Causal Network Inference by Optimal Causation Entropy[*]

Jie Sun[†], Dane Taylor[‡], and Erik M. Bollt[†]

Abstract. The broad abundance of time series data, which is in sharp contrast to limited knowledge of the underlying network dynamic processes that produce such observations, calls for a rigorous and efficient method of causal network inference. Here we develop mathematical theory of causation entropy, an information-theoretic statistic designed for model-free causality inference. For stationary Markov processes, we prove that for a given node in the network, its causal parents form the *minimal set of nodes that maximizes causation entropy*, a result we refer to as the *optimal causation entropy principle*. Furthermore, this principle guides us in developing computational and data efficient algorithms for causal network inference based on a two-step discovery and removal algorithm for time series data for a network-coupled dynamical system. Validation in terms of analytical and numerical results for Gaussian processes on large random networks highlights that inference by our algorithm outperforms previous leading methods, including conditional Granger causality and transfer entropy. Interestingly, our numerical results suggest that the number of samples required for accurate inference depends strongly on network characteristics such as the density of links and information diffusion rate and not necessarily on the number of nodes.
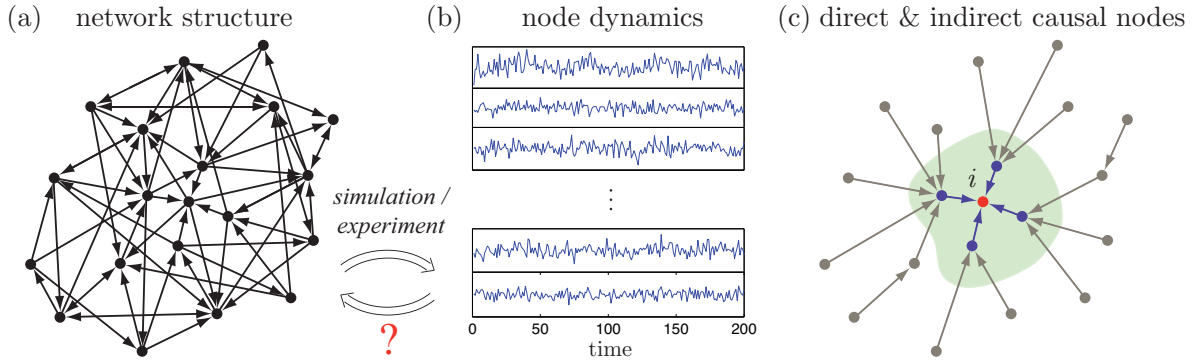
1. Introduction. Research of dynamic processes on large-scale complex networks has attracted considerable interest in recent years with exciting developments in a wide range of disciplines in social, scientific, engineering, and medical fields [48, 49, 74]. One important line of research focuses on exploring the role of network structure in determining the dynamic properties of a system [6, 17, 18, 19, 27, 55, 67, 79] and utilizing such knowledge in controlling network dynamics [15, 70] and optimizing network performance [13, 38, 50, 56, 72]. In applications such as the study of neuronal connectivity or gene interactions, it is nearly impossible to directly identify the network structure without severely interfering with the underlying system, whereas time series measurements of the individual node states are often more accessible [68]. From this perspective, it is crucial to reliably infer the network structure that shapes the dynamics of a system from time series data. It is essential to account for directed "cause and effect" relationships, which often offer deeper insight than nondirected relationships (e.g., correlations) [53, 62, 66]. In particular, causal network inference is consid-

(a)   network structure   (b)   node dynamics   (c) direct & indirect causal nodes

Figure 1. *Network dynamics, time series, and the causal network inference problem. Modern scientific approaches such as simulation, experiments, and data mining have produced an abundance of high-dimensional time series data describing dynamic processes on complex networks (a→b). Given empirical observations, an important problem is to infer the causal network structure that underlies the observed time series. As shown in (c), for each node i, the goal is to identify its "causal parents," the nodes that directly influence its dynamics (nodes in shaded region), while pruning away the nodes that do not (nodes outside the shaded region), thus recovering the direct links to node i in the causal network. The key to efficiently and accurately differentiating direct causal links from noncausal links is to follow an algorithm involving tests for independence via judiciously selected conditioning sets. The main goal of this paper is to develop and validate such algorithms for stationary Markov processes.*

ered a central problem in the research of social perception [35], epidemiological factors [57], neural connectivity [11, 12], economic impacts [34], and basic physical relationships of climatological events [60, 61]. Evidently, understanding causality is a necessary and important precursory step toward the goal of effectively controlling and optimizing system dynamics (e.g., medical intervention of biological processes and policy design for economic growth and social development).

In a network dynamic process involving a large number of nodes, causal relationships are inherently difficult to infer. For example, the fact that a single node can potentially be influenced by many (if not all) others through network interactions makes it challenging to untangle the direct causal links from indirect and erroneous links (see Figure 1 for illustration). Granger recognized the crucial role played by conditioning and defines a causal relationship based on two basic principles [29, 30]:

(i) The cause should occur before the effect.

(ii) The cause should contain information about the caused that is not available otherwise. A relationship that fulfills both requirements is unambiguously defined as causal. In practice, although the first requirement is straightforward to examine when temporal ordering of the data is available, it is difficult to check the second as it involves the consideration of *all* available information (time series data from all variables). Tradeoffs are often made, by either restricting to small-scale networks with no time delay and just a handful of variables [32, 71], or partially removing the second requirement and therefore reducing the accuracy of network inference [76]. Inferring large-scale networks from time series data remains a relatively open problem [41, 68].

The classical Granger causality test was designed for linear regression models [29, 30],

although several extensions have been proposed to nonlinear models, including local linear approximations [14] and partial functional expansion via radial basis functions [2]. Information-based causality inference measures represent a systematic way of overcoming the model-dependent limitation in the linear Granger causality test. In particular, Schreiber proposed transfer entropy as a measure of information flow, or effective coupling, between two processes regardless of the actual functional relationship between them [63]. The transfer entropy from process Y to process X measures the uncertainty reduction of the future states of X as a result of knowing the past of Y given that the past of X is already known, and it is essentially the mutual information between the future of X and history of Y *conditioning* on the history of X [37, 51]. Because of its ability to associate temporal and spatial directionality with coupling, transfer entropy has quickly started to gain popularity in a broad range of disciplines, including bioinformatics, neuroscience, and climatology, as a tool to infer effective pairwise coupling that underlies complex dynamic processes [9, 76]. However, transfer entropy, which was introduced specifically for detecting the directionality of information flow between two processes, has fundamental limitations when applied in a multivariate setting to the inference of networks [65, 71]. In particular, without proper conditioning, inference based on transfer entropy tends to produce systematic errors due to, for example, the effects of indirect influences and dominance of neighbors [71]. As shown in Figure 1(c), the main purpose of this work is to identify for each node i its "causal parents" that directly influence node i, while not falsely inferring indirect (i.e., noncausal) nodes.

Proper conditioning can distinguish between direct and indirect causal relationships, and it is thus unsurprising that conditioning is widely adopted as a key ingredient in many network inference methods [21, 32, 37, 51, 60, 61, 65, 66, 71]; however, even within such a general theme, the inference of networks requires a theoretically sound approach that is also algorithmically reliable and efficient. For example, one must develop a strategy for choosing which potential links to examine and which nodes to condition on. Thus we note two essential steps in causal network inference: (1) adopting a statistic for the inference of a causal relationship, and (2) developing an algorithm that iteratively employs step (1) to learn the causal network. Whereas accuracy, tractability, and generality of the chosen statistic are often the priority for (1), various challenges arise regarding (2). In particular, these often include minimizing the computational cost by reducing the number of statistics that needs to be computed, as well as reducing the error incurred by finite-sized data by keeping the size of the conditioning sets (i.e., the dimension of the estimation problem) as small as possible. In general, the inaccuracy when estimating statistical measures from finite data grows rapidly with dimensionality, making the dimensionality of the problem a priority for any networks containing more than a couple nodes.

One approach for network inference is to test each candidate causal link conditioned on all other variables [32]. That is, a directed link $j \to i$ is inferred if such a relationship remains effective when conditioning on all other variables in the system. Although intuitive and correct in theory, this method requires computing a statistic in a sample space as high dimensional as the entire system and therefore falls short when applied to a large network. The PC algorithm [66] overcomes this difficulty by repeated testing of the candidate causal link conditioned on subsets of the remaining variables [60, 61]. To be more specific, a link $j \to i$ is disqualified as a candidate causal relationship if it is insignificant when conditioned on some subset of the nodes. The advantage of the PC algorithm is that it reduces the dimensionality of

the sample space in the test of independence to be proportional to the size of the conditioning set (which in some cases can be much smaller than the system size). However, unless the maximum degree of the nodes is known a priori, the algorithm in principle needs to search over combinations of subsets (as conditioning sets) up to the size of the entire network. In this respect, regardless of the dimensionality of the sample space, the combinatorial search itself can be computationally infeasible for moderate to large networks. In practice, a tradeoff needs to be made between an algorithm's computational cost and data efficiency (in terms of the estimation of the test statistic).

In this paper we develop theory of causation entropy—a type of conditional mutual information designed for causal network inference. In particular, we prove the optimal causation entropy principle for Markov processes: the set of nodes that directly cause a given node is the unique minimal set of nodes that maximizes causation entropy. This principle allows us to convert the problem of causality inference into the optimization of causation entropy. We further show that this optimization problem, which appears to be combinatorial, can in fact be solved by simple greedy algorithms, which are both computational efficient and data efficient. We verify the effectiveness of the proposed algorithms through analytical and numerical investigations of Gaussian processes on various network types, including trees, loops, and random networks. Somewhat surprisingly, our results suggest that it is the density of links and the information diffusion rate rather than the number of nodes in a network that determine the minimal sample size required for accurate inference.

2. Stochastic process and causal network inference. We begin by introducing a theoretical framework for inferring causal networks from high dimensional time series. This framework is general in that it is applicable to both linear and nonlinear systems.

Consider a network (graph) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = \{1, 2, \ldots, n\}$ being the set of nodes and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V} \times \mathbb{R}$ being the set of weighted links (or edges). The *adjacency matrix* $A = [A_{ij}]_{n \times n}$ is defined as

$$(2.1) \qquad A_{ij} = \begin{cases} \text{weight of the link } j \to i & \text{if } j \to i \text{ in the network,} \\ 0 & \text{otherwise.} \end{cases}$$

We use $\chi_0(A)$ to denote the corresponding unweighted adjacency matrix defined entrywise by $\chi_0(A)_{ij} = 1$ iff $A_{ij} \neq 0$ and $\chi_0(A)_{ij} = 0$ iff $A_{ij} = 0$. We define the set of *causal parents* of i as

$$(2.2) \qquad N_i = \{j | A_{ij} \neq 0\} = \{j | \chi_0(A)_{ij} = 1\}.$$

For a subset of nodes $I \subset \mathcal{V}$, we similarly define its set of causal parents as

$$(2.3) \qquad N_I = \cup_{i \in I} N_i.$$

We consider stochastic network dynamics in the following form (for each node i):

$$(2.4) \qquad X_t^{(i)} = f_i\big(A_{i1}X_{t-1}^{(1)}, A_{i2}X_{t-1}^{(2)}, \ldots, A_{ij}X_{t-1}^{(j)}, \ldots, A_{in}X_{t-1}^{(n)}, \xi_t^{(i)}\big),$$

where $X_t^{(i)} \in \mathbb{R}^d$ is a random variable representing the state of node i at time t, $\xi_t^{(i)} \in \mathbb{R}^d$ is the random fluctuation (noise) on node i at time t, and $f_i : \mathbb{R}^{d \times (n+1)} \to \mathbb{R}^d$ models the

functional dependence of the state of node i on the past states of nodes j with $A_{ij} \neq 0$. Note that other than the noise term $\xi_t^{(i)}$, the state $X_t^{(i)}$ only depends (stochastically) on the past states of its causal parents, $X_{t-1}^{(j)}$ ($j \in N_i$).

For a subset $K = \{k_1, k_2, \ldots, k_q\} \subset \mathcal{V}$, we define

$$(2.5) \qquad X_t^{(K)} \equiv [X_t^{(k_1)}, X_t^{(k_2)}, \ldots, X_t^{(k_q)}]^\top.$$

If $K = \mathcal{V}$, we simplify the notation and denote

$$(2.6) \qquad X_t \equiv X_t^{(\mathcal{V})} = [X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(n)}]^\top.$$

2.1. Problem of causal network inference and challenges. Given quantitative observations of the dynamic states of individual nodes, often in the form of time series, a central problem is to infer its (causal) system dynamics, which involves the inference of (1) the causal network topology, $\chi_0(A)$; (2) the link weights, $\{A_{ij}\}$; and (3) the specific forms of functional dependence between nodes, $\{f_i\}$. These problems are interrelated and all challenging. We focus on the first problem: inferring the causal network topology $\chi_0(A)$, which serves as the skeleton of the actual network dynamics. See Figure 1 as a schematic illustration. In particular, the problem of causal network inference can be cast mathematically as

$$(2.7) \qquad \begin{cases} \text{Given:} & \text{Samples of the node states } x_t^{(i)} \ (i = 1, 2, \ldots, n; \ t = 1, 2, \ldots, T). \\ \text{Goal:} & \text{Infer the structure of the underlying causal network,} \\ & \text{i.e., find } \operatorname{argmin}_{\hat{A}} \|\chi_0(A) - \hat{A}\|_0, \text{ where } \|M\|_0 \equiv \sum_{i,j} |M_{ij}|^0. \end{cases}$$

One key challenge is that in many applications, the number of nodes n is often *large* (usually hundreds at least), but the sample size T is much smaller than needed for reliable estimation of the $(n \times d)$-dimensional joint distribution. We propose that a practical causation inference method should fulfill the following three requirements:

1. *Model-free.* The method should not rely on assumptions about either the form or parameters of a model that underlie the process.
2. *Computational efficient.* The method should be computationally efficient.
3. *Data efficient.* The method should achieve high accuracy with a relatively small number of samples (i.e., convergence in probability needs to be fast).

In this paper we address the model-free requirement by utilizing information-theoretic measures and, in particular, by using causation entropy. On the other hand, our theoretical development of the optimal causation entropy principle enables us to develop algorithms that are both computationally efficient and data efficient.

2.2. Markov assumptions. We study the system in a probabilistic framework assuming stationarity and existence of a continuous distribution. We further make the following assumptions regarding the conditional distributions $p(\cdot|\cdot)$ arising from the stationary process

given by (2.4). For every node $i \in \mathcal{V}$ and time indices t, t',

$$
(2.8) \quad
\begin{cases}
(1) \ \textit{Temporally Markov}: \\
\qquad p(X_t|X_{t-1}, X_{t-2}, \dots) = p(X_t|X_{t-1}) = p(X_{t'}|X_{t'-1}). \\
(2) \ \textit{Spatially Markov}: \\
\qquad p(X_t^{(i)}|X_{t-1}) = p(X_t^{(i)}|X_{t-1}^{(N_i)}). \\
(3) \ \textit{Faithfully Markov}: \\
\qquad p(X_t^{(i)}|X_{t-1}^{(K)}) \neq p(X_t^{(i)}|X_{t-1}^{(L)}) \ \text{whenever} \ (K \cap N_i) \neq (L \cap N_i).
\end{cases}
$$

Throughout the paper, the relationship between two probability density functions p_1 and p_2 is denoted as "$p_1 = p_2$" iff they equal *almost everywhere* and "$p_1 \neq p_2$" iff there is a set of positive measure on which the two functions do not equal.

