Inferring connectivity in networked dynamical systems: Challenges using Granger causality

Bethany Lusch, Pedro D. Maia, and J. Nathan Kutz
Department of Applied Mathematics, University of Washington, Seattle, Washington 98195-3925, USA
(Received 5 May 2016; published 27 September 2016)

Determining the interactions and causal relationships between nodes in an unknown networked dynamical system from measurement data alone is a challenging, contemporary task across the physical, biological, and engineering sciences. Statistical methods, such as the increasingly popular Granger causality, are being broadly applied for data-driven discovery of connectivity in fields from economics to neuroscience. A common version of the algorithm is called pairwise-conditional Granger causality, which we systematically test on data generated from a nonlinear model with known causal network structure. Specifically, we simulate networked systems of Kuramoto oscillators and use the Multivariate Granger Causality Toolbox to discover the underlying coupling structure of the system. We compare the inferred results to the original connectivity for a wide range of parameters such as initial conditions, connection strengths, community structures, and natural frequencies. Our results show a significant systematic disparity between the original and inferred network, unless the true structure is extremely sparse or dense. Specifically, the inferred networks have significant discrepancies in the number of edges and the eigenvalues of the connectivity matrix, demonstrating that they typically generate dynamics which are inconsistent with the ground truth. We provide a detailed account of the dynamics for the Erdős-Rényi network model due to its importance in random graph theory and network science. We conclude that Granger causal methods for inferring network structure are highly suspect and should always be checked against a ground truth model. The results also advocate the need to perform such comparisons with *any* network inference method since the inferred connectivity results appear to have very little to do with the ground truth system.

DOI: 10.1103/PhysRevE.94.032220

I. INTRODUCTION

In 1956, Norbert Wiener proposed a statistical notion of causality [1]: Y causes X if knowing the past of Y improves the prediction of X (as compared to using the past of X alone). In 1969, the Nobel Prize winning econometrician Clive Granger formalized this concept in the context of linear autoregressive modeling [2]. The resulting method is now commonly referred to as *Granger causality* (GC). The importance of understanding causal relationships in complex, dynamical networks from time-series measurements alone is clear: it becomes a fundamental tool for data-driven scientific discovery [3–5]. Methods to infer causality are the source of much debate and require entirely different statistical models from those used in associational inference [6]. Complicating the methodology is the fact that correlation does not imply causation. So, despite numerous methods for computing correlation, they only serve a limited role in understanding if there is an underlying causal relationship. In this manuscript, we consider a popular and commonly used form of GC to infer the connectivity in a known, networked system of Kuramoto oscillators. Our goal is to evaluate GC as a tool for data-driven scientific discovery. We demonstrate that the method is highly suspect, inferring connectivity and dynamics that are significantly different than the known ground truth model. With the ever-increasing demand to understand connectivity in dynamic networks, we hope that the results from this study will serve as a strong cautionary note to the broader scientific community using such statistical techniques for data-driven network inference.

Following Wiener's statistical innovations, the seminal work of Granger was originally defined in terms of two variables X and Y. However, it was quickly generalized to larger sets of variables where *pairwise-conditional Granger causality* could be computed among the variables. By checking for causal links between each pair of variables, the aim was to infer the most probable directed graph structure. Figure 1 illustrates this idea: each node in the dynamical network generates its own time series data that is influenced by interactions with other nodes. In practice, we are usually limited to individual noisy recordings without knowledge of the underlying network connectivity—which is precisely what GC attempts to determine. This mathematical framework became popular in the economics community [7] for determining how nodes of a financial network might be influencing each other. For example, Hamilton [8] used GC as evidence that oil shocks were a contributing factor to recessions. More recently, it has risen in popularity in neuroscience [9] where Bressler *et al.* [10] used it to justify that activity in certain areas of the frontal and parietal lobes can predict visual processing activity before an anticipated visual stimulus. More broadly, pairwise-conditional GC is currently being used to infer networks of connectivity in many applications [11–16].

The method is highly attractive in such systems due to the fact that there may be no other way to understand the underlying network of causal relationships. Attempts to infer causality have also led to numerous other statistical innovations for determining causality [3–5], including those leveraging independent component analysis [17] and network structure [18], for instance. A seminal recent contribution by Sugihara *et al.* [19] called convergent cross mapping (CCM) tests for causation by measuring the extent to which the historical record of Y values can reliably estimate states of X. The CCM method looks for the signature of X in Y's time series by seeing whether there is a correspondence between the attractor manifold built from Y and points in the X manifold, where the two manifolds are constructed from lagged (time-delay) coordinates of the time-series variables.

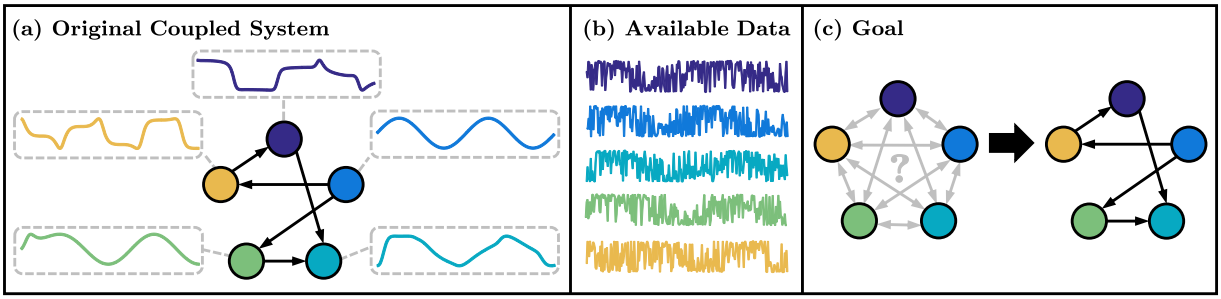

FIG. 1. Inferring network connectivity via Granger causality analysis. (a) Schematics of a coupled dynamical system where the directed network architecture plays an important role. The time series generated by each node is influenced by its connectivity to other nodes. (b) In several applications, the connectivity structure is unknown, but noisy measurements from each node are available. (c) We use Granger causality to infer the original network structure from the noisy data.

This is a promising avenue especially for systems displaying a dynamical attractor. More recently, a formulation by Wahl *et al.* [20] has employed local linear models in the GC framework to resolve causal relationships in distinct regions of state space, leading to a promising technique for resolving overall GC structure.

As is still the case today, Granger's definition was met with controversy. Concerns have ranged from philosophical matters [21] to conceptual limitations [22] to analytical and practical implementation issues [19,23]. Granger responded to criticism in 1980 [21] by arguing that although there is no consensus for the concept of causality, it is still worth choosing a specific and operational definition for the context of a written work or lecture. He suggested that GC should be viewed merely as evidence in a Bayesian sense. In 2003, he acknowledged in his Nobel Lecture that because his definition was pragmatic and easy to apply, "of course, many ridiculous papers appeared" (see [24]). Several concerns have led to variations in the methodology which we describe in Sec. II. We will primarily consider the version called *pairwise-conditional* GC. We do not address theoretical or philosophical concerns with Granger causality. Instead, we accept it as a technical definition and evaluate its efficacy in inferring network structure. We use data generated from a known network of nonlinear Kuramoto coupled oscillators [25]. This is a canonical choice for studying synchronizable systems, such as power grids, pacemaker cells in the heart, pedestrian crowds, and coupled cortical neurons (see [26] and references therein). We generate random networks to reconstruct, sampling from the Erdős-Rényi family [27]. This is a well-studied network model [28] and provides a practical way to generate random networks with a large range of edge densities. We calculate the GC structure primarily using the Multivariate Granger Causality (MVGC) Matlab Toolbox [29]. MVGC is a popular implementation of pairwise-conditional GC written with neuroscience data in mind [11–13,30–32], but we also consider other numerical implementations of GC in order to cross-validate the results.

The outline of the paper is as follows. Sections II and III provide all necessary background information for the GC framework and Kuramoto systems, respectively. We describe our methodology in Sec. IV and present a comprehensive list of results in Sec. V. We summarize our conclusions in Sec. VI.

II. BACKGROUND: GRANGER CAUSALITY

Granger causality (GC) is defined in the context of linear autoregressive modeling, which computes the relationship of a time series with its own past. One important model that is used for multivariate stochastic processes is called the vector autoregressive (VAR) model. Let $\boldsymbol{X_t}$ be a vector-valued stochastic process with mean zero (averaging at each time t over the realizations). A VAR model for $\boldsymbol{X_t}$ is a sequence of $n \times n$ real matrices $\mathbf{A}_k$ and an n-dimensional white noise process (independently and identically distributed and serially uncorrelated) $\boldsymbol{\epsilon}_t$ such that

$$\boldsymbol{X_t} = \sum_{k=1}^{p} \mathbf{A}_k \boldsymbol{X_{t-k}} + \boldsymbol{\epsilon}_t. \tag{1}$$

The $\mathbf{A}_k$ matrices (called the *regression coefficients*) describe how $\boldsymbol{X_t}$ depends on its past and represent the predictable behavior of the process. The $\boldsymbol{\epsilon}_t$ process (called the *residuals*) represents the unpredictable behavior. We call p the *model order*. Note that fitting a VAR model to data does not imply that the data was generated by a VAR process.

