PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

Research

Cite this article: Freilich MA, Rebolledo R, Corcoran D, Marquet PA. 2020 Reconstructing ecological networks with noisy dynamics. *Proc. R. Soc. A* **476**: 20190739. http://dx.doi.org/10.1098/rspa.2019.0739

Received: 31 October 2019
Accepted: 2 April 2020

Subject Areas:
complexity, statistics, environmental engineering

Keywords:
food webs, ecological networks, network inference, stochastic model

Author for correspondence:
Mara A. Freilich
e-mail: maraf@mit.edu

One contribution to a special feature 'A generation of network science' organized by Danica Vukadinovic-Greetham and Kristina Lerman.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4950948.

THE ROYAL SOCIETY
PUBLISHING

Reconstructing ecological networks with noisy dynamics

Mara A. Freilich[1], Rolando Rebolledo[2], Derek Corcoran[3] and Pablo A. Marquet[3,4,5,6]

[1]Massachusetts Institute of Technology-Woods Hole Oceanographic Institution Joint Program, Cambridge, MA, USA
[2]Instituto de Ingeniería Matemática, Facultad de Ingeniería, Universidad de Valparaíso, Valparaíso, Chile
[3]Departamento de Ecología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago, Chile
[4]Instituto de Ecología y Biodiversidad (IEB), Santiago, Chile
[5]The Santa Fe Institute, Santa Fe, NM, USA
[6]Instituto de Sistemas Complejos de Valparaíso (ISCV), Valparaíso, Chile

MAF, 0000-0003-0487-8518; PAM, 0000-0001-6369-9339

Ecosystems functioning is based on an intricate web of interactions among living entities. Most of these interactions are difficult to observe, especially when the diversity of interacting entities is large and they are of small size and abundance. To sidestep this limitation, it has become common to infer the network structure of ecosystems from time series of species abundance, but it is not clear how well can networks be reconstructed, especially in the presence of stochasticity that propagates through ecological networks. We evaluate the effects of intrinsic noise and network topology on the performance of different methods of inferring network structure from time-series data. Analysis of seven different four-species motifs using a stochastic model demonstrates that star-shaped motifs are differentially detected by these methods while rings are differentially constructed. The ability to reconstruct the network is unaffected by the magnitude of stochasticity in the population dynamics. Instead, interaction between the stochastic and deterministic parts of the system determines the path that the whole system takes to equilibrium and shapes the species covariance. We highlight the effects of long transients on the path to equilibrium and suggest a path forward for developing more ecologically sound statistical techniques.

1. Introduction

Species interaction networks offer a quantitative method for understanding the structure and dynamics of complex ecological systems, e.g. [1,2]. However, in high biodiversity environments, such as microbial communities, it may be difficult to directly observe species interactions. Instead, it may be easier to observe temporal and spatial patterns that result from species interactions. Based on this premise, co-occurrence of species in space and time is increasingly used to infer networks of species interactions [3–5].

While there is a plethora of metrics to construct networks based on observations of species distributions in time and space (see [6] for a review of metrics for time series), very few attempts have been made to verify these metrics, e.g. [7,8] and even fewer to place them on sound ecological *and* mathematical footing. Some of the theoretical issues that are largely unaddressed are how the network architecture affects the ability to detect interactions, what types of interactions and network motifs are best detected, and whether biological systems fit the statistical assumptions of the metrics used [9]. A recent study of a number of different tools for detecting interaction networks found that the tools produced different edges for the same real and simulated data and that the power to detect interactions depends on the distributions of species abundance [9,10].

Covariance and correlation are basic methods to quantify pairwise associations. Assuming that all species have Gaussian distributions, the covariance and the mean value is a complete descriptor of the species distributions. However, the Gaussian assumption is often not satisfied and there is not a direct link between correlation or covariance and interactions [7,8]. One way to relax the assumption of a Gaussian distribution while ensuring that the statistical technique is not making assumptions about missing data are to use an entropy maximizing method [11,12]. These methods also reduce the effects of indirect interactions or large environmental trends on the inferred associations. There is a similarity between the maximum entropy metric proposed by Lezon *et al*. [11] and the statistical techniques used in press experiments in ecology [13]. An entropy maximizing method does not measure pairwise associations, but instead infers the most probable whole network structure; the ability to detect relationships involving many interconnected members is pertinent to the use of this metric.