In (2.8), condition (1) states that the underlying dynamics is a time-invariant Markov process.[1] Condition (2) is often referred to as the (local) Markov property [44], which we call spatially Markov here to differ from temporally Markov. This condition guarantees that in determining the future state of a node, if knowledge about the past states of all its causal parents N_i (as defined in (2.2)) is given, information about the past of any other node becomes irrelevant. Finally, condition (3) ensures that the set of causal parents is unique and that every causal parent presents an observable effect regardless of the information about other causal parents.[2]

The conditional independence between two random variables X and Y given Z is denoted by $(X \perp\!\!\!\perp Y \mid Z)$, i.e.,

$$
(2.9) \qquad\qquad (X \perp\!\!\!\perp Y \mid Z) \iff p(X|Y,Z) = p(X|Z).
$$

The following results regarding conditional independence will be useful in later sections and are direct consequences of the basic axioms of probability theory [31, 44, 53]:

$$
(2.10) \quad
\begin{cases}
\text{Symmetry: } (X \perp\!\!\!\perp Y \mid Z) \iff (Y \perp\!\!\!\perp X \mid Z). \\
\text{Decomposition: } (X \perp\!\!\!\perp YW \mid Z) \implies (X \perp\!\!\!\perp Y \mid Z). \\
\text{Weak union: } (X \perp\!\!\!\perp YW \mid Z) \implies (X \perp\!\!\!\perp Y \mid ZW). \\
\text{Contraction: } (X \perp\!\!\!\perp Y \mid Z) \wedge (X \perp\!\!\!\perp W \mid ZY) \implies (X \perp\!\!\!\perp YW \mid Z). \\
\text{Intersection: } (X \perp\!\!\!\perp Y \mid ZW) \wedge (X \perp\!\!\!\perp W \mid ZY) \implies (X \perp\!\!\!\perp YW \mid Z).
\end{cases}
$$

Here "$\wedge$" denotes the logical operation "and" (the symbol "$\vee$" is used later for "or"), and YW denotes a joint random variable of Y and W.

---

[1]If the process is Markov but with higher order, our approach is to convert it into a first-order process as illustrated in Appendix A and then apply the theory and algorithms in the main body of the paper to the resulting first-order process.

[2]Note that without condition (3), the "true positive" statement in Theorem 2.2 is no longer valid. One simple example is given in Appendix B to illustrate this point.

2.3. Causation entropy as an information-theoretic measure. We review several fundamental concepts in information theory, leading to causation entropy, a model-free information-theoretic statistic that can be used to infer direct causal relationships [71].

Originally proposed by Shannon as a measure of uncertainty and complexity, the (differential) *entropy* of a continuous random variable $X \in \mathbb{R}^n$ is defined as [16, 64][3]

$$(2.11) \qquad h(X) = -\int p(x) \log p(x) dx,$$

where $p(x)$ is the probability density function of X. The joint and conditional entropies between two random variables X and Y are defined as (also see Figure 2(a))

$$(2.12) \qquad \begin{cases} \text{Joint entropy: } h(X,Y) \equiv h(Y,X) \equiv -\int p(x,y) \log p(x,y) dx dy. \\ \text{Conditional entropies: } \begin{cases} h(X|Y) \equiv -\int p(x,y) \log p(x|y) dx dy, \\ h(Y|X) \equiv -\int p(x,y) \log p(y|x) dx dy. \end{cases} \end{cases}$$

For more than two random variables, the entropies are similarly defined (as above) by grouping the variables into two classes, one acting as X and the other as Y.

The *mutual information* between two random variables X and Y (conditioning on Z) can be interpreted as a measure of the deviation from independence between X and Y (conditioning on Z). The corresponding unconditioned and conditional mutual information are defined respectively as

$$(2.13) \qquad \begin{cases} \text{Mutual information: } I(X;Y) \equiv h(X) - h(X|Y) \equiv h(Y) - h(Y|X). \\ \text{Conditional mutual information:} \\ \qquad I(X;Y|Z) \equiv h(X|Z) - h(X|Y,Z) \equiv h(Y|Z) - h(Y|X,Z). \end{cases}$$

The mutual information among three variables X, Y, and Z is defined as[4]

$$(2.14) \quad I(X;Y;Z) \equiv I(X;Y) - I(X;Y|Z) \equiv I(Y;Z) - I(Y;Z|X) \equiv I(X;Z) - I(X;Z|Y).$$

The mutual information between two variables is always nonnegative; $I(X;Y) \geq 0$ with equality iff X and Y are independent. Similarly, $I(X;Y|Z) \geq 0$ with equality iff X and Y are independent when conditioned on Z. Interestingly, for three or more variables, such an inequality does not hold: the mutual information $I(X;Y;Z)$ can be either positive, negative, or zero [47]. Figure 2(a) visualizes the relationships between entropy, conditional entropy, and mutual information.

To measure the directionality of information flow between two random processes, Schreiber proposed a specific type of conditional mutual information called *transfer entropy* [63]. For a

---

[3]We follow the convention in [16] to use $h(\cdot)$ for the entropy of a continuous random variable and reserve $H(\cdot)$ for the entropy of a discrete random variable. In the discrete case, we need to replace the integral by summation and probability density by probability mass function in the definition.

[4]This quantity is often referred to as *interaction information* [47] or *co-information* [7]. Another multivariate generalization of mutual information is total correlation [78] (also known as multivariate constraint [24] or multi-information [69]).

(a)

(b)
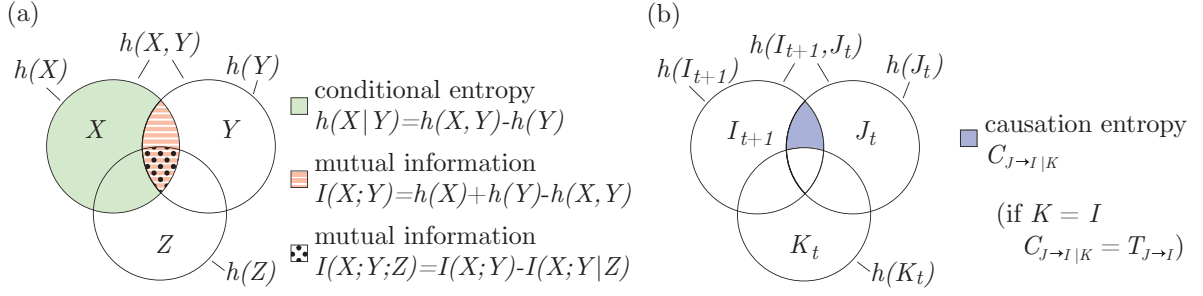
Figure 2. *Venn-like diagrams for information-theoretic measures.* (a) *Visualization of the relationships between entropy, conditional entropy, and mutual information.* (b) *Visualization of the relationships between conditional entropy, causation entropy, and transfer entropy. In* (b)*, letters I, J, and K are used to denote* $X^{(I)}$, $X^{(J)}$, *and* $X^{(K)}$, *respectively.*

stationary first-order Markov process such as that given by (2.4), the transfer entropy from j to i can be expressed as

$$(2.15) \qquad T_{j \to i} \equiv h(X_{t+1}^{(i)} | X_t^{(i)}) - h(X_{t+1}^{(i)} | X_t^{(i)}, X_t^{(j)}),$$

where $h(\cdot | \cdot)$ denotes conditional entropy [16]. Since $h(X_{t+1}^{(i)} | X_t^{(i)})$ measures the uncertainty of $X_{t+1}^{(i)}$ given information about $X_t^{(i)}$ and $h(X_{t+1}^{(i)} | X_t^{(i)}, X_t^{(j)})$ measures the uncertainty of $X_{t+1}^{(i)}$ given information about *both* $X_t^{(i)}$ *and* $X_t^{(j)}$, the transfer entropy $T_{j \to i}$ can be interpreted as the *reduction of uncertainty* about future states of $X^{(i)}$ when the current state of $X^{(j)}$ is provided *in addition to that of* $X^{(i)}$.

Networks of practical interest inevitably contain (many) more than two nodes. As we will show later, without appropriate conditioning, transfer entropy fails to distinguish between direct and indirect causality in networks. To overcome the pairwise limitation of transfer entropy, we define *causation entropy*. The relationships between entropy, transfer entropy, and causation entropy are illustrated in Figure 2(b).

Definition 2.1 (causation entropy [71]). The causation entropy *from the set of nodes J to the set of nodes I conditioning on the set of nodes K is defined as*[5]

$$(2.16) \qquad C_{J \to I | K} = h(X_{t+1}^{(I)} | X_t^{(K)}) - h(X_{t+1}^{(I)} | X_t^{(K)}, X_t^{(J)}),$$

where I, J, K *are all subsets of* $\mathcal{V} = \{1, 2, \ldots, n\}$. *In particular, if* $J = \{j\}$ *and* $I = \{i\}$, *we simplify the notation as* $C_{j \to i | K}$. *If the conditioning set* $K = \varnothing$, *we often omit it and simply write* $C_{J \to I}$.

Remark 2.1. Causation entropy is a natural generalization of transfer entropy from measuring pairwise causal relationships to network relationships of many variables. In particular, if $j \in K$, then the causation entropy $C_{j \to i | K} = 0$ as j does not carry extra information (compared to that of K). On the other hand, if $K = \{i\}$, causation entropy recovers transfer

---

[5]Note that the definitions in (2.15) and (2.16) can be extended for asymptotically stationary processes by taking the limit of $t \to \infty$, although the proofs in this paper do not directly apply to such a general scenario.

entropy; i.e.,

$$(2.17) \qquad C_{j \to i|i} = T_{j \to i}.$$

Interestingly, in this framework we see that transfer entropy assumes that nodes are self-causal, whereas causation entropy relaxes this assumption. Preliminary exploration of the differences between the two measures can be found in [71].

Remark 2.2. We note that in addition to [71], the conditional mutual information between time-lagged variables has been proposed as a statistic for network inference in a few previous studies [21, 60, 61, 75] (although not referred to as transfer or causation entropy).

Remark 2.3. It seems plausible to conjecture that if two subsets of the nodes satisfy $K_1 \subset K_2$, then $C_{j \to i|K_1}$ would be no less than $C_{j \to i|K_2}$. We remark that this statement about monotonicity is false (see the two examples below).

Example 1. Consider the stochastic process

$$(2.18) \qquad X_t^{(1)} = X_{t-1}^{(2)} + X_{t-1}^{(3)},$$

where $X_t^{(k)}$ are independent and identically distributed (i.i.d.) Bernoulli variables: $P(X_t^{(k)} = 0) = P(X_t^{(k)} = 1) = 0.5$ ($k = 2, 3$). Let $i = 1$, $j = 2$, $K_1 = \varnothing$, and $K_2 = \{3\}$. It follows that

$$(2.19) \qquad \begin{cases} C_{2 \to 1|\varnothing} = \frac{3}{2}\log 2 - \log 2 = \frac{1}{2}\log 2 \\ C_{2 \to 1|\{3\}} = \log 2 - 0 = \log 2 \end{cases} \Rightarrow C_{2 \to 1|\varnothing} < C_{2 \to 1|\{3\}}.$$

Example 2. Consider the stochastic process

$$(2.20) \qquad X_{t+1}^{(1)} = X_t^{(3)}, \ X_{t+1}^{(2)} = X_t^{(3)},$$

where $X_t^{(3)}$ are Bernoulli variables with $P(X_t^{(3)} = 0) = P(X_t^{(3)} = 1) = 0.5$. Let $i = 1$, $j = 2$, $K_1 = \varnothing$, and $K_2 = \{3\}$. It follows that

$$(2.21) \qquad \begin{cases} C_{2 \to 1|\varnothing} = \log 2 - 0 = \log 2 \\ C_{2 \to 1|\{3\}} = 0 - 0 = 0 \end{cases} \Rightarrow C_{2 \to 1|\varnothing} > C_{2 \to 1|\{3\}}.$$

The seemingly paradoxical observation that $C_{j \to i|K_1}$ can be either larger or smaller than $C_{j \to i|K_2}$ despite the fact that $K_1 \subset K_2$ can be understood as follows: When $K_1 \subset K_2$, $C_{j \to i|K_1} - C_{j \to i|K_2}$ corresponds to the mutual information among the three variables $X_{t+1}^{(i)}|X_t^{(K_1)}$, $X_{t+1}^{(i)}|X_t^{(j)}$, and $X_{t+1}^{(i)}|X_t^{(K_2 - K_1)}$ (see Figure 2). Contrary to the two-variable case, where mutual information is always nonnegative, the mutual information among three (or more) variables can either be positive, negative, or zero [47].

2.4. Theoretical properties of causation entropy and the optimal causation entropy principle. In the following we show that analysis of causation entropy leads to exact network inference for the network stochastic process given by (2.4) subject to the Markov assumptions in (2.8).

We start by exploring basic analytical properties of causation entropy, which is presented in Theorem 2.2 and also summarized in Figure 3.
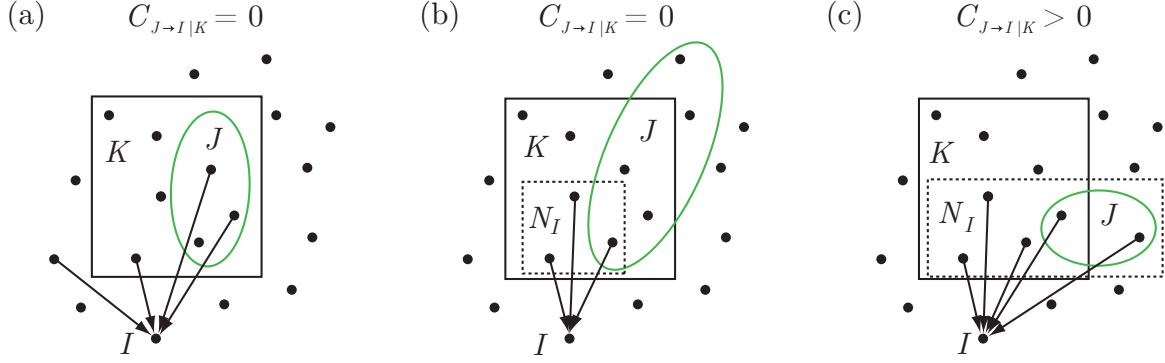
Figure 3. *Basic analytical properties of causation entropy (Theorem 2.2) allowing for the inference of the causal parents N_I of a set of nodes I. (a) Redundancy: If J is a subset of the conditioning set K ($J \subset K$), then the causation entropy $C_{J \to I|K} = 0$. (b) No false positive: If N_I is already included in the conditioning set K ($N_I \subset K$), then $C_{J \to I|K} = 0$. (c) True positive: If a set J contains at least one causal parent of I that does not belong to the conditioning set K, i.e., $(J \subset N_I) \wedge (J \not\subset K)$, then $C_{J \to I|K} > 0$.*

Theorem 2.2 (basic analytical properties of causation entropy). *Suppose that the network stochastic process given by (2.4) satisfies the Markov assumptions in (2.8). Let $I \subset \mathcal{V}$ be a set of nodes and N_I be its causal parents. Consider two sets of nodes $J \subset \mathcal{V}$ and $K \subset \mathcal{V}$. The following results hold:*

(a) *(Redundancy) If $J \subset K$, then $C_{J \to I|K} = 0$.*
(b) *(No false positive) If $N_I \subset K$, then $C_{J \to I|K} = 0$ for any set of nodes J.*
(c) *(True positive) If $J \subset N_I$ and $J \not\subset K$, then $C_{J \to I|K} > 0$.*
(d) *(Decomposition) $C_{J \to I|K} = C_{(K \cup J) \to I} - C_{K \to I}$.*

Proof. Under the temporal Markov condition in (2.8), there is no time dependence of the distributions. For notational simplicity we denote the joint distribution $p(X_{t+1}^{(I)} = i, X_t^{(J)} = j, X_t^{(K)} = k)$ by $p(i, j, k)$ and use similar notation for the marginal and conditional distributions. It follows that

$$C_{J \to I|K} = h(X_{t+1}^{(I)}|X_t^{(K)}) - h(X_{t+1}^{(I)}|X_t^{(K)}, X_t^{(J)}) = -\int p(i, j, k) \log\left[\frac{p(i|k)}{p(i|j, k)}\right] di\, dj\, dk$$

$$\geq -\log \int p(i, j, k) \frac{p(i|k)}{p(i|j, k)} di\, dj\, dk \quad \text{(by Jensen's inequality [58])}$$

$$(2.22) \qquad = -\log \int p(j, k) \frac{p(i, k)}{p(k)} di\, dj\, dk = -\log(1) = 0,$$

where equality holds iff $p(i|k) = p(i|j, k)$ *almost everywhere*. The above inequality is also known as the *Gibbs inequality* in statistical physics [26].