The above formulation is often written as a first-order VAR model of the form $\tilde{\boldsymbol{X}}_t = \boldsymbol{A}\tilde{\boldsymbol{X}}_{t-1} + \tilde{\boldsymbol{\epsilon}}_t$, where

$$\tilde{\boldsymbol{X}}_t = \begin{bmatrix} \boldsymbol{X_t} \\ \boldsymbol{X_{t-1}} \\ \vdots \\ \boldsymbol{X_{t-(p-1)}} \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \mathbf{A_1} & \mathbf{A}_2 & \cdots & \mathbf{A}_p \\ \mathbf{I}_n & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots \\ 0 & 0 & \mathbf{I}_n & 0 \end{bmatrix},$$

and $\tilde{\boldsymbol{\epsilon}}_t = [\boldsymbol{\epsilon}_t, 0, \ldots, 0]^T$ with $\mathbf{I}_n$ being an $n \times n$ identity matrix. The spectral radius of a VAR model is defined to be the spectral radius of $\mathbf{A}$, $\rho(\mathbf{A})$. Recall that the spectral radius of a matrix $\mathbf{A}$ is defined as $\rho(\mathbf{A}) = \max\{|\lambda_1|, \ldots, |\lambda_n|\}$, where $\{\lambda_i\}$ are the eigenvalues of $\mathbf{A}$. The stability criteria for a VAR model is analogous to those for difference equations: $x^{(k+1)} = Tx^{(k)}$ is stable if and only if $\rho(T) < 1$. Thus a VAR model is stable if and only if $\rho(\mathbf{A}) < 1$.

The statistical basis of GC can be stated as follows: Y causes X if the past of Y improves the prediction of X as compared to only using the past of X. Specifically, if a stochastic process $\mathbf{Y}_t$ is used to predict $\mathbf{X}_t$, this can be written as

$$\mathbf{X}_t = \sum_{k=1}^{p} \mathbf{A}'_k \mathbf{X}_{t-k} + \sum_{k=1}^{p} \mathbf{B}_k \mathbf{Y}_{t-k} + \boldsymbol{\epsilon}'_t. \tag{2}$$

Then we say that Y *Granger causes* X if Eq. (2) is a "better" prediction of X than Eq. (1). In particular, Y Granger causes X if the variance of $\boldsymbol{\epsilon}'_t$ is statistically significantly lower than the variance of $\boldsymbol{\epsilon}_t$.

There are many variations on the original definition. Most formulations rely on representing data as a VAR model, although some differ significantly. Extensions include blockwise GC [33], partial GC [34], and piecewise GC [35]. Improvements for nonlinear time series are studied in [36–39].

We focus on the version called *pairwise-conditional* GC, specifically as implemented in the MVGC Toolbox [29]. GC might wrongly infer that Y Granger causes X if there is a third, latent, confounding variable Z that influences both X and Y. To minimize this effect, pairwise-conditional GC involves "conditioning out" any other variables for which a time series is available. Let Z be a third variable. Then conditioning out is done by changing Eqs. (1) and (2) to

$$\boldsymbol{X_t} = \sum_{k=1}^{p} \mathbf{A}_k \boldsymbol{X_{t-k}} + \sum_{k=1}^{p} \mathbf{B}_k \boldsymbol{Y_{t-k}} + \sum_{k=1}^{p} \mathbf{C}_k \boldsymbol{Z_{t-k}} + \boldsymbol{\epsilon}_t, \quad (3)$$

$$\boldsymbol{X_t} = \sum_{k=1}^{p} \mathbf{A}'_k \boldsymbol{X_{t-k}} + \sum_{k=1}^{p} \mathbf{C}'_k \boldsymbol{Z_{t-k}} + \boldsymbol{\epsilon}'_t. \quad (4)$$

Now, when fitting each of these VAR models, the idea is that Y causes X if the past of Y improves the prediction of X as compared to only using the past of X and Z. Thus, if Z is a confounding variable, Y does not Granger cause X because it does not carry additional predictive information beyond what Z contributed.

We are considering the null hypothesis that $\mathbf{B}_1 = \mathbf{B}_2 = \cdots = \mathbf{B}_p = 0$. We calculate the G *causality* by considering the log-likelihood ratio

$$\mathcal{F}_{Y\to X|Z} := \ln \frac{|\Sigma'|}{|\Sigma|},$$

where $\Sigma = \mathrm{Cov}(\boldsymbol{\epsilon}_t)$ and $\Sigma' = \mathrm{Cov}(\boldsymbol{\epsilon}'_t)$. Thus, to check the causality between a pair of variables, we can condition out the other $n-2$ variables. In particular, if $\boldsymbol{U}$ is composed of n processes $U_{1t}, \ldots, U_{nt}$, we can compute *pairwise-conditional* causalities $\mathcal{G}_{i,j}(\boldsymbol{U}) := \mathcal{F}_{U_j\to U_i|\boldsymbol{U}_{[ij]}}$, where $\boldsymbol{U}_{[ij]}$ denotes omitting U_i and U_j, and perform a statistical test to determine which values $\mathcal{G}_{i,j}$ are large enough to represent a causal relationship between U_j and U_i. In our network context, this is a directed edge from node j to node i. In the MVGC Toolbox [29], this is calculated using multiple representations of a VAR model. It computes causality both in temporal and frequency domains and returns an error message if the results do not match. See [29] for details.

Not all data sets lend themselves to GC analysis. The coefficients $\mathbf{A}_k$ of the fitted VAR model, for instance, must be square summable and stable [29]. Square summability implies $\sum_{k=1}^{p} \|A_k\|^2 < \infty$, which is trivially true for finite p. However, some stochastic processes may only be fit by a VAR with $p = \infty$. The MVGC Toolbox [29] does not provide a practical way to check this criterion, but mentions that violations may occur if the data contains a strong, slow moving average component. This may trigger a warning or an error.

According to [29], there are five likely reasons for problems with using GC on time series data as follows. (i) *Colinearity*: if there are linear or nearly linear relationships between time series, the VAR representation will be ambiguous. This is likely to be detected by the toolbox and reported, stopping with an error. (ii) *Stationarity*: the data must be covariance stationary. If the spectral radius of the estimated VAR model is larger than one, the GC analysis stops with an error. (iii) *Long-term memory*: if the autocorrelation does not decay exponentially, the data is unsuited to VAR modeling since it may silently yield spurious results. This may be detected when computing the autocovariance sequence where long-term memory typically manifests itself as power-law behavior. The sequence should decay exponentially when the process has a spectral radius less than one. However, there is a limit to how far the sequence is calculated, and if the spectral radius is close to one, it may not decay below a specified tolerance within that length. In that case, the results may be inaccurate and a warning may be issued. (iv) *Moving average*: if the data contains a strong, slow moving average component, the coefficients might not be square-summable, the analysis may be invalid, and the toolbox will typically report warnings or errors. (v) *Heteroscedasticity*: if the variance of the residual terms depends on the values of the process, then the statistical inference is likely to suffer. It can invalidate standard statistical significance tests or confound G-causal inference. The toolbox does not offer any way to test or counteract this effect. All the results in this paper were attained after running all of the diagnostic tests recommended in [29]. The toolbox did not return any errors in our runs. The only warnings given were from the autocovariance sequence not decaying sufficiently quickly, which we carefully annotated.

III. BACKGROUND: KURAMOTO OSCILLATORS

Coupled oscillators have been of long-standing interest in the scientific community due to their ability to describe canonical phenomena such as synchronization. Yoshiki Kuramoto proposed one of the most well-studied systems modeling nonlinear coupled oscillators, the *Kuramoto oscillators*:

$$\dot{\theta}_i = \omega_i + \frac{K}{n} \sum_{j=1}^{n} A_{ij} \sin(\theta_j - \theta_i), \quad i = 1, \ldots, n. \quad (5)$$

In this model, the dynamics of the ith oscillator is governed by θ_i, which has a natural frequency ω_i. The n oscillators are coupled in a network with adjacency matrix $\mathbf{A}$ and coupling strength K. Depending on the parameters of the model, the oscillators may synchronize or exhibit chaotic dynamics. Kuramoto defined an order parameter to describe these different potential dynamics:

$$r(t) = \frac{1}{n} \left| \sum_{j=1}^{n} e^{i\theta_j(t)} \right|, \quad (6)$$

where $r(t)$ varies from $O(1/\sqrt{n})$ to unity when synchronization occurs. When K increases, so does the average order parameter r. Figure 2 depicts the synchronization as a function of strength and probability of connection in a 12-node Erdős-Rényi network.

Figure 3 depicts several two-oscillator examples. Synchronization occurs if both oscillators converge to the same frequency. Depending on the network structure, they may

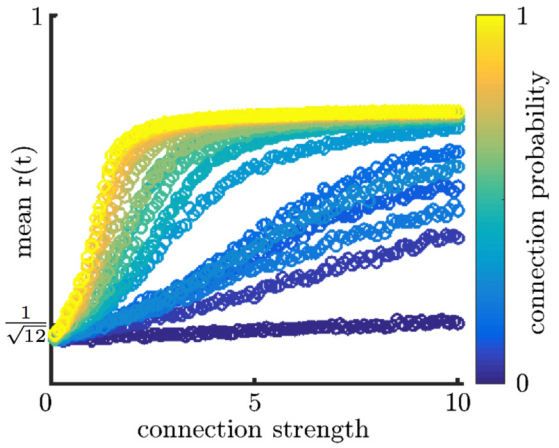

FIG. 2. Increase in synchrony as connection strength increases. We generate random 12-node Erdős-Rényi networks with a range of connection probabilities. We then solve the Kuramoto model on each network for varying connection strengths. For each network and connection strength pair, we calculate the average order parameter $r(t)$ [Eq. (6)]. We see that as the connection strength increases, the synchrony also increases. However, for sparse networks, the network remains unsynchronized ($r(t) \approx \frac{1}{\sqrt{n}}$) and for dense networks, the network synchronizes for moderate connection strength. This data was generated in Experiment C1 (see Sec. V).

