Directed Information Graphs

Christopher J. Quinn, Student Member, IEEE, Negar Kiyavash, Senior Member, IEEE,
and Todd P. Coleman, Senior Member, IEEE

Abstract—We propose two graphical models to concisely represent causal influences between agents in a network. The first, the minimal generative model graph, reflects a minimal state space description of relationships. The second, the directed information graph, is a statistical approach similar to conventional graphical models and uses directed information to generalize Granger causality. Although they are motivated differently, we show that under minimal assumptions, the graphs are equivalent.

In order to identify the underlying graph, we present several algorithms. In general, joint statistics of the whole network are needed. We present an algorithm that uses the minimal-dimension statistics necessary when upper bounds on the in-degrees are known. In the event that the upper-bounds are not valid, the result is nonetheless an optimal approximation.

The algorithms require calculations of directed information. For the setting when directed information is estimated from data, we characterize the sample-complexity of two directed information estimators. Their performance is similar to standard results for statistical estimation with iid data. When point estimates of directed information are not reliable, we compute confidence intervals. Furthermore, we propose an algorithm that uses confidence intervals to identify the best bounded in-degree graph approximation that is robust to estimation error.

Lastly, we demonstrate the effectiveness of the proposed algorithms through simulations and by identifying which news agencies influence which users in the Twitter network by analyzing only tweet times. The algorithms determine influences with high precision.

Index Terms—Graphical models, network inference, causality, generative models, directed information.

I. INTRODUCTION

Research in many disciplines, including biology, economics, social sciences, computer science, and physics involves studying large networks of interacting agents. A major topic in many of these disciplines is identifying the network topology—which agents interact with which other agents. For instance, neuroscientists seek to determine which neurons communicate with which other neurons. Financial investors want to learn fluctuations in which stocks affect their portfolios. Computer security experts attempt to uncover which computers in a network infected others with malicious software. A common theme in all these cases is that the researcher can observe time-series of agent activity, such as neural spikes, stock prices, and network traffic, to infer causal influences between agents. This work develops a theoretically grounded framework for analyzing stochastic processes to identify the underlying causal influences between agents.

For concreteness, consider the following example. Online social networks have become increasingly popular in recent years. As these communities become more active, advertisement agencies are interested in leveraging word-of-mouth advertising within these networks [1]. An important, but difficult, issue is determining who influences whom in the network. While it is often comparatively easy to learn the “friend” or “follower” graphs of social networks, identifying influence is more elusive.

A natural question is whether influences between agents in a network can be learned by analyzing their activity. For instance, suppose an advertiser wants to target a specific population of users in a micro-blogging network such as Twitter. The advertiser sees that many of the users follow several major news companies and celebrities. See Fig. 1. The advertiser wants to identify which of the companies or celebrities have strong influence on the users in order to decide whom to pay to advertise. In a general sense, for word-of-mouth advertising, a celebrity influences a user if the celebrity’s messaging of a topic or content inspires the user to do likewise, either as a duplicate (such as a retweet in Twitter) or a novel message.

The advertiser can observe time-series of activities, such as message times (see Fig. 2), and using that data wants to calculate a measure of influence. Important aspects include that the target population follows many companies and celebrities, so the algorithms must be efficient. Also, the advertiser needs reliable methods to compute statistics from data. Finally, the resulting graph of influences must be well-defined. For instance, if the advertiser infers that both celebrities A and B influence user Y, the advertiser must be sure both are direct, and not that A influences B who in turn influences user Y.

This work proposes a framework to address these issues. It introduces two approaches to graphically represent which agents influence which others. In both, agents are depicted by nodes and directed edges represent inferred influences. The first graphical model is motivated by generative models. If the system had a known generative model for how the past activity in the network effected the agents’ future activity, it would be straightforward to draw a graph of the dependencies. The

1Here “causal” is used in a colloquial sense related to Granger causality—the past activity of one agent influences the future activity of another.


C. J. Quinn was supported by the Department of Energy Computational Science Graduate Fellowship, which is provided under Grant DE-FG02-97ER25308. This work was supported by AFOSR Grant FA9550-11-1-0016, MURI under AFOSR Grant FA9550-10-1-0573, NSF Grant CCF-1065352, and NSF Grant CCF-0939370.

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

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

T. P. Coleman is with the Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093 tpcoleman@ucsd.edu
first approach operates when a generative model is unknown or does not exist.

The second graphical model uses a generalization of Granger causality. Clive Granger, a Nobel laureate, proposed a methodology for deciding when, in a statistical sense, one stochastic process $X$ causally influences another process $Y$ in a network [2]. His methodology is based on quantifying how much having the past of $X$ as side information helps in sequentially predicting the future of $Y$. Researchers have applied his framework in a number of fields, including biology, economics, and social sciences [3]–[8]. We use an information-theoretic quantity known as directed information [9], which generalizes the concept of Granger causality, to measure influences between processes.

Many networks of interest have large numbers of nodes, such as millions of people in an online social network and billions of neurons in the brain. We will consider several different algorithms to recover the graphical models, trading off computational efficiency for assumptions on the topology. For instance, in microblogging networks such as Twitter, some users follow hundreds or thousands of accounts. Advertising agencies would only be interested in the two or five most influential accounts. Using knowledge of in-degree bounds can lead to algorithms that require joint statistics of dimensions much smaller than in general.

We also discuss estimation techniques for directed information. In many real-world applications of interest, required statistics must be estimated from observed time-series. It is important to assess how much data is needed for reliable estimation. In situations where data is too limited to compute point estimates, it is valuable to have means of computing confidence intervals.

Although the proposed framework is theoretically-grounded, we demonstrate its utility by identifying influences in the Twitter network. Specifically, we record and analyze activity of 16 news corporations and 48 user accounts to infer which news corporations influenced which user accounts in tweeting about recent events in the Middle East. We develop models for user activity, such as how the likelihood of one user's tweeting depends on recent tweeting of other accounts. Without knowledge of the follower graph in the analysis, the algorithms infer influences with high precision.

We next describe the specific contributions of this work and then discuss related work.

A. Our Contribution

Our first contribution is to propose two graphical models for representing networks of stochastic processes. The first, the minimal generative model graph, is based on reduced factorizations of the joint distribution of the network over time. It is motivated by simplifying sets of known, coupled differential equations. The graph is well-defined even absent a generative model.

The other, the directed information graph, is motivated by Granger causality. Directed information quantifies Granger causality in a general prediction framework with a logarithmic loss function. Influences between each pair of processes are directly queried using directed information. We show that under appropriate assumptions, the two graphical models are equivalent.

Our second contribution is to propose algorithms to infer the graph. Two algorithms are described without assumptions on the network. Another, more efficient, algorithm is discussed that recovers the graph when there are known upper bounds on the in-degrees of the nodes. We show this algorithm returns an optimal approximation when the upper-bounds are not valid. We also show a modified version returns a bounded-degree approximation that is robust to estimation errors.

Our third contribution involves identifying sample complexity and confidence intervals for plug-in empirical and parametric estimators of directed information. We use these to compute confidence intervals and develop the robust bounded-degree approximation algorithm.

We also show that even if our assumptions of strict causality are invalid, our results still hold for the strictly causal part of...
the system dynamics. Lastly, we demonstrate the algorithms using simulations and real-world data. We model activity of accounts in Twitter. Our algorithms infer with high precision which users’ activity is influenced by which news corporations. Our analysis only uses message times of topical tweets, not the content nor knowledge of the “follower” graph.

B. Related Work

We next discuss related work on directed information, Granger causality, and graphical models for networks of processes.

1) Directed information: In the information theory community, directed information was first introduced by Marko [9] and independently rediscovered by Kamitake et al. [10] and Rissanen and Wax [11]. Rissanen and Wax proposed their work as an extension of Granger’s framework [2]. Solo showed that Marko’s and Kamitake’s formulas were equivalent [12]. Massey modified Marko’s work for the setting of communication channels [13]. Kramer extended Massey’s work to multiple processes, introducing causal conditioning [14]. Independently in the physics community, Schreiber introduced transfer entropy [15].

Directed information has been used in a variety of settings, e.g., to characterize the capacity of channels with feedback [14], [16]–[20], to quantify achievable rates for source encoding with noiseless feed-forward [21], [22], and for feedback control [23]–[27]. Permuter et al. explored its relevance to gambling, hypothesis testing, and portfolio theory [28].

Applications of directed information include Marko’s study of primate social networks [9], neuroscience studies such as Quinn et al. [29] and [30]–[32], analysis of gene regulatory data [33], [34], and video recordings [35]. Ver Steeg and Galstyan estimated pairwise transfer entropy between Twitter users and their known followers using message timings [36] and between pairs using message content [37].

2) Granger causality: Granger proposed a framework for testing causal relationships between two time-series in a network [2], [38]. It was based on earlier work by Wiener [39]. Sims proposed an alternative test, equivalent to Granger’s [40]. Later works developed general forms of Granger causality based on conditional independencies [41], [42].

Geweke extended Granger’s work, using the logarithm of model errors [43], [44]. Bouissou et al. used log likelihoods of empirical estimates [45]. Gourieroux et al. later proposed directed information as a test statistic for the two-process setting, independent of Marko’s work [46].

In our preliminary work [47], we explored several ways of quantifying Granger causality in a generalized prediction setup with causal side information. We showed that directed information is the value of a prediction problem when predictions are probability measures and expected, cumulative logarithmic loss is used.

3) Graphical models for networks of processes: Dalhaus and Eichler proposed graphical models for autoregressive processes using linear Granger causality [48]–[50]. Quinn et al. proposed a graphical model using directed information, an algorithm to learn the network, and a consistent estimation procedure for directed information [29]. Amblard and Michel [51] proposed extending [48] using directed information as a test statistic. Their work was independent of [29] and did not discuss algorithms, estimation, or graphical properties. In the preliminary work Quinn et al. [52], we proposed two graphical models, one based on generative models and the other using directed information, independent of [51] and with weaker assumptions with [29]; the two graphical models were shown to be equivalent.

Eichler [53] extended his work in the linear setting [48]–[50] for general processes using conditional independence tests based on [41], [42]. The graphs obtained by [53] are identical to the ones obtained in this work and separation criteria identified in [53] hold for the present work. However, note that in [53], no quantitative methods are proposed for performing the conditional independence tests, no algorithms are suggested for efficiently recovering the graphs or approximating them, and there is no measure of the “strength” of the influences.

Other works have investigated linear dynamical systems. Materassi and Innocenti [54] proposed an algorithm analogous to Chow and Liu’s [55], for the case when the network is a tree. Materassi and Salapaka [56] extended [54] to larger classes of topologies. Tan also investigated learning tree structured networks of linear dynamical systems [57]. An alternative approach to identifying sparse networks of stochastic processes used group Lasso [58].

There is also work on an extension of Bayesian networks to the setting of processes, known as dynamic Bayesian networks [59]. Each variable in each process is represented as a node. We note that the variable dependencies underlying our graphical models are dynamic Bayesian networks.

C. Paper Organization

In Section II, we establish definitions and notations. In Section III, we discuss two approaches for graphically representing influences between agents in a network. One focuses on a minimal state space description of relationships. The other is similar to conventional graphical models and is motivated by Granger causality. We demonstrate that under mild assumptions, these different approaches are graphically equivalent. In Section IV, we propose algorithms that identify the graph when no assumptions about the topology are made. Two algorithms also operate without any assumptions on the joint dynamics, but in general require computing the full joint distribution. In Section V, we describe a procedure that uses knowledge of in-degree bounds to more efficiently identify the exact graph—the procedure uses the minimal dimension of statistics needed to do so, as compared to the full joint. We also show that if the in-degree bounds are invalid, the algorithm nonetheless recovers an optimal approximation. In Section VI, we evaluate the sample complexity of two plug-in estimators for directed information. The complexity is similar.
to standard results for statistical estimation with iid data. This allows us to compute confidence intervals for situations where point estimates are not reliable. We then propose an algorithm to identify the best bounded in-degree approximation that is robust to estimation errors. In Section VII, we demonstrate the effectiveness of the algorithms by analyzing simulated autoregressive networks. The algorithms perform well, identifying most of the edges that are influential for the dynamics. In Section VIII we show that our proposed framework can be applied in practical settings. We record, model, and analyze activity of news corporations and users in the Twitter network. The algorithms all identify which news corporations influence which users with high precision. In the appendices, we provide the proofs of the lemmas and theorems in the paper.

II. BACKGROUND

A. Notation and Information-Theoretic Definitions

- For a sequence $a_1, a_2, \ldots, a_i$ denote $(a_i, \ldots, a_j)$ as $a_i^j$ and $a^i := a_i^j$.
- Denote $[m] = \{1, \ldots, m\}$ and the power set $2^{[m]}$ on $[m]$ to be the set of all subsets of $[m]$.
- For any Borel space $Z$, denote its Borel sets by $B(Z)$ and the space of probability measures on $(Z, B(Z))$ as $\mathcal{P}(Z)$.
- Consider two probability measures $P$ and $Q$ on $\mathcal{P}(Z)$. $P$ is absolutely continuous with respect to $Q$ ($P \ll Q$ if $Q(A) = 0$ implies $P(A) = 0$ for all $A \in B(Z)$). If $P \ll Q$, denote the Radon-Nikodym derivative as the random variable $\frac{dP}{dQ} : Z \rightarrow \mathbb{R}$ that satisfies

$$P(A) = \int_{z \in A} \frac{dP}{dQ}(z)Q(dz), \ A \in B(Z).$$

- The Kullback-Leibler divergence between $P \in \mathcal{P}(Z)$ and $Q \in \mathcal{P}(Z)$ is defined as

$$D(P || Q) := \mathbb{E}_P \left( \log \frac{dP}{dQ}(z) \right) = \int_{z \in Z} \log \frac{dP}{dQ}(z)P(dz)$$

if $P \ll Q$ and $\infty$ otherwise.

- Throughout this paper, we will consider $m$ random processes where the $i$th (with $i \in \{1, \ldots, m\}$) random process at time $t$ (with $t \in \{1, \ldots, n\}$), takes values in a Borel space $X$. Denote the $i$th random variable at time $t$ by $X_{i,t} : \Omega \rightarrow X$, the $i$th random process as $X_i = (X_{i,1}, \ldots, X_{i,n})^T$, the whole collection of all $m$ random processes as $X = (X_1, \ldots, X_m)^T$, and a subset of $K$ processes indexed by $A \subseteq [m]$ as $X_A = (X_A(1), \ldots, X_A(K))^T$.