To prove (a), we note that $J \subset K$ implies that $p(i|k) = p(i|j, k)$, and therefore equality holds (rather than inequality) in (2.22).

To prove (b), it suffices to show that for $J \not\subset K$, $C_{J \to I|K} = 0$. Since $J \not\subset K$ and $N_I \subset K$, based on the spatial Markov condition in (2.8), we have

$$(2.23) \qquad p(X_{t+1}^{(I)}|X_t) = p(X_{t+1}^{(I)}|X_t^{(K \cup J)}) = p(X_{t+1}^{(I)}|X_t^{(K)}) = p(X_{t+1}^{(I)}|X_t^{(N_I)}).$$

Therefore, $p(i|j,k) = p(i|k)$, and equality holds in (2.22).

To prove (c), we use the faithfully Markov condition in (2.8). Since $J \subset N_I$ and $J \not\subset K$, it follows that

$$(2.24) \qquad p(X_{t+1}^{(I)}|X_t^{(K)}) = p(X_{t+1}^{(I)}|X_t^{(K \cap N_I)}) \neq p(X_{t+1}^{(I)}|X_t^{(K)}, X_t^{(J)}).$$

Thus, $p(i|j,k) \neq p(i|k)$, and strict inequality holds in (2.22).

Finally, part (d) follows directly from the definition of C. ■

Theorem 2.2 allows us to convert the problem of causal network inference into the problem of estimating causation entropy among nodes. In particular, for a given set of nodes I, each node j can *in principle* be checked independently to determine whether or not it is a causal parent of I via either of the following two equivalent criteria (proved in Theorem 2.3(a) below):

$$(2.25) \qquad \begin{cases} \text{(1) node } j \in N_I \text{ iff there is a set } K \supset N_I, \text{ such that } C_{j \to I|(K-\{j\})} > 0; \\ \text{(2) node } j \in N_I \text{ iff for any set } K \subset \mathcal{V}, C_{j \to I|(K-\{j\})} > 0. \end{cases}$$

Practical application of either criterion to infer large networks is challenging. Criterion (1) requires a conditioning set K that contains N_I as its subset. Since N_I is generally unknown, one often must use $K = \mathcal{V}$. When the network is large ($n \gg 1$), this requires the estimation of causation entropy for very high dimensional random variables from limited data, which is inherently unreliable [60, 61]. Criterion (2), on the other hand, requires a combinatorial search over all subsets, making it computationally infeasible.

In the following we prove the two inference criteria in (2.25). Furthermore, we show that *the set of causal parents is the minimal set of nodes that maximizes causation entropy*, which we refer to as the *optimal causation entropy principle.*

Theorem 2.3 (optimal causation entropy principle for causal network inference). *Suppose that the network stochastic process given by (2.4) satisfies the Markov properties in (2.8). Let $I \subset \mathcal{V}$ be a given set of nodes and N_I be the set of I's causal parents, as defined in (2.3). It follows that*

(a) *(Direct inference) Node $j \in N_I$ iff $\Leftrightarrow \exists K \supset N_I$ such that $C_{j \to I|(K-\{j\})} > 0 \Leftrightarrow \forall K \subset \mathcal{V}, C_{j \to I|(K-\{j\})} > 0$.*

(b) *(Partial conditioning removal) If there exists $K \subset \mathcal{V}$ such that $C_{j \to I|(K-\{j\})} = 0$, then $j \notin N_I$.*

(c) *(Optimal causation entropy principle) The set of causal parents is the minimal set of nodes with maximal causation entropy.*

Define the family of sets with maximal causation entropy as

$$(2.26) \qquad \mathcal{K} = \{K | \forall K' \subset \mathcal{V}, C_{K' \to I} \leq C_{K \to I}\}.$$

Then the set of causal parents is given by

$$(2.27) \qquad N_I = \cap_{K \in \mathcal{K}} K = \text{argmin}_{K \in \mathcal{K}} K.$$

Proof. First we prove part (a). If $j \in N_I$, then for every $K \subset \mathcal{V}$, $C_{j \to I|(K-\{j\})} > 0$ following Theorem 2.2(c). This proves both $\Rightarrow$'s. On the other hand, suppose that $\forall K \subset$

$\mathcal{V}, C_{j \to I|(K - \{j\})} > 0$; then for $K = \mathcal{V} \supset N_I$, it follows that $C_{j \to i|(\mathcal{V} - \{j\})} > 0$. Node $j \in N_I$ since otherwise $(\mathcal{V} - \{j\}) \supset N_I$, which would imply that $C_{j \to i|(\mathcal{V} - \{j\})} = 0$ from Theorem 2.2(b). Therefore, the two $\Leftarrow$'s are also proven.

Next, part (b) follows directly from the contrapositive of Theorem 2.2(c).

Finally, we prove part (c). Note that if $N_I \not\subset K$, then $J = N_I - K \neq \varnothing$, and so $C_{(K \cup J) \to I} - C_{K \to I} = C_{J \to I|K} > 0$. Therefore, $K \in \mathcal{K} \Rightarrow N_I \subset K$. This implies $N_I \subset \cap_{K \in \mathcal{K}} K$. On the other hand, if $\exists j \in \cap_{K \in \mathcal{K}} K$ with $j \notin N_I$, let $K \in \mathcal{K}$ and $L = K - \{j\}$. Since $j \notin N_I$, we have $N_I \subset L \subset K$, and therefore $C_{K \to I} - C_{L \to I} = C_{j \to I|L} = 0$, where the second equality follows from Theorem 2.2(c). This shows that $L \in \mathcal{K}$, contradicting $j \in \cap_{K \in \mathcal{K}} K$. So $j \in \cap_{K \in \mathcal{K}} K \Rightarrow j \in N_I$, which implies that $\cap_{K \in \mathcal{K}} K \subset N_I$. Since $\mathcal{K}$ is finite, it follows that $\cap_{K \in \mathcal{K}} K = \operatorname{argmin}_{K \in \mathcal{K}} K.$ ∎

Based on the optimal causation entropy principle, it seems straightforward to solve the minimax optimization for the inference of N_I by enumerating all subsets of $\mathcal{V}$ with increasing cardinality (starting from $\varnothing$) and terminating when a set K is found to have maximal causation entropy among all subsets of cardinality $|K| + 1$ (i.e., adding any node j to set K does not increase the causation entropy $C_{K \to I}$). Based on Theorem 2.3, the set $K = N_I$. However, this brute-force approach requires $\mathcal{O}(n^{|N_I|})$ causation entropy evaluations, which is computationally inefficient and therefore infeasible for the inference of real-world networks, which often contain large numbers of nodes ($n \gg 1$). Such limitation is removed only when the number of causal parents is moderately small, $|N_I| = \mathcal{O}(1)$. In the following section we develop additional theory and algorithms to efficiently solve this minimax optimization problem for causal network inference.
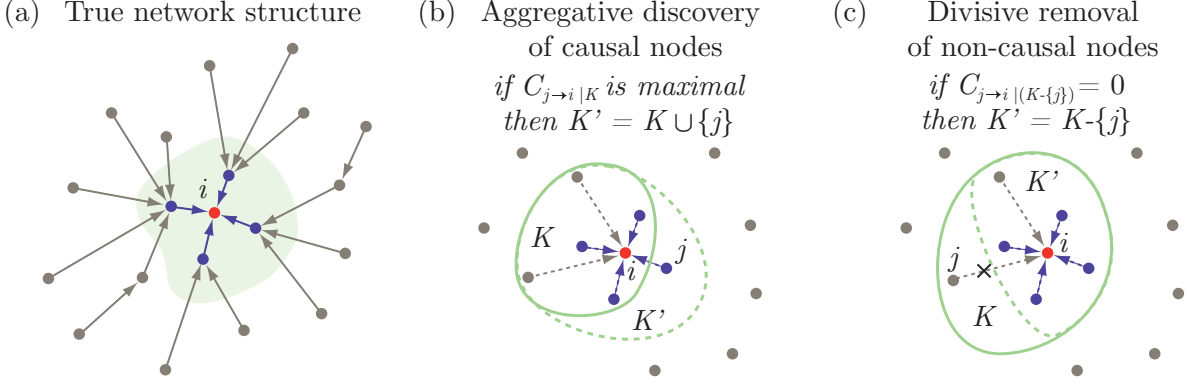
2.5. Computational causal network inference. Algorithmically, causal network inference via the optimal causation entropy principle should require as few computations as necessary (computational efficiency) and as few data samples as possible while retaining accuracy (data efficiency). We introduce two such algorithms that jointly infer the causal network. For a given node i, the goal is to infer its causal parents, as illustrated by nodes in the shaded region of Figure 4(a). Algorithm 2.1 aggregatively identifies nodes that form a superset of the causal parents, $K \supset N_i$ (proven by Lemma 2.4 and illustrated in Figure 4(b)). Starting from a set $K \supset N_i$, Algorithm 2.2 prunes away noncausal nodes from K, leaving only the causal parents N_i (proven by Lemma 2.5 and illustrated in Figure 4(c)).

Lemma 2.4 (aggregative discovery of causal nodes). *Suppose that the network stochastic process given by* (2.4) *satisfies the Markov properties in* (2.8). *Let $I \subset \mathcal{V}$ and N_I be its causal parents. Define the sequences of numbers $\{x_1, x_2, \dots\}$, nodes $\{p_1, p_2, \dots\}$, and nested sets $\{K_0, K_1, K_2, \dots\}$ as $K_0 = \varnothing$, and*

$$(2.28) \qquad \begin{cases} x_i = \max_{x \in (\mathcal{V} - K_{i-1})} C_{x \to I|K_{i-1}}, \\ p_i = \operatorname{argmax}_{x \in (\mathcal{V} - K_{i-1})} C_{x \to I|K_{i-1}}, \\ K_i = \{p_1, p_2, \dots, p_i\} \end{cases}$$

for every $i \geq 1$. There exists a number q, with $|N_I| \leq q \leq n$, such that the following hold:

(a) *The numbers $x_i > 0$ for $1 \leq i \leq q$ and $x_i = 0$ for $i > q$.*

(b) *The set of causal parents $N_I \subset K_q = \{x_1, x_2, \dots, x_q\}$.*

(a)   True network structure   (b)   Aggregative discovery   (c)   Divisive removal
                                        of causal nodes   of non-causal nodes
                                  $\textit{if } C_{j\to i\,|K} \textit{ is maximal}$ $\textit{if } C_{j\to i\,|(K-\{j\})} = 0$
                                  $\textit{then } K' = K \cup \{j\}$ $\textit{then } K' = K\text{-}\{j\}$

Figure 4. *Causal network inference by optimal causation entropy.* (a) *Causal parents and noncausal nodes of a node i. Causal network inference corresponds to identifying the causal parents N_i (nodes in shaded region) for every node $i \in \mathcal{V}$.* (b) *Nodes are added to the set K in an aggregative fashion, maximizing causation entropy at each step (see Algorithm* 2.1*).* (c) *Starting from a set $K \supset N_i$ (K obtained by Algorithm* 2.1*), noncausal nodes are progressively removed from K if their causation entropy to node i conditioned on the rest of K is zero (see Algorithm* 2.2*).*

Algorithm 2.1. Aggregative discovery of causal nodes.

Input: Set of nodes $I \subset \mathcal{V}$
Output: K (which will include N_I as its subset)

1: Initialize: $K \leftarrow \varnothing$, $x \leftarrow \infty$, $p \leftarrow \varnothing$.
2: **while** $x > 0$ **do**
3:     $K \leftarrow K \cup \{p\}$
4:     **for** every $j \in (\mathcal{V} - K)$ **do**
5:         $x_j \leftarrow C_{j\to I|K}$
6:     **end for**
7:     $x \leftarrow \max_{j\in(\mathcal{V}-K)} x_j$, $p \leftarrow \operatorname{argmax}_{j\in(\mathcal{V}-K)} x_j$
8: **end while**

Proof. If $N_I = \varnothing$, the lemma holds trivially. Suppose that $|N_I| \geq 1$ and so $x_1 > 0$.

To prove (a), we define $q \equiv \min_{x_i=0}(i-1)$ (if all $x_i > 0$, define $q \equiv n$). By construction, $x_i > 0$ when $i \leq q$ and $x_{q+1} = 0$. This implies that $N_I \subset K_q$ since otherwise there is a node j with $C_{j\to I|K_q} > 0 \Rightarrow x_{q+1} > 0$. For any $i > q$, $N_I \subset K_q \subset K_{i-1}$, and thus $C_{j\to I|K_{i-1}} = 0$ for all $j \in (\mathcal{V} - K_{i-1})$, which implies that $x_i = 0$.

To prove (b), we note that if there is a node $j \in N_I$ such that $j \notin K_q$, then by the definition of x_i and Theorem 2.2(c), it follows that $x_{q+1} \geq C_{j\to I|K_q} > 0$. This is in contradiction with the fact that $x_i = 0$ for all $i > q$. Therefore, $N_I \subset K_q$. ∎

Algorithm 2.1 recursively constructs the set $K_q \supset N_I$ (further denoted as K) as described by Lemma 2.4 and illustrated in Figure 4(b). To remove indirect and spurious nodes in K that do not belong to N_I, we apply the result of Theorem 2.2(c), $C_{j\to I|(K-\{j\})} = 0 \Rightarrow j \notin N_I$. This gives rise to Lemma 2.5 and Algorithm 2.2.

Algorithm 2.2. Progressive removal of noncausal nodes.