More than 30 years ago, Peter Yodzis [13] proposed an indeterminacy principle wherein available statistical techniques were insufficient to experimentally determine species interactions in natural communities for a large suite of ecological reasons, including indirect interactions and weak links in species networks. The uncertainty about the ability to detect interactions extends to null models, another commonly used ecological technique [14,15]. Even with modern statistical techniques, relationships involving more than three members of a community seem to be nearly impossible to detect reliably [9,16,17]. We present an updated indeterminacy principle based on both mathematical and ecological insights. In this study, we use a first principle stochastic species interaction model to investigate the impact of food web structure on both the dynamics of communities and the propagation of stochasticity through them and hence on the ability to detect the network structure of species interactions. We examine the power of three metrics; covariance, Pearson’s correlation, and inverse correlation to detect species interactions across a wide parameter space. These metrics are at the heart of most of the network reconstruction techniques. Unlike previous work [10], we focus on small community modules to gain an understanding of both the ecological and statistical reasons for trends in performance.

2. Methods

(a) Stochastic species interaction model

We model food webs (interaction networks) using a stochastic differential equation, which is a generalized Lotka–Volterra model with stochasticity (equation (2.1)). The generalized Lotka–Volterra model represents the dynamics of biomass in the main system, while the diffusion

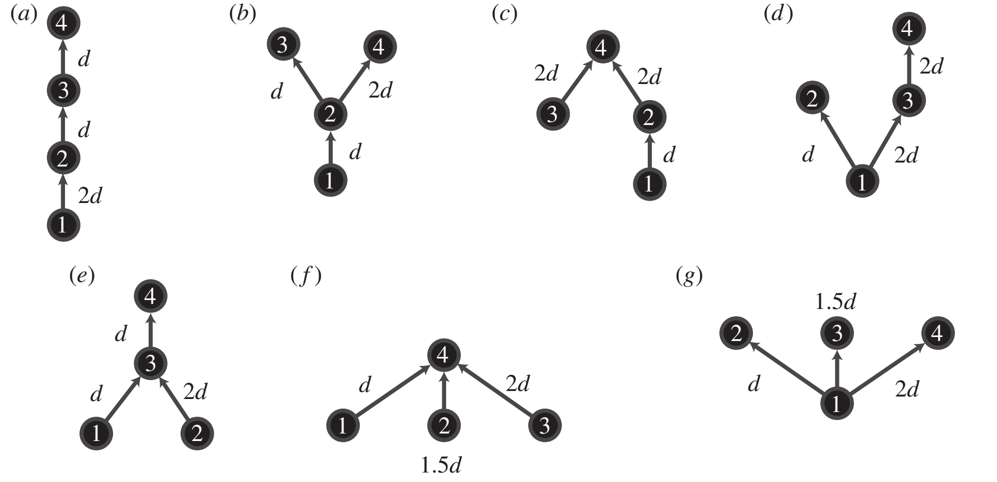

Figure 1. Possible predator–prey networks with four species, when distinguishing between trophic levels. The top predators are the uppermost species in the network. The arrows show movement of biomass from prey to predators. The values near each edge indicate the rate of this interspecific interaction. The parameter *d* is varied through the simulations.

coefficient represents biomass fluctuations associated with the interactions between the main system and the environment where it is embedded. This includes not only environmental variability, but everything that we do not model explicitly in the main system, such as interaction with species outside the main system and non-trophic interactions with species in the main system [18]. We modelled the biomass dynamics in the main system assuming that each species has resources outside the food web that allow it to grow, which prevents extinction at long times in the stochastic simulation:

$$\begin{cases} \mathrm{d}X_i(t) = \underbrace{X_i(t)\left(r_i + \sum_{j=1}^{N} D_{ij}X_j\right)\mathrm{d}t}_{\text{generalized Lotka–Volterra}} + \underbrace{\sqrt{\gamma_i X_i(t)(1-X_i(t))}\mathrm{d}W(t)}_{\text{diffusion}} \\ X_i(0), \text{ with a given probability distribution } \mu \end{cases} \tag{2.1}$$

where $X_i(t)$ represents the proportion of biomass in species i (or stochastic abundance of species i) and r_i its *per capita* growth rate in the absence of interspecies and intraspecific interactions. In addition, let a_i denote the strength of interaction. Species interactions are represented by a symmetric N by N matrix $D = (D_{i,j})_{i,j=1}^{N}$, which, for all $i \neq j$ represents the impact of predation upon the focal species as a biomass loss rate for prey and gain rate for predator. We define the intraspecific interactions as $D_{i,i} = -a_i$, for all $i = 1, \ldots, N$. The interspecific interactions can be represented as a network (as in figure 1). The magnitude of the interspecific interaction rates is given by d. The second term in the right-hand part of equation (2.1) represents interactions of the main system with the environment, where $W(t)$ is a Wiener or Brownian motion, γ_i is the intensity of the fluctuation for all $i = 1, \ldots, N$. And one assumes that the initial distribution μ is known.