- The probability measure $P$ thus induces a joint distribution on $X$ given by $P_X(\cdot) \in \mathcal{P}(X^n)$, a joint distribution on $X_i$ given by $P_{X_i}(\cdot) \in \mathcal{P}(X^n)$, and a marginal distribution on $X_{i,t}$ given by $P_{X_{i,t}}(\cdot) \in \mathcal{P}(X)$. With slight abuse of notation, denote $X \equiv X_j$ for some $j$ and $Y \equiv X_i$ for some $i \neq j$ and denote the conditional distribution and causally conditioned distribution [14] of $Y$ given $X$ as

$$P_{Y|X=x}(dy) \triangleq P_{Y|X}(dy|x) = \prod_{t=1}^{n} P_{Y_t|Y_{t-1}, X^n}(dy_t|y_{t-1}^t, x^n) \quad (1)$$

$$P_{Y||X=x}(dy) \triangleq P_{Y|X}(dy||x) = \prod_{t=1}^{n} P_{Y_t|Y_{t-1}, X^n}(dy_t|y_{t-1}^t, x^{t-1}) \quad (2)$$

Note the similarity between (1) and (2), though in (2) the future ($x^n_t$) is not conditioned on.

- With slight abuse of notation, $\mathcal{W} \equiv \mathcal{X}_A$ for some $A \subseteq [m] \setminus \{i\}$ with $\mathcal{W} = \mathcal{X}_A^n$. Consider two sets of causally conditioned distributions $\{P_{Y|W=w} \in \mathcal{P}(Y) : w \in \mathcal{W}\}$ and $\{Q_{Y|W=w} \in \mathcal{P}(Y) : w \subseteq \mathcal{W}\}$ along with a marginal distribution $P_W \in \mathcal{P}(\mathcal{W})$. Then the conditional KL divergence is given by

$$D(P_{Y|W||Q_{Y|W=w}}) = \int_{\mathcal{W}} D(P_{Y|W=w}||Q_{Y|W=w})P_W(dw).$$

The following Lemma will be useful throughout:

Lemma 2.1: $D(P_{Y|W||Q_{Y|W=w}}) = 0$ if and only if $P_{Y|W=w}(dy) = Q_{Y|W=w}(dy)$ with $P_W$ probability one.

- Let $X \equiv X_i$ for some $i$, $Y \equiv X_k$ for some $k$ and $\mathcal{W} \equiv \mathcal{X}_A$ for some $A \subseteq [m] \setminus \{i, k\}$. The mutual information, directed information [9], and causally conditioned directed information [14] are given by

$$I(X; Y) \triangleq D(P_{Y|X}||P_Y|P_X)$$

$$= \sum_{t=1}^{n} I(X^n_t; Y_{t-1}^t)$$

$$I(X \rightarrow Y) \triangleq D(P_{Y|X}||P_Y|P_X)$$

$$= \sum_{t=1}^{n} I(X^{t-1}; Y_{t-1}^t)$$

$$I(X \rightarrow Y|W) \triangleq D(P_{Y|X,W}||P_Y|P_X,W)$$

$$= \sum_{t=1}^{n} I(X^{t-1}; Y_{t-1}^t, W_{t-1}) \quad (4)$$

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. We later justify this statement showing that directed information is a general formulation of Granger causality. Note that $I(X; Y) = I(Y; X)$, but $I(X \rightarrow Y) \neq I(Y \rightarrow X)$ in general.

Remark 1: Note that in (2), there is no conditioning on the present $x_t$. This follows Marko’s definition [9] and is consistent with Granger causality [2]. Massey [13] and Kramer [14] later included conditioning on $x_t$ for the specific setting of communication channels. In such settings, since the directions of causation (e.g., that $X$ is the input and $Y$ is the output) are known, it is convenient

$^4$The “$\equiv$” is used to relabel for notational simplicity. Thus “$X \equiv X_j$” means denote $X_j$ by $X$. 
to work with synchronized time, for which conditioning on \(x_t\) is meaningful. Note, however, that by conditioning on the present \(x_t\) in (2), that in a binary symmetric channel (for example) with input \(X\), output \(Y\), and no feedback, \(I(Y \rightarrow X) > 0\), even though \(Y\) does not influence \(X\).

As a consequence of Lemma 2.1 and (4), we have the following corollary.

**Corollary 2.2:** \(I(X \rightarrow Y \| W) = 0\) if and only if \(P_{Y|X=x, W=w}(dy) = P_{Y|W=w}(dy),\ P_X W - a.s.\)

In this case, we say \(Y\) is causally independent of \(X\) causally conditioned on \(W\). We also denote \(I(X \rightarrow Y \| W) = 0\) as \(X \rightarrow W \rightarrow Y\), a causal\(^{5}\) Markov chain.

This is analogous to the Markov chain, denoted \(W\rightarrow X\rightarrow Y\). We also denote \(I(X; Y | W) = 0\) if and only if \(Y\) is independent of \(X\) conditioned on \(W\).

\[
P_{Y|X=x, W=w}(dy) = P_{Y|W=w}(dy),\ P_X W - a.s.
\]

- Let \(G = (V, E)\) denote a directed graph. For each edge \((u, v) \in E\), \(u\) is called the parent and \(v\) is the child.

### III. Minimal Generative Model Graphs and Directed Information Graphs

We now consider the problem of graphically representing causal\(^{6}\) relationships in stochastic dynamical systems.

#### A. Minimal Generative Model Graphs

Stochastic dynamical systems have a natural representation, the coupled differential equations that characterize the dynamics of the system over time. Such a representation explicitly describes inter-dependencies.

**Example 1:** Let \(x_t, y_t, \) and \(z_t\) be three processes comprising a dynamical system. Suppose the differential equations are known and are simplified by removing unnecessary dependencies. For small \(\Delta\), the system becomes

\[
\begin{align*}
x_{t+\Delta} &= x_t + \Delta f(x_t, y_t) \quad (5a) \\
y_{t+\Delta} &= y_t + \Delta g(x_t, y_t) \quad (5b) \\
z_{t+\Delta} &= z_t + \Delta h(y_t, z_t). \quad (5c)
\end{align*}
\]

A natural graphical representation simply depicts the remaining dependencies. See Fig. 3. Note the property that for sufficiently small \(\Delta\), (5) is strictly causal (e.g. \(x_{t+\Delta}\) depends only \(y_t\) but not \(y_{t+\Delta}\)).

We will define a graphical model for stochastic dynamical systems motivated by this simple example. The graphical model is for the more general context where there is no generative model or it might be unknown.

Consider a stochastic dynamical system \(X\) of \(m\) processes with joint distribution \(P_X\). The dynamics of the system are fully described by \(P_X\). First factorize \(P_X\) over time.

\[
P_X(dx) = \prod_{t=1}^{n} P_{X_t|X_{t-1}}(dx_t | x_{t-1}).
\]

5This is “causal” in the sense of Kramer’s causal conditioning (2); \(Y_t\) is independent of \(X_{t-1}\) given \(W_{t-1}\) and \(Y_{t-1}\).

6We will consider two definitions of “causal” in this section. Both are based on observed time-series.

---

**Fig. 3.** A graphical model of the causal influences in the stochastic dynamical system of Example 1.

If \(P_X\) is strictly causal, then like the difference equations (5) in Example 1, it can be further factorized over the indices of the processes,

\[
P_X(dx) = \prod_{i=1}^{m} \prod_{t=1}^{n} P_{X_{i,t} | X_{i,t-1}}(dx_{i,t} | x_{i,t-1}).
\]

Equivalently, using causal conditioning notation (2),

\[
P_X(dx) = \prod_{i=1}^{m} P_{X_{i} | X_{\{m\} \setminus \{i\}}}(dx_{i} | x_{\{m\} \setminus \{i\}}). \quad (6)
\]

By factorizing over time first, each \(X_i\) is still conditioned on the full past of every other process. We will assume that \(P_X\) is both non-degenerate and strictly causal.

**Definition 3.1:** A distribution \(P_X\) is called non-degenerate\(^7\) if there exists a measure \(\phi\) such that \(P_X\) is absolutely continuous with respect to \(\phi\) (\(P_X \ll \phi\)) and \(\frac{dP_X}{d\phi}(x) > 0\) for all \(x\) in the support of \(P_X\).

**Assumption 1:** For the remainder of this paper, we only consider joint distributions that are non-degenerate and strictly causal, satisfying (6).

**Remark 2:** The assumption of non-degeneracy is to avoid cases that arise with purely deterministic relationships between variables. Granger argued that strict causality is a valid assumption if the sampling rate is high enough [2], [61] and relevant processes are observed. A treatment of latent processes is outside the scope of this paper. Furthermore, we will show in Appendix A that even if \(P_X\) is not strictly causal, our algorithms will still recover the strictly causal part of \(P_X\).

We now remove unnecessary dependencies between processes in (6). For each process \(X_i\), let \(A(i) = [m] \setminus \{i\}\) denote a subset of other processes. Define the corresponding induced probability measure \(P_A\).

\[
P_A(dx) = \prod_{i=1}^{m} P_{X_i|X_{A(i)}}(dx_i | x_{A(i)}).
\]

We want to pick the parent sets \(\{A(i)\}_{i=1}^{m}\) so that their cardinalities are small, while still capturing the full dynamics of \(P_X\).\(^8\)

\[
D(P_X \| P_A) = 0. \quad (7)
\]

In Example 1, the \(A(i)\)'s would correspond to \(\{Y\}, \{X\}\), and \(\{Y\}\) for \(X, Y, \) and \(Z\), respectively.

---

\(^{7}\)We follow the notation in [60].

\(^{8}\)The \(A(i)\)'s are defined over the whole time horizon. The \(A(i)\)'s could be defined over sliding windows of time, but that is outside the scope of this work.
Definition 3.2: Under Assumption 1, for a joint distribution $P_X$, a minimal generative model is a function $A : [n] \to 2^{[m]}$ where the cardinalities of the parent sets $\{|A(i)\}|_{i=1}^m$ are minimal such that (7) holds.

Note that the $A(i)$’s together must satisfy (7), suggesting that the choices for $A(i)$ and $A(j)$ might be coupled. However, by non-negativity of the KL divergence and the product form of (6), (7) corresponds to

$$D\left(P_X \| X_{\{m\} \setminus \{i\}} \right| \left. P_X \| X_{A(i)} \right| P_X \| X_{\{m\} \setminus \{i\}} = 0$$

for all $i \in [n]$. Thus, the sets can be chosen separately to satisfy this condition. Furthermore, under our assumption, minimal generative models are unique.

Lemma 3.3: For any distribution $P_X$ satisfying Assumption 1, the minimal generative model is unique.

The proof appears in Appendix B. We now define a corresponding graphical model.

Definition 3.4: A minimal generative model graph is a directed graph for a minimal generative model, where each process is represented by a node, and there is a directed edge from $X_k$ to $X_i$ for $i, k \in [n]$ iff $i \in A(k)$.

Remark 3: Note that while minimal generative model graphs are motivated by the setting where an explicit generative model is known, they are well defined for any $P_X$ satisfying Assumption 1. There need not be a generative model or it could be unknown.

The $\{A(i)\}$ are the smallest parent sets sufficient to capture the full dynamics, satisfying (7). A natural question is whether they are also necessary. Could larger parent sets, $\{W(i)\}$, that do not contain the $\{A(i)\}$, $A(i) \not\subset W(i)$, also capture the full dynamics? The following lemma shows this is not the case. The proof is in Appendix C.

Lemma 3.5: Let $P_X$ be a distribution satisfying Assumption 1. For any subset $B(i)$ containing the parent set $A(i)$, $A(i) \subseteq B(i) \subseteq [m] \setminus \{i\}$, and any subset $W(i) \subseteq [m] \setminus \{i\}$, the causal Markov chain $X_{W(i)} \rightarrow X_{B(i)} \rightarrow X_i$ holds,

$$I(X_{W(i)} \rightarrow X_i ; X_{B(i)}) = 0,$$

and

$$I(X_{W(i)} \rightarrow X_i) \leq I(X_{B(i)} \rightarrow X_i),$$

with equality iff $A(i) \subseteq W(i)$.

Minimal generative model graphs represent reduced factorizations of the joint distribution of the system. They encode causal relationships by only depicting as parents in the graph those subsets of processes that are necessary and sufficient to describe the full dynamics. We next propose an alternative graphical model based on the framework of Granger causality, which directly tests for causal influences between each pair of processes. We show that causally conditioned directed information captures Granger’s principle and use it as an edge selection criterion.

B. Granger Causality and Directed Information Graphs

In 1969, motivated by earlier work by Wiener [39], Nobel laureate Clive Granger proposed a framework for identifying when one process statistically “causes” another [2];

“We say that $X$ is causing $Y$ if we are better able to predict [the future of] $Y$ using all available information than if the information apart from [the past of] $X$ had been used.”

While this definition is general, its previous formulations have mostly been restricted to specific classes of models, such as autoregressive linear models. Specifically, Granger’s setup was as follows [2]. Suppose we have three processes $X, Y, Z$. Testing if $X$ causes $Y$ given knowledge of $Z$ involves finding the least squares estimates for the autoregressive models

$$Y_t = \sum_{\tau > 0} a_{\tau} Y_{t-\tau} + b_{\tau} Z_{t-\tau} + c_{\tau} X_{t-\tau} + E_t$$

$$Y_t = \sum_{\tau > 0} \tilde{a}_{\tau} Y_{t-\tau} + \tilde{b}_{\tau} Z_{t-\tau} + \tilde{E}_t.$$ 

Let $\sigma^2$ and $\tilde{\sigma}^2$ denote the average variances of $E^n$ and $\tilde{E}^n$ respectively. If $\tilde{\sigma}^2 / \sigma^2$ is larger than one in a statistically significant sense, then $X$ is said to cause $Y$ given knowledge of $Z$. This test, and the related $\log \tilde{\sigma}^2 / \sigma^2$ proposed by Geweke [43], have been widely used.