Input: Sets of nodes $I \subset \mathcal{V}$ and $K \subset \mathcal{V}$
Output: $\hat{N}_I$ (inferred set of causal parents of I)
1: **for** every $j \in K$ **do**
2:     **if** $C_{j \to I | (K - \{j\})} = 0$ **then**
3:         $K \leftarrow K - \{j\}$
4:     **end if**
5: **end for**
6: $\hat{N}_I \leftarrow K$

Lemma 2.5 (*progressive removal of noncausal nodes*). *Suppose that the network stochastic process given by* (2.4) *satisfies the Markov properties in* (2.8)*. Let* $I \subset \mathcal{V}$ *and* N_I *be its causal parents. Let* $K = \{p_1, p_2, \ldots, p_q\}$ *such that* $K \supset N_I$*. Define the sequence of sets* $\{K_0, K_1, K_2, \ldots, K_q\}$ *by* $K_0 = K$*, and*

$$(2.29) \qquad K_i = \begin{cases} K_{i-1} & \text{if } C_{p_i \to I | (K_{i-1} - \{p_i\})} > 0, \\ K_{i-1} - \{p_i\} & \text{if } C_{p_i \to I | (K_{i-1} - \{p_i\})} = 0 \end{cases}$$

for every $1 \leq i \leq q$*. Then* $K_q = N_I$*.*

Proof. By definition, $K_0 = K \supset N_I$. We prove that $K_q \supset N_I$ by induction. Suppose that $K_{i-1} \supset N_I$. If node $p_i \in N_I$, then $C_{p_i \to I | (K_{i-1} - \{p_i\})} > 0$ by Theorem 2.2(c), and therefore $K_i = K_{i-1} \supset N_I$. If node $p_i \notin N_I$, then $K_i \supset K_{i-1} - \{p_i\} \supset N_I$.

Next we prove that $K_q \subset N_I$. Suppose that node $p_i \notin N_I$. Since $K_{i-1} \supset N_I$, the causation entropy $C_{p_i \to I | (K_{i-1} - \{p_i\})} = 0$ by Theorem 2.2(b), and so $K_i = K_{i-1} - \{p_i\}$. Therefore, $p \notin K_i \supset K_q$, which implies that $K_q \subset N_I$ (contrapositive). ■

Algorithm 2.2 iteratively removes nodes that are not causal parents from a set K until the set converges to N_I as described by Lemma 2.5 and illustrated in Figure 4(c).

Jointly, Algorithms 2.1 and 2.2 can be applied to identify the causal parents of each node, thus inferring the entire causal network.[6]

Remark 2.4. There exists a number of algorithms for the problem of network inference, and we will comment on the two most relevant techniques. First, we note that the ARACNE algorithm [46] attempts to infer a (noncausal) interaction network based on mutual information. The ARACNE algorithm first computes the mutual information between all pairs of nodes/variables, filtering out the insignificant ones, and then enumerates through all triplets and removes links based on the data processing inequality [16]. It was proven to correctly infer the undirected network under the assumptions that (i) mutual information is estimated without error, and (ii) the network is a tree [46]. Second, the PC algorithm developed by Spirtes, Glymour, and Scheines removes noncausal links by potentially testing all combinations of conditioning subsets and was proven to correctly infer general causal networks if the conditional independence between the variables can be perfectly examined [66]. Runge et al.

---

[6]Numerically estimated causation entropy is always positive due to finite sample size and numerical precision. In practice, one needs to use a statistical test (e.g., permutation test as described in section 4) to examine the conditions $x > 0$ in Algorithm 2.1 and $C_{j \to I | (K - \{j\})} = 0$ in Algorithm 2.2.

[60, 61] recently utilized the PC algorithm to infer causal networks by establishing the conditional dependence/independence via estimation of appropriately defined conditional mutual information between time-lagged variables. We note that whereas we utilize Algorithm 2.2 for the divisive step in network inference, an alternative would be to utilize the PC algorithm for the divisive step. Although the accuracy versus efficiency tradeoff for such a modification has yet to be tested, we expect that it may be helpful specifically for inferring the causal parents for nodes with large degree, suggesting that in practical applications one may wish to switch back and forth between Algorithm 2.2 and the PC algorithm for the divisive step, depending on a node's degree.

3. Application to Gaussian process: Analytical results. In this section we make analytical comparisons among three approaches to causal network inference: causation entropy, transfer entropy [63], and conditional Granger causality [29, 30]. The next section will be devoted to the exploration of the numerical properties of these approaches for general random networks.

While information-theoretic approaches including causation entropy do not require stringent model assumptions, a linear model must be assumed to offer a fair comparison with the conditional Granger causality. As a benchmark example, we focus on the following linear discrete stochastic network dynamics:

$$(3.1) \qquad X_t^{(i)} = \sum_{j \in N_i} A_{ij} X_{t-1}^{(j)} + \xi_t^{(i)} \ \big(\text{or, in matrix form, } X_t = A X_{t-1} + \xi_t\big).$$

Here $X_t^{(i)} \in \mathbb{R}$ represents the state of node i at time t ($i \in \{1, 2, \ldots, n\}, t \in \mathbb{N}$), $\xi_t^{(i)} \in \mathbb{R}$ represents noise, and $A_{ij} X_{t-1}^{(j)}$ models the influence of node j on node i. Equation (3.1) finds application in a broad range of areas, including time series analysis (as a multivariate linear autoregressive process [10]), information theory (as a network communication channel [16]), and nonlinear dynamical systems (as a linearized stochastic perturbation around equilibrium states [43]). It is straightforward to check that (3.1) is a special case of the general network stochastic process, (2.4), and asymptotically (as $t \to \infty$) satisfies the Markov assumptions in (2.8).

3.1. Analytical properties of the solution. We begin by reviewing the well-known statistical properties of time series data generated by (3.1) [20], which will serve as a framework for us to compare methods for causal network inference.

3.1.1. Solution formula. Defining $X_0 = \xi_0$ for convenience, the solution to (3.1) can be expressed as

$$(3.2) \qquad\qquad\qquad X_t = \sum_{k=0}^{t} A^k \xi_{t-k}.$$

We assume that $\xi_t^{(i)}$ are i.i.d. Gaussian random variables with zero mean and finite nonzero variance, denoted as $\xi_t^{(i)} \sim N(0, \sigma_i^2)$ with $\sigma_i > 0$. Therefore,

$$(3.3) \qquad\qquad\qquad \xi_t \sim N(0, S),$$

where the covariance matrix S is defined by $S_{ij} = \delta_{ij}\sigma_i^2$ with δ denoting the Kronecker delta. It follows that

$$(3.4) \qquad \begin{cases} \mathrm{E}[\xi_t^{(i)}] = 0, \\ \mathrm{Cov}(\xi_t^{(i)}, \xi_\tau^{(j)}) = \delta_{ij}\delta_{t\tau}. \end{cases}$$

Note that a random variable obtained by an affine transformation of a Gaussian variable is also Gaussian. For example, if $Y = [Y_1; Y_2]$ is Gaussian, the distribution of Y_1 conditioned on Y_2 is also Gaussian [20]. The proposition below follows from expressing random variables via appropriate affine transformations of ξ_t's.

Proposition 3.1. *Let I and K be any subsets of $\mathcal{V}$. Let $t \in \mathbb{N}$ and $\tau \in \{0\} \cup \mathbb{N}$. The conditional distribution of $X_{t+\tau}^{(I)}$ given $X_t^{(K)}$ is Gaussian.*

3.1.2. Covariance matrix. Under an affine transformation from Gaussian variable Y to Z as $Z = CY + d$, the mean and covariance of Y and Z are related by $\mu_Z = C\mu_Y + d$ and $\Sigma_Z = C\Sigma_Y C^\top$ [20]. We consider covariance matrices $\Phi(\tau, t)$, where the (i, j)th entry of $\Phi(\tau, t)$ is defined as

$$(3.5) \qquad \Phi(\tau, t)_{ij} \equiv \mathrm{Cov}[x_{t+\tau}^{(i)}, x_t^{(j)}].$$

It follows from (3.2) and (3.3) that

$$(3.6) \qquad X_t \sim N(0, \Phi(0, t)), \text{ where } \Phi(0, t) = \sum_{k=0}^{t} A^k S (A^k)^\top.$$

In the following we prove a sufficient condition for the converge of the covariance matrix $\Phi(0, t)$ as time $t \to \infty$. Denote the *spectral radius* of a square matrix M by

$$(3.7) \qquad \rho_M \equiv \max\{|\lambda| : \lambda \text{ is an eigenvalue of } M\}.$$

Note that $\rho_M = \rho_{M^\top}$ since a square matrix and its transpose have the same set of eigenvalues. For the dynamical system defined by (3.1), matrices A with $|\rho_A| < 1$ are the only matrices for which the underlying system poses a stable equilibrium in the absence of noise. We refer to these matrices as stable.

Definition 3.2 (stable matrix). *Matrix M is stable if $\rho_M < 1$.*

The following is a known result from classical matrix theory [36].

Theorem 3.3 (convergence of matrix series [36]). *The matrix series $\sum_{k=0}^{\infty} M_k$ converges if the scalar series $\sum_{k=0}^{\infty} \|M_k\|$ under any induced norm $\|\cdot\|$ converges.*

Note that it is possible for the matrix series $\sum_{k=0}^{\infty} M_k$ to be convergent while the corresponding scalar series $\sum_{k=0}^{\infty} \|M_k\|$ diverges, analogous to the possibility of a scalar series that is convergent but not absolutely convergent. Next we state and prove a sufficient condition under which the matrix series in (3.6) converges.

Proposition 3.4 (convergence of the covariance). *The series $\sum_{k=0}^{\infty} A^k S (A^k)^\top$ converges if A is stable.*

Proof. Let $\|\cdot\|$ be any induced norm. Then $\|A^k S (A^k)^\top\| \leq \|A^k\| \cdot \|S\| \cdot \|(A^\top)^k\|$ for any $k \in \mathbb{N}$. Gelfand's formula (see [25]) implies that

$$(3.8) \qquad \lim_{k\to\infty} \|A^k\|^{1/k} = \lim_{k\to\infty} \|(A^\top)^k\|^{1/k} = \rho_A.$$

On the other hand, $\lim_{k\to\infty} \|S\|^{1/k} = 1$. Therefore,

$$\lim_{k\to\infty} \|A^k S(A^k)^\top\|^{1/k} \leq \lim_{k\to\infty} \left(\|A^k\| \cdot \|S\| \cdot \|(A^\top)^k\|\right)^{1/k} = \rho_A^2 < 1,$$

where the last inequality follows from the fact that A is stable. Hence the scalar series $\sum_{k=0}^\infty \|A^k S(A^k)^\top\|_2$ is convergent. The proposition follows by Theorem 3.3. ∎

For the remainder of this section, it will be assumed that A is stable in (3.1). As $t \to \infty$, we drop the second argument in $\Phi(0,t)$ and define the *asymptotic covariance matrix*

$$(3.9) \qquad\qquad \Phi(0) \equiv \lim_{t\to\infty} \Phi(0,t) = \sum_{k=0}^\infty A^k S(A^k)^\top.$$

It follows that $\Phi(0)$ satisfies an algebraic equation given by the proposition below.

Proposition 3.5 (asymptotic covariance matrix). *Assume that A is stable. The asymptotic covariance matrix $\Phi(0) = \sum_{k=0}^\infty A^k S(A^k)^\top$ satisfies the equation*

$$(3.10) \qquad\qquad A\Phi(0)A^\top - \Phi(0) + S = 0.$$

Proof. Since A is stable, both of the two matrix series below converge:

$$\begin{cases} \Phi(0) = S + ASA^\top + A^2 S(A^2)^\top + A^3 S(A^3)^\top + \cdots \\ A\Phi(0)A^\top = ASA^\top + A^2 S(A^2)^\top + A^3 S(A^3)^\top + \cdots. \end{cases}$$

Subtracting the two equations gives the result of the proposition. ∎

Equation (3.10) is a (discrete) Lyapunov equation which often appears in stability analysis and optimal control problems [59]. Using "$\otimes$" as the Kronecker product and "vec" for the operation of transforming a square matrix to a column vector by stacking the columns of the underlying matrix in order, (3.10) can be converted into

$$(3.11) \qquad\qquad (I_{n^2} - A \otimes A)\,\text{vec}(\Phi(0)) = \text{vec}(S),$$

where I_{n^2} denotes the identity matrix of size n^2-by-n^2. Matrix $\Phi(0)$ can be computed either by solving (3.10) through iterative methods (see [5]) or by directly solving (3.11) as a linear system. In practice, we found the iterative approach to be numerically more efficient and stable compared to direct inversion.

Covariance matrices are in general positive semidefinite [20]. For the network dynamics defined in (3.1), we show that they are indeed positive definite.

Proposition 3.6 (positive definiteness of the covariance matrix). *The covariance matrix $\Phi(0,t)$ is positive definite for any $t \in \mathbb{N}$. The asymptotic covariance matrix $\Phi(0)$ is also positive definite.*

Proof. For any unit vector $v \in \mathbb{R}^n$, $v^\top A\Phi(0,0)A^\top v = (A^\top v)^\top A^\top v \geq 0$. From (3.2) and (3.3), for any $t \in \mathbb{N}$, $\Phi(0,t) = A\Phi(0,t-1)A^\top + S$. By induction,

$$v^\top \Phi(0,t)v = v^\top A\Phi(0,t-1)A^\top v + v^\top Sv$$
$$(3.12) \qquad\qquad \geq \left(A^\top v\right)^\top \Phi(0,t-1)\left(A^\top v\right) + \min_i \sigma_i^2 \geq \min_i \sigma_i^2 > 0.$$

This shows that $\Phi(0,t)$ is positive definite (indeed, we have $\rho_{\Phi(0,t)} \geq \min_i \sigma_i^2 > 0$). Taking $t \to \infty$ in the above estimate also shows that $\Phi(0)$ is positive definite. ∎

3.1.3. Time-shifted covariance matrices. We define the time-shifted covariance matrix $\Phi(\tau, t)$ for each $t \in \mathbb{N}$ (time) and $\tau \in \mathbb{N}$ (positive time shift between states). If A is stable, then the covariance matrix $\Phi(t, \tau)$ converges for each time shift τ as $t \to \infty$. The (asymptotic) covariance matrices with different time shifts are related by a simple algebraic equation given in the following proposition.

Proposition 3.7 (relationship between time-shifted covariance matrices). *Assume that A is stable. For each $\tau \in \mathbb{N}$, the following limit exists:*

$$\lim_{t \to \infty} \Phi(\tau, t) = \Phi(\tau),$$

where matrix $\Phi(\tau)$ satisfies

$$(3.13) \qquad \Phi(\tau) = A\Phi(\tau - 1) = A^2 \Phi(\tau - 2) = \cdots = A^\tau \Phi(0).$$

Proof. For every $\tau \in \mathbb{N}$ and $t \in \mathbb{N}$, it follows that

$$(3.14) \qquad \Phi(\tau, t)_{ij} = \mathrm{E}\left[\sum_{k=1}^{n} a_{ik} x_{t+\tau-1}^{(k)} + \xi_{t+\tau}^{(i)}, x_t^{(j)}\right] = \sum_{k=1}^{n} a_{ik} \Phi(\tau - 1, t)_{kj}.$$

Therefore, the matrix $\Phi(\tau, t)$ satisfies

$$(3.15) \qquad \Phi(\tau, t) = A\Phi(\tau - 1, t) = A^2 \Phi(\tau - 2, t) = \cdots = A^\tau \Phi(0, t).$$

Taking the limit as $t \to \infty$ and making use of the fact that A is stable, we reach the conclusion of the proposition. ∎

3.2. Analytical expressions of causation entropy. Here we provide analytical expressions for causation entropy of the Gaussian process described in (3.1). Because causation entropy can be interpreted as a generalization of both transfer entropy and conditional Granger causality under the appropriate selection of nodes i and j and the conditioning set K, these results also provide analytical expressions for transfer entropy and conditional Granger causality.

