



An Information Criterion for Inferring Coupling of Distributed Dynamical Systems

Oliver M. Cliff^{1*}, Mikhail Prokopenko² and Robert Fitch^{1,3}

¹ Australian Centre for Field Robotics, The University of Sydney, Sydney, NSW, Australia, ² Complex Systems Research Group, The University of Sydney, Sydney, NSW, Australia, ³ Centre for Autonomous Systems, University of Technology Sydney, Sydney, NSW, Australia

The behavior of many real-world phenomena can be modeled by non-linear dynamical systems whereby a latent system state is observed through a filter. We are interested in interacting subsystems of this form, which we model by a set of coupled maps as a synchronous update graph dynamical system. Specifically, we study the structure learning problem for spatially distributed dynamical systems coupled via a directed acyclic graph. Unlike established structure learning procedures that find locally maximum posterior probabilities of a network structure containing latent variables, our work exploits the properties of dynamical systems to compute globally optimal approximations of these distributions. We arrive at this result by the use of time delay embedding theorems. Taking an information-theoretic perspective, we show that the log-likelihood has an intuitive interpretation in terms of information transfer.

OPEN ACCESS

Edited by:

Michael Wibral,
Goethe University Frankfurt, Germany

Reviewed by:

Robin A. A. Ince,
University of Manchester, UK
Raul Vicente,
Max Planck Society, Germany

*Correspondence:

Oliver M. Cliff
o.cliff@acfr.usyd.edu.au

Specialty section:

This article was submitted to
Computational Intelligence,
a section of the journal
Frontiers in Robotics and AI

Received: 19 August 2016

Accepted: 31 October 2016

Published: 28 November 2016

Citation:

Cliff OM, Prokopenko M and Fitch R
(2016) An Information Criterion for
Inferring Coupling of Distributed
Dynamical Systems.
Front. Robot. AI 3:71.
doi: 10.3389/frobt.2016.00071

Keywords: complex networks, structure learning, dynamic Bayesian networks, graph dynamical systems, information theory, dynamical systems, state space reconstruction

1. INTRODUCTION

Complex systems are broadly defined as systems that comprise interacting non-linear components (Boccaletti et al., 2006). Discrete-time complex systems can be represented using graphical models such as *graph dynamical systems (GDSs)* (Mortveit and Reidys, 2001; Wu, 2005), where spatially distributed dynamical units are coupled via a directed graph. The task of learning the structure of such a system is to infer directed relationships between variables; in the case of dynamical systems, these variables are typically hidden (Kantz and Schreiber, 2004). In this paper, we study the *structure learning* problem for complex networks of non-linear dynamical systems coupled via a directed acyclic graph (DAG). Specifically, we formulate synchronous update GDSs as dynamic Bayesian networks (DBNs) and study this problem from the perspective of information theory.

The structure learning problem for distributed dynamical systems is a precursor to inference in systems that are not fully observable. This case encompasses many practical problems of known artificial, biological, and chemical systems, such as neural networks (Lizier et al., 2011; Vicente et al., 2011; Schumacher et al., 2015), multi-agent systems (Xu et al., 2013; Gan et al., 2014; Cliff et al., 2016; Umenberger and Manchester, 2016), and various others (Boccaletti et al., 2006). Modeling a partially observable system as a dynamical network presents a challenge in synthesizing these models and capturing their global properties (Boccaletti et al., 2006). In addressing this challenge, we draw on probabilistic graphical models (specifically Bayesian network (BN) structure learning) and non-linear time series analysis (differential topology).

In this paper, we exploit the properties of discrete-time multivariate dynamical systems in inferring coupling between latent variables in a DAG. Specifically, the main focus of this paper is to analytically derive a measure (score) for evaluating the fitness of a candidate DAG, given data. We assume the data are generated by a certain family of multivariate dynamical system and are thus able to overcome the issue of latent variables faced by established structure learning algorithms. That is, under certain assumptions of the dynamical system, we are able to employ time delay embedding theorems (Stark et al., 2003; Deyle and Sugihara, 2011) to compute our scores.

Our main result is a tractable form of the log-likelihood function for synchronous GDSs. Using this result, we are able to directly compute the *Bayesian information criterion (BIC)* (Schwarz, 1978) and *Akaike information criterion (AIC)* (Akaike, 1974) and thus achieve globally optimal approximations of the posterior distribution of the graph. Finally, we show that the log-likelihood and log-likelihood ratio can be expressed in terms of *collective transfer entropy* (Lizier et al., 2010; Vicente et al., 2011). This result places our work in the context of effective network analysis (Sporns et al., 2004; Park and Friston, 2013) based on information transfer (Honey et al., 2007; Lizier et al., 2011; Cliff et al., 2013, 2016) and relates to the information processing intrinsic to distributed computation (Lizier et al., 2008).

2. RELATED WORK

We are interested in classes of systems whereby dynamical units are coupled via a graph structure. These types of systems have been studied under several names, including complex dynamical networks (Boccaletti et al., 2006), spatially distributed dynamical systems (Kantz and Schreiber, 2004; Schumacher et al., 2015), master-slave configurations (or systems with a skew product structure) (Kocarev and Parlitz, 1996), and coupled maps (Kaneko, 1992). Common to each of these definitions is that the multivariate state of the system comprises individual subsystem states, the dynamics of which are given by a set of either discrete-time maps or first-order ordinary differential equations (ODEs), called a flow. We assume the discrete-time formulation, where a map can be obtained numerically by integrating differential equations or recording experimental data (observations) at discrete-time intervals (Kantz and Schreiber, 2004). The literature on coupled dynamical systems is often focused on the analysis of characteristics such as stability and synchrony of the system. In this work, we draw on the fields of BN structure learning and non-linear time series analysis to infer coupling between spatially distributed dynamical systems.

BN structure learning comprises two subproblems: *evaluating* the fitness of a graph and *identifying* the optimal graph given this fitness criterion (Chickering, 2002). The evaluation problem is particularly challenging in the case of graph dynamical systems, which include both latent and observed variables. A number of theoretically optimal techniques exist for the evaluation problem for BNs with complete data (Bouckaert, 1994; Lam and Bacchus, 1994; Heckerman et al., 1995), which have been extended to DBNs (Friedman et al., 1998). With incomplete data, however, the common approach is to resort to approximations that find local

optima, e.g., expectation-maximization (EM) (Friedman et al., 1998; Ghahramani, 1998). An additional caveat with respect to structure learning is that algorithms find an equivalence class of networks with the same Markov structure, and not a unique solution (Chickering, 2002).