Autoregressive models can capture linear relationships between continuously valued processes. However, many areas of research involve discrete or non-linear processes. For these cases, the linear test has limited applicability. Rissanen and Wax [11] and Bouissou et al. [45] independently proposed comparing log-likelihoods of conditional probabilities. Such tests can be applied to a much larger range of processes than the linear test can. The former work [11] was based on information theoretic methods and the latter work [45] obtained test statistics for the generalized versions of Granger’s work using conditional independencies [41], [42].

Note that the aforementioned works entail using directed information as a test statistic for Granger causality [11], [45]. Later works showed that the value of directed information is the same as the value of Geweke’s statistic $\log \tilde{\sigma}^2 / \sigma^2$ when the processes are jointly Gaussian [62], [63]. These results were for the two processes setting. Other works used causally conditional directed information for Granger causality for networks of processes, such as [29], [30], [51], and [52] (the basis for this work). They were motivated by equivalence in the linear test for jointly Gaussian processes and the relationship of the formula of directed information and the conditional independencies in [41], [42]. Specifically, $I(X^n \rightarrow Y^n) = 0$ if $Y_t$ is independent of $X^{t-1}$ given $Y^{t-1}$.

We now show a stronger connection between directed information and Granger’s framework. We will show that for any class of distributions, the directed information explicitly quantifies Granger’s statement in the setting of sequential prediction with causal side information. Furthermore, the value has meaning not only as a statistic to accept or reject a hypothesis, for which any value above a threshold has the same result. The value measures how strong the influence is in a predictive sense, measured in bits.

Let $Y^n \equiv X_i$ denote the stochastic process being predicted. Let $X^n \equiv X_k$ be another process. We will quantify how much the causal side information of $X^n$ helps in sequentially predicting $Y^n$. (See [64] for an overview of sequential prediction.)
Denote $F_t$ to be the sigma-algebra pertaining to information about the past of all processes, and $\tilde{F}_t$ to be the sigma-algebra pertaining to information about the past of all processes excluding $X^n$:
\[
F_t = \sigma (X_{j,\tau} : j \in [m] \setminus \{k\}, \tau < t; X_{k,\tau} : \tau < t)
\]
\[
\tilde{F}_t = \sigma (X_{j,\tau} : j \in [m] \setminus \{k\}, \tau < t)
\]

Remark 4: Note that if $|m| = 1, 2$, this reduces to
\[
F_t = \sigma (X^{t-1}, Y^{t-1})
\]
\[
\tilde{F}_t = \sigma (Y^{t-1})
\]

At time $t$, one predictor knows the full past of all the processes and specifies a prediction $q_t \in Q$ about $y_t$ that is $F_t$-measurable. The other predictor knows the past of all the processes except $X^n$ and specifies a prediction $\tilde{q}_t \in Q$ about $y_t$ that is $\tilde{F}_t$-measurable. Define the spaces of candidate predictors as

\[
A_t = \{ q : \Omega \to Q \text{ s.t. } q \text{ is } F\text{-measurable} \}
\]
\[
\tilde{A}_t = \{ \tilde{q} : \Omega \to Q \text{ s.t. } \tilde{q} \text{ is } \tilde{F}\text{-measurable} \}
\]

Subsequently, $y_t$ is revealed, and a loss function $l : Q \times Y \to \mathbb{R}^+$ assesses the loss $l(p, y_t)$ for a prediction $p \in Q$ given the outcome $y_t$. Thus, one predictor incurs loss $l(q_t, y_t)$ and the other incurs $l(\tilde{q}_t, y_t)$.

The reduction in loss
\[
r_t(q_t, \tilde{q}_t, y_t) = l(\tilde{q}_t, y_t) - l(q_t, y_t)
\]
characterizes how much the side information of $X^{t-1}$ helps.

Let the space of predictors be the space of probability measures over $Y$ that form a density with respect to $\mu$,
\[
Q = \{ p \in \mathcal{P}(Y) : p \ll \mu \}
\]

A natural loss function for probability measures is the logarithmic loss
\[
l(q, y) = -\log \frac{dq}{d\mu}(y) \quad q \ll \mu.
\]

There are multiple ways of comparing $A_t$ and $\tilde{A}_t$. We consider the expected cumulative reduction in loss between the predictors in $A_t$ and $\tilde{A}_t$ respectively whose expected cumulative loss is minimal. This is analogous to how Granger’s test compares only the linear models with smallest mean-square error. Thus, we focus on

\[
q^*_t = \arg \min_{q_t \in A_t} E_{P_X} [l(q_t, Y_t)] \tag{10}
\]
\[
\tilde{q}^*_t = \arg \min_{\tilde{q}_t \in \tilde{A}_t} E_{P_X} [l(\tilde{q}_t, Y_t)] \tag{11}
\]

The expected cumulative reduction in loss is
\[
\tilde{R}(q^*, \tilde{q}^*) \triangleq E_{P_X} \left[ \sum_{t=1}^{n} r_t(q^*_t, \tilde{q}^*_t, Y_t) \right]
\]

We now state our main lemma, showing that the optimal predictors $q^*$ and $\tilde{q}^*$ are the true conditional distributions and that the reduction in expected loss is precisely the causally conditioned directed information.

Lemma 3.6: The optimal solutions to (10) and (11) are given by
\[
q^*_t = \hat{P}_{Y|X_i}, \quad \tilde{q}^*_t = \hat{P}_{Y|X_i}.
\]
The expected cumulative reduction in loss is given by the causally conditioned directed information
\[
\tilde{R}(q^*, \tilde{q}^*) = I \left( X^n \to Y^n \| X_{[m] \setminus \{i,k\}} \right).
\]

Proof: Note that
\[
q^*_t = \arg \min_{q_t \in A_t} \mathbb{E} P_X \left[ -\log \frac{dq_t}{d\mu}(Y_t) \right]
\]
\[
= \arg \min_{q_t \in A_t} \mathbb{E} P_X \left[ -\log \frac{dP_{Y|X_i}}{d\mu}(Y_t) + \log \frac{dP_{Y|\tilde{F}_t}}{d\mu} \right]
\]
\[
= \arg \min_{q_t \in A_t} \mathbb{D} (P_{Y|X_i} \| q_t) \tag{12}
\]
where (12) follows from the definition of divergence and that the left-hand term in the expectation does not affect the arg min. Moreover, note that clearly $P_{Y|X_i}$ is $F_t$-measurable and thus from the non-negativity of the KL divergence, $q^*_t = \hat{P}_{Y|X_i}$. Similarly, $\tilde{q}^*_t = \hat{P}_{Y|\tilde{F}_t}$.

We now discuss using Granger’s notion of “better” to address the two predictors. Since clearly $q^*_t \ll \tilde{q}^*_t$, note that the reference measure $\mu$ disappears and the reduction in loss becomes a log-likelihood ratio
\[
\tilde{r}_t(q^*_t, \tilde{q}^*_t, y_t) = \log \frac{dq^*_t}{d\mu}(y_t) = \log \frac{dP_{Y|X_i}}{dP_{Y|\tilde{F}_t}}(y_t).
\]

Thus,
\[
\tilde{R}(q^*, \tilde{q}^*) = \mathbb{E} P_X \left[ \sum_{t=1}^{n} \tilde{r}_t(q^*_t, \tilde{q}^*_t, Y_t) \right]
\]
\[
= \mathbb{E} P_X \left[ \sum_{t=1}^{n} \log \frac{dP_{Y|X_i}}{dP_{Y|\tilde{F}_t}}(Y_t) \right]
\]
\[
= I \left( X^n \to Y^n \| X_{[m] \setminus \{i,k\}} \right).
\]

Lemma 3.6 states that in sequentially predicting $Y_t$, the expected cumulative reduction in loss due to the causal side information $X^{t-1}$, which is a general formulation of Granger’s statement, is precisely the directed information when the predictors are probability measures and the loss is the logarithmic loss. Thus, in this setting, we can interpret the value of directed information as quantifying the “strength” of the influence in the reduction in bits. In the preliminary work [47], we explore other sequential prediction settings, such as minimax, where the value of Granger’s statement is a different quantity.

We now define a graphical model using directed information.

Definition 3.7: For a set of random processes $X$, the directed information graph is a directed graph where each node represents a process and there is a directed edge from process $X \equiv X_k$ to process $Y \equiv X_i$ (for $i, k \in [m]$) iff
\[
I(X \to Y \| X_{[m] \setminus \{i,k\}}) > 0.
\]
In this model, there is a directed edge from $X$ to $Y$ if and only if causal knowledge of $X$ still influences $Y$, even with causal knowledge of all other processes. Since edges are found separately, directed information graphs are unique. Also, directed cycles are possible.

Remark 5: Note that this graphical model is similar to the model in [53] (which was independent of our earlier, preliminary work [52]). However, while [53] only tests for the presence of edges, we use the strength of the influence, as measured by directed information. This will be necessary to approximate the graph (see Section V-B).

Minimal generative model graphs and directed information graphs are alternative graphical models to characterize the relationships in stochastic dynamical systems. We now investigate their relation.

Theorem 3.8: For any joint distribution $P_X$ satisfying Assumption 1, the corresponding minimal generative model graph and directed information graph are equivalent.

The proof is in Appendix D. That directed information graphs and minimal generative model graphs are the same suggests their utility—if they produced different graphs, then it would be necessary to assess which was “correct” for the edges where they differed.

IV. Graphical Model Identification — General

In this section, we discuss algorithms to identify the graph when no assumptions about the topology are made. They will take as inputs directed informations. Efficiency will correspond to the dimension of the statistics that will be necessary, such as only needing joint statistics of pairs of processes as compared to the full joint distribution. Also note that by Theorem 3.8, the algorithms will learn the network by either identifying the directed information graph or the minimum generative model graph.

A. Algorithm 1 — parent set search

Identifying the minimal generative model graphs by Def. 3.4 involves determining, for each process $X_i$, the minimal cardinality parent set $A(i)$ that satisfies (8). No search order is prespecified. Since the goal is to find the smallest $A(i)$, one approach is to test increasing sizes of subsets of potential parents. For instance, first the empty set $\emptyset$ is tested, then individual processes, then pairs of processes, etc. This would require calculating an exponential number of causally conditioned directed informations (4).

An alternative method is motivated by Markov chains and Lemma 3.5. To find process $X_i$’s parents, start with the subset of all other processes as a trivial Markov blanket and sequentially test each process’s contribution, removing insignificant ones. This method is formally described in Algorithm 1.

Let $D\mathcal{I}_{\text{MGM}}$ denote the set of all causally conditioned directed information values from one process to another, causally conditioned on a subset of the rest,

$$D\mathcal{I}_{\text{MGM}} = \left\{ I(X_k \rightarrow X_i; X_{B(i)}): k, i \in [m], B(i) \subseteq [m] \backslash \{i, k\} \right\}.$$ 

Algorithm 1. MGMconstruct

Input: $D\mathcal{I}_{\text{MGM}}$, $m$

1. For $i \in [m]$
2. $A(i) \leftarrow [m]\backslash\{i\}$
3. For $k \in A(i)$
4. $B(i) \leftarrow A(i)\backslash\{k\}$
5. If $I(X_k \rightarrow X_i; X_{B(i)}) = 0$
6. $A(i) \leftarrow B(i)$
7. Return $\{A(i)\}_{i=1}^{m}$

Lemma 4.1: Let $P_X$ be a distribution satisfying Assumption 1. Algorithm 1 recovers the minimal generative model graph.

The proof follows from Lemma 3.5 and the uniqueness of minimal generative model graphs, Lemma 3.3. Algorithm 1 requires the full joint distribution. However, it only uses $O(m^2)$ tests. Note that the tests used in line 5 are adaptive, using the current $B(i)$. Next consider an alternative algorithm following the definition of directed information graphs (Def. 3.7).

B. Algorithm 2 — edge tests

Directed information graphs are identified by testing each edge separately. Testing an edge entails computing a directed information from one process to another, causally conditioned on all other processes. This is described in Algorithm 2. Let $D\mathcal{I}_{\text{DI}}$ denote that set of causally conditioned directed informations

$$D\mathcal{I}_{\text{DI}} = \{ I(X_k \rightarrow X_i; X_{[m]\backslash\{i, k\}}): i, k \in [m] \}.$$ 

Algorithm 2. DIconstruct

Input: $D\mathcal{I}_{\text{DI}}$, $m$

1. For $i \in [m]$
2. $A(i) \leftarrow \emptyset$
3. For $i, k \in [m]$
4. If $I(X_k \rightarrow X_i; X_{[m]\backslash\{i, k\}}) > 0$
5. $A(i) \leftarrow A(i) \cup \{k\}$
6. Return $\{A(i)\}_{i=1}^{m}$

Lemma 4.2: Let $P_X$ be a distribution satisfying Assumption 1. Algorithm 2 recovers the directed information graph.

The proof follows from Definition 3.7. Unlike Algorithm 1, Algorithm 2 uses each of the $O(m^2)$ elements in $D\mathcal{I}_{\text{DI}}$. Line 4 could be executed in parallel for every possible edge. The number of causally conditioned directed information tests is the same as Algorithm 1, though the tests themselves are different.

C. Algorithm 3 — adaptive

In [29], we proposed an adaptive algorithm which identifies parents by first using pairwise tests, then conditioning on one process, then two, etc. Thus, the initial tests are easy to compute. We include the algorithm, Algorithm 3, for completeness. The input is

$$D\mathcal{I}_{\text{GenAdp}} = \left\{ I(X_{B(i)} \rightarrow X_i): i, j \in [m], B \subseteq [m]\backslash\{i, j\} \right\}.$$
For networks that are sparse but do not have known in-degree bounds, Algorithm 3 could potentially run quickly with the design of identifying non-parents using directed information with few processes. However, the following assumption is required for Algorithm 3 to identify the exact topology.

Assumption 2: For distribution $P_X$, for all $X_i, X_j \in X$ and $S \subseteq [m] \setminus \{i, j\}$,

$I(X_j \rightarrow X_i | X_S) > 0 \implies I(X_j \rightarrow X_i | X_S) > 0$.