3.2.1. Joint entropy expressions. Let Σ be the covariance matrix of a multivariate Gaussian variable $X \in \mathbb{R}^n$ (i.e., $X \sim N(\boldsymbol{\mu}, \Sigma)$); it follows that [1]

$$(3.16) \qquad h(X) = \frac{1}{2} \log[\det(\Sigma)] + \frac{1}{2} n \log(2\pi e).$$

Note that the right-hand side of the above is actually an upper bound for a general random variable (i.e., the equality "=" becomes inequality "≤" [16]). Therefore, a Gaussian variable maximizes entropy among all variables of equal covariance.

The random variable X_t is Gaussian and converges to $N(0, \Phi(0))$ as $t \to \infty$. For an arbitrary subset of the nodes $K = \{k_1, k_2, \ldots, k_\ell\}$. The joint entropy is

$$(3.17) \qquad h(X^{(K)}) = \lim_{t \to \infty} h(X_t^{(K)}) = \frac{1}{2} \log(\det(\Phi_{KK}(0))) + \frac{1}{2} \ell \log(2\pi e).$$

Here we have introduced the notation

$$(3.18) \qquad \Phi_{IJ}(0) \equiv P(I)\Phi(0)P(J)^\top,$$

where, for a set $K = \{k_1, k_2, \ldots, k_\ell\}$, $P(K)$ is the ℓ-by-n projection matrix defined as

$$(3.19) \qquad P(K)_{ij} = \delta_{k_i, i}.$$

3.2.2. Causation entropy. For the Gaussian process given by (3.1), we obtain the analytical expression of causation entropy as

$$(3.20) \qquad C_{J \to I|K} = \frac{1}{2} \log \left(\frac{\det \left[\Phi(0)_{II} - \Phi(1)_{IK} \Phi(0)_{KK}^{-1} \Phi(1)_{IK}^{\top} \right]}{\det \left[\Phi(0)_{II} - \Phi(1)_{I,K \cup J} \Phi(0)_{K \cup J, K \cup J}^{-1} \Phi(1)_{I,K \cup J}^{\top} \right]} \right).$$

If $J = \{j\}$ and $I = \{i\}$, this equation simplifies to

$$(3.21) \qquad C_{j \to i|K} = \frac{1}{2} \log \left(\frac{\Phi(0)_{ii} - \Phi(1)_{iK} \Phi(0)_{KK}^{-1} \Phi(1)_{iK}^{\top}}{\Phi(0)_{ii} - \Phi(1)_{i,K \cup \{j\}} \Phi(0)_{K \cup \{j\}, K \cup \{j\}}^{-1} \Phi(1)_{i,K \cup \{j\}}^{\top}} \right).$$

3.2.3. Transfer entropy. Recall that causation entropy recovers transfer entropy when $K = \{i\}$. Letting $K = \{i\}$ in the formula above gives the transfer entropy (with single time lag) for multivariate Gaussian variables:

$$T_{j \to i} = C_{j \to i|i} = \frac{1}{2} \log \left(1 + \frac{\alpha_{ij}}{\beta_{ij} - \alpha_{ij}} \right),$$

$$(3.22) \qquad \text{where} \begin{cases} \alpha_{ij} \equiv \left(\Phi(0)_{ii} \Phi(1)_{ij} - \Phi(0)_{ij} \Phi(1)_{ii} \right)^2, \\ \beta_{ij} \equiv \left(\Phi(0)_{ii}^2 - \Phi(1)_{ii}^2 \right) \left(\Phi(0)_{ii} \Phi(0)_{jj} - \Phi(0)_{ij}^2 \right). \end{cases}$$

It follows that $\beta_{ij} \geq \alpha_{ij} \geq 0$, and therefore $T_{j \to i} \geq 0$ ($T_{i \to i} = 0$). Furthermore,

$$(3.23) \qquad T_{j \to i} = 0 \iff \alpha_{ij} = 0 \iff \sum_{k=1}^{n} A_{ik} \left(\Phi(0)_{ii} \Phi(0)_{kj} - \Phi(0)_{ij} \Phi(0)_{ki} \right) = 0.$$

3.2.4. Conditional Granger causality. As shown in [3], when the random variables are Gaussian, expression of Granger causality is equivalent to that of transfer entropy (and also causation entropy introduced here). In fact, for Gaussian variables, the Granger causality from j to i without conditioning equals $2C_{j \to i}$, while the conditional Granger causality (with full conditioning) equals $2C_{j \to i|(\mathcal{V} - \{j\})}$.

3.3. Analytical results for directed linear chain, directed loop, and directed trees. We derive expressions of transfer entropy and causation entropy for several classes of networks, including directed linear chains, directed loops, and directed trees. These results highlight that although transfer entropy may indicate the direction of information flow between two nodes, its application to causal network inference is often unjustified as it cannot distinguish between direct and indirect causal relationships (unless appropriate conditioning is adopted as in causation entropy).

3.3.1. Directed linear chain. Denote a directed linear chain of n nodes as

$$(3.24) \qquad\qquad 1 \to 2 \to 3 \cdots \to n.$$

For simplicity we assume that all links have the same weight $w = 1$. Consequently, the corresponding adjacency matrix $A = [A_{ij}]_{n \times n}$ is given by

$$(3.25) \qquad\qquad A_{ij} = \delta_{i,j+1}.$$
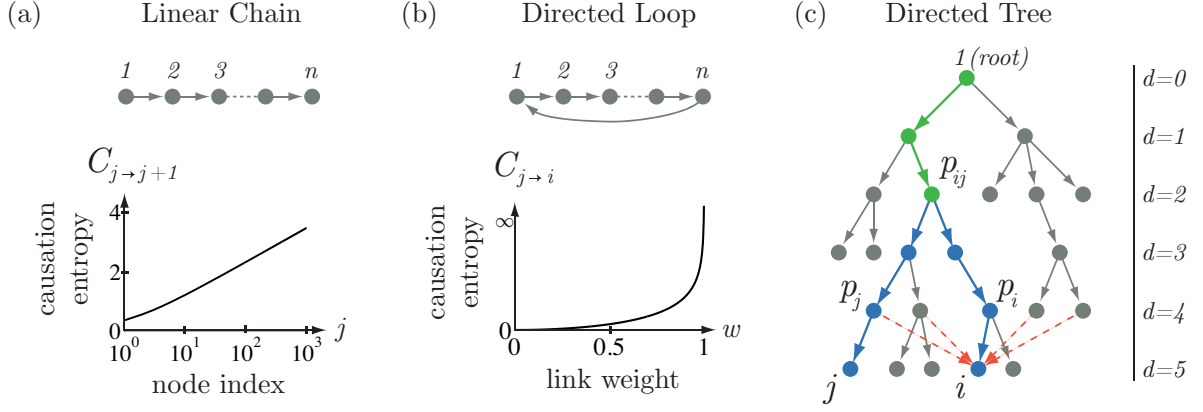
Figure 5. *Causation entropy and transfer entropy for a Gaussian process on three classes of networks. (a) For directed linear chains, both causation entropy and transfer entropy correctly identify the network as $C_{j\to i} = T_{j\to i} > 0$ iff $i = j + 1$ (otherwise $C_{j\to i} = T_{j\to i} = 0$). The dependence of $C_{j\to j+1}$ on node index j is given by (3.27) and plotted. (b) For directed loops, causation entropy and transfer entropy again correctly identify the network topology with $C_{j\to i} = T_{j\to i} > 0$ iff $j \to i$. The dependence of $C_{j\to i}$ on link weight w is given by (3.33) as shown. (c) For directed trees, causation entropy given by (3.41) correctly identifies the network topology based on (3.43). In contrast, transfer entropy without appropriate conditioning infers many links that do not exist in the actual network (red dashed lines), as described by (3.42).*

It follows that $\rho_A = 0$, and therefore A is stable. By inverting the lower-triangular matrix $(I_{n^2} - A \otimes A)$ in (3.11) and applying (3.13), we obtain that

$$(3.26) \qquad \begin{cases} \Phi(0)_{ij} = \delta_{ij} \sum_{k=1}^{j} \sigma_k^2, \\ \Phi(1)_{ij} = \delta_{i,j+1} \sum_{k=1}^{j} \sigma_k^2. \end{cases}$$

Letting $K = \varnothing$ and $K = \{i\}$, respectively, in (3.21) and (3.22), it follows that

$$(3.27) \qquad C_{j\to i} = T_{j\to i} = \frac{1}{2}\delta_{i,j+1} \log\left(1 + \frac{\sum_{k=1}^{j} \sigma_k^2}{\sigma_i^2}\right).$$

Therefore, for the directed linear chain defined in (3.25), transfer entropy $T_{j\to i} = C_{j\to i}$, and it is positive iff there is a direct link $j \to i$, i.e.,

$$(3.28) \qquad C_{j\to i} = T_{j\to i} > 0 \ \Leftrightarrow \ A_{ij} = 1, \quad \text{and} \quad C_{j\to i} = T_{j\to i} = 0 \ \Leftrightarrow \ A_{ij} = 0.$$

Interestingly, both causation entropy $C_{j\to j+1}$ and transfer entropy $T_{j\to j+1}$ increase monotonically as a function of j, and the values depend only on part of the chain from the top node (node 1) to node $j + 1$ and not on the rest of the network. Interpreting the monotonicity in terms of the network structure, the closer node j is to the end of the chain, effectively the more information is transferred through the directed link $j \to j + 1$. Figure 5(a) illustrates this via a network of $n = 1000$ nodes.

3.3.2. Directed loop. Consider now a directed loop with n nodes, denoted as

$$(3.29) \qquad 1 \to 2 \to 3 \cdots \to n \to 1.$$

Let $w > 0$ be the uniform link weight. It follows that $\rho_A = w$. Thus, for the adjacency matrix A to be stable, we must have $w < 1$. To keep the symmetry of the problem, we further assume that the variance of noise is the same at each node; therefore,

$$(3.30) \qquad \sigma^2 \equiv \sigma_1^2 = \sigma_2^2 = \cdots \sigma_n^2.$$

The entries in $\Phi(0, t)$ satisfy

$$(3.31) \qquad \Phi(0, t)_{ij} = w^2 \Phi(0, t - 1)_{p_i, p_j} + \delta_{ij}\sigma^2,$$

where p_i denotes the unique node that directly links to node i. Taking the limit as $t \to \infty$ and solving the resulting recursive equations, we obtain that

$$(3.32) \qquad \begin{cases} \Phi(0)_{ij} = \delta_{ij}\sigma^2/(1 - w^2), \\ \Phi(1)_{ij} = \delta_{p_i, j}\sigma^2 w/(1 - w^2), \end{cases}$$

where the second equation is obtained through $\Phi(0)_{ij}$ and (3.13). Letting $K = \varnothing$ and $K = \{i\}$, respectively, in (3.21) and (3.22), we conclude that

$$(3.33) \qquad C_{j \to i} = T_{j \to i} = \frac{1}{2}\delta_{p_i, j} \log\left(\frac{1}{1 - w^2}\right).$$

Note that causation entropy and transfer entropy equal and do not depend on the noise variation σ^2, and they are positive iff there is a direct link $j \to i$; i.e.,

$$(3.34) \qquad C_{j \to i} = T_{j \to i} > 0 \;\Leftrightarrow\; A_{ij} = 1, \quad \text{and} \quad C_{j \to i} = T_{j \to i} = 0 \;\Leftrightarrow\; A_{ij} = 0.$$

By symmetry, causation entropy and transfer entropy through each directed link are the same. As the link weight w increases in $(0, 1)$, both increase monotonically in $(0, \infty)$. The larger the link weight w is, the larger the amount of information is that is transferred via each directed link, as intuitively expected. Also see Figure 5(b) as an illustration.

3.3.3. Directed trees.
We now consider directed tree networks with uniform link weight $w = 1$ and unit node variance[7]

$$(3.35) \qquad \sigma_1^2 = \sigma_2^2 = \cdots \sigma_n^2 = 1.$$

A directed tree has one root (indexed as node 1 without loss of generality), and each nonroot node i ($i \neq 1$) has exactly one *ancestor*, denoted by p_i. The corresponding adjacency matrix $A = [A_{ij}]_{n \times n}$ thus satisfies

$$(3.36) \qquad A_{ij} = (1 - \delta_{i1})\delta_{i, p_i}.$$

It can be shown that $\rho_A = 0$. For $i \neq 1$, we denote the directed path from 1 to i by

$$(3.37) \qquad 1 = p_i^{(d_i)} \to p_i^{(d_i - 1)} \to \cdots \to p_i^{(1)} \equiv p_i \to p_i^{(0)} \equiv i,$$

---

[7]Similar results hold for trees with general link weights and node variances, but the corresponding equations are too cumbersome to list.

where d_i is the *depth* of node i in the tree (for node 1, we define its depth $d_1 = 0$). Thus, the highest node in the tree is the root, and the lowest nodes have the greatest depth. For any two nodes (i, j), we denote their *lowest common ancestor* by p_{ij}, i.e.,

$$(3.38) \qquad p_{ij} = \underset{\{k | \exists \ell, m \geq 0, s.t., p_i^{(\ell)} = p_j^{(m)}\}}{\arg \max} d_k.$$

The covariance matrix $\Phi(0, t)$ satisfies

$$(3.39) \qquad \Phi(0)_{ij} = \delta_{1i}\delta_{1j}\sigma_1^2 + (1 - \delta_{1i})(1 - \delta_{1j})[\Phi(0)_{p_i, p_j} + \delta_{ij}].$$

We solve these recursive equations to obtain

$$(3.40) \qquad \begin{cases} \Phi(0)_{ij} = \delta_{d_i, d_j}(d_{p_{ij}} + 1), \\ \Phi(1)_{ij} = (1 - \delta_{i1})\delta_{d_i, d_j + 1}(d_{p_{ij}} + 1), \end{cases}$$

where p_{ij} is defined in (3.38) and $\Phi(1)_{ij}$ is obtained by $\Phi(1) = A\Phi(0)$.

We calculate causation entropy and transfer entropy through (3.21) and (3.22):

$$(3.41) \qquad C_{j \to i} = T_{j \to i} = \frac{1}{2}\delta_{d_i, d_j + 1} \log \frac{(d_i + 1)(d_j + 1)}{(d_i + 1)(d_j + 1) - (d_{p_{ij}} + 1)^2}.$$

Note that, in general, $0 \leq d_{p_{ij}} \leq \min\{d_i, d_j\}$. Thus, $C_{j \to i} = T_{j \to i} \leq \frac{1}{2}\log(1 + d_i)$ with equality iff j is the ancestor of i (i.e., $j = p_i = p_{ij}$). Therefore, we have

$$(3.42) \qquad \begin{cases} T_{j \to i} > 0 \iff d_i = d_j + 1 \impliedby A_{ij} = 1 \text{ (but } T_{j \to i} > 0 \nRightarrow A_{ij} = 1), \\ T_{j \to i} = 0 \iff d_i \neq d_j + 1 \implies A_{ij} = 0 \text{ (but } A_{ij} = 0 \nRightarrow T_{j \to i} = 0). \end{cases}$$

In other words, transfer entropy being positive (without appropriate conditioning) corresponds to a superset of the links that actually exist in a directed tree, and the inferred network using this criterion will potentially contain many false positives. See Figure 5(c) for an example. On the other hand, for a given node $i \neq 1$, we have

$$(3.43) \qquad \begin{cases} p_i = \arg\max_j C_{j \to i}, \\ C_{j \to i | \{p_i\}} = 0. \end{cases}$$

Therefore, for each node i, the node j that maximizes causation entropy $C_{j \to i}$ among all nodes is inferred as the causal parent of i. Conditioned on this node, the causation entropy from any other node to i will become zero, indicating no other directed links to node i. This causation entropy based procedure allows for exact and correct inference of the underlying causal network, a directed tree.