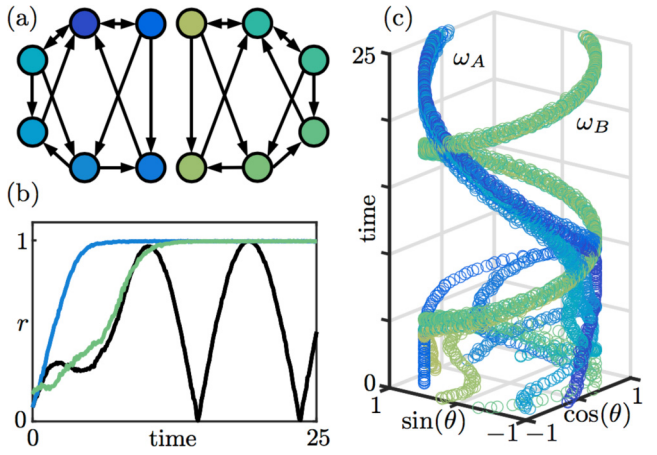

FIG. 4. Example of synchronicity in structured Kuramoto networks. (a) We have two disjoint subnetworks. The blue oscillators have frequencies with average -0.2 while the green oscillators have average frequency 0.5. As we see in the right panel (c), the individual trajectories collapse. The blue nodes synchronize to frequency ω_A and the green nodes synchronize to frequency ω_B. In the lower left panel (b), we see that the measure of synchronicity for each community, $r(t)$ [Eq. (6)], approaches one but at different synchronization times. However, when $r(t)$ is evaluated on the entire network, we do not achieve total synchronicity because the two communities are not connected.

converge to one oscillator's natural frequency or an average of the two. Notice how the cases with exactly one edge appear to match Granger's definition of causality; the dominating oscillator predicts itself, but the other oscillator is strongly influenced by it.

Figure 4 exemplifies a Kuramoto system with twelve nodes connected in two disjoint communities. The blue community synchronizes to a slow frequency ω_A and the green community synchronizes to a fast frequency ω_B. The order parameter $r(t)$ considers synchronization across the whole network, making it difficult to interpret (black line). If we evaluate $r(t)$ on each community separately (the blue and green lines) we see that each community synchronizes with itself.

In this paper, we simulate the Kuramoto model and use Granger causality (GC) to infer the adjacency matrix $\mathbf{A}$. As demonstrated in Figs. 3 and 4, the network structure influences the system dynamics. We expect the dynamics to preserve signatures of the network architecture and for GC to potentially discover these connections. Because we know the ground-truth data, our model guarantees that there are no external or hidden variables influencing the system. However, as we will see in Sec. V, GC will consistently fail to recover the known

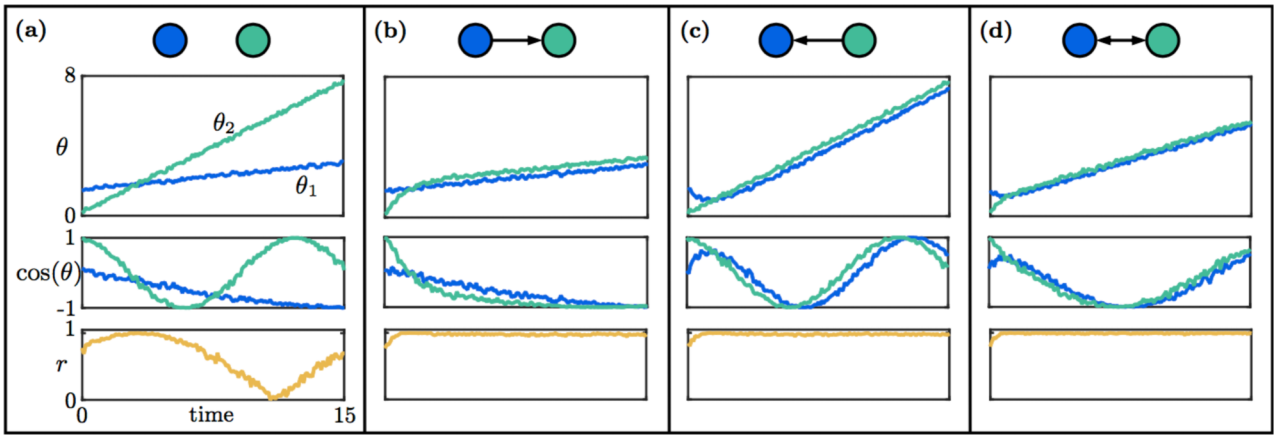

FIG. 3. Pair of coupled Kuramoto oscillators with distinct natural frequencies. We show the four possible network architectures in panels (a)–(d). We first plot θ_1 and θ_2, the solution of the differential equations in Eq. (5). We then plot $\cos(\theta_1)$ and $\cos(\theta_2)$, the more natural way to view oscillators. In panel (a), the oscillators are uncoupled, so they merely oscillate with their natural frequency. However, in panels (b)–(d), we see cases leading to synchronization. The overall synchronization of the network can be summarized by the parameter $r(t)$ with full synchronization achieved when $r(t) = 1$ [see Eq. (6)].

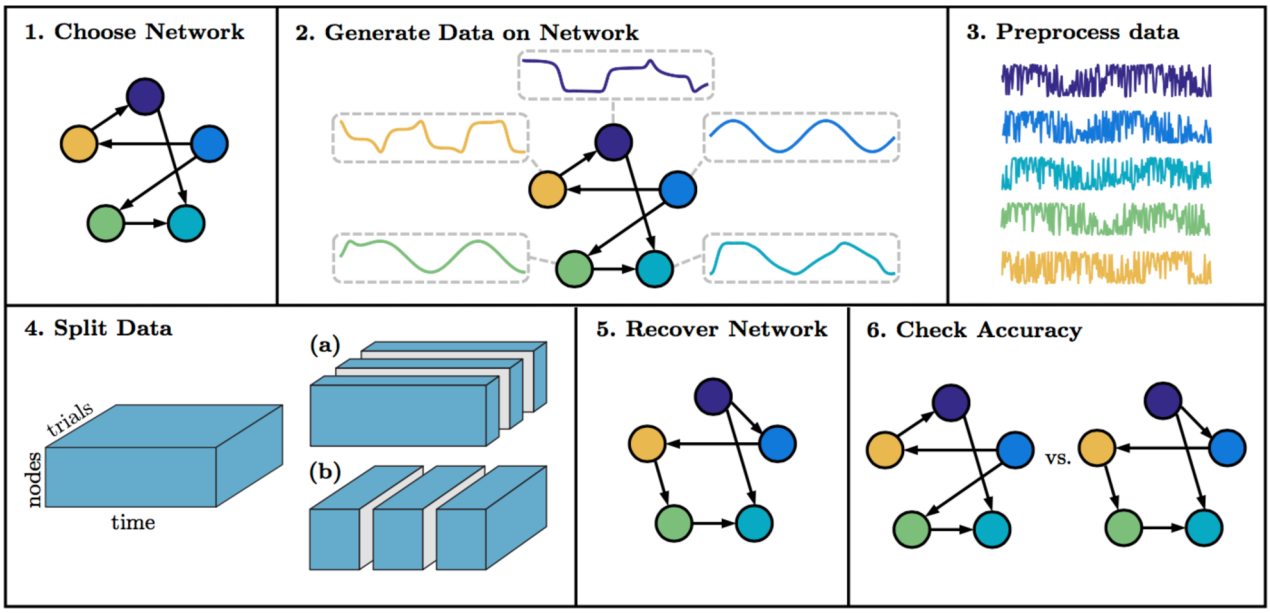

FIG. 5. Overview of steps in our methodology. We will experiment with varying the decisions made in each step—see Sec. IV.

connectivity. We are not the first to apply GC methods to Kuramoto systems. Angelini *et al.* [40,41] develop a version of GC that does not use VAR modeling and is specialized for circular variables, using Kuramoto oscillators as an example. Wu *et al.* [42] develop an algorithm for inferring a network of Kuramoto oscillators using piecewise GC [35] followed by a pruning of edges.

IV. NUMERICAL EXPERIMENTS

We test Granger causality (GC) by applying it to data generated from the Kuramoto model Eq. (5) with the goal of reconstructing the network adjacency matrix A. We split our methodology into six steps; see Fig. 5 for a schematic overview. We explore several options at each step to avoid limiting ourselves to the best or worst cases for GC performance. However, as we will show in Sec. V, network reconstruction is consistently poor, usually without warnings from the toolbox. The following specific steps are taken in our evaluation algorithm.

(1) *Choose network*. We set up a network with n nodes that we wish to reconstruct. Our default value ($n = 12$) yields a sizable network while performing simulations in a timely manner. We tried other values for comparison ($n = 2$, 6, and 24). See Exp. A1–A2 and C2–C3 for details.

In most experiments, we generate Erdős-Rényi networks; each potential edge is included with constant probability p [27]. We vary $p = 0.05, 0.1, \ldots, 1$ to address how the density of the network affects GC results.

(2) *Generate data on network*. We simulate several Kuramoto systems with a variety of parameters as follows.

(i) Connection strength K. By default, we consider $K = 0.5, 1, 2, 4, 8$ to span GC reconstructions ranging from underestimation to overestimation of edges (see Fig. 9). Experiments A1–A2 and C1 display a wider range of K values.

(ii) Initial conditions θ^0: randomly sampled from uniform distributions. We reset them for each trial. This is reasonable for real data and additionally helps the data have a constant mean when averaging over trials (a requirement for being covariance stationary). Our default distribution is $[0, 2\pi]$ since we will apply cosine to the data, which has a period of 2π. In Exp. C4–C5, we shift this distribution for comparison.

(iii) Natural frequencies ω: randomly sampled from uniform distributions. We reset them for each trial. A uniform distribution of $[-1, 1]$ is used in some studies of Kuramoto oscillators [43,44]. However, when we used that range of natural frequencies, the toolbox gave many warnings (see Experiment C7). We, therefore, shifted the distribution to $[0, 2]$ for most experiments. See Exp. C6–C7 for comparisons to other distributions.

(iv) Number of trials N (each from solving the Kuramoto model once with random initial conditions and natural frequencies). Our default is $N = 100$, but we consider other values in Exp. C8–C9.

(v) Data sampling rate: $[0, T]$ with time step Δt, giving $m = T/\Delta t$ time points. Our default values are $\Delta t = 0.1$, $T = 25$, and $m = 250$. These were chosen by searching the parameter space for cases with no warnings and low error. We compare to other values in Exp. C10–C20.

TABLE I. Summary of our four classes of numerical experiments. Full details are in Table VI.