Algorithm 3. GenStructAdapt

Input: $DL_{GenAdp}, m$

1. $\text{For } i \in [m]$
2. $A(i) \leftarrow [m] \setminus \{i\}$
3. $K \leftarrow 1$
4. $\text{While } K \leq |A(i)|$
      5. $B \leftarrow \{B': B' \subseteq A(i), |B'| = K\}$
      6. $\text{For } B \in B$
         7. For $j \in B$
            8. If $I(X_j \rightarrow X_i | X_B) = 0$
               9. $A(i) \leftarrow A(i) \cup \{j\}$
      10. $K \leftarrow K + 1$
11. Return $\{A(i)\}_{i=1}^m$

Lemma 4.3: Let $P_X$ be a distribution satisfying Assumptions 1 and 2. Algorithm 3 recovers the directed information graph.

This can be seen as all non-parents $j \notin A(i)$ will eventually be tested in line 8 while conditioning on the parents, with $B = A(i) \cup \{j\}$. Every such instance will evaluate as true, removing that non-parent. For each process $X_i$, $K$ will increment until it is at most the size of the parent set plus one.

V. GRAPHICAL MODEL IDENTIFICATION – SIDE INFORMATION

We now discuss an algorithm to identify the graph when side information is known, specifically bounds on the in-degrees. We will then show even if the side information is invalid, the resulting graph is an optimal approximation.

A. Algorithm 4 — Bounded In-Degree

Algorithms 1 and 2 both require inputs computed using the whole joint distribution of the system. In general, if $X_i$ has $K = |A(i)|$ parents, then at least $(K+1)$-dimensional statistics are needed to guarantee recovery. Algorithm 3 only used $(K+2)$-dimensional statistics, but required an assumption. We next show that when there are known upper bounds $\{K(i)\}_{i=1}^m$ on the in-degrees, $(K(i)+1)$-dimensional statistics can be used to identify $X_i$’s parents.

Lemma 3.5 showed that for each process $X_i$, only its parents carry the influence. Among all sets of $K(i)$ processes, those that contain $X_i$’s parents will have maximal directed information to $X_i$. We can then take the intersection of these sets to get precisely $X_i$’s parents. Algorithm 4 formally describes this method. Let $DL_{BndInd}$ denote a set of directed information values, such that for each $i \in [m]$, $DL_{BndInd}$ contains directed information values from each $K(i)$ sized subset of processes to $X_i$.

$DL_{BndInd} = \{I(X_{B(i)} \rightarrow X_i) : i \in [m], B(i) \subseteq [m] \setminus \{i\}, |B(i)| = K(i)\}$.

Algorithm 4. BoundedInDegree

Input: $DL_{BndInd}, K, m$

1. $\text{For } i \in [m]$
2. $A(i) \leftarrow \emptyset$
3. $B \leftarrow \{B' : B' \subseteq [m] \setminus \{i\}, |B'| = K(i)\}$
4. $B_{max} \leftarrow \arg \max_{B \in B} I(X_B \rightarrow X_i)$
5. $A(i) \leftarrow \bigcap_{B \in B_{max}} B$
6. Return $\{A(i)\}_{i=1}^m$

Theorem 5.1: Let $P_X$ be a distribution satisfying Assumption 1. Algorithm 4 recovers the directed information graph for a given $P_X$ if for each $i \in [m]$, $K(i) \geq |A(i)|$.

The proof follows from Lemma 3.5.

Algorithm 4 finds the graph using only statistics of the dimension of the bound of the in-degree. Algorithm 4 uses all of the elements in $DL_{BndInd}$, which are $\sum_{i=1}^m \binom{m-1}{K(i)}$ values. Note that if the upper bounds $\{K(i)\}_{i=1}^m$ do not grow with $m$, then the algorithm performs $O(m^{K+1})$ directed information tests, where $K' = \max_{i \in [m]} K(i)$. While more tests are used than in Algorithms 1 and 2, only $(K'+1)$-wise statistics are used. Thus, the time to compute or estimate the causally conditioned directed information values for $DL_{BndInd}$ could be significantly less than that for $DL_{MGM}$ or $DL_{DI}$.

Remark 6: In practice, one should test the output $\{A(i)\}_{i=1}^m$ of Algorithm 4 to make sure the returned parent sets do carry maximal influence. If the output is correct, then the following equality holds,

$I(X_{A(i)} \rightarrow X_i) = \arg \max_{B \in B} I(X_B \rightarrow X_i)$.

A large inequality could be due to invalid bounds or noisy estimates. In the next section we address invalid bounds, and in Section VI-C we address noisy estimates.

B. Bounded In-Degree Approximations

Algorithm 4 requires bounds on the in-degrees. A natural question is whether the output is useful if the bounds are invalid. We next show that by modifying line 5 in Algorithm 4 to return any set $B \in B_{max}$, the result is an optimal approximation, regardless of the validity of the bounds.

Consider approximating $P_X$ with

$\hat{P}_X(dx) \triangleq \prod_{i=1}^m P_X | x_{\hat{A}(i)} dx_i \mid x_{\hat{A}(i)}$.

(13)

Note that in (13), the conditional marginals are exact, but the parent sets $\{A(i)\}_{i=1}^m$ are approximate. The divergence $D(\hat{P}_X \| P_X)$ measures how close $\hat{P}_X$ is to $P_X$. Let $\hat{P}_K$
denote the set of all approximations of the form (13) with \(|\hat{A}(i)| = K(i)|). The optimal approximation is thus
\[
P^*_X \triangleq \arg\min_{P_X \in \hat{P}_K} D(P_X \| \hat{P}_X).
\]

For the setting where \(\hat{P}_X\) is constrained to be a directed tree, Quinn et al. [65] show that \(\hat{P}_X\) is the directed tree with the maximal sum of directed informations along its edges. We now show an analogous result in the setting where \(\hat{P}_X\) is constrained to have specified in-degrees but not required to be connected.

**Theorem 5.2:**
\[
\arg\min_{P_X \in \hat{P}_K} D(P_X \| \hat{P}_X) = \arg\max_{I \in \hat{P}_K} \sum_{i=1}^{m} \max_{\hat{A}(i)|A(i)|=K(i)} I(X_{\hat{A}(i)} \to X_i). \tag{14}
\]
The proof is similar to that in [65]. For that setting, \(K = 1\). The same proof structure can be used here to yield
\[
\arg\min_{P_X \in \hat{P}_K} D(P_X \| \hat{P}_X) = \arg\max_{I \in \hat{P}_K} \sum_{i=1}^{m} I(X_{\hat{A}(i)} \to X_i).
\]

Note that in this setting, the parent sets do not depend on each other. The \(\arg\max\) can be brought inside to give (14). Also note that with any \(\hat{A}(i) \in \arg\max_{B \in B} I(X_B \to X_i)\), the resulting \(\hat{P}_X\) will be an optimal approximation for those in-degrees.

## VI. Estimation of Directed Information

In this section, we discuss estimation of directed information. We first review previous results. We then examine sample complexity and confidence bounds for two plug-in estimators. We will present the two-process case (the results generalize). Lastly, we consider a version of Algorithm 4 that is robust to estimation errors.

There have recently been several estimators proposed for directed information. Rao et al. studied gene expression used a consistent partitioning method based on entropy estimation [33], [34]. Liu et al. presented an alternative adaptive partitioning based method [66]. Vicente et al. [67] adapted a k-nearest neighbors mutual information estimator from Kraskov et al. [68]. However, these works do not discuss consistency. A consistent parametric estimation technique was independently proposed by Quinn et al. [29] and Kim et al. [30]. Jiao et al. developed a universal estimation scheme for directed information in the finite-alphabet setting using context-tree weighting [69].

Although two of the estimation techniques [29], [69] are consistent asymptotically, for the setting of limited data it is important to know the sample complexity of the estimators and to be able to compute confidence intervals. Jiao et al. provides sample complexity results for the universal estimator [69]. We next identify sample complexity and confidence bounds for two plug-in estimators, the first a finite alphabet empirical estimator and the second a parametric estimator. For simplicity of presentation, we focus on estimating directed information between pairs of processes, such as \(I(X_j \to X_i)\). The results extend for more general cases.

**Assumption 3:** We assume the network \(\mathbf{X}\) is jointly stationary, ergodic, and Markov of finite order \(l\). We further assume that each pair of processes \(\{X_i, X_j\}\) are Markov order \(l\).

Under Assumption 3, the pairwise directed information is
\[
I(X_j \to X_i) = \frac{1}{n} \sum_{t=1}^{n} I(X_{i,t}; X_{j,t-1}^{t-1} X_{j,t-2}^{t-1}) \tag{15}
\]
\[
= I(X_{i,l+1}; X_{j,l+1}^{l+1}) \tag{16}
\]
\[
= \sum_{x_j^{l+1} \in X_j^{l+1}} P_{X_j, X_i^{l+1}}(x_j^{l+1}) \times \log \frac{P_{X_j, X_i^{l+1}}(x_j^{l+1})}{P_{X_i^{l+1}}(x_i^{l+1})}. \tag{17}
\]

Eq. (15) follows from Markovity, (16) follows from stationarity, and (17) follows from the definition of mutual information (3).

We first estimate the pairwise distributions \(\hat{P}_{X_j, X_i^{l+1}}\), and then plug those into (17) to obtain a directed information estimate \(\hat{I}(X_j \to X_i)\). The confidence interval is set as
\[
\hat{I}(X_j \to X_i) \triangleq [\hat{I}(X_j \to X_i) - \delta, \hat{I}(X_j \to X_i) + \delta]\]
for a given constant \(\delta > 0\). Let \(R = m(m-1)\) count the number of ordered pairs \((i, j)\). Let \(B_\delta\) denote the event that all pairwise directed informations are within their respective intervals, i.e.,
\[
B_\delta \triangleq \{ \forall r \in \{1, \ldots, R\}, |\hat{I}(X_{j_r} \to X_{i_r}) - I(X_{j_r} \to X_{i_r})| < \delta \}.
\]

We next examine the sample complexity of these estimators to characterize \(\mathbb{P}(B_\delta)\) as a function of \(n\).

### A. Non-parametric Empirical Estimator for Finite Alphabets

In the setting of a finite alphabet \(\mathbf{X}\), we will use the “empirical” distribution for a non-parametric estimator. For each ordered pair \((i, j)\) we compute a distribution \(\hat{P}_{X_j, X_i^{l+1}}\), where for each possible realization \(\{x_j^{l+1}, x_i^{l+1}\} \in X^{l+1}\) of \(\{X_j, X_i^{l+1}\}\), the estimates are
\[
\hat{P}_{X_j, X_i^{l+1}}(x_j^{l+1}, x_i^{l+1}) \triangleq \frac{1}{n} \sum_{t=1}^{n} 1 \{ (x_j^{t-l-1}, x_i^{t-l-1}) = (x_j^{t-l+1}, x_i^{t-l+1}) \}. \tag{18}
\]

To ensure convergence of the empirical estimator (18), we require an additional assumption on the mixing time of the processes. Denote the state of the network from time \(t - l\) to time \(t\) by \(\mathbf{V}_t \triangleq \mathbf{X}_{m(l+l)}\). Then \(\{\mathbf{V}_t\}_{t=1}^{n}\) forms a first-order Markov chain. We assume this chain satisfies the following condition which is related to uniform ergodicity [70].

**Assumption 4:** There exists a probability measure \(\phi\) on \(X^{l+1}\), a constant \(0 < \lambda \leq 1\), and an integer \(d \geq 2\) such that for all \(v_1 \in X^{m(l+1)}\), \(\mathbb{P}(V_{2t} = v | V_1 = v_1) \geq \lambda \phi(v)\).

\(^9\)For the general case of directed informations with \(N\) processes, each \(N\)-tuple of processes \(\{X_{i_1}, \ldots, X_{i_N}\}\) must be Markov order \(l\). The pairwise Markovity is required to have a simplified, time-independent formula (15).

\(^{10}\)For notational simplicity, shift the time indexing to start at \(t = 0\) so for \(t = 1\) there is a length \(l\) history.
If the Markov chain converges to a stationary distribution \( \pi \), then Assumption 4 can be applied with \( \phi = \pi \) and \( \lambda \approx 1 \) for sufficiently large \( d \). There is a trade-off between decreasing \( d \) and increasing \( \lambda \).

**Theorem 6.1:** Under Assumptions 3 and 4, for a given \( \delta > 0 \), \( \mathbb{P}(B_3) \geq 1 - \rho \), where \( \epsilon \) is chosen so that \( \delta = -4|X|^{2d+1}e\log \epsilon \) and

\[
\rho = 8R|X|^{2d+1} \exp\left(-\frac{(ne - 2d/\lambda)^2}{2nd^2/\lambda^2}\right).
\]  

(19)

For any \( \epsilon' > 0 \), the sample complexity of Algorithm 1 is

\[ \delta = \mathcal{O}(n^{-1/2} + \epsilon') \]

and the graph sample complexity is \( n = \mathcal{O}(\log n) \).

The proof appears in Appendix E.

**B. Parametric estimator**

Parametric models are widely used for modeling time-series in economics, biology, and other fields. We next identify the sample complexity for parametric plug-in estimators. We consider a network of stochastic processes whose conditional distribution \( P_{X_i|X_{i-1}, \theta} \) is characterized by a parameter vector \( \theta^* \). We next discuss conditions for the maximum likelihood estimate (MLE) \( \hat{\theta}_n \) to exist.

Let \( F_i \) denote the \( \sigma \)-field generated by \( X_i \). Let \( Q = \{ \theta^* \} \) and suppose \( \theta^* \) is in the interior of \( \Theta \), a compact subset of \( \mathbb{R}^q \). Let \( q \) index the parameter vector \( \theta = \{ \theta_q \}_{q=1}^Q \in \Theta \). Denote the conditional log-likelihood of \( X_i \) parameterized by \( \theta \) as

\[ L_i(\theta) \triangleq \log P_{X_i|X_{i-1}; \theta}(X_i|X_{i-1}). \]

Define the matrices \( A_i(\theta) \) and \( G_i(\theta) \) evaluated at \( \theta^* \) as

\[
A_i(\theta^*) = \left[ -\frac{\partial^2 L_i(\theta)}{\partial \theta_q \partial \theta_r} \right]_{\theta = \theta^*} 1 \leq q, r \leq Q
\]

\[
G_i(\theta^*) = \left[ \frac{\partial L_i(\theta)}{\partial \theta_q} \right]_{\theta = \theta^*} 1 \leq q, r \leq Q
\]

**Assumption 5:** \( L_i(\theta) \) is almost surely and continuously twice differentiable in terms of \( \theta \), \( \mathbb{E}[\sup_{\theta \in \Theta} |L_i(\theta)|] < \infty \) and \( \mathbb{E}[L_i(\theta)] \) has a unique maximizer at \( \theta^* \). The vector \( \left[ \frac{\partial L_i(\theta)}{\partial \theta_q} \right]_{\theta = \theta^*} 1 \leq q \leq Q \) is a martingale differentiable in terms of \( F_i \) with \( \mathbb{E}[G_i(\theta^*)] \) finite and positive definite. \( \mathbb{E}[A_i(\theta^*)] \) is positive definite and \( \mathbb{E}[\sup_{\theta \in \Theta} ||\theta^* - \theta||_2] < \infty \) for some \( \eta > 0 \).