One may write the previous equation in vector and integral notation as follows. Let denote $X(t)$ the column vector with components $X_i(t)$, similarly, call r the vector with coordinates r_i, $\sigma(x)$ the diagonal matrix with components $\sqrt{\gamma_i x_i(1-x_i)}$, where $x \in [0,1]^N$, (i.e. each $x_i \in [0,1]$). Moreover, let us denote $\bullet$ the Schur (or component-wise) product of vectors. Equation (2.1) becomes:

$$X(t) = X(0) + \int_0^t X(s) \bullet (r + DX(s))\,\mathrm{d}s + \int_0^t \sigma(X(s))\,\mathrm{d}W(s). \tag{2.2}$$

This model provides considerable insight into the way in which species covary and correlate with each other. Past work has used similar generalized Lotka–Volterra models, but with competition rather than predation, to draw conclusions about the influence of niche and neutral processes in ecological communities [19,20].

A proof of the existence and uniqueness of the solution in distribution P_θ to this general equation is given in theorem 1 of [21]. P_θ is a probability defined on the set $C(\mathbb{R}_+, \mathbb{R}^N)$ of continuous functions, depending on the parameter $\theta = (r, D, \gamma)$, where γ is the vector with components γ_i. Notice that here we assume a form of the functional diffusion coefficient $\sigma(x)$ depending on γ only, which models the rescaled limit of birth and death processes that appear in a number of ecological and genetics models [18]. This assumption introduces a simplification of the statistical inference problem.

(b) Model parameters

Predator–prey interactions are modelled using a linear functional response in equation (2.1). The numerical simulation is performed with forward Euler time stepping. Since the predation coefficients are symmetric, the loss to the prey in biomass is equivalent to the gain to the predator in biomass and there is no biomass loss from the system, except due to stochastic fluctuations (an 'open system'). We systematically explore the parameter space of the model by varying the amount of stochasticity from the Brownian term and the ratio of interspecific to intraspecific interactions.

(i) Interaction motifs

We simulate species interactions for all possible directed food webs with four species. These motifs are the building blocks from which larger networks are built. Interaction topologies are shown in figure 1. In the deterministic case, the species grow to an equilibrium in which all species coexist. The parameters a_i and r_i are chosen such that all species grow to the same biomass at equilibrium when there are no interactions. The value of the equilibrium depends on both intraspecific and interspecific interactions associated with the strength of predation, in other words, the amount of energy transferred between prey and predators. In order to systematize the simulations across network geometries, we assign the coefficients $a_i = 0.05$ and $r_i = 0.5$ to the top predator in the simulation (any species that has no predators). We assign the same coefficients as above $a_i = 0.1$ and $r_i = 1$ to basal species and intermediate consumers. As a result, all species have the same carrying capacity in the absence of interactions ($r_i/a_i = 0.1$), but the prey species grow towards this carrying capacity more quickly. If there is more than one species with the same interactions, we assign different interaction rates so that the species dynamics differ between these species. In order to explore the effects of the parameter choices, we vary γ, the coefficient for the stochasticity, between 0 and 0.05. All species have the same stochastic growth coefficient γ, but the noise is independent. We vary the interspecific interaction coefficient between 0.001 and 0.25. We simulated 10 000 time series for each parameter combination. We present results from time step 150 to time step 350. There are 20 possible undirected networks for the model with four species and three links between species. While the food web models are directed, the networks inferred by correlation and covariance are necessarily undirected. The LIMITS algorithm can detect directed networks.

(ii) Niche model trophic network

In order to evaluate the implications of the results obtained by studying the dynamics of motifs in detail, we simulate larger networks using the niche model for trophic networks (food webs) [22]. This model is a stochastic model that constructs trophic networks that share many of the properties of observed trophic networks. We specify that the networks have 100 species and are sparse, with a connectivity of 0.05. There is a continuous spectrum of trophic levels in the model so we assign the same intraspecific interaction parameters to all species, $r_i = a_i = 0.09$. We vary the interspecific interaction strength from 0.01 to 1. We assign the same interaction strength to every link present. We also vary the stochasticity γ. We perform 1000 simulations for each parameter combination, each with a different randomly generated trophic network.