4. Application to Gaussian process: Numerical results. In this section, we illustrate that causal network inference by optimal causation entropy is reliable and efficient for the Gaussian process, (3.1), on large random networks.

4.1. Random network model and time series generation. We consider signed Erdős–Rényi networks, which are a generation of their original model [8]. In particular, each network consists of n nodes ($\mathcal{V} = \{1, 2, \ldots, n\}$) such that each directed link $j \to i$ is formed independently with equal probability p, giving rise to a directed network with an expected number $n^2 p$ of directed links. For generality, we allow the link weight of each link $j \to i$ to be either positive ($A_{ij} = w$) or negative ($A_{ij} = -w$) with equal probability. Recalling that the network adjacency matrix A is defined entrywise by $A_{ij} \in \{w, -w\}$ iff there exists a directed link $j \to i$ (otherwise $A_{ij} = 0$), the link weight w may be selected to tune the spectral radius $\rho(A)$ of matrix A.

We generate time series from the stochastic equation, (3.1), where matrix A is obtained from the network model and random variables $\xi_t \sim \mathcal{N}(0, S)$, where the covariance matrix S is taken to be the identity matrix of size $n \times n$. To reduce transient effects, for a given sample size T we solve (3.1) for $10T$ time steps and use only the final 10% of the resulting time series.

To summarize, our numerical experiments contain the following parameters: n (network size), p (connection probability), $\rho(A)$ (spectral radius of A), and T (sample size).

4.2. Practical considerations for network inference. We have established by Theorems 2.2 and 2.3 and Lemmas 2.4 and 2.5 that, in theory, exact network inference can be achieved by optimal causation entropy, which involves implementing Algorithms 2.1 (aggregative discovery) and 2.2 (progressive removal) to correctly identify the set of causal parents N_i for each node $i \in \mathcal{V}$.

In practice, the success of our optimal causation entropy approach (and, in fact, any entropy-based approaches) depends crucially on reliable estimation of the relevant entropies in question from data. This leads to two practical challenges.

(1) Entropies must be *estimated* from *finite* time series data. While there are several techniques for estimating entropies for general multivariate data, such estimations are increasingly inaccurate for small sample sizes and high-dimensional random variables [52]. In this research, we side-step this computational complexity by using knowledge of the asymptotic functional form for the entropy of the Gaussian process, where the covariance matrices $\Phi(0)$ and $\Phi(1)$ in (3.20) and (3.21) are estimated directly from the time series data.

(2) Application of the theoretical results rely on determining whether the causation entropy $C_{j \to i|K} > 0$ or $C_{j \to i|K} = 0$. However, the estimated value of $C_{j \to i|K}$ based on sample covariances is necessarily positive given finite sample size and finite numerical precision. Therefore, a statistical test must be used to assess the significance of the observed positive causation entropy. We here adopt a widely used approach in nonparametric statistics called the *permutation test*.[8] Specifically, we propose the following permutation test based on the null hypothesis that causation entropy $C_{j \to i|K} = 0$: first, we perform r random temporal permutations of the time series $\{X_t^{(j)}\}$, leaving the rest of the data unchanged; we then construct an empirical cumulative distribution $\hat{F}(x)$ of the estimated causation entropy from the permuted

---

[8]The idea of a permutation test is to perform (a large number of) random permutations of a subset of the data, leaving the rest unchanged, giving rise to an empirical distribution of the static of interest. The observed statistic from the original data is then located on this empirical distribution in order to associate its statistical significance [28].

time series;[9] finally, given a prescribed significance threshold θ, the observed $C_{j \to i|K} = c$ is declared *significant* (i.e., the null hypothesis is rejected at level $1 - \theta$) if $\hat{F}(c) > \theta$.

To summarize, the inference algorithms contain two parameters to be used in the permutation test: r (number of random permutations) and θ (significance threshold).

4.3. Comparing optimal causation entropy, conditional Granger, and transfer entropy.
Here we compare the performance of three approaches of causal network inference: conditional Granger (see, for example, [23, 32]), transfer entropy (see [76] and the references therein), and optimal causation entropy (oCSE). In particular, the conditional Granger and transfer entropy approaches under consideration both estimate the entropy $C_{j \to i|K}$ for each pair of nodes (i, j) independently, with the choice of $K = \mathcal{V} - \{j\}$ in the case of conditional Granger and $K = \{i\}$ in the case of transfer entropy. In both approaches, a causal link $j \to i$ is inferred if the observed $C_{j \to i|K} > 0$ is assessed as significant under the permutation test. The oCSE approach combines Algorithms 2.1 and 2.2, and the permutation test is used once per each iteration (line 2 of both algorithms).

The performance of the three approaches is quantified by two types of inference error: false negative ratio, denoted as ε_- and defined as the fraction of links in the original network that are not inferred, and false positive ratio, denoted as ε_+ and defined as the fraction of nonexisting links in the original networks that are inferred. In terms of the adjacency matrix A of the original network and that of the inferred network $\hat{A}$, these ratios can be computed as

$$(4.1) \quad \begin{cases} \varepsilon_- \equiv \dfrac{\text{number of } (i,j) \text{ pairs with } \chi_0(A)_{ij} = 1 \text{ and } \chi_0(\hat{A})_{ij} = 0}{\text{number of } (i,j) \text{ pairs with } \chi_0(A)_{ij} = 1}, \\[2ex] \varepsilon_+ \equiv \dfrac{\text{number of } (i,j) \text{ pairs with } \chi_0(A)_{ij} = 0 \text{ and } \chi_0(\hat{A})_{ij} = 1}{\text{number of } (i,j) \text{ pairs with } \chi_0(A)_{ij} = 0}. \end{cases}$$

For the random networks considered here, we found that Algorithm 2.1 achieves almost the same accuracy as the combination of Algorithms 2.1 and 2.2. We therefore present results which are based on the numerical application of Algorithm 2.1 alone, leaving detailed numerical study of Algorithm 2.2 to future work.

Figure 6(a)–(b) shows that although the conditional Granger approach is theoretically correct and works well for small network size with sufficient samples, it suffers from increasing inference error as the network size increases, and it becomes extremely inaccurate when the network size n starts to surpass the sample size T. Such limitations are overcome by the oCSE approach, where both the false positive and false negative ratios remain close to zero as the network size increases. The reason that oCSE is accurate even as n increases is that it builds the causal parent set in an *aggregative* manner, therefore relying only on estimating entropy in relatively low dimensions (here roughly the same dimension as the number of causal parents per node). In sharp contrast, the conditional Granger approach requires the estimation of entropy in the full n-dimensional space and therefore requires many (potentially exponentially) more samples to achieve the same accuracy when n becomes large.

---

[9]The accuracy of this empirical distribution and therefore the power of the permeation test increases with an increasing number of permutations r. However, as r increases, the computational complexity also increases, scaling roughly as a linear function of r.
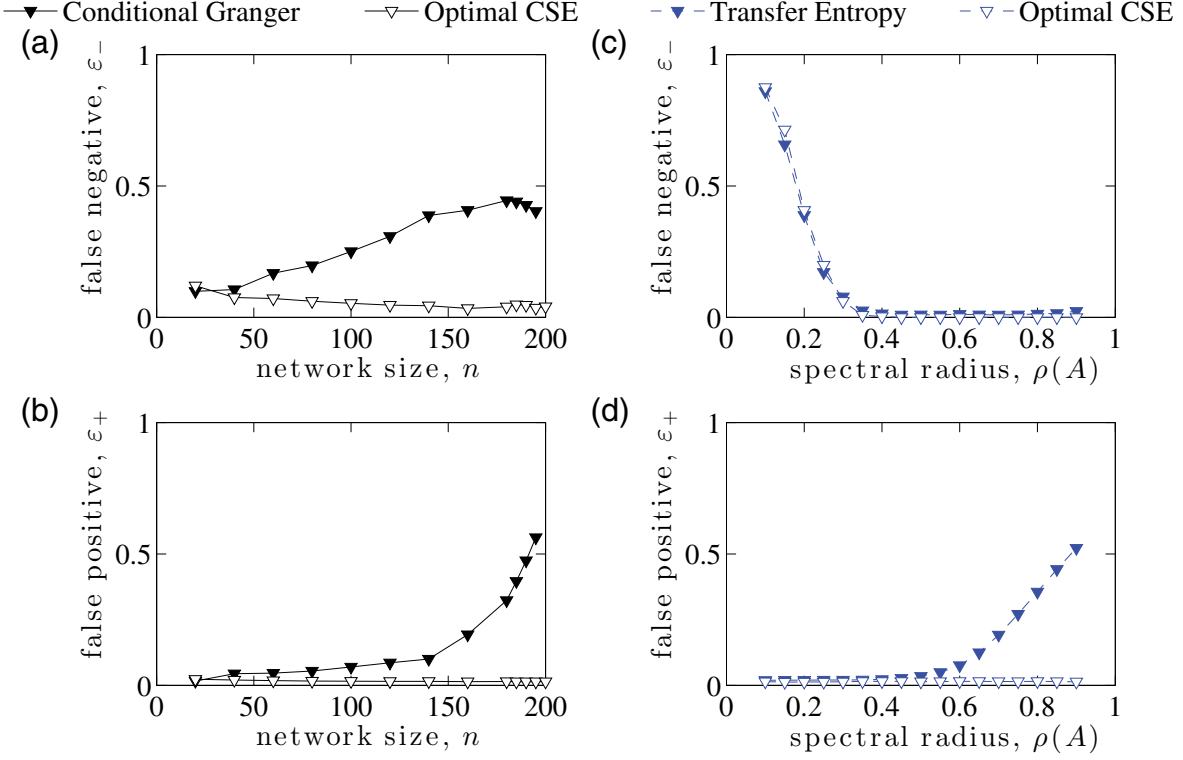
Figure 6. *Comparison of causal network inference approaches: Conditional Granger, transfer entropy, and oCSE. The time series are generated from the Gaussian process defined in (3.1) using signed Erdős–Rényi networks (see section 4.2 for details). Two types of inference error are examined—false negative and false positive ratios, defined in (4.1). (a)–(b) Inference error as a function of network size n using conditional Granger versus oCSE approaches. Here the networks have a fixed average degree $np = 10$ and spectral radius $\rho(A) = 0.8$. Sample size is $T = 200$. (c)–(d) Inference error as a function of the spectral radius $\rho(A)$ using transfer entropy versus oCSE approaches. Here the networks have a fixed number of nodes $n = 200$ and average degree $np = 10$. Sample size is $T = 2000$. For all three approaches we apply the permutation test using $r = 100$ permutations and significance threshold $\theta = 99\%$. Each data point is obtained from averaging over 20 independent simulations of the network dynamics, (3.1).*

Figure 6(c)–(d) shows that even for a sufficient number of samples, the transfer entropy approach without appropriate conditioning can lead to considerable inference error and is therefore inherently unsound for causal network inference. In particular, although inference by both transfer entropy and oCSE gives similar false negatives in the regime of $\rho(A) \approx 0$, where the dynamics is dominated by noise and not the causal dependencies, transfer entropy yields increasing false positives when the causal links dominate, $\rho(A) \to 1$. This is mainly due to the fact that as $\rho(A) \to 1$, indirect causal nodes become increasingly difficult to distinguish from direct ones without appropriate conditioning [71]. oCSE, on the other hand, consistently yields nearly zero false positive ratios in the entire range of $\rho(A)$. Interestingly, the spectral radius $\rho(A)$ can be interpreted as the *information diffusion rate* on networks and is found to be very close to criticality (i.e., $\rho(A) \approx 1$) in neuronal networks [39, 42].

These numerical experiments highlight that whereas the conditional Granger approach
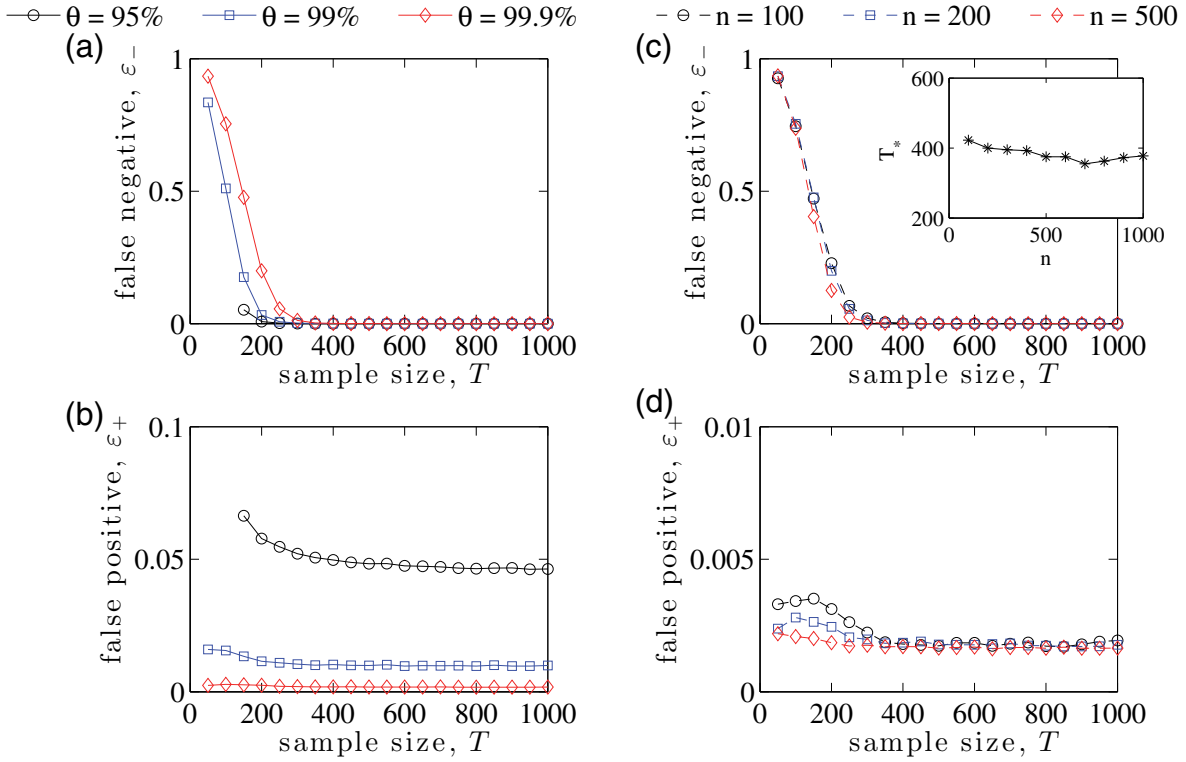
Figure 7. *Performance of the oCSE approach for causal network inference with different significance thresholds for networks of various sizes. The time series are generated from the Gaussian process defined in* (3.1) *using signed Erdős–Rényi networks (see section* 4.2 *for details). False negative ratio (upper row) and false positive ratio (lower row) are defined in* (4.1). *(a)–(b) Inference error as a function of sample size T for various significance levels θ used in the permutation test. Here networks have $n = 200$ nodes with expected average degree $np = 10$ and information diffusion rate $\rho(A) = 0.8$. (c)–(d) Inference error as a function of sample size T for various network sizes. Here networks have the same expected average degree $np = 10$ and information diffusion rate $\rho(A) = 0.8$, and we use $r = 1000$ permutations in the permutation test with $\theta = 0.999$. Note that all three false negative curves in* (c) *appear to converge for $T \approx 300$. The critical sample size T_* (defined as the minimum T for which $\varepsilon_- < 1 - \theta$) as a function of the network size n is shown in the inset of* (c), *suggesting the absence of scaling of T_* in terms of n. Each data point is obtained from averaging over* 20 *independent simulations of the network dynamics,* (3.1).*

is inaccurate for $T \lesssim n$ and the transfer entropy approach is inaccurate when $\rho(A) \lesssim 1$, the proposed oCSE approach overcomes both limitations and yields almost exact network inference even for rather limited sample size.