Ref.	Class	Figures	Tables
A1–A2	Two-node networks	3, 6	II, VI
B1–B3	Two independent communities	4, 7, 8	VI
C1–C30	Erdős-Rényi[a]	2, 9, 10, 11, 12, 13, 16	IV, V, VI
D1–D2	Erdős-Rényi[b]	14, 15	VI

[a]Changing parameters.
[b]Changing implementation.

TABLE II. Performance (%) of Granger causality in two-node Kuramoto oscillator example. Confusion matrix. The table summarizes reconstruction results for Experiment A1: four different true networks and 100 values of connection strength. As we will continue to see in larger networks, the performance is weakest when the network is not extremely sparse or dense. The specific parameter choices for this experiment are in Table VI.

Truth \ Estimate				
	93%	4%	2%	1%
	9%	2%	0%	89%
	7%	0%	2%	91%
	4%	0%	2%	94%

(3) *Preprocess data*. We add noise to our simulations since measurement errors are expected in most applications, and it helps the data be more covariance stationary. Specifically, we add white Gaussian noise of strength s, i.e., a constant power spectral density of s^2. Each one of the N random trials will have different noise realizations. Our default value of s is 2.5, based on experiments to minimize the error in the results, but other values are compared in Exp. C21–C22.

Next, we usually apply cosine: instead of using $\theta_1, \ldots, \theta_n$, we use $\cos(\theta_1), \ldots, \cos(\theta_n)$. This is a natural way to view oscillations (see Fig. 3) and remove linear trends. We explore alternatives in Exp. C23–C24.

(4) *Split data*. At this point we already generated a "cube" of data with N random trials, each with a time series of length m for each of the n oscillators. As pictured in step 4 of Fig. 5, we could apply GC to the whole cube of data at once. However, we have the option to split the data cube into smaller cubes by (a) splitting trials into smaller sets or (b) splitting the time into smaller time intervals. We apply GC to each one of the smaller cubes, letting them "vote" for edges. We include a directed edge if at least half of the voting networks include it.

Barnett and Seth [29] suggest splitting the data into smaller time intervals for making it covariance stationary. This is also the idea behind piecewise GC [35]. Splitting trials may reduce error if the subsets are each sufficiently large for reasonable

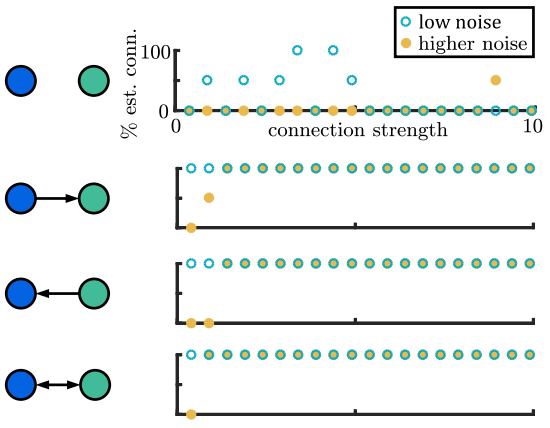

FIG. 6. Percentage of inferred edges as the connection strength varies. We try all four possible two-node networks (0, 1, or 2 edges), and we vary the connection strength across the horizontal axis. We also try two amounts of noise. The teal circles are for Experiment A1 (low noise) and the orange closed circles are for Experiment A2 (higher noise). The specific parameter choices for these experiments are in Table VI. The error is higher for low noise. As the connection strength increases, so does the number of inferred edges, a pattern that will continue for larger networks.

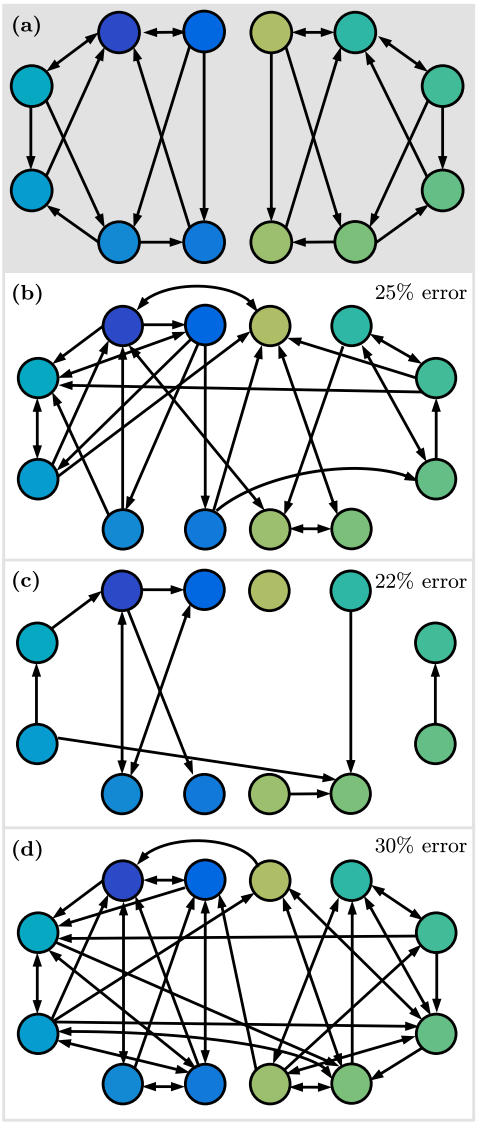

FIG. 7. Results of Granger causality inference on the two-community network. Panel (a) depicts the true network. The resulting network from Experiment B1 in panel (b) has many extra connections and even connects the two separate communities, but the MVGC Toolbox [29] provides warnings. In Experiment B2, we increase the noise and try again, producing the network in panel (c) without warnings. This network is missing many edges but also connects the two communities. In Experiment B3, we keep the higher level of noise but halve the time step, resulting in the network in panel (d) without warnings. We again have vast overestimation of edges and the community structure is lost. The specific parameter choices for these experiments are in Table VI.

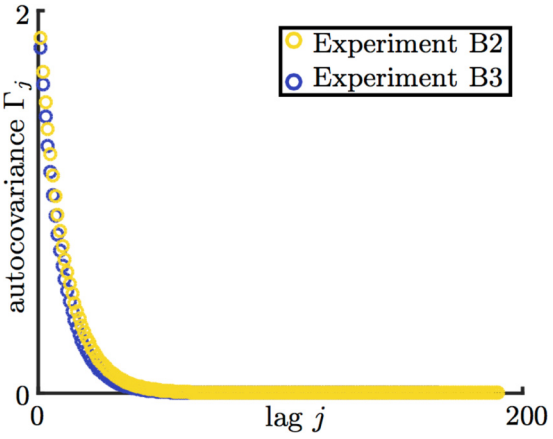

FIG. 8. Autocovariance decay for Experiments B2 and B3. In these two experiments, the toolbox does not provide warnings, but there are significant errors (Fig. 7). Here we plot the autocovariance sequence for each experiment to demonstrate that it decays exponentially, as required.

inference. Then the rationale is that the process would become more robust when considering each vote.

We experiment with splitting data and voting in Exp. C25–C30.

(5) *Recover network with Granger causality*. We use the MVGC Toolbox to recover a network from our data. Alternative implementations of GC are explored in Sec. V C.

(6) *Check accuracy*. Finally, we compare the GC estimated network with the ground truth. Our standard error metric is the percentage of wrong edges. For n nodes, there are $n^2 - n$ potential directed edges. We add the number of false-positive and false-negative edges and divide by $n^2 - n$. We consider other error metrics in Sec. V C.

We list all parameter choices of the exhaustive computational exploration in Table VI.

V. NETWORK RECONSTRUCTION RESULTS

We summarize our four classes of numerical experiments in Table I. In Sec. V A we consider a pair of oscillators, as pictured in Fig. 3. In Sec. V B, we consider the network structure with two independent communities from Fig. 4. Finally, in Sec. V C, we generate random Erdős-Rényi networks. For each experiment, we make choices for all six steps described in Sec. IV, which are detailed in Table VI. For purposes of reproducibility, all MATLAB codes constructed are available online [45].

A. Two-node networks

Experiments A1–A2 investigate a simple, two oscillator system. This could be an example in economics, such as the relationship between oil shocks and recessions. We try all possible two-node networks (see Fig. 3) and vary the parameters of the system Eq. (5) with $n = 2$. The parameter choices for our experiments are summarized in Table VI. We present the results from Experiment A1 as a confusion matrix in Table II. Each row shows the distribution of output networks for a given true network. If the method perfectly recovers

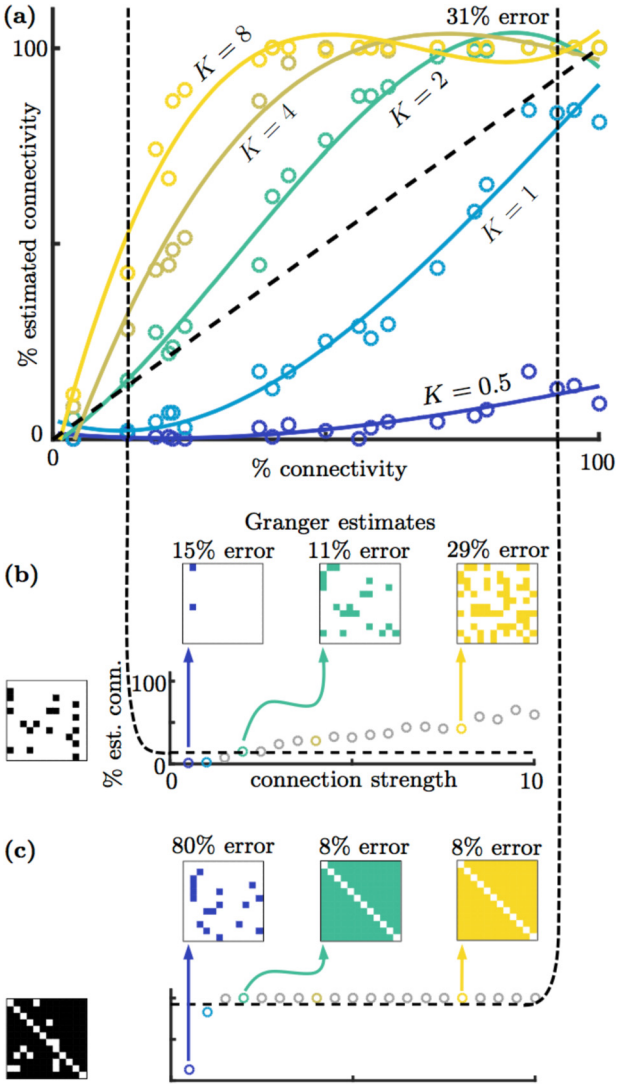

FIG. 9. Results from Experiment C1. Here $n = 12$ and twenty different Erdős-Rényi networks are generated while varying the percentage of connections. We also vary the connection strength K. In panel (a), we plot the true percentage of connections versus the estimated percentage of connections for five values of K. If the percentage of connections was correct, our points would be on the dashed diagonal line. However, they may still have the wrong edges even if the correct number are inferred. For the sparse and dense cases, a varying connection strength K is considered in panels (b) and (c). The correct percentage of connections is plotted as a horizontal dashed line for reference. In the inset plots, we see some examples of the inferred networks. These are colored visualizations of the adjacency matrices. White squares denote zeros (no edge) and colored squares denote ones (an edge).