**Remark 7:** Several classes of autoregressive (AR) time-series models such as threshold AR models, bilinear AR moving averages, GARCH models, and random coefficient AR models have been shown to satisfy Assumption 5 when they are stationary and ergodic [71].

Define the covariance matrix

\[ \Sigma \triangleq [\mathbb{E}[A_i(\theta^*)]]^{-1} [\mathbb{E}[G_i(\theta^*)]] [\mathbb{E}[A_i(\theta^*)]]^{-1}. \]

(20)

**Lemma 6.2:** Under Assumptions 3 and 5,

\[ \sqrt{n}(\hat{\theta}_n - \theta^*) \rightarrow \mathcal{N}(0, \Sigma) \] in distribution.

(21)

This follows from [71]. Lemma 6.2 extends to functions of the parameters. Let \( g_r(\theta) \) denote the directed information of the \( r \)th pair \( (i_r, j_r) \) computed with \( \theta^* \).

\[ g_r(\theta^*) \triangleq I(X_{j_r} \rightarrow X_{i_r}). \]

Using the \( Q \times Q \) parameter covariance matrix \( \Sigma = (\sigma_{q,q'}) \) (20), define the \( R \times R \) covariance matrix \( \Sigma' = (\sigma'_{r,r'}) \) for the directed information estimates as \( \sigma'_{r,r'} = \sum_{q=1}^Q \sigma_q q' \frac{\partial \theta_q}{\partial \theta_{q'}} \). Under Assumptions 3 and 5.

**Lemma 6.3:** Under Assumptions 3 and 5,

\[ \sqrt{n} \left( (g_1(\hat{\theta}_n) - g_1(\theta^*)), \ldots, (g_R(\hat{\theta}_n) - g_R(\theta^*)) \right) \]

\[ \rightarrow \mathcal{N}(0, \Sigma') \] in distribution.

(22)

This follows by the multivariate delta method (see Theorem 5.4.6 of [72]). Note that under Assumptions 3 and 5, the unknown covariance matrices \( \Sigma \) and \( \Sigma' \) in (21) and (22) respectively can be consistently estimated by using \( \hat{\theta}_n \) in place of the unknown \( \theta^* \) [71]. We next identify the sample complexity results.

**Theorem 6.4:** Under Assumptions 3 and 5, the sample complexity is \( \delta = \mathcal{O}(n^{-1/2}) \) and the graph sample complexity is \( n = \mathcal{O}(\log n) \).

The proof appears in Appendix F.

**Remark 8:** Calculating \( \Sigma' \) might be difficult in some cases. For practical implementation, confidence intervals for directed information can be approximated as follows. Separately fit the conditional marginals \( P_{Y_i|Y_{i-1}^i, \hat{\theta}_n} \) and \( P_{Y_i|Y_{i-1}^i, \hat{\theta}_n} \) to compute confidence intervals and compute the directed information for each sample to estimate the confidence interval for \( I(X \rightarrow Y) \).

**C. Robust Graph Identification**

Algorithms 1-4 require calculations or point estimates of directed information to recover the graph. When point estimates \( \{I(X_{A(i)} \rightarrow X_i)\} \) are not reliable, one might prefer to work with confidence intervals, denoted as \( \{I(X_{A(i)} \rightarrow X_i)\} \). For example, the 95\% confidence region corresponds to

\[ \mathbb{P}( \bigcup_{(i,A(i))} \{I(X_{A(i)} \rightarrow X_j) \in \hat{I}(X_{A(i)} \rightarrow X_j)\} \geq 0.95. \]

(23)

The intersection in (23) is over all \( m(m-1) \) pairs of \( i \) and \( A(i) \subseteq [m] \setminus \{i\} \) with \( |A(i)| = K \). Denote the Cartesian product of confidence intervals as

\[ S \triangleq \bigotimes_{(i,A(i))} \hat{I}(X_{A(i)} \rightarrow X_i). \]

Note that \( S \) is a subset of \( \mathbb{R}^m(m-1) \). Each element \( s \in S \) is a length \( m(m-1) \) vector. We refer to each \( s \in S \) as a scenario.

---

11In this work, we consider the practical setting of using a set of simultaneous confidence intervals for each parameter, resulting in a rectangular, joint confidence region. (See Ch. 5 of [73].) In general, multidimensional confidence regions need not be rectangular.

12In this section, we use a constant in-degree \( K \) for notational simplicity; the results generalize.
Each scenario selects an estimate value $\hat{I}_s(\mathbf{X}_{A(i)} \rightarrow \mathbf{X}_i) \in \hat{I}(\mathbf{X}_{A(i)} \rightarrow \mathbf{X}_i)$ for all pairs $(i, A(i))$.

We now modify Algorithm 3, specifically the optimal approximation version in Section V-B, to select approximate parent sets so that the approximation will be robust to estimation errors. Recall from (14) that the quality of the approximation corresponds to the summation of directed informations from parents to children. For a given scenario $s \in \mathcal{S}$ and approximation $\hat{P}_X$ (13), following from Theorem 5.2, the goodness of the approximation is characterized by its weight, $W(\hat{P}_X, s) \triangleq \sum_{i=1}^{m} \hat{I}_s(\mathbf{X}_{\hat{A}(i)} \rightarrow \mathbf{X}_i)$. The best approximation for a particular scenario $s$ is $\hat{P}_X^*(s) \triangleq \arg \max_{\hat{P}_X \in \hat{P}_K} W(\hat{P}_X, s)$.

Algorithm 5 can solve (24) to give the best parent sets for a given $s$, those parents might perform poorly in a different scenario $s' \in \mathcal{S}$ compared to $\hat{P}_X^*(s')$. A natural question is whether there is a $\hat{P}_X$ that performs well under all scenarios. In particular, we want to select the “robust” approximation $\hat{P}_{rob}$ that attains the minimax regret,

$$\hat{P}_{rob} \triangleq \arg \min_{\hat{P}_X \in \hat{P}_K} \max_{s \in \mathcal{S}} \{W(\hat{P}_X^*(s), s) - W(\hat{P}_X, s)\}.$$  

Algorithm 5. RobustBoundedIn-Degree

Input: $\mathcal{S}, K, m$

1. For $i \in [m]
2. \quad \hat{A}(i) \leftarrow \emptyset
3. \quad \mathcal{B} \leftarrow \{ B' : B' \subseteq [m] \setminus \{i\} \}, |B'| = K
4. For $j \in \{1, \cdots, |\mathcal{B}|\}$
5. \quad $B_j \leftarrow B(j)$
6. $M(B_j) \leftarrow \text{midpoint}(\hat{I}(\mathbf{X}_{B_j} \rightarrow \mathbf{X}_i))$
7. $H(B_j) \leftarrow \max(\hat{I}(\mathbf{X}_{B_j} \rightarrow \mathbf{X}_i))$
8. $L(B_j) \leftarrow \min(\hat{I}(\mathbf{X}_{B_j} \rightarrow \mathbf{X}_i))$
9. $j_1 \leftarrow \arg \max_{j} H(B_j)$
10. $j_2 \leftarrow \arg \max_{j} L(B_j)$
11. $j_3 \leftarrow \arg \max_{j \neq j_1, j_2} H(B_j)$
12. If $M(B_{j_1}) \geq \frac{1}{4} (H(B_{j_1}) + L(B_{j_2}))$
13. \quad $\hat{A}(i) \leftarrow B_{j_1}$
14. Else
15. \quad $\hat{A}(i) \leftarrow B_{j_2}$
16. Return $\{ \hat{A}(i) \}_{i=1}^{m}$

Algorithm 5 extends the optimal approximation version of Algorithm 3 to the setting of using confidence intervals.

Theorem 6.5: Under Assumption 1, Algorithm 5 identifies $\hat{P}_{rob}$.

The proof appears in Appendix G.

VII. Simulations

In this section, we demonstrate the performance of the algorithms presented in Section IV through analysis of simulated networks of autoregressive processes.

A. Exact recovery simulations – Algorithms 2, 3, and 4

1) Setup: We first analyzed the performance of Algorithms 2, 3, and 4 using network simulations of autoregressive (AR) processes. The simulations used Markov order-1 models

$$X_t = CX_{t-1} + N_t$$

for a given $m$ by $m$ coefficient matrix $C$ and noise vector $N_t$.

Two network sizes $m \in \{6, 15\}$ were tested. For each size $m$, there were 200 trials. In each trial, the parent sets and the corresponding AR coefficients were generated. For each node, the number of parents was chosen at uniform between 0 and 3 for $m = 6$ and between 0 and 6 for $m = 15$. The non-zero AR coefficients were drawn i.i.d. from a standard normal distribution. The AR coefficient matrix was then scaled so that the largest magnitude of its eigenvalues was 0.9. The noise process $\{N_t\}_{t=1}^n$ had i.i.d. entries drawn from a normal distribution with mean zero and variance $1/4$. Data was generated for $n = 750$ time-steps.

Each of Algorithms 2, 3, and 4 ran on the simulated data. Performance was measured by two quantities. The first is the proportion of edges correctly identified as present or absent. The second is the proportion of the sum of directed information of the estimated parents to their children as compared to that for the true parents

$$\frac{\sum_{i=1}^{m} I(\hat{X}_{\hat{A}(i)} \rightarrow X_i)}{\sum_{i=1}^{m} I(X_{A(i)} \rightarrow X_i)},$$

where $A(i)$ and $\hat{A}(i)$ denote the true and inferred parent sets respectively. This second measure characterizes how much of the dynamics are captured by the estimated parent sets. Performance is averaged over the trials.

Each of the algorithms require directed information estimates. The estimate for directed informations of the form $I(\mathbf{X} \rightarrow \mathbf{Y} \| \mathbf{Z})$ was computed as follows. Least square estimates for the coefficients in

$$Y_t = b_1 Y_{t-1} + b_2 Z_{t-1} + b_3 X_{t-1} + N_t$$

were computed. Let $\sigma$ and $\sigma'$ denote std($N_t$) and std($N'_t$) respectively. From Theorem 8.4.1 of [75], the entropy $H(Y \| \mathbf{Z}, \mathbf{X})$ is $1/2 \log_2(2\pi e \sigma'^2)$. The directed information is then $\log\sigma'/\sigma$. Algorithms 2 and 3 make decisions based on whether certain directed informations are zero or positive. To avoid small values due to over-fitting, we use the minimum description length (MDL) penalty [76], $J * \log_2(n)/(2n)$, where $J$ is the number of parameters. Thus, the first model (26)’s total complexity is $H(Y \| \mathbf{Z}, \mathbf{X}) + 3 \log_2(n)/(2n)$ and the second model (27)’s total complexity is $H(Y \| \mathbf{Z}) + 2 \log_2(n)/(2n)$. To select edges, Algorithms 2 and 3 test if the difference in total complexity is positive

$$\hat{I}(\mathbf{X} \rightarrow \mathbf{Y} \| \mathbf{Z}) > \frac{(3 - 2)}{2n} \log_2(n).$$

Multivariate AR time-series have a limiting stationary distribution iff the largest magnitude of its eigenvalues is less than one (see pg. 88 in [74]).
Relatedly, for Algorithm 4, \( B_{\text{max}} \) initially consisted of a single parent set, denote as \( B_{\text{max}} \). To resolve which other parent sets had the same maximal influence except for numerical discrepancies, we again used the MDL penalty. We set
\[
B_{\text{max}} = \left\{ B : I(\hat{X}_B \rightarrow X_i) > I(\hat{X}_{B_{\text{max}}} \rightarrow X_i) - \frac{\log_2(n)}{2n} \right\}.
\]
Eq. (28) uses the property that if the true parent set is \( A \), then from over-fitting, \( I(\hat{X}_B \rightarrow X_i) \) would have value up to \( I(\hat{X}_B \rightarrow X_i) + (|B| - |A|) \log_2(n)/(2n) \). Though instead of a search over possible values of \( (|B| - |A|) \), for simplicity we only look at value one. Letting \( A \) denote the inferred parent set, we check that it is valid by testing
\[
I(\hat{X}_A \rightarrow X_i) > I(\hat{X}_{B_{\text{max}}} \rightarrow X_i) - \frac{(|B_{\text{max}}| - |\hat{A}|) \log_2(n)}{2n}.
\]
Otherwise, the difference cannot be explained through over-fitting and a mistake might have occurred, in which case we simply default to accepting \( B_{\text{max}} \) as the parent set. We also note that for \( m = 6 \), the in-degree bound for Algorithm 4 was set at \( K = 4 \) and for \( m = 15 \) it was set at \( K = 8 \).

2) Results: The results are shown in Fig. 4. Standard error bars are drawn. The algorithms all performed well. Algorithm 2 and 4 were the best, and their performances were almost identical. For \( m = 6 \), the proportions of dynamics preserved (25) by Algorithms 2, 3, and 4 respectively are \{99.6\%, 96.1\%, 99.6\%\}. The proportion of errors correctly identified as present or absent are \{94.2\%, 89.8\%, 94.1\%\} respectively. For \( m = 15 \), the proportion of dynamics preserved by Algorithms 2, 3, and 4 are \{99.3\%, 92.2\%, 99.2\%\}. The proportion of edges correctly identified are \{95.4\%, 91.0\%, 95.3\%\} respectively.

It is noteworthy that the increase in \( m \) does not result in any significant performance changes. Also, Algorithms 2 and 4 capture almost all the dynamics though make some mistakes on classifying edges. This suggests that the edges missed are weak influences.