(c) Network reconstruction

We calculated species covariance matrices from the time series generated by the stochastic simulations as cov = $\overline{(X_i - \overline{X_i})(X_j - \overline{X_j})}$ where $\bar{\cdot}$ represents temporal averaging. We also calculated species correlation matrices using Pearson's correlation coefficient corr = cov$/S_{X_i}S_{X_j}$ where S_{X_i} is the standard deviation in time of species i. The maximum entropy technique we used is computed as the inverse Pearson correlation coefficient matrix [11]. This is a more complete descriptor if the species distributions are Gaussian.

For the four species motifs, we list the top three inferred connections for each network construction technique (covariance, correlation, and inverse correlation [MaxEnt]). The links with magnitudes that are above a certain threshold are selected when using the covariance and inverse correlation techniques. The links with the lowest p-values are selected when using correlation. For the 100 species network we only present the results using correlation to determine the links. We selected all links with a p-value below 0.01 when using correlation. We then use a binomial approximation to assign confidence intervals to the probability of detecting each possible network.

LIMITS [23] is one of the most successful network reconstruction techniques for time-series data [24]. LIMITS is designed to reconstruct Lotka–Volterra networks, which are the type of interactions used in this study. LIMITS generates directed networks while the other metrics generate undirected networks. We implement this metric using MATHEMATICA code provided by the authors. We use a threshold for inferring an interaction of 0.01. The results from LIMITS provide a baseline for success of reconstruction using the other metrics.

(d) Statistical analysis

We used Generalized Linear Models (GLM) [25] to find the relationship between the probability of finding the true configuration of the network (*prob*) and its predictive variables the strength of the relationships (*d*), the level of noise (*gamma*), the network motif (*network*), and the method used to infer the network structure (*Method*). We performed logistic regression using each of the predictive variables in isolation and in combination to predict the probability of detecting the true network. This quantifies the importance of each of the above factors on the network reconstruction.

3. The statistical problem

Before presenting the results, we outline some expectations about the performance of the statistical tools from a theoretical perspective. The statistical inference problem consists of the identification of the probability P_θ that rules the dynamics of the open system described by (2.1). That probability provides the answer to the query on the network structure as well as a complete dynamical picture of biomass exchanges between species. This is a hard problem since P_θ is a probability on an infinite-dimensional space. P_θ represents the state of the whole open system. In our equation (2.1), the network architecture is carried by the matrix D. That is, one defines a graph $G = (V, E)$, where $V = \{1, \ldots, N\}$ is the set of species and E the set of edges: $\{i, j\} \in E$ if and only if $D_{i,j} \neq 0$. What statistical techniques can reliably infer $D_{i,j}$?

Statistical data have the form of matrices: $(X(\omega_k, t_\ell) : k = 1, \ldots, K;\ \ell = 1, \ldots, L)$. Notice that one needs to observe different trajectories (ω_k) at any arbitrary finite sequence of times t_ℓ). This is a first difficulty for the application of time-series methods based on correlation and maximum entropy to estimate P_θ.

The covariance metric is limited as a measure of the network structure because of the influence of the noise terms. The expectation under P_θ of $X(t)$ from that equation is given by

$$E_\theta(X(t)) = E_\theta(X(0)) + \int_0^t E_\theta\,[X(s) \bullet (r + DX(s))]\,\mathrm{d}s. \tag{3.1}$$

The covariances between components $X_i(t)$ and $X_j(t)$ at a given time t are then

$$\begin{aligned} C(X_i(t), X_j(t)) &= E_\theta\left[(X_i(t) - E_\theta(X_i(t)))(X_j(t) - E_\theta(X_j(t)))\right] \\ &= \int_0^t E_\theta(\sigma_i(X(s))\sigma_j(X(s)))\,\mathrm{d}s. \end{aligned} \tag{3.2}$$

Therefore, if i and j are not connected nodes, that is $D_{i,j} = 0$ one may have $C(X_i(t), X_j(t)) \neq 0$ if $\sigma_i(X(s))$ and $\sigma_j(X(s))$ are not orthogonal in the space $L^2(\Omega \times [0, t], \mathrm{d}P_\theta \otimes \mathrm{d}s)$.