the connectivity of all of the networks, this matrix would have entries of 100% along the diagonal. Instead we see that networks with one edge are rarely recovered correctly. The method has a tendency to overestimate the number of edges. We will see in later experiments that this pattern continues as the size of the network increases; performance is weakest when the number of edges is not extremely low or extremely high.

It may be argued that there was too much noise on the data for accurate connectivity reconstruction. We decrease the

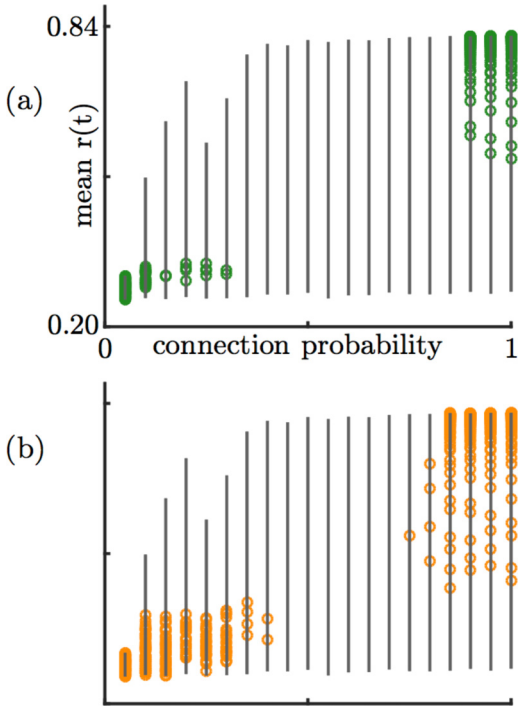

FIG. 10. Optimal bands of synchrony. We consider the results of Experiment C1 in terms of average r, the synchrony measure in Eq. (6). For each connection probability p across the horizontal axis, we plot a gray line showing the range of average synchrony r attained as we varied the connection strength K. We then plot green circles in panel (a) for the values of r for which the error was less than 10%. We also plot orange circles in panel (b) for the values of r for which the error was less than 20%.

noise strength to 0.5 for Experiment A2, since in this case, the low noise does not cause warnings in the MVGC Toolbox. Figure 6 compares the results from these two experiments. We find that, perhaps counterintuitively, the error is higher with lower noise. In particular, the lower noise results in even more overestimation of edges. Another pattern that will persist for larger networks is that as the connection strength increases, so does the number of edges inferred.

B. Two-community example

For Experiments B1–B3, we return to the two-community example in Fig. 4. In Experiment B1, we try to reconstruct a 12-node network [Eq. (5) with $n = 12$] using Granger causality on the same data plotted in the right panel of Fig. 4. The full parameter choices for this experiment are given in Table VI. The resulting network is shown in Fig. 7(b). There are many extra edges and some missing edges, resulting in an error of 25%. Note that the community structure is lost even though it is clear from the plot of the data in Fig. 4 that the blue and green nodes synchronize separately. An error of 25% may sound reasonable, but visually comparing the two networks suggests that the error is significant. Similar error percentages are used as evidence of a Granger causality variation working well in papers such as [42].

Addressing warnings. The MVGC Toolbox did produce warnings for Experiment B1, so for Experiment B2, we

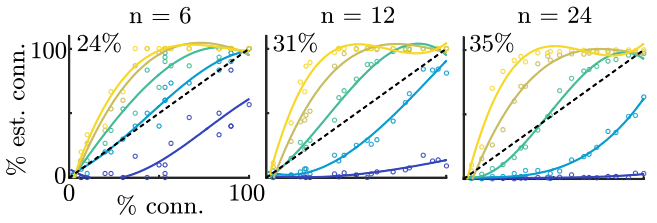

FIG. 11. Varying number of oscillators. Here we plot the percent connectivity vs estimated percent connectivity and five values of connection strength K for three numbers of oscillators. Left to right, we compare six, twelve, and twenty-four oscillators. The general pattern is consistent, but the average error seems to grow with the number of oscillators.

increased the noise to a strength of $s = 2.5$, leaving the rest of the parameters the same. The new data did not cause any warnings, and the resulting network is shown in Fig. 7(c). This network had an error of 22%. It is missing many edges but also added some, including connecting the two communities.

Varying time sampling. In Experiment B3, we tried solving the Kuramoto model again but after halving the step size Δt. Generally, we hope that algorithms are stable, i.e., that small changes in the input will lead to small changes in the output. However, changing the time sampling led to a vastly different estimated network. Again, the data did not cause any warnings,

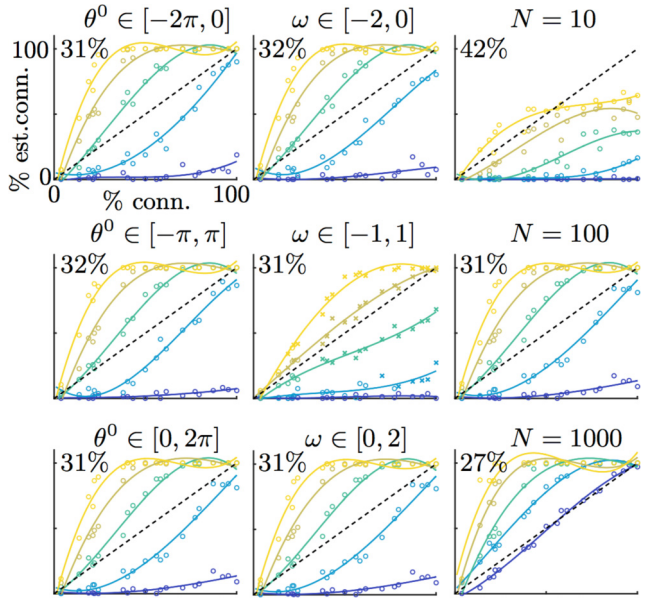

FIG. 12. Varying system parameters. We plot the percent connectivity vs estimated percent connectivity and five values of connection strength K while changing the distribution of initial conditions, the distribution of natural frequencies, and the number of trials. In the first column, we vary the distributions of random initial conditions, and in the second column, we vary the distributions of random natural frequencies. In the third column, we change the number of trials. The general pattern is consistent except when the number of trials is varied; the number of edges inferred grows as the number of trials grows. Results accompanied by a warning are marked with an "×" instead of a circle.

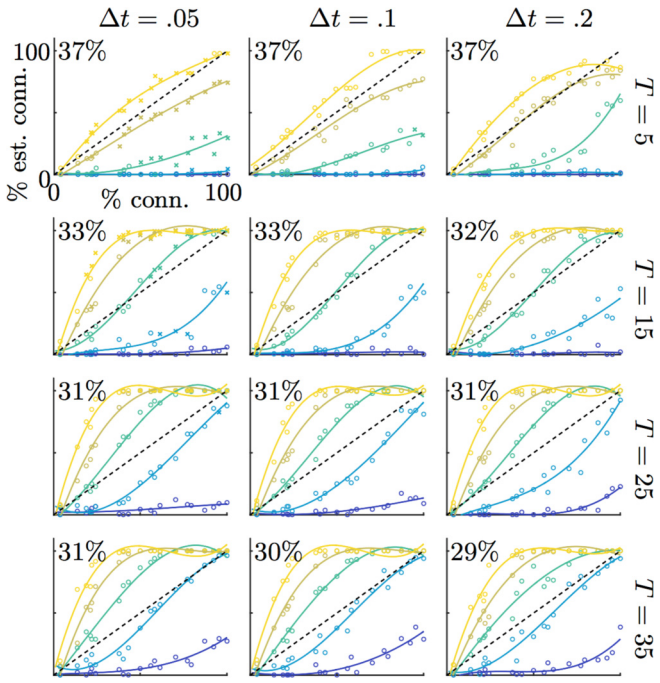

FIG. 13. Varying data sampling in time. We plot the percent connectivity vs estimated percent connectivity and five values of connection strength K while changing the data sampling in time. We use data from time 0 to T, where T varies down the rows. We use a time step of Δt where Δt varies across the columns. The general pattern is consistent except when the end time T is small. Results accompanied by a warning are marked with an "×" instead of a circle.

but this time, the number of edges were vastly overestimated, as shown in Fig. 7(d). This network has an error of 30%.

Checking autocovariance decay. If the autocovariance sequence does not decay exponentially, the data is not suitable for VAR modeling. This should be detected by the toolbox, but as a verification, we plot the required exponential decay in Fig. 8. The parameter choices for Experiments B1–B3 are summarized in Table VI.