4.4. Performance of optimal causation entropy approach for causal network inference.
Having established the advantages of the oCSE approach, we now examine its performance under various parameter settings.

First, we examine the effect of the significance threshold θ on the inference error. As shown in Figure 7(a)–(b), the false negative ratio ϵ_- does not seem to depend on θ and converges to zero as sample size T increases. On the other hand, as $T \to \infty$, the false positive ratio saturates at the level $\epsilon_+ \sim (1 - \theta)$, which is consistent with the implementation of the

permutation test which rejects the null hypothesis at θ. This observation suggests that in order to achieve higher accuracy given sufficient sample size, one should choose θ as close to 1 as possible. The tradeoff in practice is that reliable implementation using larger θ requires an increasing number of permutations and therefore increases the computational complexity of the inference algorithms.

Next, we investigate the effect of sample size T on the inference error for networks of different sizes. The results are shown in Figure 7(c)–(d). As expected, when T increases, the false negative ratio decreases toward zero. Somewhat unexpectedly, the false positive ratio stays close to zero (in fact, close to the significance level θ) even for relatively small sample size (T as small as 50 for networks of up to 500 nodes). Furthermore, it appears that for networks of different sizes but the same average degree and information diffusion rate, the false negative ratios drop close to zero almost at the same sample size. To better quantify these effects, we define the *critical sample size* T_* as the smallest number of samples for which the false negative ratio falls below $1 - \theta$. As shown in the inset of Figure 7(c), for networks with the same average degree and information diffusion rate, the critical sample size T_* remains mostly constant despite the increase of the network size. This result is unexpected. Traditionally, the network size n represents a lower bound on sample size T as any covariance matrix (e.g., application of the conditional Granger requires that $T > n$ for the invertibility of the covariance matrices). Our result surprisingly indicates that sample size T does not need to scale with network size n for accurate network inference and highlights the fact that the oCSE approach is *scalable* and *data efficient*, with accuracy depending *not* on the size of the network but rather on other network characteristics such as the density of links and spectral radius (information diffusion rate).

To strengthen our claim that for Erdős–Rényi networks, performance of the causal inference by the oCSE approach depends on the density of links as measured by average degree and information diffusion rate as measured by the spectral radius rather than network size, we further investigate the dependence of inference error on these two additional parameters, np and $\rho(A)$. As shown in Figure 8(a), for networks of the same size $n = 200$ with fixed $\rho(A) = 0.8$, the larger the average degree np, the larger the number of samples required to reduce the false negative ratio to zero. In fact, as shown in the inset of Figure 8(a), the critical sample T_* to reach $\varepsilon_- < 1 - \theta$ appears to scale *linearly* as a function of the average degree np, but not the network size (see the inset of Figure 7(c)). On the other hand, Figure 8(c)–(d) shows that the information diffusion rate, $\rho(A)$, seems to pose a harder constraint on accurate network inference: the smaller it is, the more samples are needed for accuracy. In particular, as shown in the inset of Figure 8(c), the critical sample size appears to increase *exponentially* as $\rho(A)$ decreases toward zero. Interestingly, as shown in Figure 8(b),(d), the false positive ratios in both cases remain close to its saturation level around $1 - \theta = 10^{-3}$ even for very small sample size ($T \sim 50$), and this holds across networks with different average degree and different size (also see Figure 7(d)).

To briefly summarize these numerical experiments, we found that for the Gaussian process, practical causal network inference by the proposed oCSE overcomes fundamental limitations of previous approaches including conditional Granger and transfer entropy. One important advantage of the oCSE approach, as suggested by the numerical results, is that it often requires a relatively small number of samples to achieve high accuracy, making it a data-efficient
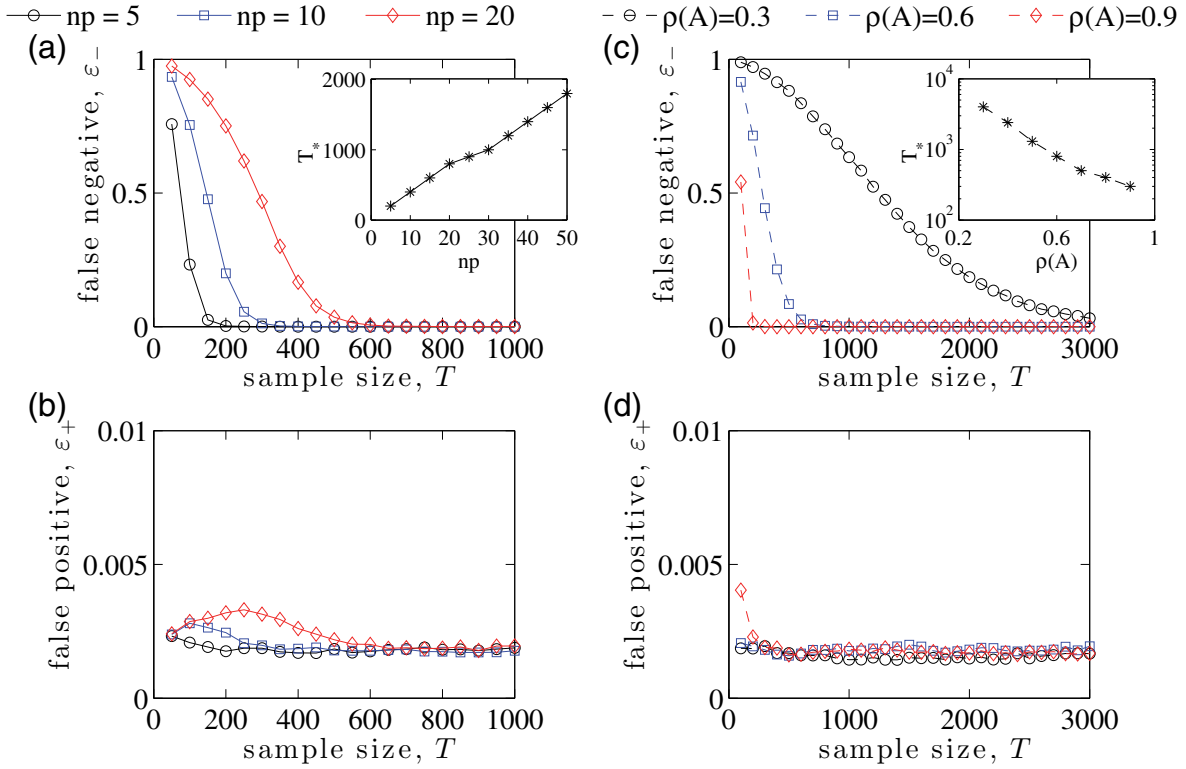
Figure 8. *Performance of the oCSE approach for causal network inference for networks with different average degree and spectral radius. The time series are generated from the Gaussian process defined in (3.1) using signed Erdős–Rényi networks (see section 4.2 for details). False negative ratio (upper row) and false positive ratio (lower row) are defined in (4.1). (a)–(b) Inference error as a function of sample size for networks with various average degree np. Here the networks have the same size $n = 200$ and spectral radius $\rho(A) = 0.8$. The inset shows the critical sample size T_* (see text) as a function of np. (c)–(d) Inference error as a function of sample size for networks with various special radii $\rho(A)$. Here the networks have the same size $n = 200$ and average degree $n = 10$. The permutation test used for the data in all panels involve $r = 1000$ permutations with the significance threshold $\theta = 0.999$. Each data point is obtained from averaging over 20 independent simulations of the network dynamics, (3.1).*

method to use in practice. In fact, we found that for Erdős–Rényi networks, the critical number of samples required for the false negatives to vanish does not depend on the network size but rather depends on the density of links (as measured by average degree) and the information diffusion rate (as measured by the spectral radius of the network adjacency matrix). This is somewhat surprising because traditionally the network size poses as an absolute lower bound for the sample size in order for proper inversion of the covariance matrix (recent advances such as *Lasso* have partially resolved this issue by making specific assumptions about the model form and utilizing l_1 optimization techniques [22, 73]). On the other hand, our numerical results also suggest that only a very small number of samples is needed for the false positives to reach saturation level. This level is inherently set by the significance threshold used in the permutation test rather than other network characteristics and can be systematically reduced by increasing the significance threshold and the number of permutations.

5. Discussion and conclusion. Although time series analysis is broadly utilized for scientific research, the inference of large networks from relatively short times series data, and in particular causal networks describing "cause-and-effect" relationships, has largely remained unresolved. The main contribution of this paper includes the theoretical development of causation entropy, an information-theoretic statistic designed for causality inference. Causation entropy can be regarded as a type of conditional mutual information which generalizes the traditional, unconditioned version of transfer entropy. When applied to Gaussian variables, causation entropy also generalizes Granger causality and conditional Granger causality. We proved that for a general network stochastic process, the causal parents of a given node are exactly the minimal set of nodes that maximizes causation entropy, a key result which we refer to as the *optimal causation entropy principle* (see Definition 2.1). Based on this principle, we introduced an algorithm for causal network inference called oCSE, which utilizes two steps to jointly infer the set of causal parents of each node.

The effectiveness and data efficiency of the proposed oCSE approach were illustrated through numerical simulation of a Gaussian process on large-scale random networks. In particular, our numerical results show that the proposed oCSE approach consistently outperforms previous conditional Granger (with full conditioning) and transfer entropy approaches. Furthermore, inference accuracy using the oCSE approach generally requires fewer samples and fewer computations due to its aggregative nature: the conditioning set encountered in entropy estimation remains relatively low-dimensional for sparse networks. The number of samples required for the desired accuracy does not appear to depend on network size but rather on the density of links (or, equivalently, the average degree of the nodes) and the spectral radius (which measures the average rate at which information transfers across links). This makes oCSE a promising tool for the inference of networks, in particular large-scale sparse causal networks, as found in a wide range of real-world applications [6, 19, 48, 49]. Therefore, we wish to emphasize that among all the details we presented herein, our oCSE-based algorithmic development (aggregative discovery jointly with progressive removal) is the most central contribution, serving as a computational tool to systematically infer casual relationships from data generated by a complex interrelated process. In principle, we expect our two-step process given by Algorithms 2.1 and 2.2 to also be effective for network inference when the statistic is not necessarily causation entropy.

We conclude by noting several problems that remain to be tackled. First, for general stochastic processes, exact expression of entropy is rarely obtainable. Practical application of the oCSE therefore requires the development of nonparametric statistics for estimating causation entropy for general multidimensional random variables. An ideal estimation method should rely on as few assumptions about the form of the underlying variable as possible and be able to achieve the desired accuracy even for relatively small sample size. Several existing methods, including various binning techniques [62] and k-nearest neighbor estimates [40], seem promising, but further exploration is necessary to examine their effectiveness [33]. Second, temporal stationarity assumptions are often violated in real-world applications. It is therefore of critical importance to divide the observed time series data into stationary segments [77], allowing for the inference of causal networks that are *time-dependent* [45]. Finally, information causality suggests physical causality, but they are not necessarily equivalent [33, 53]. It is our goal to put this notion onto a more rigorous footing and further explore their relationships.

Appendix A. Causal inference of finite-order Markov processes. The main body of the paper deals with causal inference of a first-order stationary Markov process. Such a framework can in fact be extended to any finite-order stationary Markov processes. The idea is to convert a finite-order process to a first-order process and define nodes in the causal network to be variables at different time layers.

Consider a stationary Markov process $\{Z_t\}$ of order τ, which satisfies

$$(A.1) \qquad\qquad p(Z_t|Z_{t^-}) = p(Z_t|Z_{t-1},\ldots,Z_{t-\tau}),$$

where $Z_{t^-} = [Z_{t-1}, Z_{t-2}, \ldots]$ denotes the infinite past of Z_t. Define a delay vector

$$(A.2) \qquad\qquad X_t = [Z_t, \ldots, Z_{t-\tau+1}].$$

Then, for every $x_t = [z_t, z_{t-1}, \ldots, z_{t-\tau+1}]$ and x_{t^-},

$$
\begin{aligned}
p(X_t = x | X_{t^-} = x_{t^-}) &= p(X_t = x_t | Z_{t-1} = z_{t-1}, Z_{t-2} = z_{t-2}, \ldots) \\
&= p(X_t = x_t | Z_{t-1} = z_{t-1}, Z_{t-2} = z_{t-2}, \ldots, Z_{t-\tau} = z_{t-\tau}) \\
(A.3) \qquad &= p(X_t = x_t | X_{t-1} = x_{t-1}),
\end{aligned}
$$

where the last step follows from (A.1) and the definition of X_t. See Figure 9 for an example with $\tau = 2$. This shows that the process $\{X_t\}$ is indeed a first-order Markov process. The inference of the causal network is therefore converted into the identification of the causal parents of the nodes corresponding to $\{Z_t\}$ in the equivalent first-order process, for which the results in the main body of the paper apply so long as the conditions in (2.8) are met.

In practice, if the order of the underlying Markov process is *unknown*, then one needs to estimate it before being able to turn the process into a first-order process. The determination of Markov order has been a long-standing problem and is traditionally addressed by performing hypothesis tests based on computing a χ^2 statistic [4]. The main disadvantage is that the χ^2 distribution is only valid in the infinite-sample limit. A breakthrough was made recently by Pethel and Hahs [54], who developed a relatively efficient procedure for surrogate data generation which yields an exact test statistic valid for arbitrary sample size at the expense of increased computational burden.