B. Optimal Approximation – Algorithm 4

We also characterized the performance of the optimal approximation version of Algorithm 4, discussed in Section V-B, using AR network simulations.

1) Setup: The setup is similar to Section VII-A1. Networks of sizes \( m = 6 \) and \( m = 15 \) were simulated for \( n = 750 \) time-steps. Unlike Section VII-A1, the in-degrees were not constrained. Edges were picked i.i.d. with probability 1/2. The non-zero AR coefficients were drawn i.i.d. from a standard normal distribution, then scaled for stationarity, as in Section VII-A1. There were 200 trials for each \( m \). For each trial for \( m = 6 \), in-degree bounds \( K \in \{1, 2, 3, 4\} \) were used. For \( m = 15 \), bounds \( K \in \{2, 4, 6, 8, 10\} \) were used. The same performance measures in Section VII-A were used here.

2) Results: The results are shown in Fig. 5. For \( m = 6 \), the proportion of dynamics preserved by \( K \in \{1, 2, 3, 4\} \) respectively were \{70.3\%, 91.5\%, 98.3\%, 99.4\%\}. The proportion of edges correctly identified as present or absent were \{67.4\%, 76.6\%, 74.4\%, 64.0\%\}. For \( m = 15 \), the proportion of dynamics preserved by \( K \in \{2, 4, 6, 8, 10\} \) were \{65.8\%, 89.0\%, 97.0\%, 99.1\%, 99.6\%\} respectively. The proportion of edges correctly identified were \{64.2\%, 76.1\%, 81.3\%, 77.6\%, 69.6\%\}.

The proportion in dynamics kept by the approximation increases monotonically with the in-degree bound. Note, however, the percentage of edges correctly identified is concave. The peak is near the expected number of parents per node, 2.5 parents for \( m = 6 \) and 7 parents for \( m = 15 \). This optimal approximation algorithm does not try to remove edges which do not significantly contribute, though such variations could be done. Note, however, that for an optimal parent set \( B \), if \( l \) parents are removed, the resulting parent set is not necessarily the optimal set with \( |B| - l \) parents.

VIII. Social network analysis

We also demonstrated the utility of Algorithms 2, 3, and 4 by inferring which news sources influenced which users in the online micro-blogging network Twitter. All of the news sources covered major events in the Middle East during late 2013. By analyzing only the times of relevant posts from the news outlets and the users, the algorithms identified which news agency accounts influenced which user accounts with high precision.

A. Setup

1) Data: We analyzed activity on the micro-blogging platform Twitter. Accounts on Twitter can view content and post content. Blog posts are called “tweets” and are limited to 140 characters. Users specify which accounts to view tweets from, to “follow.” Note that this relationship is one-way. Each user
14

(a) The proportion of dynamics preserved for $m = 6$.

(b) The proportion of edges correctly identified for $m = 6$.

(c) The proportion of dynamics preserved for $m = 15$.

(d) The proportion of edges correctly identified for $m = 15$.

Fig. 5. These figures depict the performance for the optimal approximation version of Algorithm 4 using randomly generated networks of AR processes. There were 200 networks for $m = 6$ and $m = 15$. Performance was measured by the ratio between estimated and true parent sets for the sum of directed information from the parent sets to children (25). Also, the proportion of edges correctly identified as present or absent was computed.

Fig. 6. A snapshot of a user’s Twitter feed. The posts are in reverse chronological order from the top.

sees the tweets, both original and reposted tweets (“retweets”), of the accounts they follow. The interface is shown in Fig. 6. Tweets are in reverse-chronological order from the top. Users can retweet content, and it will become visible to all their followers, but the original poster’s name, not the reposter’s, will be shown. For instance, in Fig. 6, the user follows NBA, and sees a post by ATLHawks that was retweeted by NBA.

Table 1. The 16 news sources and their Twitter account handles analyzed. Three corporations, ABC, BBC, and Reuters, had multiple accounts which re-tweeted each other.

<table>
<thead>
<tr>
<th>News Agency</th>
<th>Twitter Account Handle</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC News</td>
<td>ABC</td>
</tr>
<tr>
<td>ABC World News</td>
<td>ABCWorldNews</td>
</tr>
<tr>
<td>Agence France-Presse</td>
<td>AFP</td>
</tr>
<tr>
<td>Al Jazeera English</td>
<td>AJEnglish</td>
</tr>
<tr>
<td>The Associated Press</td>
<td>AP</td>
</tr>
<tr>
<td>Al Arabiya English</td>
<td>AlArabia_Eng</td>
</tr>
<tr>
<td>BBC Breaking News</td>
<td>BBCBreaking</td>
</tr>
<tr>
<td>BBC News (World)</td>
<td>BBCWorld</td>
</tr>
<tr>
<td>Drudge Report</td>
<td>DRUDGE_REPORT</td>
</tr>
<tr>
<td>Fox News</td>
<td>FoxNews</td>
</tr>
<tr>
<td>The Jerusalem Post</td>
<td>Jerusalem_Post</td>
</tr>
<tr>
<td>NBC News</td>
<td>NBCNews</td>
</tr>
<tr>
<td>Reuters Top News</td>
<td>Reuters</td>
</tr>
<tr>
<td>Reuters World</td>
<td>ReutersWorld</td>
</tr>
<tr>
<td>CNN Breaking News</td>
<td>cnbrkr</td>
</tr>
<tr>
<td>The New York Times</td>
<td>nytimes</td>
</tr>
</tbody>
</table>

For data collection, 16 Twitter accounts corresponding to major news corporations were selected (see Table 1). All of these accounts published content in English. Major events in the Middle East were covered by most/all of the news corporations. Also, three corporations, ABC, BBC, and Reuters, had multiple accounts which re-tweeted each other. We did not especially mark retweeted content from the news sources.

A Python package\(^\text{14}\) was used to call the Twitter REST API\(^\text{15}\) to retrieve account time lines. For each tweet in the time lines of the news-source, if the tweet contained particular words, then the API was called to identify up to 100 users who retweeted that tweet. The keywords used were \{‘Syria’, ‘Strike’, ‘Assad’, ‘Chemical’, ‘Intervention’, ‘Iraq’, ‘Afghanistan’, ‘Iran’, ‘Terrorist’\}. A large pool of users were thus identified.

The API limited access to only the 3200 most recent tweets of an account. The earliest date when tweets from all of the news sources were available was October 10, 2013. A group of 48 users was picked who had at least five retweets with a keyword from one news source since that date. For the remainder of this section, we ignored all tweets of all accounts before that date. Data was collected until Dec. 10, 2013.

Fig. 7 shows a 48 hour snapshot of activity from two news sources and two user accounts. Tweets containing a relevant keyword are represented by long black lines. Other tweets are depicted with short green lines. User BoneToBone_ retweeted content from BBCBreaking, and user hrblock_21 retweeted content from FoxNews. Note the long periods of inactivity of user hrblock_21. Also note that the first tweets of BoneToBone_ and hrblock_21 containing a keyword were after BBCBreaking and FoxNews tweeted using the keywords.

2) Ground-truth: To assess the ground-truth for whether a news source X influences a particular user Y, the following

---

\(^{14}\)https://pypi.python.org/pypi/twitter

\(^{15}\)https://dev.twitter.com/docs/api
conditions were used. User $Y$ must have had at retweeted at least 5 tweets of $X$ that contained a keyword. Of all of $Y$’s tweets that contained a keyword, at least 15% must have been retweets from $X$. If both conditions were met, then $X$ was said to have influenced $Y$. The Twitter API included the identity of originating account for retweets, so it was unambiguous whether a user’s tweet was a retweet or not.

B. Modeling

We next discuss how the tweet histories were converted to time-series and the models used for analysis. Note that activity over communication networks such as Twitter are a form of point processes in terms of messages arriving over time. The values of time-series are positive and discrete, counting how many messages arrived in a given interval. Thus, linear autoregressive modeling, as specified by traditional Granger causality, is not appropriate for this type of data. We used a logistic model for how a user’s tweeting activity in the previous depended on the past. We first discuss how time was discretized.

Although tweets of accounts that user $Y$ follows are received continuously, $Y$ might access Twitter intermittently. For example, user $Y$ might receive a tweet of interest early in the morning but not log into Twitter and read it until the afternoon. Note that in Fig. 7, FoxNews tweets frequently, whereas user hrblock_21 has periods of inactivity.

We used the times that $Y$ tweets, with or without the keywords, to estimate when $Y$ was active. If $Y$ tweeted at a time $t$, then we modeled $Y$ as being active at least one minute before. To model how long $Y$ was active after sending a tweet, we used $Y$’s median retweet delay. See Fig. 8 for a diagram. Specifically, suppose $X$ posted a tweet with a keyword at time $t_1$, and the first tweet of $Y$ after $t_1$ was at time $t_2$, and $Y$ later retweeted $X$’s post at time $t_3$. We considered the delay $t_3 - t_2$. Thus, while $t_3 - t_1$ might have been large because $Y$ was logged out, the delay $t_3 - t_2$ could have been small. We modeled $Y$ as being active after a tweet for $Y$’s median delay\(^\text{16}\) plus three minutes. For each of $Y$’s tweets, an active period was thus calculated. Overlapping active periods were merged. See Fig. 9.

We next describe how each user $Y$’s activity was modeled, given news sources $X$ and $Z$. See Fig. 10 for a diagram. Active periods of $Y$ were divided up into intervals. The width of each interval was the median retweet time for user $Y$. Each period of time between active periods was also considered an interval. For $Y$, each interval during an active period was modeled as a binary variable, with value one if $Y$ had a tweet, possibly a retweet, during that interval which contained a keyword. That variable was conditioned on the tweets of $X$, $Y$, and $Z$ during the previous interval. Let $\bar{N}_X(\Delta t)$ denote the number of tweets containing a keyword that were sent by $X$ during the interval $\Delta t$. Define $\bar{N}_X(\Delta t)$ as

$$
\bar{N}_X(\Delta t) \triangleq \begin{cases} 
N_X(\Delta t) & \text{if } N_X(\Delta t) \leq 2, \\
1 + \lceil \log(1 + N_X(\Delta t)) \rceil & \text{o/w.}
\end{cases}
$$

We used $\bar{N}_X(\Delta t)$ to describe the past of $X$ during interval $\Delta t$. If $Y$ was inactive for a long period, then $X$ might have had many tweets with a keyword during that period. However, due to the Twitter feed interface having reverse chronological ordering, it would take greater effort for $Y$ to find older, unread tweets. Thus, $\bar{N}_X(\Delta t)$ was used as a simple method to capture this effect.

Let $\Delta t$ denote a time interval during which $Y$ was active. Let $\Delta t'$ denote the preceding time interval. We modeled $Y$’s tweeting activity during $\Delta t$ as depending on the past of $X$.

\(^{16}\)In our data-set, many users had short delays for most retweets. However, many also had a few large delays, which by the nature of the Twitter feed suggests they found these tweets through search or another means. Thus, to model regular activity, we used the median.
The entropy was then estimated as
\[
\hat{H}(Y \mid Z) = \frac{3 \log_2(n')}{2n'} > \hat{H}(Y \mid X, Z) + \frac{4 \log_2(n')}{2n'}
\]
\[
\hat{I}(X \to Y \mid Z) > \frac{\log_2(n')}{2n'}.
\]

This check for statistical significance was used for Algorithms 2, 3, and 4. For Algorithm 2, directed informations of the form \(I(X_{k} \to X_{i} \mid X_{m} \setminus \{i,k\})\) were computed for each ordered pair \((i, k)\). If \(I(X_{k} \to X_{i} \mid X_{m} \setminus \{i,k\}) > \frac{\log_2(n')}{2n'}\), where \(n'\) counted the periods that \(X_{i}\) was active, then the edge \(X_{k} \to X_{i}\) was accepted. For Algorithm 3, for each test in line 8, the threshold \(\frac{\log_2(n')}{2n'}\) was also used. In Algorithm 4, a maximum was taken. In line 4, the maximum is over directed informations \(I(X_{B} \to X_{i})\). In our implementation, we took the maximum over \(I(X_{B} \to X_{i}) - (|B| + 1) \frac{\log_2(n')}{2n'}\). This value described how informative conditioning on the past of \(X_{B}\) was above the amount expected from over-fitting. We then set \(B\) as the set of all \(B\)'s with values within 90% of that maximum.

The coefficients \(\alpha_{i}\) in the logistic model (29) could have been positive or negative. A positive coefficient for \(N_{X}(\Delta t)\) corresponded to \(Y\) having an increased likelihood of posting a tweet or retweet with a keyword, if \(X\) posted one or more tweets with keywords in the previous period. The MDL complexity threshold was used to distinguish between meaningful influences and cases of over-fitting. However, a negative coefficient was not interpretable and was likely due to over-fitting. In Algorithms 2, 3, and 4, any process \(X\) that had corresponding negative coefficient in the logistic model (29) was rejected. In Algorithm 4, if a set \(B\) had negative coefficients, those processes were removed from \(B\) and another fit on the remaining processes was performed. This was repeated until a subset of \(B\) with only positive coefficients was obtained.