In non-linear time series analysis, the problem of inferring coupling strength and causality in complex systems has received significant attention recently (Schreiber, 2000; Hoyer et al., 2009). Early work by Granger defined causality in terms of the predictability of one system linearly coupled to another (Granger, 1969). Although this measure is popular for identifying coupling, it requires systems are linear statistical models and is considered insufficient for inferring coupling between dynamical systems due to inseparability (Sugihara et al., 2012). Another method popular in neuroscience is transfer entropy, which was introduced to quantify the information transfer between non-linear (finite-order Markov) systems (Schreiber, 2000). Transfer entropy has been used to recover interaction networks in numerous fields such as multi-agent systems (Cliff et al., 2016) and effective networks in neuroscience (Lizier et al., 2011; Vicente et al., 2011; Lizier and Rubinov, 2012). More recently, researchers have used the additive noise model (Hoyer et al., 2009; Peters et al., 2011) to infer unidirectional cause and effect relationships with observed random variables and find a unique DAG (as opposed to an equivalence class). These studies have been extended by exploring weakly additive noise models for learning the structure of systems of observed variables with non-linear coupling (Gretton et al., 2009).

A recent approach to inferring causality is convergent cross-mapping (CCM), which is based on Takens theorem (Takens, 1981) and tests for causation (predictability) by considering the history of observed data of a hidden variable in predicting the outcome of another (Sugihara et al., 2012). Using a similar approach, Schumacher et al. (2015) used Stark's bundle delay embedding theorem (Stark, 1999; Stark et al., 2003) to predict one subsystem from another using Gaussian processes. This algorithm can thus be used to infer the driving systems in spatially distributed dynamical systems in a similar manner to our work. However, both papers do not consider the problem of inference over the entire network structure, or formally derive the measures used therein. In our work, we provide a rigorous proof based on established structure learning procedures and discuss the problem of inference within a distributed dynamical system.

3. BACKGROUND

This section summarizes relevant technical concepts used throughout the paper. First, a stochastic temporal process X is defined as a sequence of random variables (X_1, X_2, \dots, X_N) with a realization (x_1, x_2, \dots, x_N) for countable time indices $n \in \mathbb{N}$. Consider a collection of M processes, and denote the i th process X^i to have associated realization x_n^i at temporal index n , and x_n as all realizations at that index $x_n = \langle x_n^1, x_n^2, \dots, x_n^M \rangle$. If X_n^i is a discrete random variable, the number of values the variable can take on is denoted $|X_n^i|$. The following sections collect results from DBN literature, attractor reconstruction, and information theory that are relevant to this work.

3.1. Dynamic Bayesian Networks

DBNs are a general graphical representation of a temporal model, representing a probability distribution over infinite trajectories of random variables (Z_1, Z_2, \dots) compactly (Friedman et al., 1998). These models are a more expressive framework than the hidden Markov model (HMM) and Kalman filter model (KFM) (or linear dynamical system) (Friedman et al., 1998). In this work, we denote $Z_n = \{X_n, Y_n\}$ as the set of hidden and observed variables, respectively, where $n \in \{1, 2, \dots\}$ is the temporal index.

BNs $B = (G, \Theta)$ represent a joint distribution $p(z)$ graphically and consist of: a DAG G and a set of conditional probability distribution (CPD) parameters Θ . DBNs $B = (B_1, B_{\rightarrow})$ extend the BN to model temporal processes and comprise two parts: the prior BN $B_1 = (G_1, \Theta_1)$, which defines the joint distribution $p_{B_1}(z_1)$; and the *two-time-slice Bayesian network (2TBN)* $B_{\rightarrow} = (G_{\rightarrow}, \Theta_{\rightarrow})$, which defines a first-order Markov process $p_{B_{\rightarrow}}(z_{n+1}|z_n)$ (Friedman et al., 1998). This formulation allows for a variable to be conditioned on its respective parent set $\Pi_{G_{\rightarrow}}(Z_{n+1}^i)$ that can come from the preceding time slice or the current time slice, as long as G_{\rightarrow} forms a DAG. The 2TBN probability distribution factorizes according to G_{\rightarrow} with a local CPD p_D estimated from an observed dataset. That is, given a set of stochastic processes (Z_1, Z_2, \dots, Z_N) , the realization of which constitutes a dataset $D = (z_1, z_2, \dots, z_N)$, we obtain the 2TBN distribution as

$$p_{B_{\rightarrow}}(z_{n+1}|z_n) = \prod_i p_{B_{\rightarrow}}(z_{n+1}^i | \pi_{G_{\rightarrow}}(Z_{n+1}^i)), \quad (1)$$

where $\pi_{G_{\rightarrow}}(Z_{n+1}^i)$ denotes the (index-ordered) set of realizations $\{z_o^j : Z_o^j \in \Pi_{G_{\rightarrow}}(Z_{n+1}^i)\}$.

3.2. Embedding Theory

Embedding theory refers to methods from differential topology for inferring the (hidden) state of a dynamical system from a reconstructed sequence of observations. The state of a discrete-time dynamical system is given by a point x_n confined to a d -dimensional manifold \mathcal{M} . The time evolution of this state is described by a map $f: \mathcal{M} \rightarrow \mathcal{M}$, so that the sequence of states (x_n) is given by $x_{n+1} = f(x_n)$. In many situations, we only have access to a filtered, scalar representation of the state, i.e., the measurement $y_n = \psi(x_n)$ given by some *measurement function* $\psi: \mathcal{M} \rightarrow \mathbb{R}$ (Takens, 1981; Stark, 1999). The celebrated Takens' theorem (Takens, 1981) shows that for typical f and ψ , it is possible to reconstruct f from the observed time series up to some smooth coordinate change. More precisely, fix some κ (the *embedding dimension*) and some τ (the *time delay*), then define the *delay embedding map* $\Phi_{f,\psi}: \mathcal{M} \rightarrow \mathbb{R}^\kappa$ by

$$\Phi_{f,\psi}(x_n) = y_n^{(\kappa)} = \langle y_n, y_{n-\tau}, y_{n-2\tau}, \dots, y_{n-(\kappa-1)\tau} \rangle. \quad (2)$$

In differential topology, an *embedding* refers to a smooth map $\Psi: \mathcal{M} \rightarrow \mathcal{N}$ between manifolds \mathcal{M} and \mathcal{N} if it maps \mathcal{M} diffeomorphically onto its image; therefore, $\Phi_{f,\psi}$ has a smooth inverse $\Phi_{f,\psi}^{-1}$. The implication of Takens' theorem is that for typical f and ψ , the image $\Phi_{f,\psi}(\mathcal{M})$ of \mathcal{M} is completely equivalent to

\mathcal{M} itself, apart from the smooth invertible change of coordinates given by the mapping $\Phi_{f,\psi}$. An important consequence of this theorem is that we can define a map $F = \Phi_{f,\psi} \circ f \circ \Phi_{f,\psi}^{-1}$ on $\Phi_{f,\psi}$, such that $y_{n+1}^{(\kappa)} = F(y_n^{(\kappa)})$ (Stark, 1999).