C. Erdős-Rényi networks

In our remaining experiments (Experiments C1–C30 and D1–D2) we consider random Erdős-Rényi networks. For each experiment, we vary p, the probability that a directed edge exists, and we vary K, the connection strength. See Table VI for all of the parameter choices. Experiments C2–C30 are small variations on Experiment C1. A sampling of the results for Experiment C1 are shown in Fig. 9. In panel (a), we plot the percentage of true edges against the percentage of estimated edges. If the density of edges was inferred correctly, the results should match the identity line (the diagonal dashed line). The next assessment is whether or not the edges inferred were actually the correct ones. However, we generally do not even estimate the correct number of edges. Just as we saw with the two-node case in Fig. 6 and Table II, the number of edges is most accurate for the extremes—very sparse or very dense. Another general pattern persists: for lower connection strength, pairwise-conditional GC underestimates

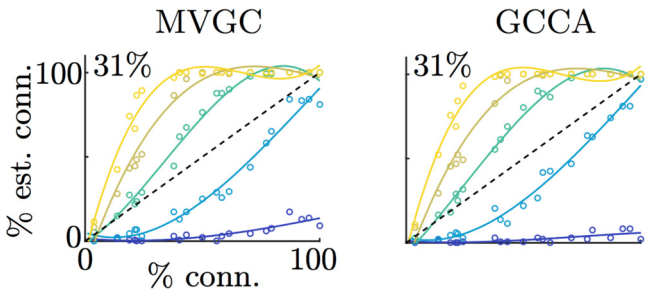

FIG. 14. Comparing implementations. Here we plot the percent connectivity vs estimated percent connectivity and five values of connection strength K. In Experiment D1, we repeat Experiment C1 with the GCCA implementation and observe very little difference.

the number of edges, and for higher connection strength, pairwise-conditional GC overestimates the number of edges.

The sparse and dense limits of connectivity are the only two regions where the inferred number of connections is somewhat consistent with the ground truth. In panels 9(b) and 9(c), we consider these two limiting network cases more closely. In particular, they are marked by two vertical dashed lines in panel 9(a). Here we plot the connection strength against the percentage estimated connectivity for a broader range of K values. The correct percentage connectivity is marked with horizontal dashed lines. We see again that for low connection strengths, the number of edges is underestimated. As K increases, our density estimates go from underestimating to overestimating the connectivity. We can visualize the exact networks inferred for three values of K, $K = 0.5$, 2, and 8. We see that even when the density of edges is approximately correct, the actual chosen edges do not match.

The Erdős-Rényi networks can also be analyzed from the viewpoint of the synchrony metric. We consider how our results relate to the strength of connection and the average order parameter $r(t)$ [Eq. (6)] for each data set. In the top plot of Fig. 10, we consider each of the 20 networks, plotted by the connection probability p used to generate them. For each network, we generated the data for 100 values of connection strength K. In general, higher connection strength K means that the network is more likely to synchronize—a higher average $r(t)$ is produced (see Fig. 2). The full range of average $r(t)$ values attained for each network is plotted as a gray line segment. We see that, for sparse networks, high synchronization was not attained for any value of K in our range. This makes sense, since $r(t)$ measures synchronization over the entire network, and a sparse network will not even be fully connected. On the other hand, we see that dense networks

TABLE III. Network inference toolboxes and methods compared in our simulations.

Method	Acronym	Ref.
Multivariate Granger causality	MVGC	[29]
Granger causal connectivity analysis	GCCA	[46]
Granger causality test	GCT	[48]
Extended Granger causality	eGC	[49]
eGC toolbox for standard GC		

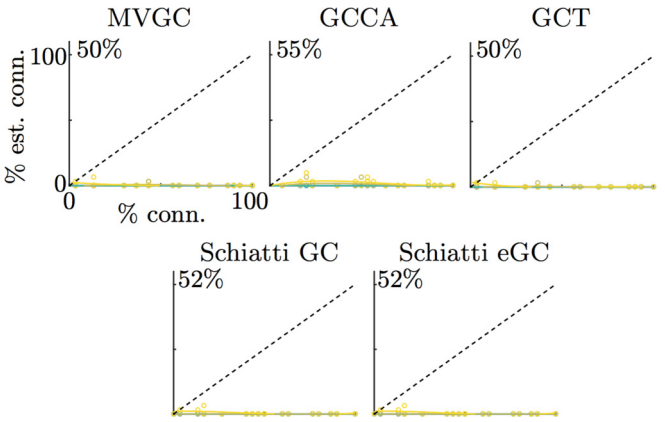

FIG. 15. Comparing implementations. Here we plot the percent connectivity vs estimated percent connectivity and five values of connection strength K. In Experiment D2, we compare the MVGC Toolbox to other implementations. We observe that each method infers very few edges when considering only one trial at a time and voting over 50 sets.

attain a wide range of synchronization as K is varied. We then checked for the cases where the percentage wrong was under ten percent and plotted them as green circles. We see that for sparse networks, the error is best when the network does not synchronize. For dense networks, the error is best when the network does synchronize strongly. For medium-density networks, the error is never below 10%. In the bottom panel of Fig. 10, we check a looser standard—plotting all cases with the error below 20%. We see that the general pattern continues.

Many variations to the experiment can be performed, including varying the number of nodes, percentage of connectivity in the Erdős-Rényi network, and the strength of connections. These variations are summarized in Table VI. Figure 11 demonstrates the connectivity results as a function of network size. The computations show that the qualitative behavior does not change with n. We can also vary the θ^0 (initial condition) distribution, the ω (natural frequency) distribution, and the number of trials N, the third dimension of our data cube. These results are shown in Fig. 12. We note that having both positive and negative natural frequencies seems to cause many warnings, perhaps due to different synchronization effects. The number of trials has a large impact on the number of edges inferred—as N increases, so does the number of edges. The other variations in the experiments do not change the qualitative shape of the results. We also modify how we sample in time. In Fig. 13, we vary the time step Δt down the rows and vary the end time T across the columns. This means that the number of observations m is different in each plot.

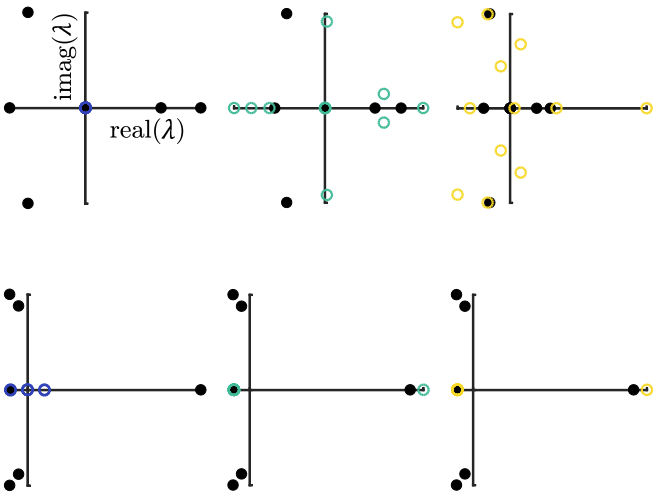

FIG. 16. Eigenvalue comparison. One measure of accuracy is how well the estimated network $\tilde{A}$ would recreate the same dynamics as the true network A. We therefore compare the eigenvalues of $\tilde{A}$ (open colored circles) and A (closed black circles) for the six estimated networks in Fig. 9. We see many cases of eigenvalues being significantly wrong. For example, in the first plot, the inferred network has only eigenvalues of about 0, missing eigenvalues for significant growth.

TABLE IV. Closeness ranking, part I. Closeness has been used to rank institutions on a network. We return to Fig. 9 and compare the closeness ranking on the first true network to the closeness rankings on three estimates. We see that the rankings formed from the estimated networks have little relationship with the true ranking.

Node	True rank	Estimated, $K = 0.5$	Estimated, $K = 2$	Estimated, $K = 8$
1	12	1	8	3
2	11	3	9	12
3	3	3	4	3
4	10	3	5	3
5	1	3	5	7
6	6	3	10	10
7	4	1	7	8
8	1	3	2	1
9	5	3	10	11
10	6	3	2	3
11	9	3	1	2
12	6	3	10	8

TABLE V. Closeness ranking, part II. Closeness has been used to rank institutions on a network. We return to Fig. 9 and compare the closeness ranking on the second true network to the closeness rankings on three estimates. We see that the rankings formed from the estimated networks have little relationship with the true ranking.

Node	True rank	Estimated, $K = 0.5$	Estimated, $K = 2$	Estimated, $K = 8$
1	6	11	1	1
2	1	9	1	1
3	1	9	1	1
4	6	3	1	1
5	1	5	1	1
6	1	6	1	1
7	6	2	1	1
8	6	1	1	1
9	11	11	1	1
10	6	6	1	1
11	12	3	1	1
12	1	8	1	1

TABLE VI. Summary of experiments. For Experiments C1–C30, gray boxes highlight any change from the usual parameters.