### C. Estimation

Directed information estimates were computed using the consistent, parametric estimation technique proposed in [29]. To estimate the directed information \(I(X \to Y \mid Z)\), we first estimated two causally conditioned entropy terms, \(H(Y \mid Z)\) and \(H(Y \mid X, Z)\). For each entropy term, a logistic model of the form (29) was fit using generalized linear regression functions in Matlab’s Statistics toolbox. To estimate the entropy \(H(Y \mid X, Z)\), the likelihood function \(l_{Y}(\Delta t)\) was computed
\[
l_{Y}(\Delta t) \triangleq \begin{cases} P(N_{X}(\Delta t) > 0 | X, Y, Z) & \text{if } N_{X}(\Delta t) > 0, \\ P(N_{X}(\Delta t) = 0 | X, Y, Z) & \text{o/w}. \end{cases}
\]
The entropy was then estimated as
\[
\hat{H}(Y \mid X, Z) \triangleq \frac{1}{n'} \sum_{\Delta t} - \log_{2} l_{Y}(\Delta t),
\]
where the summation was over all periods \(\Delta t\) when \(Y\) was active, and \(n'\) was the number of such periods. The estimate \(\hat{H}(Y \mid Z)\) was computed in the same manner. The directed information estimate was then
\[
\hat{I}(X \to Y \mid Z) = \hat{H}(Y \mid Z) - \hat{H}(Y \mid X, Z).
\]

To assess whether a positive directed information value was statistically significant, we used minimum description length (MDL) complexity penalties [76]. For a parametric entropy estimate \(\hat{H}(Y \mid X, Z)\) with \(J\) parameters and \(n'\) observations, the MDL complexity is \(J \log_{2}(n')/(2n')\). The logistic model (29) involving \(J - 1\) processes has \(J\) parameters. Thus, the directed information estimate \(\hat{I}(X \to Y \mid Z)\) was considered significant if including \(X\) helps reduce the total complexity,
\[
\hat{H}(Y \mid Z) + \frac{3 \log_{2}(n')}{2n'} > \hat{H}(Y \mid X, Z) + \frac{4 \log_{2}(n')}{2n'}
\]
\[
\hat{I}(X \to Y \mid Z) > \frac{\log_{2}(n')}{2n'}.
\]

### D. Evaluation

We now describe criteria we used to evaluate the performance of the algorithms. As discussed in Section VIII-A2, for each news source \(X\) and user \(Y\), there was an established ground-truth of whether \(X\) influenced \(Y\). Each algorithm decided for each possible pair \((X, Y)\) whether there was an edge. Thus, for a given algorithm, each edge could have been true positive (TP), false positive (FP), true negative (TN), or false negative (FN). The following criteria were used to evaluate the performance [77].

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}
\]
\[
\text{Precision} = \frac{TP}{TP + FP}
\]
\[
\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN}
\]
\[
\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN}.
\]

Accuracy measures the proportion of edges labeled correctly. Precision measures the proportion of edges picked that were
correct. True positive rate measures the proportion of influences that were identified. False positive rate is the proportion of non-influences that were labeled incorrectly. We compared our methods against the expected performance of a baseline algorithm that knew how many news sources influenced each user $Y$, but guessed them uniformly at random.

E. Results

We now discuss the performance of Algorithms 2, 3, and 4 to identify the influential news sources for each user’s activity. Overall, the algorithms performed much better than baseline. Table 2 shows the values of the performance metrics. Fig. 11 depicts the algorithms’ performance on an ROC plot. The ground-truth graph was sparse, and the average in-degree of the users was $1.3 \pm 0.5$. The maximum possible in-degree was 16, one for each news source. The algorithms had approximately 95% accuracy. They correctly identified many potential influences as TN. The baseline also had high accuracy due to the sparsity. Note that the baseline algorithm knew each user’s in-degree. If a user had a single parent, even if the baseline guessed the wrong parent, 14 of the potential influences would have been correctly identified as TN.

The algorithms performed comparably, as can be seen in Fig. 11. Each had a very low FPR, meaning they were highly conservative. They selected few edges, but selected correctly. Hence they had low TPR and FPR, but high precision. There was some variation amongst them. Alg. 2 was the most conservative and Alg. 4 with $K = 3$ was the least. This is reflected in the monotonic decrease in precision and increase in FPR. Note that the differences in precision between the algorithms were larger than the increases in FPR because the ground-truth graph is sparse.

For some users, all of the algorithms correctly agreed on the influence. Fig. 12 depicts a subset of the network for which all of the users’ parents were selected by all the algorithms. Note that by construction, Alg. 4 with $K = 1$ made mistakes on any user with more than one parent. The majority of errors for the algorithms were false negatives, though there were cases of false positives. Fig. 13 depicts a subset of the network with users which some of the algorithms either did not pick a parent or picked the wrong parent. The ground truth is shown in Fig. 13(a). The networks inferred by the algorithms are shown in Fig. 13(b) through 13(e). Note, for instance, that for user balt_swag, only Alg. 2 selected a parent and did so correctly. For user Jacob Malin, Alg. 2 picked neither parent, Alg. 4 with $K = 1$ picked one, and the others correctly picked both.

IX. Conclusion and Future Directions

Methods that characterize which agents causally influence which other ones in a network could significantly bolster research in a number of diverse disciplines, including social sciences, economics, biology, and physics. We proposed a widely-applicable framework to address this issue. No assumptions on classes of processes were needed. The framework included meaningful graphical representations for networks of interacting agents, multiple algorithms to identify the underlying graph—in some cases using side information to improve efficiency, and procedures to estimate required statistics from data, as well as robust algorithms when the estimates were not reliable. We demonstrated the practical utility of the framework by identifying which news agencies influenced which users in the Twitter network with high precision.

There are a number of directions for future research. One is to improve estimation techniques of causally conditioned directed information. As discussed in this paper, there are already several estimation techniques [29], [30], [32], [69]. Computational feasibility of current methods is an aspect that will need to be further explored, especially for data-rich applications. Also, the performance of these estimators in the small-sample regime is not well characterized.

Another major avenue of future research involves time-varying graphical models and estimation procedures. The graphical models proposed here model dynamically interacting agents, but assumes the graph itself is time-invariant. Likewise, the estimation techniques rely on stationarity assumptions for how the future of processes depends on the past of others, in order to establish convergence. Especially for long-term studies of social networks, biological networks, and economic networks, graphical models and estimation techniques that can handle dynamic topologies would be greatly beneficial.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Acc.</th>
<th>Prec.</th>
<th>TPR</th>
<th>FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alg. 2</td>
<td>95</td>
<td>100</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>Alg. 3</td>
<td>94</td>
<td>81</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>Alg. 4 $K = 1$</td>
<td>94</td>
<td>74</td>
<td>46</td>
<td>1</td>
</tr>
<tr>
<td>Alg. 4 $K = 3$</td>
<td>94</td>
<td>70</td>
<td>51</td>
<td>2</td>
</tr>
<tr>
<td>Random</td>
<td>86</td>
<td>9</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 2: The performance of Alg. 2, 3, and 4 for identifying which of the 16 news sources influenced which of the 48 users is shown. Percentages are rounded to nearest integer. Alg. 3 was run with $K = 1$ and $K = 3$. For comparison, the expected performance of a baseline algorithm “Random” is shown. It selected parents uniformly at random; it knew how many parents each node had.
Fig. 12. This figure shows a subset of the network. The influences shown here were correctly identified by all of the algorithms.

APPENDIX A

INFORMATION DECOMPOSITION

This work assumes strict causality (see Assumption 1). Granger discussed that strict causality is a valid assumption if the sampling rate is high enough and all relevant processes are observed [2], [61]. Here we provide further motivation for why it is meaningful to analyze the causal interactions without accounting for the instantaneous correlations.

Under Assumption 1, \( P_X \) has the form \( \prod_{i=1}^{m} P_{X_i|X_{\{m\}\setminus\{i\}}} \). The KL divergence \( D(P_X||\prod_{i=1}^{m} P_{X_i|X_{\{m\}\setminus\{i\}}} ) \) quantifies how bad this assumption is. This divergence can be evaluated using total correlation. Total correlation, a generalization of mutual information, was introduced by Watanabe [78]. For a set of random variables \( \{ A, B, C, D \} \), it has the form

\[
I(A; B; C; D) \triangleq D(P_{A,B,C,D}\|P_A P_B P_C P_D).
\]

Eq. (32) factorizes over time, (33) follows from the definition of total correlation (30), and (34) uses the notation (31).

\[
\sum_{t=1}^{n} \bar{I}(X_1; \ldots; X_m, t | X_{t-1}^m).
\]

This is the sum over time of the total correlation of all the processes at time \( t \), conditioned on the full past. We now characterize how close \( P_X \) is to being strictly causal.

**Lemma A.1:**

\[
D\left( P_X \left| \prod_{i=1}^{m} P_{X_i|X_{\{m\}\setminus\{i\}}} \right) \right) = \bar{I}(X_1; \ldots; X_m).
\]

**Proof:**

\[
D\left( P_X \left| \prod_{i=1}^{m} P_{X_i|X_{\{m\}\setminus\{i\}}} \right) \right) = \mathbb{E}_{P_X} \left[ \log \prod_{i=1}^{m} P_{X_i|X_{\{m\}\setminus\{i\}}} \right] = \sum_{t=1}^{n} \mathbb{E}_{P_X} \left[ \log \prod_{i=1}^{m} P_{X_i|X_{\{m\}\setminus\{i\}}} \right] = \sum_{t=1}^{n} \bar{I}(X_1; \ldots; X_m, t | X_{t-1}^m) = \bar{I}(X_1; \ldots; X_m).
\]

Eq. (32) factorizes over time, (33) follows from the definition of total correlation (30), and (34) uses the notation (31).

**Corollary A.2:**

\[
D\left( P_X || \hat{P}_X \right) = \bar{I}(X_1; \ldots; X_m) + \sum_{i=1}^{m} D\left( P_{X_i|X_{\{m\}\setminus\{i\}}} || \hat{P}_{X_i|X_{\{m\}\setminus\{i\}}} \right).
\]

The following result shows that the divergence \( D\left( P_X || \hat{P}_X \right) \) decomposes into the sum of a common penalty for \( P_X \) violating strict causality and a second penalty measuring the quality of the approximation.
Proof:

\[
D\left( P_X \parallel P_X \right) = E_{P_X} \left[ \log \frac{P_X(X)}{P_X(X)} \right] = E_{P_X} \left[ \log \frac{P_X(X)}{\prod_{i=1}^{m} P_X(X_m)_{\{i\}}(X_i | X_m(\{i\})} \right) \]

(35)

Thus, by Lemma B.1, Eq. (47) implies (44).

We can now prove Lemma 3.3.

Proof: Suppose not. Let A and B be two, distinct minimal generative models for \( P_X \). Let \( Y \equiv X_i \) for any \( i \in [m] \) where \( A(i) \neq B(i) \). By (8),

\[
D\left( P_Y \parallel X_{m(\{i\})} P_Y | X_{m(\{i\})} \right) = 0 \quad \text{and} \quad D\left( P_Y \parallel X_{m(\{i\})} P_Y | B_{(i)} \right) = 0.
\]

Thus, by Lemma B.1,

\[
D\left( P_Y \parallel X_{m(\{i\})} P_Y | X_{k(\{i\})} \cap B_{(i)} \right) = 0.
\]

This is a contradiction, as \(|A(i) \cap B(i)| < |A(i)| \) but \(|A(i)| \) is minimal by definition.

Appendix C

Proof of Lemma 3.5

The proof is based on applying the chain rule in different ways, the non-negativity of directed information, and Lemma B.1.

Proof: Combining (8) in the definition of minimal generative models and (4) yields

\[
0 = I(X_{m(\{i\})} \to X_{i}) = I(X_{B(i)} \to X_{A(i)}) + I(X_{W(i)} \to X_{i}) + I(\{X_{m(\{i\})} \cap W(i) \to B(i)} \to X_{i} \cap X_{W(i)} \cap B(i)) \]

(48)

\[\text{A simple example with four variables} \quad \{X, Y, W, Z\} \text{ is as follows.} \]

If \( P_{X, Y, W, Z} \) is non-degenerate, and both \( I(X; Y | W, Z) = 0 \) and \( I(X; W | Y, Z) = 0 \) hold, then \( I(X; Y, W | Z) = 0 \).
where (48) uses the chain rule with $A(i) \subseteq B(i)$. Since directed information is non-negative, each term in (48) must be zero. Using the chain rule again,

$$I(X_{W(i)} \rightarrow B(i) \rightarrow X_i)$$

where (49) and (50) apply the chain rule in different ways and (51) uses that $I(X_{W(i)} \rightarrow X_i) = I(X_{B(i)} \rightarrow X_i)$ must be zero from (48).

Consequently, (49) and (51) imply that

$$I(X_{W(i)} \rightarrow X_i) \leq I(X_{B(i)} \rightarrow X_i).$$

This is the inequality (9) in Lemma 3.5. Equality occurs when $I(X_{B(i)} \rightarrow X_i) = 0$. Using the chain rule,

$$0 = I(X_{B(i)} \rightarrow X_i)$$

where (52) adds a zero term from (48), (53) uses the chain rule, and (54) uses (4). Since we also have that $I(X_{W(i)} \rightarrow X_i)$ must be zero from (48).

This is a contradiction because $I(A(i))$ is minimal. Therefore, there is no edge in the minimal generative model graph either. ■

**APPENDIX D**

**PROOF OF THEOREM 3.8**

*Proof:* We will show absence of an edge in one graph implies absence of an edge in the other. We consider two cases. Let $\{A(i)\}_{i=1}^m$ denote the parent sets in the minimal generative model, and let $Y \equiv X_i$ for some $i \in [m]$.

First suppose that there is no edge from $X_k$ to $Y$ in the minimal generative model graph, $k \notin A(i)$. Thus, where (55) uses (8) from the definition of minimal generative models, (56) uses the definition of causally conditioned directed information (4), (57) uses the chain rule with the fact that $k \notin A(i)$, and (58) implies that there is no edge from $X_k$ to $Y$ in the directed information graph either.

Now suppose that there is no edge from $X_k$ to $Y$ in the directed information graph. Then

$$0 = I(X_k \rightarrow Y)$$

where (59) is from Def. 3.7 and (60) is from (4). For the minimal generative model, by (8),

$$D\left(P_Y \| X_{W(\{i\})}, P_Y \| X_{A(\{i\})}, P_{X_{m(\{i\})}} \right) = 0.$$

Suppose $k \in A(i)$. By Lemma B.1, (60) and (61) together imply that

$$D\left(P_Y \| X_{m(\{i\})}, P_Y \| X_{A(\{i\})}, P_{X_{m(\{i\})}} \right) = 0.$$

This is a contradiction because $|A(i)|$ is minimal. Therefore, there is no edge in the minimal generative model graph either.

**APPENDIX E**

**PROOF OF THEOREM 6.1**

*Proof:* We first obtain concentrations on the empirical distribution from the Hoeffding and union bounds. We then use an $L_1$ bound on entropy to translate concentrations on entropies to ones on the directed information estimates.