There are technical assumptions for Takens' theorem (and the generalized versions employed herein) to hold. These assumptions require: (f, ψ) to be generic functions (in terms of Baire space), a restricted number of periodic points, and distinct eigenvalues at each neighborhood of these points (Takens, 1981; Stark, 1999; Stark et al., 2003; Deyle and Sugihara, 2011).

3.3. Information Theoretic Measures

Conditional entropy represents the uncertainty of a random variable X after taking into account the outcomes of another random variable Y by

$$H(X|Y) = - \sum_{x,y} p(x,y) \log_2 p(x|y). \quad (3)$$

Multivariate *transfer entropy* is a measure that computes the information transfer from a set of source processes to a set of destination process (Lizier et al., 2011). In this work, we use the formulation of collective transfer entropy (Lizier et al., 2010), where the information transfer from m source processes $V = \{Y^1, Y^2, \dots, Y^m\}$ to a single destination process Y can be decomposed as a sum of conditional entropy terms:

$$T_{V \rightarrow Y} = H(Y_{n+1} | Y_n^{(\kappa)}) - H(Y_{n+1} | Y_n^{(\kappa)}, \langle Y_n^{i,(\kappa^i)} \rangle), \quad (4)$$

where $Y_n^{i,(\kappa^i)} = \langle Y_n^i, Y_{n-\tau^i}^i, Y_{n-2\tau^i}^i, \dots, Y_{n-(\kappa^i-1)\tau^i}^i \rangle$ for some κ^i and τ^i , and similarly for $Y_n^{(\kappa)}$.

4. REPRESENTING NON-LINEAR DYNAMICAL NETWORKS AS DBNs

We express multivariate dynamical systems as a synchronous update GDS to allow for generic maps. With this model, we can express the time evolution of the GDS as a stationary DBN, and perform inference and learning on the subsequent graph. We formally state the network of dynamical systems as a special case of the sequential GDS (Mortveit and Reidys, 2001) with an observation function for each vertex.

Definition 1. Synchronous graph dynamical system (GDS). A *synchronous GDS* is a tuple $(G, x_n, y_n, \{f^i\}, \{\psi^i\})$ that consists of:

- a finite, directed graph $G = (\mathcal{V}, \mathcal{E})$ with edge-set $\mathcal{E} = \{E^i\}$ and M vertices comprising the vertex set $\mathcal{V} = \{V^i\}$;
- a multivariate state $x_n = \langle x_n^i \rangle$, composed of states for each vertex V^i confined to a d^i -dimensional manifold $x_n^i \in \mathcal{M}^i$;
- an M -variate observation $y_n = \langle y_n^i \rangle$, composed of scalar observations for each vertex $y_n^i \in \mathbb{R}$;
- a set of local maps $\{f^i\}$ of the form $f^i: \mathcal{M} \rightarrow \mathcal{M}^i$, which update synchronously and induce a global map $f: \mathcal{M} \rightarrow \mathcal{M}$; and
- a set of local observation functions $\{\psi^i\}$ of the form $\psi^i: \mathcal{M}^i \rightarrow \mathbb{R}$.

Without loss of generality, we can use local functions to describe the time evolution of the subsystems:

$$x_{n+1}^i = f^i(x_n^i, \langle x_n^{ij} \rangle_j) + v_{f^i} \quad (5)$$

$$y_{n+1}^i = \psi^i(x_{n+1}^i) + v_{\psi^i}. \quad (6)$$

Here, v_{f^i} is i.i.d. additive noise and v_{ψ^i} is noise that is either i.i.d. or dependent on the state, i.e., $v_{\psi^i}(x_{n+1}^i)$. The subsystem dynamics [equation (5)] are therefore a function of the subsystem state x_n^i and the subsystem parents' state $\langle x_n^{ij} \rangle_j$ at the previous time index such that $f^i : \mathcal{M}^i \times_j \mathcal{M}^{ij} \rightarrow \mathcal{M}^i$. Each subsystem observation is given by equation (6). We assume the functions $\{f^i\}$ and $\{\psi^i\}$ are invariant w.r.t. time and thus the graph G is stationary.

The time evolution of a synchronous GDS can be modeled as a DBN. First, each subsystem vertex $V^i = \{X_n^i, Y_n^i\}$ has an associated state variable X_n^i and observation variable Y_n^i ; the parents of subsystem V^i are denoted $\Pi_G(V^i)$. Since the graph G_{\rightarrow} is stationary and synchronous, parents of X_{n+1}^i come strictly from the preceding time slice, and additionally $\Pi_{G_{\rightarrow}}(Y_{n+1}^i) = X_{n+1}^i$. Thus, we can build the edge set $\mathcal{E} = \{E^1, E^2, \dots, E^M\}$ in the GDS by means of the DBN. That is, each edge subset E^i is built by the DBN edges

$$E^i = \{V^j \rightarrow V^i : X_n^j \in \Pi_{G_{\rightarrow}}(X_{n+1}^i) \wedge V^j \in \mathcal{V} \setminus V^i\},$$

so long as G forms a DAG. As an example, consider the synchronous GDS in **Figure 1A**. The subsystem V^3 is coupled to both subsystem V^1 and V^2 through the edge set $\mathcal{E} = \{V^1 \rightarrow V^3, V^2 \rightarrow V^3\}$. The time-evolution of this network is shown in **Figure 1B**, where the top two rows (processes X^1 and Y^1) are associated with subsystem V^1 , and similarly for V^2 and V^3 . The distributions for the state [equation (5)] and observation [equation (6)] of M arbitrary subsystems can therefore be factorized according to equation (1):

$$p_{B_{\rightarrow}}(z_{n+1}|z_n) = \prod_{i=1}^M p_D(x_{n+1}^i|x_n^i, \langle x_n^{ij} \rangle_j) \cdot p_D(y_{n+1}^i|x_{n+1}^i). \quad (7)$$

In the rest of the paper, we use simplified notation, given this constrained graph structure. First, since our focus is on learning coupling between distributed systems, the superscripts refer to individual *subsystems*, not variables. Thus, although the 2TBN B_{\rightarrow} is constrained such that $\Pi_{G_{\rightarrow}}(Y_n^i) = X_n^i$, the notation Y_n^{ij} denotes the *measurement variable* of the j th parent of subsystem i , e.g., in **Figure 1**, an arbitrary ordering of the parents gives $Y_n^{3,1} = Y_n^1$ and $Y_n^{3,2} = Y_n^2$. Second, the scoring functions for the 2TBN network B_{\rightarrow} can be computed independently of the prior network B_1 (Friedman et al., 1998). We will assume the prior network is given, and focus on learning the 2TBN. As a result, we drop the subscript and note that all references to the network B are to the 2TBN. Since B_{\rightarrow} is stationary, learning B_{\rightarrow} is equivalent to learning the synchronous GDS.