Using a maximum entropy technique based on the correlations between $X_i(t)$ and $X_j(t)$ to discover ρ_t may not be possible in most cases. Under P_θ, one can find a probability density $\rho_t(x)$ for each fixed time t:

$$P_\theta\,(X(t) \in A) = \mu_t(A) = \int_A \rho_t(x)\,\mathrm{d}x, \tag{3.3}$$

for all measurable subset in $\mathbb{R}^N$ in the space A. Though, as it is well known, the knowledge of ρ_t for all $t \geq 0$, does not suffice to identify P_θ. One needs to know an infinite family $(\mu_{t_1,\ldots,t_n})$ of measures, where $t_1 < \ldots < t_n$ run over all finite sequences of times, and each $\mu_{t_1,\ldots,t_n})$ is a measure on the space $(\mathbb{R}^N)^n$, such that

$$\mu_{t_1,\ldots,t_{n-1},t_n}(A_1 \times \ldots \times A_{n-1} \times \mathbb{R}^N) = \mu_{t_1,\ldots,t_{n-1}}(A_1 \times \ldots \times A_{n-1}),$$

the so-called Kolmogorov's consistency relation. So, under that hypothesis one could prove the existence of a probability measure P_θ on the set of trajectories, such that

$$P_\theta\,(X(t_1) \in A_1, \ldots, X(t_n) \in A_n) = \mu_{t_1,\ldots,t_n}(A_1 \times \ldots \times A_n).x$$

The maximum entropy at each step $t_1, \ldots, t_n$ does not preserve Kolmogorov's consistency relation and so cannot be applied to each measure $\mu_{t_1,\ldots,t_n}$. However, the network underlying a Gaussian process may be relatively detectable. The probability distribution of Gaussian processes is entirely determined by the covariance kernel and the mean, and it is well known that Gaussian laws maximize the Shannon entropy among all distributions with second moments. Unfortunately, as we will demonstrate, our process $X(t)$ here is not a Gaussian one, nor are ecological species abundance distributions commonly Gaussian. Angulo *et al*. [26] have also shown that in order for a network to be reconstructed more information beyond a time series of abundances, such as information about the interaction functional forms, must be known.

We expect that the covariance, correlation and inverse correlation (maximum entropy) metrics will not perform well in detecting the whole network structure. Simulations allow us to probe the limits of the network inference and expose the diversity of ways in which the network inference depends on the underlying dynamics.

4. Numerical results

(a) Motifs

In order the explore the consequences of these statistical limitations, we perform numerical simulations of the model system (equation (2.1)). There are 20 possible networks with four species and three edges (the minimum spanning tree for a four node network), based on combinatorics. Out of these 20 possible networks, there are seven distinct directed motifs in which all four species are connected. The networks are referred to by their labels in figure 1. The power to detect the species interactions is low even for networks that are detected more often than expected by random chance. True species interactions are typically only detected around 6–10% of the time.

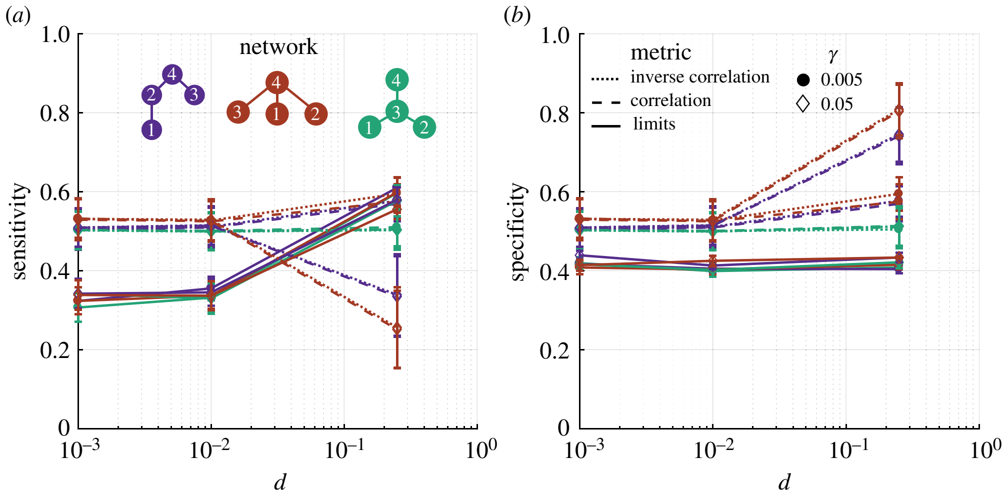

Figure 2. (*a*) Sensitivity (true positive rate) of reconstruction technique (*b*) Specificity (true negative rate) of reconstruction technique. The sensitivity and specificity are shown as a function of interaction strength *d* for a selection of network motifs (colours), reconstruction metrics (line styles) and stochasticity parameters (symbols). The uncertainty ranges are given by the standard deviation. (Online version in colour.)