	n	A	K	θ^0	ω	N	Δt	T	m	s	Prep	Voting
A1	2	All four two-node networks	0.1,0.2, ...,10	$[0,2\pi]$	$[-1,1]$	100	0.1	25	250	2.5	$\cos(\theta)$	None
A2	2	All four two-node networks	0.1,0.2, ...,10	$[0,2\pi]$	$[-1,1]$	100	0.1	25	250	0.5	$\cos(\theta)$	None
B1	12	Fig. 4	5	[a]	[b]	1	0.1	25	250	0.1	$\cos(\theta)$	None
B2	12	Fig. 4	5	[a]	[b]	1	0.1	25	250	2.5	$\cos(\theta)$	None
B3	12	Fig. 4	5	[a]	[b]	1	0.05	25	500	2.5	$\cos(\theta)$	None
C1	12	E-R, $p = 0.05,0.1,\ldots,1$	0.1,0.2, ...,10	$[0,2\pi]$	[0,2]	100	0.1	25	250	2.5	$\cos(\theta)$	None
C2	6	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.1	25	250	2.5	$\cos(\theta)$	None
C3	24	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.1	25	250	2.5	$\cos(\theta)$	None
C4	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[-2\pi,0]$	[0,2]	100	0.1	25	250	2.5	$\cos(\theta)$	None
C5	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[-\pi,\pi]$	[0,2]	100	0.1	25	250	2.5	$\cos(\theta)$	None
C6	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	$[-2,0]$	100	0.1	25	250	2.5	$\cos(\theta)$	None
C7	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	$[-1,1]$	100	0.1	25	250	2.5	$\cos(\theta)$	None
C8	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	10	0.1	25	250	2.5	$\cos(\theta)$	None
C9	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	1000	0.1	25	250	2.5	$\cos(\theta)$	None
C10	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.05	5	25	2.5	$\cos(\theta)$	None
C11	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.05	15	75	2.5	$\cos(\theta)$	None
C12	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.05	25	125	2.5	$\cos(\theta)$	None
C13	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.05	35	175	2.5	$\cos(\theta)$	None
C14	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.1	5	50	2.5	$\cos(\theta)$	None
C15	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.1	15	150	2.5	$\cos(\theta)$	None
C16	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.1	35	350	2.5	$\cos(\theta)$	None
C17	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.2	5	50	2.5	$\cos(\theta)$	None
C18	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.2	15	300	2.5	$\cos(\theta)$	None
C19	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.2	25	500	2.5	$\cos(\theta)$	None
C20	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.2	35	700	2.5	$\cos(\theta)$	None
C21	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.1	25	250	1.5	$\cos(\theta)$	None
C22	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.1	25	250	3.5	$\cos(\theta)$	None
C23	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.1	25	250	2.5	$\theta_{i+1} - \theta_i$	None
C24	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.1	25	250	2.5	Detrend	Tenths in time
C25	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	1	0.1	25	250	2.5	$\cos(\theta)$	1000 sets
C26	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	10	0.1	25	250	2.5	$\cos(\theta)$	100 sets
C27	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.1	25	250	2.5	$\cos(\theta)$	10 sets
C28	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.1	25	250	2.5	$\cos(\theta)$	Halves in time
C29	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.1	25	250	2.5	$\cos(\theta)$	Fourths in time
C30	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.1	25	250	2.5	$\cos(\theta)$	Eighths in time
D1	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	100	0.1	25	250	2.5	$\cos(\theta)$	None
D2	12	E-R, $p = 0.05,0.1,\ldots,1$	0.5,1,2,4,8	$[0,2\pi]$	[0,2]	1	0.1	25	250	2.5	$\cos(\theta)$	50 sets

[a] $\theta^0 = [10,11,6,9,5,3,8,4,0,2,7,1]\frac{2\pi}{11}$.
[b] $\omega = [0.6,0.4,0.65,0.35,0.55,0.45,-0.1,-0.3,-0.05,-0.35,-0.15,-0.25]$.

Extensive computational experiments also considered varying the noise added to the data (including adding it before or after applying cosine), differencing and detrending of the data instead of applying a cosine (recommended by [29] for making data covariance stationary), and splitting the data and weighting multiple inferences of network structure. In all these cases, the same trends as shown in the preceding figures hold, i.e., as the connection strength increases, so does the number of edges inferred. In no case does the GC method produce accurate results.

The MVGC Toolbox is a successor to the Granger Causal Connectivity Analysis (GCCA) toolbox [46]. The newer toolbox adds more diagnostic warnings and errors and is intended to be more accurate. For Experiment D1, we rerun Experiment C1 with the GCCA toolbox and obtain very similar results (Fig. 14). We also try three other implementations of Granger causality. First, we consider the implementation of the classic Granger causality test (GCT) [47] provided with the [48] paper. This implementation only accepts one trial at a time ($N = 1$). They compare it to other network inference procedures, including the MVGC toolbox [29], on data generated by three models. The first two models are simply VARs. The third adds latent and exogenous variables. Although this is not explicitly stated, it seems that all implementations correctly infer the first two network models but sometimes make mistakes on the third. The focus of the paper is on whether the methods are consistent over repeated trials. They state that the MVGC toolbox [29] is *anomalous* in its lack of compliance to Neyman-Pearson criteria. We additionally consider a version of Granger causality called *extended Granger causality* that allows instantaneous causal relationships (zero lag) [49]. The paper is accompanied by two implementations, one that includes zero-lag relationships (which we refer to as *Schiatti eGC*) and one that does not

(which we refer to as *Schiatti GC*). These implementations also only accept one trial at a time ($N = 1$). They are tested on data generated by an extended VAR model that allows for zero-lag relationships. The methods are then compared on real data where the true network structure is not known. Table III shows the methods compared along with their commonly used acronym and initial source reference.

In order to test implementations that only accept one trial at a time, in Experiment D2, we generate 50 trials, separate them into sets of $N = 1$, and then vote over the 50 estimates. We compare the MVGC [29], GCCA [46], GCT [48], Schiatti eGC, and Schiatti GC [49] implementations. We see that in all five implementations, almost no edges are kept after voting (Fig. 15). Although individual estimates contain some correct edges, the methods are not sufficiently consistent to estimate the same edge at least half of the time. This suggests that it is not sufficient to test a new version of Granger causality on data generated by a VAR model.

Thus far, we have evaluated the accuracy of our results in three ways: visually comparing the original network to the estimated network (as in Figs. 7 and 9), comparing the percentage connnectivity to the percentage estimated connectivity, and calculating an error—the percentage of potential edges that are correctly labeled as an edge or not an edge. Perhaps in some applications, what is important is reconstructing a network that would produce similar dynamics. Thus we may be concerned with comparing the eigenvalues of the estimated network $\tilde{\mathbf{A}}$ to the original network $\mathbf{A}$. We return to the six estimated networks in Fig. 9: two true networks and the corresponding estimates when $K = 0.5$, 2, and 8. We plot the eigenvalues of the true network with the eigenvalues of the estimated network in Fig. 16. We see that even when the densities are relatively correct, the dynamics produced by the connectivity matrix would be significantly wrong.

In [50], Billio *et al.* use a form of Granger causality to infer a network of financial institutions. Then they propose various econometric measures of connectedness to assign ranks to institutions and predict financial loss. One metric used to rank is closeness: the average distance from node j to the remaining nodes, where unconnected nodes are defined to have the maximum distance, $n - 1$. In Tables IV and V, we use this metric to assign ranks to the twelve nodes in our examples from Fig. 9. We see that a ranking formed from the estimated networks has little relationship with the true ranking. This once again shows that the GC method fails to capture meaningful results concerning the ground truth network.

VI. CONCLUSIONS

The inference of causal structure from time series measurements remains one of the most challenging tasks in data-driven discovery across the sciences. It has become especially important in the emerging area of network science for understanding how different dynamical nodes of a system interact to produce overall network functionality. A variety of statistical methods have been instrumental in developing mathematical architectures for inferring connections between nodes. These methods often make assumptions about the physical processes generating the data and the form of the connections (e.g., linear). Foremost among these methods is pairwise-conditional Granger causality as it has been used extensively across a variety of disciplines [7–16]. We consider a nonlinear, networked dynamical system of Kuramoto oscillators as a ground-truth test model for inferring network connectivity and demonstrate that, without exception, Granger causality gives highly inaccurate results for the inferred causal relations and the eigenvalues of the connectivity matrix. This is consistent with an additional study of the quantitative accuracy of the GC method [22]. This is an important assessment of the statistical efficacy of the GC method, and it further suggests that it should be carefully validated before use with any networked time series data.

The Kuramoto oscillator model is chosen for consideration for this study as it has become a canonical model in networked dynamical systems. It has simple oscillatory behavior that is influenced by its interaction structure. Both synchronization, partial and complete, and chaotic behavior is possible in the network. Given that we can specify a ground truth connectivity structure, the GC method can be used to test the efficacy of the inference method. We observe that as the connection strength or number of trials increases, we transition from underestimating the number of edges to overestimating the number of edges, quickly surpassing the correct number. This pattern is consistent over variations in parameters, and individual networks inferred are not consistent with the ground truth (see Fig. 7). This suggests that the algorithm is not stable; small changes in the input data lead to large changes in the estimated network. Accuracy is best on very sparse or very dense networks, although arguably still not sufficient. Perhaps the errors are controlled in very sparse networks because the network overall does not synchronize even for high connection strength, mitigating confusion from synchronized but unconnected nodes. (See Fig. 10.) Similarly, perhaps the errors are limited in dense networks because the networks become fairly synchronized overall, thus implying many connections.

The limitations of GC—a chief inference method for connectivity in networked dynamical systems—may have serious implications for the neuroscience community. In fact, Bullmore and Sporns [51] emphasize the distinction between structural (anatomical) links and effective (causal-functional) relationships between the network elements. Our simulations suggest that causal inference alone might provide unsatisfactory results, and thus one should try to maximize the usage of anatomical and biophysical constraints in conjunction with the method. For instance, in Experiments B1–B3 (see Fig. 7), GC inference failed to detect a simple two-community structure in a relatively small network; in some cases, the inferred network missed many edges and, in others, it overestimated their number. More importantly, it inferred nonexistent connections between the disjoint communities. Neuroscientists should try to incorporate this sort of prior information into the algorithm and carefully validate if the inferred results are within an anatomically plausible range of parameters. Some may argue that the goal in connectomic inference is not to exactly reconstruct links but to only capture similar statistical or functional-dynamical properties of the system. Figure 16 suggests that they might not hold for GC since there are significant discrepancies between the eigenvalues of the original matrix and of the inferred ones, which would

lead to qualitatively different dynamics. Tables IV and V also show mismatches between the closeness ranking, pointing out that estimated networks may have little relationship with the true ranking. Furthermore, neural networks exhibit not only excitatory connections but also inhibitory ones. This would likely complicate the inference problem since the number of possible wrong links would increase dramatically. Finally, our analysis was limited to the Kuramoto model, which exhibits a simpler repertoire of collective states compared to more complex dynamical models such as spiking neural networks.