5. LEARNING SYNCHRONOUS GDSs FROM DATA

In this section, we develop the theory for learning the synchronous update GDS from data. We will focus on techniques for learning graphical models using the *score and search* paradigm, the objective of which is to find a DAG G^* that maximizes a score $g(B:D)$. Given such a score, we can then employ established search procedures to find the optimal graph G^* . Thus, we can state that our main goal is to derive a tractable scoring function $g(B:D)$ for synchronous GDSs that gives a parsimonious model for describing the data.

To derive the score, we use the DBN formulation of synchronous GDSs (Sec. 4) to show that we cannot directly compute the posterior probability of the network structure (Sec. 5.1). By making some assumptions about the system, however, we are able to compute scores for GDSs by use of attractor reconstruction methods (Sec. 5.2). We conclude this section by giving an interpretation of the log-likelihood in terms of information transfer (Sec. 5.3).

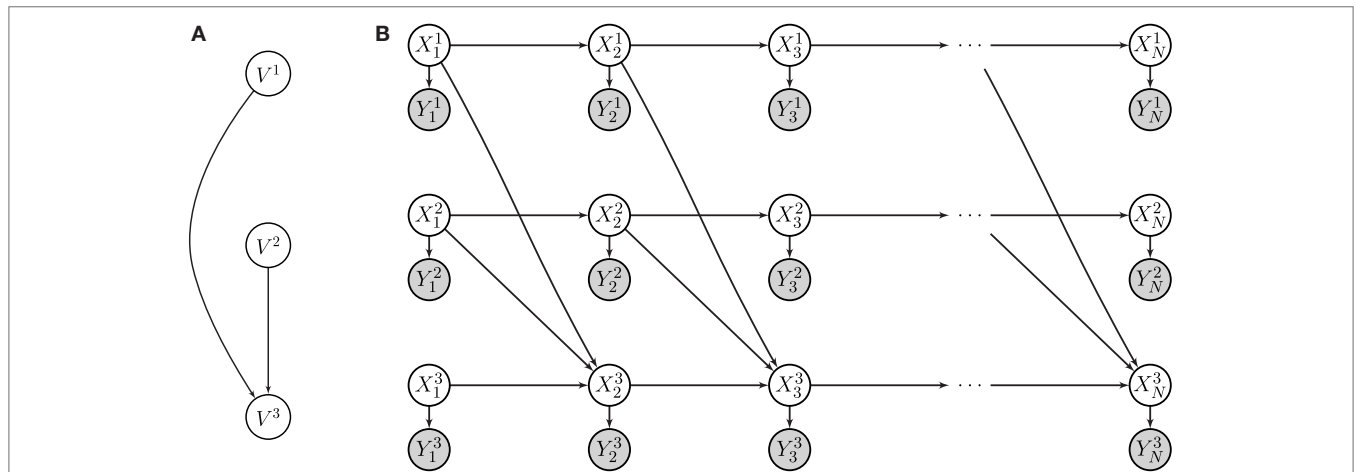


FIGURE 1 | Representation of (A) the synchronous GDS with three vertices (V^1 , V^2 and V^3), and (B) the rolled-out DBN of the equivalent structure. Subsystem V^3 is coupled to both subsystems V^1 and V^2 by means of the edges between latent variables $X_n^1 \rightarrow X_{n+1}^3$ and $X_n^2 \rightarrow X_{n+1}^3$.

5.1. Structure Learning for DBNs

Ideally, we want to be able to compute the posterior probability of the network structure G , given data D . Using Bayes' rule, we can express this distribution as $p(G|D) \propto p(D|G)p(G)$, where $p(G)$ encodes any prior assumptions we want to make about the network G . Thus, the problem becomes that of computing the likelihood of the data, given the model, $p(D|G)$. The likelihood can be written in terms of distributions over network parameters (Friedman et al., 1998):

$$p(D|G) = \int p(D|G, \Theta)p(\Theta|G)d\Theta, \tag{8}$$

where we denote $\ell(\hat{\Theta}_G : D) = \log p(D|G, \hat{\Theta}_G)$ as the log-likelihood function for a choice of parameters $\hat{\Theta}_G$ that maximize $p(D|G, \Theta)$, given a graph G .

A common approach to compute equation (8) in closed form is by using Dirichlet priors. This leads to the BD (Bayesian–Dirichlet) score and variants (Heckerman et al., 1995; Friedman et al., 1998). However, to obtain this analytic solution, we require counts of the tuples $(z_n^i, \pi_G(Z_n^i))$, which involve hidden variables. We will instead use Schwarz's (Schwarz, 1978) asymptotic approximation of the posterior distribution, which states that

$$\lim_{N \rightarrow \infty} \log p(D|G) \approx \ell(\hat{\Theta}_G : D) - \frac{\log N}{2} C(G) + \mathcal{O}(1), \tag{9}$$

where $C(G)$ is the model dimension (i.e., number of parameters needed for the graph G (Friedman et al., 1998)) and $\mathcal{O}(1)$ is a constant bounded by the number of potential models. The approximation of the posterior [equation (9)] requires that data come from an exponential family of likelihood functions with conjugate priors over the model G , and the parameters given the model Θ_G (Schwarz, 1978).

Akaike (1974) gives a similar criterion by approximating the KL-divergence of any model from the data. We can compute both criteria in terms of the log-likelihood function $\ell(\hat{\Theta}_G : D)$ and the model dimension $C(G)$, and thus the problem can be generalized to that of deriving an information criterion for scoring the graph of the form

$$g(B : D) = \ell(\hat{\Theta}_G : D) - f(N) \cdot C(G). \tag{10}$$

When $f(N) = 1$, we have the AIC score (Akaike, 1974); $f(N) = \log(N)/2$ yields the BIC score (Schwarz, 1978), and $f(N) = 0$ gives the maximum likelihood score.

5.2. Deriving the Scores for Synchronous GDSs

To calculate the information criterion [equation (10)], we require tractable expressions for the log-likelihood function $\ell(\hat{\Theta}_G : D)$ and the model dimension $C(G)$. The form of the CPD in equation (7) specifies these functions, and for equation (9) to hold, this distribution must come from an exponential family (Schwarz, 1978). We do not assume the underlying model is linear-Gaussian or other known distributions, and thus express the log-likelihood as the maximum likelihood estimate for multinomial distributions

(Friedman et al., 1998). From equation (7), the log-likelihood then decomposes as

$$\begin{aligned} \ell(\hat{\Theta}_G : D) = & -N \sum_{i=1}^M \sum_{x_{n+1}^i} \sum_{\langle x_n^i \rangle_j} p_D(x_{n+1}^i, x_n^i, \langle x_n^i \rangle_j) \log p_D(x_{n+1}^i | x_n^i, \langle x_n^i \rangle_j) \\ & -N \sum_{i=1}^M \sum_{x_{n+1}^i} \sum_{y_{n+1}^i} p_D(y_{n+1}^i, x_{n+1}^i) \log p_D(y_{n+1}^i | x_{n+1}^i) \end{aligned} \tag{11}$$

Note that although we describe the states and observations as discrete in equation (11), we assume the data are generated by a continuous and stationary process. In theory, it is conceivable to have access to an infinite dataset containing realizations of all potential states and observations. In practice, we have a limited dataset and therefore must implement a discretization scheme. Modeling the dynamical systems with non-parametric techniques requires that the number of parameters scales linearly in the size of the data, and thus $C(G)$ scales linearly with N . Instead, later, we will assume the observation data are discretized, such that there are $|Y_n^i|$ possible outcomes for an observed random variable Y_n^i .

