information. Consequently, as long as we know an upper bound L to the number of parents that $\mathbf{Y}$ has, we can calculate the directed information from all L-sized subsets of processes to $\mathbf{Y}$. There would be at least one L-sized subset with the parents. If we check all L-sized subsets, then only the parents would be common to the subsets that had maximal directed information to $\mathbf{Y}$. This procedure is formally described in Algorithm 2.

Algorithm 2. StructureRecovery

Input: set of all second order distributions $\{P_{\mathbf{X}_i,\mathbf{X}_l}\}_{i,l\in[m]}$, upper bound on number of parents $maxparents(1\dots m)$.

1. Initialize $parents(1\dots m) \leftarrow \emptyset$
2. **For** $i \in [m]$
3. $\quad L \leftarrow maxparents(i)$
4. $\quad \boldsymbol{\mathcal{I}}_L \leftarrow \{\mathcal{I} : \mathcal{I} \subseteq [m]_i,\ |\mathcal{I}| = L\}$
5. $\quad$ **For** $\mathcal{I}_l \in \boldsymbol{\mathcal{I}}_L$
6. $\quad\quad$ Compute $\mathrm{I}(\underline{\mathbf{X}}_{\mathcal{I}_l} \to \mathbf{X}_i)$
7. $\quad \boldsymbol{\mathcal{I}}_{Lmax} \leftarrow \arg\max_{\mathcal{I}_j \in \boldsymbol{\mathcal{I}}_L} \mathrm{I}(\underline{\mathbf{X}}_{\mathcal{I}_j} \to \mathbf{X}_i)$
8. $\quad parents(i) \leftarrow \bigcap_{\mathcal{I}_l \in \boldsymbol{\mathcal{I}}_{Lmax}} \mathcal{I}_l$

We will now show the correctness of the algorithm. $maxparents(1\dots m)$ denotes the upperbounds on the number of parents for each process in $[m]$.

Theorem 5.3: Algorithm 2 recovers the minimal generative model structure for a given $P_{\underline{\mathbf{X}}}$ if $maxparents(1\dots m) \leq m-2$.

Proof: The proof is analogous to the proof for Theorem 4.2. It requires Lemma 5.2, since now there are sets of processes. ∎

Like Algorithm 1, finding the parents of one process is done independently of finding the parents of other processes. Thus, the procedure can easily be distributed. Also, Algorithm 2 does not need the full statistics. It only uses K^{th} order statistics where K is the upperbound of the indegree. Thus, Algorithm 2 provides an efficient procedure to recover general structures when upperbounds on the indegree are known.

References

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

[2] C. Quinn, N. Kiyavash, and T. Coleman, "Equivalence between minimal generative model graphs and directed information graphs," in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*.

[3] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Information Theory Application (ISITA-90)*, 1990, pp. 303–305.

[4] J. Rissanen and M. Wax, "Measures of mutual and causal dependence between two time series (Corresp.)," *IEEE Transactions on Information Theory*, vol. 33, no. 4, pp. 598–601, 1987.

[5] C. Quinn, T. Coleman, N. Kiyavash, and N. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of Computational Neuroscience*, pp. 1–28, 2010.

[6] A. Rao, A. Hero III, D. States, and J. Engel, "Motif discovery in tissue-specific regulatory sequences using directed information," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, pp. 1–13, 2007.

[7] S. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, 2001.

[8] J. Banavar, F. Colaiori, A. Flammini, A. Maritan, and A. Rinaldo, "Topology of the fittest transportation network," *Physical Review Letters*, vol. 84, no. 20, pp. 4745–4748, 2000.

[9] E. Bullitt, K. Muller, I. Jung, W. Lin, and S. Aylward, "Analyzing attributes of vessel populations," *Medical image analysis*, vol. 9, no. 1, pp. 39–49, 2005.

[10] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[11] C. Quinn, T. Coleman, and N. Kiyavash, "Causal Dependence Tree Approximations of Joint Distributions for Multiple Random Processes," *Information Theory, IEEE Transactions on*, 2011, submitted, Arxiv preprint arXiv:1101.5108.

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

[13] M. Choi, V. Tan, A. Anandkumar, and A. Willsky, "Learning latent tree graphical models," *Journal of Machine Learning Research*, vol. 12, pp. 1771–1812, 2011.

[14] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction, and search*. The MIT Press, 2000.

[15] A. Bolstad, B. Van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *Signal Processing, IEEE Transactions on*, vol. 59, no. 6, pp. 2628 –2641, June 2011.

[16] G. Kramer, "Directed information for channels with feedback," Ph.D. dissertation, University of Manitoba, Canada, 1998.