We require the following concentrations on the empirical probability distributions. For every pair $(i,j)$, for every possible realization $\{x^i_j, x^{i+1}_j\} \in X^{2i+1}$, and, for a given $\epsilon > 0$ which we will later fix as a function of $\delta$,

$$|\hat{P}_{X^i_j, X^{i+1}_j}(x^i_j, x^{i+1}_j) - P_{X^i_j, X^{i+1}_j}(x^i_j, x^{i+1}_j)| < \epsilon$$

and

$$\|\hat{P}_{X^i_j, X^{i+1}_j}(x^i_j, x^{i+1}_j) - P_{X^i_j, X^{i+1}_j}(x^i_j, x^{i+1}_j)| < \epsilon$$

From the Hoeffding inequality generalized to uniformly ergodic Markov chains [70], under Assumption 4, for any $(i,j)$ and any realization $\{x^i_j, x^{i+1}_j\} \in X^{2i+1}$,

$$P \left( |\hat{P}_{X^i_j, X^{i+1}_j}(x^i_j, x^{i+1}_j) - P_{X^i_j, X^{i+1}_j}(x^i_j, x^{i+1}_j)| \geq \epsilon \right)$$

$$\leq 2 \exp \left( \frac{-n \epsilon^2}{2nd^2/\lambda^2} \right).$$

Applying the union bound to (63), the four inequalities in (62) hold for each of the $|X|^{2i+1}$ realizations for each of the $R$ pairs of processes $\{(i_r, j_r)\}$ with probability $\rho$, given in (19).

We next find the value of $\epsilon$ that corresponds to the event $B$. For simplicity, denote $\{X^i_j, X^{i+1}_j\}$ by $Z$. We want a concentration on $\|\hat{H}(Z) - H(Z)\|$. First note that

$$\|\hat{P}_Z - P_Z\|_1 \triangleq \sum_{Z \in X^{2i+1}} |\hat{P}_Z - P_Z| \leq |X|^{2i+1} \epsilon,$$

where (64) follows from (62).

Using an $L_1$ bound on entropy, if $\|\hat{P}_Z - P_Z\|_1 \leq \frac{1}{2}$, then

$$\|\hat{H}(Z) - H(Z)\| \leq -\|\hat{P}_Z - P_Z\|_1 \frac{\|\hat{P}_Z - P_Z\|_1}{|X|^{2i+1}}.$$

The bound is of the form $-b \log b^{-1}$, which is concave in $b$ and maximized at $b = c/e$. With $\epsilon \leq 1/e$, the upper bound in (64), $|X|^{2i+1} \epsilon$, is in the interval $(0, |X|^{2i+1}/e]$ where the bound (65) is increasing. Thus, (65) can be bounded using (64)

$$\|\hat{H}(Z) - H(Z)\| \leq -|X|^{2i+1} \epsilon \log |X|^{2i+1} \epsilon$$

$$= -|X|^{2i+1} \epsilon \log |X|^{2i+1} \epsilon.$$
Note that the directed information (16) decomposes into a linear combination of entropies,
\[
I(X_{i},X_{j}; X_{i}^{j}|X_{i}^{j}) = H(X_{i}^{j+1}) - H(X_{i}^{j}) \\
- H(X_{i}^{j+1}, X_{i}^{j}) + H(X_{i}^{j}, X_{i}^{j}).
\]
Applying the triangle inequality to (67) with (66) gives that for all \( R \) pairs \((i_{r}, j_{r})\),
\[
\tilde{I}(X_{j_{r}} \to X_{i_{r}}) - I(X_{j_{r}} \to X_{i_{r}}) \leq -4|X|^{2l+1} \epsilon \log \epsilon.
\]
Setting \( \delta = -4|X|^{2l+1} \epsilon \log \epsilon \) would conclude the proof. However, to obtain an analytic expression for how \( \epsilon \) depends on \( \delta \), we will bound \( \epsilon \log \epsilon \) with a polynomial expression. The function \( -\epsilon \log \epsilon \) has a maximum value of \( 1/\epsilon \) on the interval \( \epsilon \in (0,1) \). That value is attained at \( \epsilon = 1/\epsilon \). For all \( 0 < \epsilon < 1 \),
\[
-\epsilon \log \epsilon = \frac{1}{\epsilon} \left( -\epsilon \log \epsilon \right) \epsilon^{-1-a} \leq \frac{1}{ae} \epsilon^{-1-a}.
\]
For large \( \epsilon \), the bound with larger \( \epsilon \) is tighter; for small \( \epsilon \), the bound with smaller \( \epsilon \) is tighter. For all \( 0 < \epsilon < 1 \) and all \( r \),
\[
|\tilde{I}(X_{j_{r}} \to X_{i_{r}}) - I(X_{j_{r}} \to X_{i_{r}})| \leq \frac{4|X|^{2l+1}}{\epsilon e} \epsilon^{-1-a}.
\]
Setting the value of \( \epsilon \) as \( \epsilon = \left( \frac{ae}{4|X|^{2l+1}} \right)^{1/\delta} \) finishes the proof that \( P(B_{\delta}) \geq 1 - \rho \).

Note that for a fixed probability of error \( \rho \) (19), fixed \( m \), and sufficiently large \( n \), that as \( n \) increases, \( \epsilon \) decays as \( n^{-1/2} \) which implies that \( \delta = O(n^{-1/2+\epsilon'}) \) for all \( \epsilon' > 0 \). Alternatively, if \( m \) is increasing (and thus \( R \) in (19)), to maintain a fixed probability of error \( \rho \) with a fixed \( \delta \), \( n \) needs to increase as \( \log m \).

\[\text{APPENDIX F} \]
\[\text{PROOF OF THEOREM 6.4}\]

Proof: We first find the graph sample complexity. We will lower bound \( P(B_{\delta}) \). Note that the equiprobable contours of \( N(0, \Sigma') \) form ellipsoids with principal axis lengths proportional to the largest eigenvalue of \( \Sigma' \) (see pg. 108 of [80]). Let \( \sigma'' \) denote the largest eigenvalue of \( \Sigma' \). Define a new diagonal covariance matrix \( \Sigma'' \) whose entries are all \( \sigma'' \). Then the probability of any volume centered at zero under \( N(0, \Sigma'') \) will be larger than \( N(0, \Sigma') \). Also, since \( \Sigma'' \) is diagonal, the corresponding random variables are independent. Thus,
\[
\mathbb{P}(B_{\delta}) = \mathbb{P} \left( \left\{ -\delta \leq g_{r}(\tilde{\theta}_{\delta}) - g_{r}(\tilde{\theta}'') \leq \delta \right\} \right) \]
\[
= \mathbb{P} \left( \left\{ -\delta \leq \sqrt{\frac{n}{\sigma''}} (g_{r}(\tilde{\theta}_{\delta}) - g_{r}(\tilde{\theta}'')) \leq \delta \right\} \right) \]
\[
\geq \mathbb{P} \left( \left\{ -\delta \leq \frac{\sqrt{n}}{\sigma''} (g_{r}(\tilde{\theta}_{\delta}) - g_{r}(\tilde{\theta}'')) \leq \delta \right\} \right) \]
\[
= \left[ \text{erf} \left( \frac{\sqrt{n}}{\sqrt{2\sigma''}} \right) \right]^{R}.
\]
(70)

where (68) uses (22) so the distribution is \( N(0, \Sigma'') \), (69) uses the independence of the error estimates under distribution \( N(0, \Sigma'') \) and normalizes them, and (70) uses the “error” function \( \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt \). Using the first two terms of the asymptotic expansion of \( \text{erf}(x) \) with appropriate constants \( c_{1} \) and \( c_{2} \),
\[
\left[ \text{erf} \left( \frac{\sqrt{n}}{\sqrt{2\sigma''}} \right) \right]^{R} \approx \left[ 1 - \frac{c_{1}}{\sqrt{n}} e^{-c_{2}n} \right]^{R}
\]
\[
\approx 1 - \frac{m^{2}c_{1}}{n} e^{-c_{2}n}
\]
\[
= 1 - c_{1} e^{2\log(m)-c_{2}n-\frac{1}{2} \log n}.
\]
(71)

Eq. (71) uses the first two terms in the binomial expansion. Repeat these steps setting \( \sigma'' \) as the minimum eigenvalue of \( \Sigma' \) to get an upper bound in (69). Eq. (72) will have the same form. This finishes the proof for the graph sample complexity.

For the sample complexity rate, note the \( \delta \sqrt{n} \) terms in the normalized inequalities in (69). Thus \( \delta = O(n^{-1/2}) \).

\[\text{APPENDIX G} \]
\[\text{PROOF OF THEOREM 6.5}\]

Proof: We first show that the parent sets of \( \hat{P}_{\text{rob}} \) can be identified independently.
\[
\min_{\hat{P}_{X} \in \mathcal{P}_{K}} \max_{s \in S} \left[ W(\hat{P}_{X}(s), s) - W(\hat{P}_{X}) \right]
\]
\[
= \min_{\hat{P}_{X} \in \mathcal{P}_{K}} \max_{s \in S} \left[ \sum_{i=1}^{m} \tilde{I}_{s}(X_{i} | \hat{A}^{*}(i) \to X_{i}) \right] \]
\[
= \min_{\hat{P}_{X} \in \mathcal{P}_{K}} \sum_{i=1}^{m} \max_{s \in S} \left[ \tilde{I}_{s}(X_{i} | \hat{A}^{*}(i) \to X_{i}) \right]
\]
\[
= \min_{\hat{P}_{X} \in \mathcal{P}_{K}} \sum_{i=1}^{m} \max_{s \in S} \left[ \tilde{I}_{s}(X_{i} | \hat{A}^{*}(i) \to X_{i}) \right]
\]
\[
= \min_{\hat{P}_{X} \in \mathcal{P}_{K}} \sum_{i=1}^{m} \max_{s \in S} \left[ \tilde{I}_{s}(X_{i} | \hat{A}^{*}(i) \to X_{i}) \right]
\]
\[
\text{where } \{ \hat{A}^{*}(i) \}_{i=1}^{m} \text{ in (73) are the parent sets belonging to } \hat{P}_{X}(s) \text{ for the maximizing } s \in S, (74) \text{ brings the max inside, and (75) uses the property from Theorem 5.2 that finding parents for processes can be done independently for any particular scenario } s \in S. \]

If (74) holds with equality, then the parent sets of \( \hat{P}_{\text{rob}} \) can be identified independently. Note that the first \( (m-1) \) coordinates of \( S \) correspond to estimates of the form \( \{ \tilde{I}_{s}(X_{i} | \hat{A}^{*}(1) \to X_{i}) \} \) for the \( (m-1) \) choices of \( \hat{A}(1) \). The next \( (m-1) \) coordinates correspond to estimates of the form \( \{ \tilde{I}_{s}(X_{i} | \hat{A}(2) \to X_{j}) \} \), and so on. Thus, for a given \( i \in [m] \), the maximization in (74) is only over the \( i \)th set of \( (m-1) \) coordinates. Since \( S \) is rectangular, the values of the other coordinates are irrelevant. Since each of the \( m \) terms in the sum in (74) are optimizing over disjoint sets of coordinates, (74) holds with equality.

We next show that Algorithm 5 returns the individually most robust parent sets. Consider identifying robust parents for \( X_{i} \). Using the notation \( B_{j} \) in Algorithm 5, the worst case regret
for parent set $B_j$ is

$$R(B_j) \triangleq \max_{s \in S} \max_{j' \neq j} \left\{ \left( \hat{I}_s(\mathbf{X}_{B_j} \rightarrow \mathbf{X}_i) - \hat{I}_s(\mathbf{X}_{B_j} \rightarrow \mathbf{X}_i) \right) \right\} \quad (76)$$

$$\leq \max_{s \in S} \left\{ \max_{j' \neq j} \hat{I}_s(\mathbf{X}_{B_j} \rightarrow \mathbf{X}_i) \right\} \quad (77)$$

$$= \max\{0, \max_{j' \neq j} H(j') - L(B_j)\}. \quad (78)$$

The zero in (76) is for the case that there is a $j$ such that $L(B_j) > H(j')$ for all $j' \neq j$. Eq. (77) applies the max to individual terms and (78) follows from lines 6 and 8 in Algorithm 5. Since $S$ is rectangular, for any set of values $\{\hat{I}_s(\mathbf{X}_{B_j} \rightarrow \mathbf{X}_i) \}_{j=1}^{\infty}$, there exist $s \in S$ with those values. Thus, (77) holds with equality.

Note that by lines 9 and 11 in Algorithm 5,

$$R(B_j) = \begin{cases} H(B_{j1}) - L(B_j) & \text{if } j \neq j_1, \\ \max\{0, H(B_{j1}) - L(B_j)\} & \text{if } j = j_1. \end{cases} \quad (79)$$

If $j_1 = j_2$, then for all $j \neq j_1$, $H(B_{j1}) > H(B_j)$ and $L(B_{j1}) \geq L(B_j)$, so by (79),

$$R(B_{j1}) = H(B_{j1}) - L(B_{j1}) \leq H(B_{j1}) - L(B_j) \leq H(B_j) - L(B_j) = R(B_j).$$

Thus, if $j_1 = j_2$, then $B_{j1}$ is the most robust parent set.

Next consider the case that $j_1 \neq j_2$. By (79),

$$\min_{j \neq j_1} R(B_j) = \min_{j \neq j_1} [H(B_{j1}) - L(B_j)] = H(B_{j1}) - \max_{j \neq j_1} L(B_j) = H(B_{j1}) - L(B_{j2}) = R(B_{j2}).$$

Thus, if $j_1 \neq j_2$, then either $B_{j1}$ or $B_{j2}$ would be most robust. The parent set $B_{j1}$ is selected if $j_1 = j_2$ or if

$$R(B_{j1}) \leq R(B_{j2}) \quad (80)$$

$$H(B_{j1}) - L(B_{j2}) \leq H(B_{j1}) - L(B_{j2}) \quad (81)$$

$$\frac{1}{2}(H(B_{j1}) + L(B_{j2})) \leq M(B_{j1}). \quad (82)$$

Eq. (80) uses (79), (81) adds $L(B_{j2}) + L(B_{j1})$ to both sides, and (82) uses the property that $M(B_j) = \frac{1}{2}(H(B_j) + L(B_j))$. This result shows that for each node $X_i$, Algorithm 5 will return the individually most robust parents and thus $P_{rob}$. 

**References**