The log-likelihood function [equation (11)] involves distributions over latent variables, and thus we resort to state-space (attractor) reconstruction. First, Lemma 1 shows that a future observation from a given subsystem can be predicted from a sequence of past observations. Building on this result, we present a computable formulation of the 2TBN distribution $p_{B_{-}}(z_{n+1}|z_n)$ via Lemma 2. We then derive a tractable form of the log-likelihood function, presented in Lemma 1. It is then shown in Theorem 2 that these lemmas allow us to compute the information criterion equation (10).

Lemma 1. Consider a synchronous GDS $(G, x_n, y_n, \{f^i\}, \{\psi^i\})$, where the graph G is a DAG. Each subsystem state follows the dynamics $x_{n+1}^i = f^i(x_n^i, \langle x_n^i \rangle_j)$ and emits an observation $y_{n+1}^i = \psi^i(x_{n+1}^i)$; the subsystem observation can be estimated, for some map G^i , by

$$y_{n+1}^i = G^i(y_n^{i,(\kappa^i)}, \langle y_n^{j,(\kappa^j)} \rangle_j). \tag{12}$$

Proof. Consider a forced system $x_{n+1} = f(x_n, w_n)$ with forcing dynamics $w_{n+1} = h(w_n)$ and observation $y_n = \psi(x_{n+1})$. Given this type of forced system, the bundle delay embedding theorem (Stark, 1999; Stark et al., 2003) states that the delay map $\Phi_{f,h,\psi}(x_n, w_n) = y_n^{(\kappa)}$ is an embedding for generic f, ψ , and h . Stark (1999) proved this result in the case of forcing dynamics h that are independent of the state x .¹ For notational simplicity, we omit dependence on the noise process for the map $\Phi_{f,h,\psi}$; the noise can be considered an additional forcing system so long as v_f is i.i.d and v_ψ is either i.i.d or dependent on the state (Stark et al., 2003).

Given a DAG G , any ancestor of the subsystem V^i is not dependent on V^i . As such, the sequence

$$y_n^{i,(\kappa^i)} = \Phi_{f^i, \langle f^j \rangle_j, \psi^i}(x_n^i, \langle x_n^j \rangle_j) \tag{13}$$

¹Stark (1999) conjectures that the theorem should generalise to functions h that are not independent of x . To the best of our knowledge, this result remains to be proven.

is an embedding, since $\langle x_n^{ij} \rangle_j$ is independent of x_n^i . Let $\langle x_n^{ijk} \rangle_k$ be the index-ordered set of parents of node X_n^{ij} (which itself is the j th parent of the node X_n^i). Under the constraint that G is a DAG, where the state $x_{n+1}^i = f^i(x_n^i, \langle x_n^{ij} \rangle_j) + v_{f^i}$, it follows from the bundle delay embedding theorem (Stark, 1999; Stark et al., 2003) that there exists a map \mathbf{F}^i that is well defined and a diffeomorphism between observation sequences. From equation (13), we can write this map

$$\begin{aligned} y_{n+1}^{i,(\kappa^i)} &= \Phi_{f^i, \langle f^{ij} \rangle_j, \psi^i} \left(f^i \left(x_n^i, \langle x_n^{ij} \rangle_j \right), \left\langle f^{ij} \left(x_n^i, \langle x_n^{ijk} \rangle_k \right) \right\rangle_j \right) \\ &= \Phi_{f^i, \langle f^{ij} \rangle_j, \psi^i} \left(f^i \left(\Phi_{f^i, \langle f^{ij} \rangle_j, \psi^i}^{-1} \left(y_n^{i,(\kappa^i)} \right), \left\langle \Phi_{f^{ij}, \langle f^{ijk} \rangle_k, \psi^{ij}}^{-1} \left(y_n^{ij,(\kappa^{ij})} \right) \right\rangle_j \right) \right). \end{aligned} \quad (14)$$

Denote the RHS of equation (14) as $\mathbf{F}^i(y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j)$; the last $\kappa^i + \sum_j \kappa^{ij}$ components of \mathbf{F}^i are trivial. Denote the first component as $\mathbf{G}^i : \mathbb{R}^{\kappa^i} \times \mathbb{R}^{\kappa^{ij}} \rightarrow \mathbb{R}$, then we arrive at equation (12).

Lemma 2. *Given an observed dataset $D = (y_1, y_2, \dots, y_N)$ where $y_n \in \mathbb{R}^M$ are generated by a directed and acyclic synchronous GDS $(G, \mathbf{x}_n, \mathbf{y}_n, \{f^i\}, \{\psi^i\})$, the 2TBN distribution can be written as*

$$\begin{aligned} \prod_{i=1}^M p_D(x_{n+1}^i | x_n^i, \langle x_n^{ij} \rangle_j) \cdot p_D(y_{n+1}^i | x_{n+1}^i) \\ = \frac{\prod_{i=1}^M p_D(y_{n+1}^i | y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j)}{p_D(x_n | \langle y_n^{i,(\kappa^i)} \rangle)}. \end{aligned} \quad (15)$$

Proof. The generalized time delay embedding theorem (Deyle and Sugihara, 2011) states that, under certain technical assumptions, and given M inhomogeneous observation functions $\{\psi^1, \psi^2, \dots, \psi^M\}$, the map

$$\Phi_{f, \psi}(\mathbf{x}) = \langle \Phi_{f^1, \psi^1}(\mathbf{x}), \Phi_{f^2, \psi^2}(\mathbf{x}), \dots, \Phi_{f^M, \psi^M}(\mathbf{x}) \rangle \quad (16)$$

is an embedding where each subsystem (local) map $\Phi_{f^i, \psi^i} : \mathcal{M} \rightarrow \mathbb{R}^{\kappa^i}$, and, at time index n is described by

$$\begin{aligned} \Phi_{f^i, \psi^i}(\mathbf{x}_n) &= y_n^{i,(\kappa^i)} = \langle \psi^i(\mathbf{x}_n), \psi^i(\mathbf{x}_{n-\tau^i}), \\ &\quad \psi^i(\mathbf{x}_{n-2\tau^i}), \dots, \psi^i(\mathbf{x}_{n-(\kappa^i-1)\tau^i}) \rangle \end{aligned}$$