It is possible that a property of the data generated by Kuramoto oscillators makes it unsuitable for Granger causality computations. However, we used all provided tools for checking for problems. We suggest that further study is required to understand the conditions under which the results can be trusted and to provide practical ways to check those conditions. Unfortunately, of the myriad of uses made of GC in practice [8,10–16], none of the authors validate the technique against a ground truth example. There may be another version of Granger causality that can correctly infer networks from our time series data. However this remains an open challenge to the community at large. Our code is available online [45] so that our experiments can be repeated with other network inference methods. In particular, we have shown that it is important to test methods on data that is not simply generated from a VAR model and that a range of networks should be considered. It may be possible that other statistical innovations for determining causality can be used to infer network structure [3–5], including new directions leveraging independent component analysis [17], the phase slope index (PSI) [22], and/or network structure [18]. More recent innovations have considered the construction of local models of GC to infer the broader inference network [20] and finding time-delay embeddings in systems displaying attractor structures [19]. Thus open issues remain, such as whether the adopted toolboxes can be improved to solve the issues highlighted in this work, and whether innovations may provide different and more appropriate ways for GC to work in practice. Regardless of technique, this is a fundamentally difficult problem [52] requiring new ideas, innovations, and methods from the broader mathematical sciences community. Network science is here to stay and inference models will only increase in importance to the physical sciences community.

[1] N. Wiener, The theory of prediction, Mod. Math. Eng. **1**, 125 (1956).

[2] C. W. J. Granger, Investigating causal relations by econometric models and cross-spectral methods, Econometrica: J. Econ. Soc. **37**, 424 (1969).

[3] G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge University Press, Cambridge, UK, 2015).

[4] S. L. Morgan and C. Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd ed. (Cambridge University Press, Cambridge, UK, 2015).

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

[6] P. W. Holland, Statistics and causal inference, J. Am. Stat. Assoc. **81**, 945 (1986).

[7] K. D. Hoover, Causality in economics and econometrics, *The New Palgrave Dictionary of Economics* (Palgrave Macmillan, New York, NY, 2008), p. 2.

[8] J. D. Hamilton, Oil and the macroeconomy since World War II, J. Pol. Econ. **91**, 228 (1983).

[9] S. L. Bressler and A. K. Seth, Wiener-Granger causality: A well established methodology, Neuroimage **58**, 323 (2011).

[10] S. L. Bressler, W. Tang, C. M. Sylvester, G. L. Shulman, and M. Corbetta, Top-down control of human visual cortex by frontal and parietal cortex in anticipatory visual spatial attention, J. Neurosci. **28**, 10056 (2008).

[11] J. F. Alonso, S. Romero, M. À. Mañanas, and J. Riba, Serotonergic psychedelics temporarily modify information transfer in humans, Int. J. Neuropsychopharm. **18**, pyv039, (2015).

[12] A. K. Charakopoulos, T. E. Karakasidis, and A. Liakopoulos, Spatiotemporal analysis of seawatch buoy meteorological observations, Environ. Process. **2**, 23 (2015).

[13] D. Yellin, A. Berkovich-Ohana, and R. Malach, Coupling between pupil fluctuations and resting-state fmri uncovers a slow build-up of antagonistic responses in the human cortex, NeuroImage **106**, 414 (2015).

[14] F. Yang, P. Duan, S. L. Shah, and T. Chen, *Capturing Connectivity and Causality in Complex Industrial Processes* (Springer Science & Business Media, New York, 2014).

[15] G. Castagneto-Gissey, M. Chavez, and F. De Vico Fallani, Dynamic Granger-causal networks of electricity spot prices: A novel approach to market integration, En. Econ. **44**, 422 (2014).

[16] B. Zong, Y. Wu, J. Song, A. K. Singh, H. Cam, J. Han, and X. Yan, Towards scalable critical alert mining, in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, 2014), pp. 1057–1066.

[17] S. Shimizu, P. Hoyer, A. Hyvärinen, and A. Kerminen, A linear non-Gaussian acyclic model for causal discovery, J. Mach. Learn. Res. **7**, 2003 (2006).

[18] V. Pernice, B. Staude, S. Cardanobile, and S. Rotter, How structure determines correlations in neuronal networks, PLoS Comput. Biol. **7**, e1002059 (2011).

[19] G. Sugihara, R. May, H. Ye, C. Hsieh, E. Deyle, M. Fogarty, and S. Munch, Detecting causality in complex ecosystems, Science **338**, 496 (2012).

[20] B. Wahl, U. Feudel, J. Hlinka, M. Wächter, J. Peinke, and J. Freund, Granger-causality maps of diffusion processes, Phys. Rev. E **93**, 022213 (2016).

[21] C. W. J. Granger, Testing for causality: A personal viewpoint, J. Econ. Dyn. Control **2**, 329 (1980).

[22] G. Nolte, A. Ziehe, N. Krämer, F. Popescu, and K.-R. Müller, Comparison of Granger causality and phase slope index, in *NIPS Causality: Objectives and Assessment* (Citeseer, Vancouver, BC, 2010), pp. 267–276.

[23] P. A. Stokes, Ph.D. thesis, Massachusetts Institute of Technology, 2015.

[24] C. W. J. Granger, Time series analysis, cointegration, and applications, Am. Econ. Rev. **94**, 421 (2004).
[25] Y. Kuramoto, Self-entrainment of a population of coupled nonlinear oscillators, in *International Symposium on Mathematical Problems in Theoretical Physics* (Springer, New York, 1975), pp. 420–422.
[26] F. Dörfler and F. Bullo, Synchronization in complex networks of phase oscillators: A survey, Automatica **50**, 1539 (2014).
[27] P. Erdős and A. Rényi, On random graphs, Publ. Math. (Debrecen) **6**, 290 (1959).
[28] M. E. J. Newman, The structure and function of complex networks, SIAM Rev. **45**, 167 (2003).
[29] L. Barnett and A. K. Seth, The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference, J. Neurosci. Methods **223**, 50 (2014).
[30] E. Zagha, X. Ge, and D. A. McCormick, Competing neural ensembles in motor cortex gate goal-directed motor output, Neuron **88**, 565 (2015).
[31] H. Xu, E. Kroupi, and T. Ebrahimi, Functional connectivity from EEG signals during perceiving pleasant and unpleasant odors, in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference* (IEEE, New York, 2015), pp. 911–916.
[32] X. Niu, L. Shi, H. Wan, Z. Wang, Z. Shang, and Z. Li, Dynamic functional connectivity among neuronal population during modulation of extra-classical receptive field in primary visual cortex, Brain Res. Bull. **117**, 45 (2015).
[33] X. Wang, Y. Chen, S. L. Bressler, and M. Ding, Granger causality between multiple interdependent neurobiological time series: Blockwise versus pairwise methods, Int. J. Neural Syst. **17**, 71 (2007).
[34] S. Guo, A. K. Seth, K. M. Kendrick, C. Zhou, and J. Feng, Partial Granger causalityeliminating exogenous inputs and latent variables, J. Neurosci. Methods **172**, 79 (2008).
[35] X. Wu, C. Zhou, G. Chen, and J.-a. Lu, Detecting the topologies of complex networks with stochastic perturbations, Chaos **21**, 043129 (2011).
[36] Y. Chen, G. Rangarajan, J. Feng, and M. Ding, Analyzing multiple nonlinear time series with extended Granger causality, Phys. Lett. A **324**, 26 (2004).
[37] G. Wu, X. Duan, W. Liao, Q. Gao, and H. Chen, Kernel canonical-correlation Granger causality for multiple time series, Phys. Rev. E **83**, 056204 (2011).
[38] L. Faes, G. Nollo, and A. Porta, Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique, Phys. Rev. E **83**, 051112 (2011).
[39] D. Marinazzo, M. Pellicoro, and S. Stramaglia, Kernel Method for Nonlinear Granger Causality, Phys. Rev. Lett. **100**, 144103 (2008).
[40] L. Angelini, M. Pellicoro, and S. Stramaglia, Granger causality for circular variables, Phys. Lett. A **373**, 2467 (2009).
[41] L. Angelini, M. De Tommaso, D. Marinazzo, L. Nitti, M. Pellicoro, and S. Stramaglia, Redundant variables and Granger causality, Phys. Rev. E **81**, 037201 (2010).
[42] X. Wu, W. Wang, and W. X. Zheng, Inferring topologies of complex networks with hidden variables, Phys. Rev. E **86**, 046106 (2012).
[43] M. Chen, Y. Shang, Y. Zou, and J. Kurths, Synchronization in the Kuramoto model: A dynamical gradient network approach, Phys. Rev. E **77**, 027101 (2008).
[44] M. Brede, Synchrony-optimized networks of non-identical Kuramoto oscillators, Phys. Lett. A **372**, 2618 (2008).
[45] github.com/BethanyL/gc.
[46] A. K. Seth, A matlab toolbox for Granger causal connectivity analysis, J. Neurosci. Methods **186**, 262 (2010).
[47] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis* (Springer Science & Business Media, New York, 2005).
[48] K. Sameshima, D. Y. Takahashi, and L. A. Baccalá, On the statistical performance of Granger-causal connectivity estimators, Brain Inform. **2**, 119 (2015).
[49] L. Schiatti, G. Nollo, G. Rossato, and L. Faes, Extended Granger causality: A new tool to identify the structure of physiological networks, Physiol. Meas. **36**, 827 (2015).
[50] M. Billio, M. Getmansky, A. W. Lo, and L. Pelizzon, Econometric measures of connectedness and systemic risk in the finance and insurance sectors, J. Fin. Econ. **104**, 535 (2012).
[51] E. Bullmore and O. Sporns, Complex brain networks: Graph theoretical analysis of structural and functional systems, Nat. Rev. Neurosci. **10**, 186 (2009).
[52] M. T. Angulo, J. A. Moreno, A.-L. Barabási, and Y.-Y. Liu, Fundamental limitations of network reconstruction, arXiv:1508.03559.