Appendix B. Necessity of the faithfulness assumption. The faithfulness assumption is necessary for the "true positive" statement in Theorem 2.2(c) to be valid. To illustrate this point, consider a network of three nodes X, Y, and Z, and let

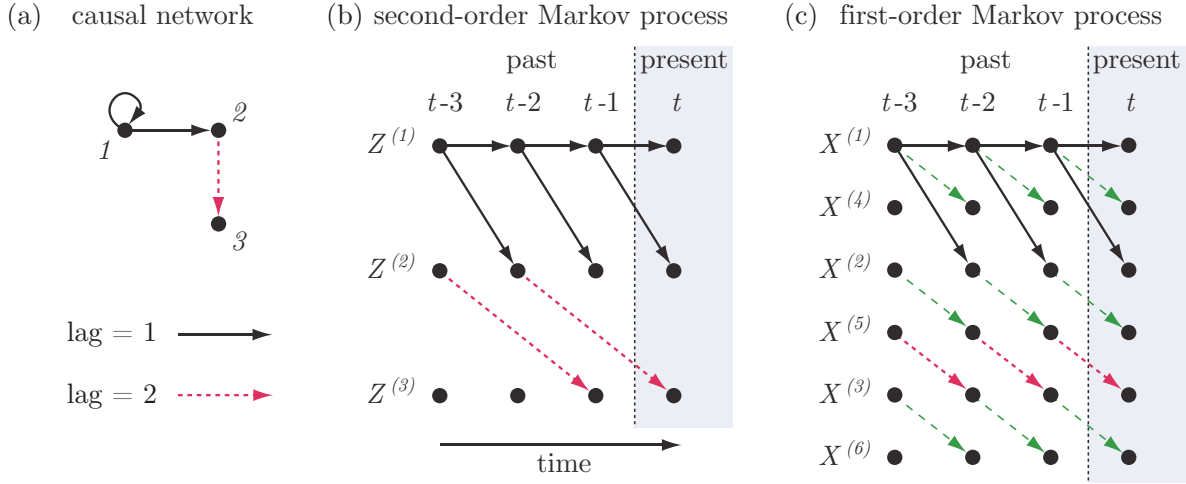$$(B.1) \qquad\qquad X_{t+1} = Y_t \oplus Z_t,$$

where $\oplus$ denotes the "exclusive or" (xor) operation and Y_t and Z_t are Bernoulli random variables with probabilities

$$(B.2) \qquad P(Y_t = 0) = P(Y_t = 1) = P(Z_t = 0) = P(Z_t = 1) = 0.5.$$

It follows that

$$(B.3) \qquad\qquad C_{Y \to X} = C_{Z \to X} = 0.$$

Figure 9. *Converting a high-order Markov process into a first-order Markov process by making multiple instances of nodes.* (a) *A second-order Markov process on $n = 3$ nodes, where causal relationships are across time lags of either 1 or 2 time steps. We denote by $Z_t^{(i)}$ the state of node i at time t.* (b) *The flow of information for the second-order Markov process. Each row corresponds to a given node $i \in \{1, 2, 3\}$, and each column corresponds to the nodes' states $\{Z_t^{(i)}\}$ at a particular time t. Solid and dotted lines denote causal relationships across a time lag of 1 and 2 time steps, respectively.* (c) *The flow of information for the equivalent first-order Markov process. Each row corresponds to a given node $i \in \{1, 2, \ldots, 2n\}$, and each column corresponds to the nodes' states $\{X_t^{(i)}\}$ at a particular time t. For $i \in \{1, 2, 3\}$, the new variables $\{X_t^{(i)}\}$ are defined by $X_t^{(i)} = Z_t^{(i)}$ and $X_t^{(n+i)} = Z_{t-1}^{(i)} = X_{t-1}^{(i)}$. For Markov processes of order τ, one can use the more general transformation $X_t^{((s-1)n+i)} = Z_{t-s+1}^{(i)}$ for nodes $i \in \{1, \ldots, n\}$ and $s \in \{1, 2, \ldots, \tau\}$.*

However,

$$(B.4) \qquad\qquad C_{(Y,Z) \to X} = \log 2 > 0.$$

This results from the fact that multiple random variables can be mutually independent but not jointly independent. Expressed in terms of causal inference, it is possible that several variables jointly cause another variable, and this causal relationship cannot be decomposed. Such occurrences are believed to be rare and are often explicitly excluded by making the faithfulness/stability assumption [48]. For instance, in our example above, it occurs only when all the discrete probabilities are exactly uniform, $p = 0.5$, a situation that is unstable under perturbations. We exclude this situation from our study by imposing condition (3) in (2.8).

REFERENCES

[1] N. A. AHMED AND D. V. GOKHALE, *Entropy expressions and their estimators for multivariate distributions*, IEEE Trans. Inform. Theory, 35 (1989), pp. 688–692.

[2] N. Ancona, D. Marinazzo, and S. Stramaglia, *Radial basis function approach to nonlinear Granger causality of time series*, Phys. Rev. E, 70 (2004), 056221.

[3] L. Barnett, A. B. Barrett, and A. K. Seth, *Granger causality and transfer entropy are equivalent for Gaussian variables*, Phys. Rev. Lett., 103 (2009), 238701.

[4] M. S. Bartlett, *The frequency goodness of fit for probability chains*, Math. Proc. Cambridge Philos. Soc., 47 (1951), pp. 86–95.

[5] A. Y. Barraud, *A numerical algorithm to solve $AXA - X = Q$*, IEEE Trans. Automat. Control, 22 (1977), pp. 883–885.

[6] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical Processes on Complex Networks*, Cambridge University Press, Cambridge, UK, 2008.

[7] A. J. Bell, *The co-information lattice*, in Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 03), Nara, Japan, 2003, pp. 921–926.

[8] B. Bollobás, *Random Graphs*, 2nd ed., Academic Press, New York, 2001.

[9] E. Bollt, *Synchronization as a process of sharing and transferring information*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 22 (2012), 1250261.

[10] P. J. Brockwell, *Time Series Analysis: Encyclopedia of Statistics in Behavioral Science*, John Wiley & Sons, Hoboken, NJ, 2005.

[11] D. S. Bassett and E. Bullmore, *Small-world brain networks*, Neuroscientist, 6 (2006), pp. 512–523.

[12] E. Bullmore and O. Sporns, *Complex brain networks: Graph theoretical analysis of structural and functional systems*, Nat. Rev. Neurosci., 10 (2009), pp. 186–198.

[13] N. Chen, *On the approximability of influence in social networks*, SIAM J. Discrete Math., 23 (2009), pp. 1400–1415.

[14] Y. Chen, G. Rangarajan, J. Feng, and M. Ding, *Analyzing multiple nonlinear time series with extended Granger causality*, Phys. Lett. A, 324 (2004), pp. 26–35.

[15] S. P. Cornelius, W. L. Kath, and A. E. Motter, *Realistic control of network dynamics*, Nat. Commun., 4 (2013), 1942.

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., John Wiley & Sons, Hoboken, NJ, 2006.

[17] G. Craciun and M. Feinberg, *Multiple equilibria in complex chemical reaction networks: Semiopen mass action systems*, SIAM J. Appl. Math., 70 (2010), pp. 1859–1877.

[18] F. Dörfler and F. Bullo, *Synchronization and transient stability in power networks and nonuniform Kuramoto oscillators*, SIAM J. Control Optim., 50 (2012), pp. 1616–1642.

[19] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, *Critical phenomena in complex networks*, Rev. Modern Phys., 80 (2008), 1275.

[20] M. L. Eaton, *Multivariate Statistics: A Vector Space Approach*, John Wiley & Sons, New York, 1983.

[21] S. Frenzel and B. Pompe, *Partial mutual information for coupling analysis of multivariate time series*, Phys. Rev. Lett., 99 (2007), 204101.

[22] J. Friedman, T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics, 9 (2008), pp. 432–441.

[23] Q. Gao, X. Duan, and H. Chen, *Evaluation of effective connectivity of motor areas during motor imagery and execution using conditional Granger causality*, NeuroImage, 54 (2011), pp. 1280–1288.

[24] W. R. Garner, *Uncertainty and Structure as Psychological Concepts*, John Wiley & Sons, New York, 1962.

[25] I. Gelfand, *Normierte Ringe*, Rech. Math. [Mat. Sbornik] N.S., 9 (51) (1941), pp. 3–24.

[26] J. W. Gibbs, *Elementary Principles in Statistical Mechanics*, Dover, New York, 1960.

[27] M. Golubitsky, I. Stewart, and A. Török, *Patterns of synchrony in coupled cell networks with multiple arrows*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 78–100.

[28] P. Good, *Permutation, Parametric and Bootstrap Tests of Hypotheses*, Springer, New York, 2005.

[29] C. W. J. Granger, *Investigating causal relations by econometric models and cross-spectral methods*, Econometrica, 37 (1969), pp. 425–438.

[30] C. W. J. Granger, *Some recent developments in a concept of causality*, J. Econometrics, 39 (1988), pp. 199–211.

[31] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Process*, 3rd ed., Oxford University Press, Oxford, UK, 2001.

[32] S. Guo, A. K. Seth, K. M. Kendrick, C. Zhou, and J. Feng, *Partial Granger causality—eliminating exogenous inputs and latent variables*, J. Neurosci. Methods, 172 (2008), pp. 79–93.

[33] D. W. Hahs and S. D. Pethel, *Distinguishing anticipation from causality: Anticipatory bias in the estimation of information flow*, Phys. Rev. Lett., 107 (2011), 128701.

[34] J. J. Heckman, *Econometric causality*, Int. Stat. Rev., 76 (2008), pp. 1–27.

[35] F. Heider, *Social perception and phenomenal causality*, Psychol. Rev., 51 (1944), pp. 358–374.

[36] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed., Cambridge University Press, Cambridge, UK, 2013.

[37] A. Kaiser and T. Schreiber, *Information transfer in continuous processes*, Phys. D, 166 (2002), pp. 43–62.

[38] J. Kleinberg, *The small-world phenomenon: An algorithmic perspective*, in Proceedings of the 32nd ACM Symposium on Theory of Computing, ACM, New York, 2000, pp. 163–170.

[39] O. Kinouchi and M. Copelli, *Optimal dynamical range of excitable networks at criticality*, Nat. Phys., 2 (2006), 348.

[40] A. Kraskov, H. Stögbauer, and P. Grassberger, *Estimating mutual information*, Phys. Rev. E, 69 (2004), 066138.

[41] O. Kuchaiev, M. Rašajski, D. J. Higham, and N. Pržulj, *Geometric de-noising of protein-protein interaction networks*, PLoS Comput. Biol., 5 (2009), e1000454.

[42] D. B. Larremore, W. L. Shew, and J. G. Restrepo, *Predicting criticality and dynamic range in complex networks: Effects of topology*, Phys. Rev. Lett., 106 (2011), 058101.

[43] A. Lasota and M. C. Mackey, *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*, 2nd ed., Springer-Verlag, New York, 1994.

[44] S. L. Lauritzen, *Graphical Models*, Oxford University Press, Oxford, UK, 1996.

[45] A. V. Mantzaris, D. S. Bassett, N. F. Wymbs, E. Estrada, M. A. Porter, P. J. Mucha, S. T. Grafton, and D. J. Higham, *Dynamic network centrality summarizes learning in the human brain*, J. Complex Networks, 1 (2013), pp. 83–92.

[46] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano, *ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context*, BMC Bioinformatics, 7 (Suppl 1) (2006), S7.

[47] W. J. McGill, *Multivariate information transmission*, Psychometrika, 19 (1954), pp. 97–116.

[48] M. E. J. Newman, *The structure and function of complex networks*, SIAM Rev., 45 (2003), pp. 167–256.

[49] M. E. J. Newman, *Networks: An Introduction*, Oxford University Press, Oxford, UK, 2010.

[50] T. Nishikawa and A. E. Motter, *Network synchronization landscape reveals compensatory structures, quantization, and the positive effect of negative interactions*, Proc. Natl. Acad. Sci. USA, 107 (2010), pp. 10342–10347.

[51] M. Paluš, V. Komárek, Z. Hrnčíř, and K. Štěrbová, *Synchronization as adjustment of information rates: Detection from bivariate time series*, Phys. Rev. E, 63 (2001), 046211.

[52] L. Paninski, *Estimation of entropy and mutual information*, Neural Comput., 15 (2003), pp. 1191–1253.

[53] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed., Cambridge University Press, Cambridge, UK, 2009.

[54] S. D. Pethel and D. W. Hahs, *Exact significance test for Markov order*, Phys. D, 269 (2014), pp. 42–47.

[55] A. Pomerance, E. Ott, M. Girvan, and W. Losert, *The effect of network topology on the stability of discrete state models of genetic control*, Proc. Natl. Acad. Sci. USA, 106 (2009), pp. 8209–8214.

[56] B. Ravoori, A. B. Cohen, J. Sun, A. E. Motter, T. E. Murphy, and R. Roy, *Robustness of optimal synchronization in real networks*, Phys. Rev. Lett., 107 (2011), 034102.

[57] K. J. Rothman and S. Greenland, *Causation and causal inference in epidemiology*, Am. J. Public Health, 95 (2005), pp. S144–S150.

[58] H. L. Royden, *Real Analysis*, 3rd ed., Prentice–Hall, Englewood Cliffs, NJ, 1988.

[59] W. J. Rugh, *Linear System Theory*, Prentice–Hall, Englewood Cliffs, NJ, 1993.

[60] J. Runge, J. Heitzig, V. Petoukhov, and J. Kurths, *Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy*, Phys. Rev. Lett., 108 (2012), 258701.

[61] J. Runge, J. Heitzig, N. Marwan, and J. Kurths, *Escaping the curse of dimensionality in estimating multivariate transfer entropy*, Phys. Rev. E, 86 (2012), 061121.

[62] K. Hlaváčková-Schindlera, M. Paluš, M. Vejmelka, and J. Bhattacharya, *Causality detection based on information-theoretic approaches in time series analysis*, Phys. Rep., 441 (2007), pp. 1–46.

[63] T. Schreiber, *Measuring information transfer*, Phys. Rev. Lett., 85 (2000), 461.

[64] C. E. Shannon, *A mathematical theory of communication*, Bell System Tech. J., 27 (1948), pp. 379–423, 623–656.

[65] D. A. Smirnov, *Spurious causalities with transfer entropy*, Phys. Rev. E, 87 (2013), 042917.

[66] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd ed., MIT Press, Cambridge, MA, 2000.

[67] D. J. Stilwell, E. M. Bollt, and D. G. Roberson, *Sufficient conditions for fast switching synchronization in time-varying network topologies*, SIAM J. Appl. Dyn. Syst., 5 (2006), pp. 140–156.

[68] G. Stolovitzky, D. Monroe, and A. Califano, *Dialogue on Reverse-Engineering Assessment and Methods: The DREAM of high-throughput pathway inference*, Ann. New York Acad. Sci., 1115 (2007), pp. 1–22.

[69] M. Studený and J. Vejnarová, *The multiinformation function as a tool for measuring stochastic dependence*, in Learning in Graphical Models, M. I. Jordan, ed., MIT Press, Cambridge, MA, 1998, pp. 261–297.

[70] J. Sun and A. E. Motter, *Controllability transition and nonlocality in network control*, Phys. Rev. Lett., 110 (2013), 208701.

[71] J. Sun and E. M. Bollt, *Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings*, Phys. D, 267 (2014), pp. 49–57.

[72] D. Taylor and J. G. Restrepo, *Network connectivity during mergers and growth: Optimizing the addition of a module*, Phys. Rev. E, 83 (2011), 066112.

[73] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B Stat. Methodol., 58 (1996), pp. 267–288.

[74] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter, *Comparing community structure to characteristics in online collegiate social networks*, SIAM Rev., 53 (2011), pp. 526–543.

[75] M. Vejmelka and M. Palus, *Inferring the directionality of coupling with conditional mutual information*, Phys. Rev. E, 77 (2008), 026214.

[76] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, *Transfer entropy—a model-free measure of effective connectivity for the neurosciences*, J. Comput. Neurosci., 30 (2011), pp. 45–67.

[77] B. Wang, J. Sun, and A. E. Motter, *Detecting structural breaks in seasonal time series by regularized optimization*, in Safety, Reliability, Risk and Life-Cycle Performance of Structures and Infrastructures, G. Deodatis, B. R. Ellingwood, and D. M. Frangopol, eds., CRC Press, Boca Raton, FL, 2014, pp. 3621–3628.

[78] S. Watanabe, *Information theoretical analysis of multivariate correlation*, IBM J. Res. Develop., 4 (1960), pp. 66–82.

[79] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature, 393 (2000), pp. 440–442.