where $\sum_i \kappa^i = 2d + 1$ (Deyle and Sugihara, 2011).² Therefore, the global map equation (16) is given by $\Phi_{f, \psi}(\mathbf{x}_n) = \langle y_n^{i,(\kappa^i)} \rangle$ and there must exist an inverse map $\mathbf{x}_n = \Phi_{f, \psi}^{-1}(\langle y_n^{i,(\kappa^i)} \rangle)$. Given Lemma 1, the existence of $\Phi_{f, \psi}^{-1}$, and

since $\forall i, \{y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j\} \subseteq \langle y_n^{i,(\kappa^i)} \rangle$, we arrive at the following equation:

$$\begin{aligned} \prod_{i=1}^M p_D \left(Y_{n+1}^i = \mathbf{G}^i \left(y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j \right) | y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j \right) \\ = p_D \left(\mathbf{X}_n = \Phi_{f, \psi}^{-1} \left(\langle y_n^{i,(\kappa^i)} \rangle \right) | \langle y_n^{i,(\kappa^i)} \rangle \right) \\ \times \prod_{i=1}^M p_D \left(X_{n+1}^i = f^i(x_n^i, \langle x_n^{ij} \rangle_j) | x_n^i, \langle x_n^{ij} \rangle_j \right) \\ \times \prod_{i=1}^M p_D \left(Y_{n+1}^i = \psi^i(x_{n+1}^i) | x_{n+1}^i \right). \end{aligned} \quad (17)$$

Rearranging equation (17) gives the equality in equation (15).

Lemma 2 shows that the distributions can be reformulated by conditioning on delay vectors. The RHS of equation (15) can be used to perform inference in the 2TBN (7). The numerator is a product of local CPDs of scalar variables, and can thus be computed by either counting (for discrete variables) or density estimation (for continuous variables). The denominator is used to compute the probability that the hidden state occurred, given an observed delay vector; fortunately, Casdagli et al. (1991) established methods to compute this CPDs for a variety of practical scenarios. Therefore, Lemma 2 provides a method to perform exact inference. Using this delay vector representation, we arrive at the following theorem.

Theorem 1. *Consider a synchronous GDS $(G, \mathbf{x}_n, \mathbf{y}_n, \{f^i\}, \{\psi^i\})$, where the graph G is a DAG. Each subsystem state follows the dynamics $x_{n+1}^i = f^i(x_n^i, \langle x_n^{ij} \rangle_j)$ and generates an observation $y_{n+1}^i = \psi^i(x_{n+1}^i)$; a complete dataset is given by the sequence of observations $D = (y_1, y_2, \dots, y_N)$. The log-likelihood of the data given a network structure can be computed in terms of conditional entropy:*

$$\begin{aligned} \ell(\hat{\Theta}_G : D) &= N \cdot H(X_n | \langle Y_n^{i,(\kappa^i)} \rangle) \\ &\quad - N \cdot \sum_{i=1}^M H(Y_{n+1}^i | Y_n^{i,(\kappa^i)}, \langle Y_n^{ij,(\kappa^{ij})} \rangle_j) \end{aligned} \quad (18)$$

Proof. Substituting equation (15) into equation (11) gives the log-likelihood $\ell(\hat{\Theta}_G : D)$ as

$$\begin{aligned} N \sum_{i=1}^M \sum_{y_{n+1}^i} \sum_{y_n^{i,(\kappa^i)}} \sum_{\langle y_n^{ij,(\kappa^{ij})} \rangle_j} p_D(y_{n+1}^i | y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j) \\ \times \log p_D(y_{n+1}^i | y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j) \\ - N \sum_{i=1}^M \sum_{\mathbf{x}_n} \sum_{\langle y_n^{i,(\kappa^i)} \rangle} p_D(x_n, \langle y_n^{i,(\kappa^i)} \rangle) \log p_D(x_n | \langle y_n^{i,(\kappa^i)} \rangle). \end{aligned} \quad (19)$$

In equation (19), we have removed arguments of the joint distributions that will be nullified when multiplied with the CPD. Expressing equation (19) in terms of conditional entropy [equation (3)], we arrive at equation (18).

²The original proof (Deyle and Sugihara, 2011) uses positive lags; however, the authors note that the use of negative lags also applies [and should be used in the case of endomorphisms (Takens, 2002)].

Theorem 2. *The information criterion [equation (10)] for synchronous GDS can be computed as:*

$$g(B : D) = -N \cdot \sum_{i=1}^M H(Y_{n+1}^i | Y_n^{i,(\kappa^i)}, \langle Y_n^{ij,(\kappa^{ij})} \rangle_j) - f(N) \cdot \sum_{i=1}^M \left(|Y_n^i|^{\kappa^i} \cdot (|Y_n^i| - 1) \cdot \prod_{V^p \in \Pi_G(V^i)} |Y_n^p|^{\kappa^p} \right). \quad (20)$$

Proof. The distributions for the first term in equation (18) do not depend on the parents of a subsystem and thus are independent of the graph G being considered. Therefore, we have the following equation for maximum log-likelihood:

$$\max_G \ell(\hat{\Theta}_G : D) = \mathcal{O}(N) - N \cdot \min_G \sum_{i=1}^M H(Y_{n+1}^i | Y_n^{i,(\kappa^i)}, \langle Y_n^{ij,(\kappa^{ij})} \rangle_j). \quad (21)$$

We can now compute the number of parameters needed to specify the model as (Friedman et al., 1998)

$$C(G) = \sum_{i=1}^M \left(|Y_n^i|^{\kappa^i} \cdot (|Y_n^i| - 1) \cdot \prod_{V^p \in \Pi_G(V^i)} |Y_n^p|^{\kappa^p} \right). \quad (22)$$

Since we are searching for the graph $G^* = \max_G g(B : D)$, holding N constant, we can substitute equation (21) and equation (22) into equation (10) and ignore the constant term $\mathcal{O}(N)$ in (21).

5.3. The Log-Likelihood and Information Transfer

To conclude our study of the scores, we look at the log-likelihood in the context of information transfer. First, rearranging the terms of collective transfer entropy [equation (4)], we can rewrite the log-likelihood function [equation (18)], leading to the following result.

Proposition 1. *The log-likelihood function for the synchronous GDS [equation (18)] decomposes as follows:*

$$\ell(\hat{\Theta}_G : D) = N \cdot H(X_n | \langle Y_n^{i,(\kappa^i)} \rangle) - N \cdot \sum_{i=1}^M H(Y_{n+1}^i | Y_n^{i,(\kappa^i)}) + N \cdot \sum_{i=1}^M T_{\langle Y^{ij} \rangle_j \rightarrow Y^i}. \quad (23)$$

Again, the first two terms in equation (23) do not depend on the proposed graph structure, and thus maximizing log-likelihood is equivalent to maximizing collective transfer entropy. This becomes clear when we consider the *log-likelihood ratio*. This ratio quantifies the gain in likelihood by modeling the data D by a candidate network B instead of the empty network B_\emptyset , i.e.,

$$\ell(\hat{\Theta}_G : D) - \ell(\hat{\Theta}_{G_\emptyset} : D) \propto \log \frac{p(B|D)}{p(B_\emptyset|D)}.$$

Recall that the empty DAG G_\emptyset is one with no parents for all vertices $\forall i, \Pi_G(V^i) = \langle Y_n^{ij,(\kappa^{ij})} \rangle_j = \emptyset$. Substituting this

definition into equation (18) [or, alternatively equation (23)] gives the following result.

Proposition 2. *The ratio of the log-likelihood [equation (18)] of a candidate DAG G to the empty network G_\emptyset can be expressed as*

$$\ell(\hat{\Theta}_G : D) - \ell(\hat{\Theta}_{G_\emptyset} : D) = N \cdot \sum_{i=1}^M T_{\langle Y^{ij} \rangle_j \rightarrow Y^i}.$$

6. DISCUSSION AND FUTURE WORK

We have presented a principled method to score the structure of non-linear dynamical networks, where dynamical units are coupled via a DAG. We approached the problem by modeling the time evolution of a synchronous GDS as a DBN. We then derived the AIC and BIC scoring functions for the DBN based on time delay embedding theorems. Finally, we have shown that the log-likelihood of the synchronous GDS can be interpreted in the context of information transfer.

The representation of synchronous GDSs as DBNs allows for inference of coupling in dynamical networks and facilitates techniques for synthesis in these systems. DBNs are an expressive framework that allows representation of generic systems, as well as a numerous general purpose inference techniques that can be used for filtering, prediction, and smoothing (Friedman et al., 1998). Our representation therefore allows for probabilistic reasoning for purposes of planning and prediction in complex systems.

Theorem 2 captures an interesting parallel between learning from complete data and learning non-linear dynamical networks. If the embedding dimension κ and time delay τ are unity, then the information criterion becomes identical to learning a DBN from complete data (Friedman et al., 1998). Thus, our result could be considered a generalization of typical structure learning procedures.

The results presented here provoke new insights into the concepts of structure learning, non-linear time series analysis, and effective network analysis (Sporns et al., 2004; Park and Friston, 2013) based on information transfer (Honey et al., 2007; Lizier et al., 2011; Cliff et al., 2013, 2016). The information-theoretic interpretation of the log-likelihood has interesting consequences in the context of information dynamics and information thermodynamics of non-linear dynamical networks. The transfer entropy terms in Propositions 1 and 2 show that the optimal structure of a synchronous GDS is immediately related to the information processing of distributed computation (Lizier et al., 2008), as well as the thermodynamic costs of information transfer (Prokopenko and Lizier, 2014).

In the future, we aim to perform empirical studies to exemplify the properties of the presented scoring functions. Specifically, the empirical studies should yield insight into the effect of weak, moderate and strong coupling between dynamical units. An important concept to consider in stochastic systems is the convergence of the shadow (reconstructed) manifold to the true manifold (Sugihara et al., 2012); we have implicitly accounted for this phenomena by using CPDs in our model, however, it is important to investigate the property of convergence with different density estimation techniques. In addition, we are interested in the effect of synchrony in these networks and the relationship to previous results

for dynamical systems coupled by spanning trees (Wu, 2005). We conjecture that approach used here will allow us to derive scoring functions without the assumption of multinomial observations, and thus afford the use of non-parametric density estimators. Parametric techniques, such as learning the parameters of dynamical systems (Ghahramani and Roweis, 1999; Hefny et al., 2015), could be considered in place of the posterior approximations.

Finally, the reconstruction theorems used in this paper typically make the assumption that the map (or flow) is a diffeomorphism (invertible in time). Thus, given any state, the past and future are uniquely determined and the time delay τ can be taken positive or negative. In certain cases, however, the time-reversed system is acausal, giving a map that is not time-invertible (an endomorphism). Ideally, we would aim to have methods to infer coupling for both endomorphisms and diffeomorphisms. Takens (2002) showed that if the map is an endomorphism, taking the delay vector of temporally *previous* observations forms an embedding. The generalized theorems in Stark (1999), Stark et al. (2003), and Deyle and Sugihara (2011), however, were established for diffeomorphisms, rather than endomorphisms; we can only conjecture that taking a delay of past observations (as we have

done throughout this paper) follows for these results. Empirical studies using the measures presented in this paper would indicate whether it is an important line of inquiry to prove the generalized reconstruction theorems for endomorphisms.

AUTHOR CONTRIBUTIONS

OC co-wrote the manuscript, derived and proved the theorems, lemmas, and propositions. MP co-wrote the manuscript, assisted with the proofs, and supervised. RF co-wrote the manuscript, assisted with the proofs, and supervised.

ACKNOWLEDGMENTS

We would like to thank Joseph Lizier, Jürgen Jost, and Wolfram Martens for many helpful discussions, particularly in regards to embedding theory. This work was supported in part by the Australian Centre for Field Robotics; the New South Wales Government; and the Faculty of Engineering & Information Technologies, The University of Sydney, under the Faculty Research Cluster Program.

REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716–723. doi:10.1109/TAC.1974.1100705
- Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: structure and dynamics. *Phys. Rep.* 424, 175–308. doi:10.1016/j.physrep.2005.10.009
- Bouckaert, R. R. (1994). “Properties of Bayesian belief network learning algorithms,” in *Proc. of AUAU UAI* (Seattle, WA), 102–109.
- Casdagli, M., Eubank, S., Farmer, J. D., and Gibson, J. (1991). State space reconstruction in the presence of noise. *Physica D* 51, 52–98. doi:10.1016/0167-2789(91)90222-U
- Chickering, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.* 2, 445–498.
- Cliff, O. M., Lizier, J. T., Wang, P., Wang, X. R., Obst, O., and Prokopenko, M. (2016). Delayed spatio-temporal interactions and coherent structure in multi-agent team dynamics. *Art. Life* 23, 1–24. doi:10.1162/ARTL_a_00221
- Cliff, O. M., Lizier, J. T., Wang, X. R., Wang, P., Obst, O., and Prokopenko, M. (2013). “Towards quantifying interaction networks in a football match,” in *RoboCup 2013: Robot World Cup XVII*, eds S. Behnke, M. Veloso, A. Visser, R. Xiong (Berlin, Heidelberg: Springer-Verlag), 1–13.
- Deyle, E. R., and Sugihara, G. (2011). Generalized theorems for nonlinear state space reconstruction. *PLoS ONE* 6:e18295. doi:10.1371/journal.pone.0018295
- Friedman, N., Murphy, K., and Russell, S. (1998). “Learning the structure of dynamic probabilistic networks,” in *Proc. of AUAU UAI* (Madison, WI), 139–147.
- Gan, S. K., Fitch, R., and Sukkarieh, S. (2014). Online decentralized information gathering with spatial-temporal constraints. *Auton. Robots* 37, 1–25. doi:10.1007/s10514-013-9369-5
- Ghahramani, Z. (1998). “Learning dynamic Bayesian networks,” in *Adaptive Processing of Sequences and Data Structures, Volume 1387 of Lecture Notes in Comp. Sci.*, eds C. L. Giles and M. Gori (Vietri sul Mare: Springer-Verlag), 168–197.
- Ghahramani, Z., and Roweis, S. T. (1999). “Learning nonlinear dynamical systems using an EM algorithm,” in *Advances in Neural Information Processing Systems 11*, eds M. J. Kearns, S. A. Solla, and D. A. Cohn (Cambridge, MA: MIT Press), 431–437.
- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424–438. doi:10.2307/1912791
- Gretton, A., Spirtes, P., and Tillman, R. E. (2009). “Nonlinear directed acyclic structure learning with weakly additive noise models,” in *Advances in Neural Information Processing Systems*, eds Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Vol. 22 (Red Hook, NY: Curran Associates, Inc.), 1847–1855.
- Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning Bayesian networks: the combination of knowledge and statistical data. *Mach. Learn.* 20, 20–197. doi:10.1023/A:1022623210503
- Hefny, A., Downey, C., and Gordon, G. J. (2015). “Supervised learning for dynamical system learning,” in *Advances in Neural Information Processing Systems*, eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Vol. 28 (Red Hook, NY: Curran Associates, Inc), 1963–1971.
- Honey, C. J., Kötter, R., Breakspear, M., and Sporns, O. (2007). Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc. Natl. Acad. Sci. U.S.A* 104, 10240–10245. doi:10.1073/pnas.0701519104
- Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). “Non-linear causal discovery with additive noise models,” in *Advances in Neural Information Processing Systems*, eds D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Vol. 21 (Red Hook, NY: Curran Associates, Inc.), 689–696.
- Kaneko, K. (1992). Overview of coupled map lattices. *Chaos* 2, 279–282. doi:10.1063/1.165869
- Kantz, H., and Schreiber, T. (2004). *Nonlinear Time Series Analysis*. eds B. Chirikov, P. Cvitanović, F. Moss, H. Swinney. Cambridge: Cambridge University Press.
- Kocarev, L., and Parlitz, U. (1996). Generalized synchronization, predictability, and equivalence of unidirectionally coupled dynamical systems. *Phys. Rev. Lett.* 76, 1816. doi:10.1103/PhysRevLett.76.1816
- Lam, W., and Bacchus, F. (1994). Learning Bayesian belief networks: an approach based on the MDL principle. *Comput. Intell.* 10, 269–293. doi:10.1111/j.1467-8640.1994.tb00166.x
- Lizier, J. T., Heinzle, J., Horstmann, A., Haynes, J.-D., and Prokopenko, M. (2011). Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity. *J. Comput. Neurosci.* 30, 85–107. doi:10.1007/s10827-010-0271-2
- Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2008). Local information transfer as a spatiotemporal filter for complex systems. *Phys. Rev. E Stat. Nonlin. Soft. Matter. Phys.* 77, 026110. doi:10.1103/PhysRevE.77.026110
- Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2010). Information modification and particle collisions in distributed computation. *Chaos* 20, 37109–37113. doi:10.1063/1.3486801
- Lizier, J. T., and Rubinov, M. (2012). *Multivariate Construction of Effective Computational Networks from Observational Data*. MIS-Preprint 25/2012. Max Planck Institute for Mathematics in the Sciences.
- Mortveit, H. S., and Reidys, C. M. (2001). Discrete, sequential dynamical systems. *Discrete Math.* 226, 281–295. doi:10.1016/S0012-365X(00)00115-1

- Park, H.-J., and Friston, K. (2013). Structural and functional brain networks: from connections to cognition. *Science* 342, 1238411. doi:10.1126/science.1238411
- Peters, J., Janzing, D., and Schölkopf, B. (2011). Causal inference on discrete data using additive noise models. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 2436–2450. doi:10.1109/TPAMI.2011.71
- Prokopenko, M., and Lizier, J. T. (2014). Transfer entropy and transient limits of computation. *Sci. Rep.* 4, 5394. doi:10.1038/srep05394
- Schreiber, T. (2000). Measuring information transfer. *Phys. Rev. Lett.* 85, 461–464. doi:10.1103/PhysRevLett.85.461
- Schumacher, J., Wunderle, T., Fries, P., Jäkel, F., and Pipa, G. (2015). A statistical framework to infer delay and direction of information flow from measurements of complex systems. *Neural Comput.* 27, 1555–1608. doi:10.1162/NECO_a_00756
- Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi:10.1214/aos/1176344136
- Sporns, O., Chialvo, D. R., Kaiser, M., and Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. *Trends Cogn. Sci.* 8, 418–425. doi:10.1016/j.tics.2004.07.008
- Stark, J. (1999). Delay embeddings for forced systems. I. Deterministic forcing. *J. Nonlin. Sci.* 9, 255–332. doi:10.1007/s003329900072
- Stark, J., Broomhead, D. S., Davies, M. E., and Huke, J. (2003). Delay embeddings for forced systems. II. Stochastic forcing. *J. Nonlin. Sci.* 13, 519–577. doi:10.1007/s00332-003-0534-4
- Sugihara, G., May, R., Ye, H., Hsieh, C.-H., Deyle, E., Fogarty, M., et al. (2012). Detecting causality in complex ecosystems. *Science* 338, 496–500. doi:10.1126/science.1227079
- Takens, F. (1981). “Detecting strange attractors in turbulence,” in *Dynamical Systems and Turbulence, Volume 898 of Lecture Notes in Math*, eds D. A. Rand and L-S. Young (Warwick: Springer-Verlag), 366–381.
- Takens, F. (2002). The reconstruction theorem for endomorphisms. *Bull. Br. Math. Soc.* 33, 231–262. doi:10.1007/s005740200012
- Umenberger, J., and Manchester, I. R. (2016). “Scalable identification of stable positive systems,” in *Proc. of IEEE CDC*, Las Vegas, NV.
- Vicente, R., Wibral, M., Lindner, M., and Pipa, G. (2011). Transfer entropy – a model-free measure of effective connectivity for the neurosciences. *J. Comput. Neurosci.* 30, 45–67. doi:10.1007/s10827-010-0262-3
- Wu, C. W. (2005). Synchronization in networks of nonlinear dynamical systems coupled via a directed graph. *Nonlinearity* 18, 1057. doi:10.1088/0951-7715/18/3/007
- Xu, Z., Fitch, R., Underwood, J. P., and Sukkarieh, S. (2013). Decentralized coordinated tracking with mixed discrete-continuous decisions. *J. Field Robot.* 30, 717–740. doi:10.1002/rob.21471

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Cliff, Prokopenko and Fitch. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.