Table 1. Parameters leading to the highest probability of detecting the network structure arranged by descending probability. The *F*-value of logistic regression of each parameter as a predictor of detecting the true network is shown below. The stochasticity γ is not a significant predictor with an *F*-value $= 3.2049$ and $Pr(> F) = 0.0740$.

	d	network	method	predicted prob
	0.25	e	Cov	0.0898
	0.25	f	Cov	0.0884
	0.25	c	Corr	0.0646
	0.25	a	Corr	0.0571
	0.25	b	Invcor	0.0568
	0.25	g	Invcor	0.0541
	0.25	d	Corr	0.0474
F-value	33.3005	90.0216	5.0344	
Pr(> *F*)	0	0	0.0068	

Since there are 20 unique networks, random detection is 5%. The overall highest probability of detection is for star networks with more basal species when using covariance to detect interactions and with a large species interaction coefficient *d* (table 1 and figure 3).

We summarize the overall performance of the metrics by calculating the sensitivity (or true positive rate), which is the probability of detecting a link when one exists, and specificity (or true negative rate), which is the probability of detecting that there is no link when there is no link (figure 2). This is a common performance metric that can be compared to other studies of network detection [8–10]. For these networks with three links and four species, if edges are chosen at random, specificity and sensitivity will, on average, equal 0.5. We find that both sensitivity and specificity are within 1 s.d. of 0.5 for almost all networks and parameter combinations when using covariance and correlation metrics. The LIMITS algorithm [23] can be used as a benchmark for the

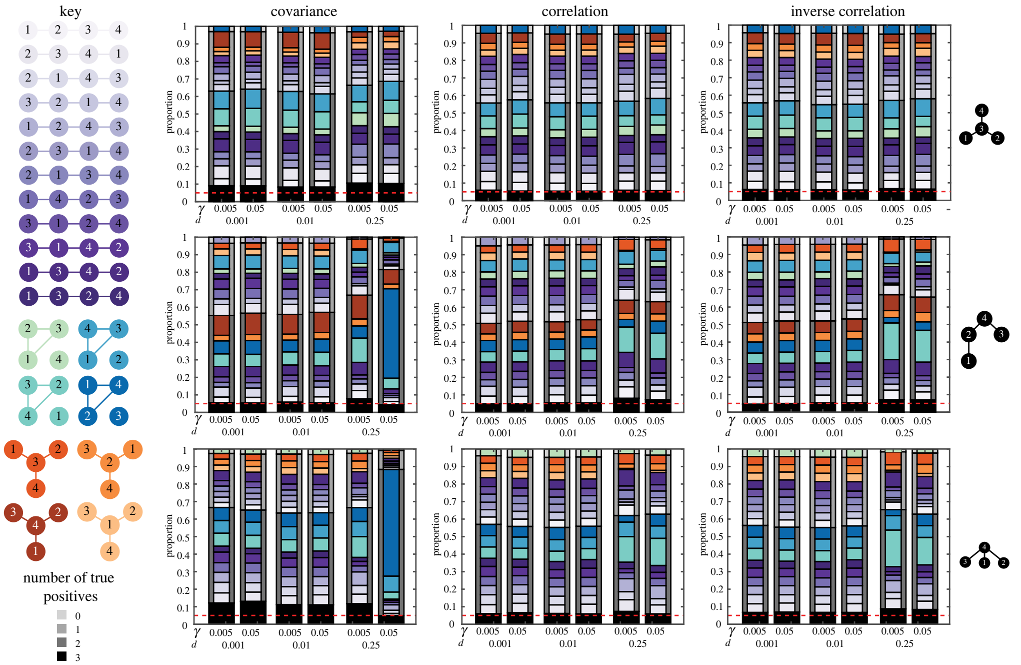

Figure 3. Proportion of trials (out of 10 000) for which each network is inferred for each parameter combination (stochasticity parameter γ and strength of predation d) for three different motifs (rows). The colour of the bar indicates the network that is inferred by reference to the key at the left. The true network is to the right of each row. The black portion of the bar at the bottom is the true network, consequently the true network is not represented as a coloured bar. The grey-shaded background bars in the background show the number of links in the true network that are correctly identified as links in the inferred network (true positives). The red dashed line indicates random detection (0.05). (Online version in colour.)

performance of the other metrics. With low interaction strengths, LIMITS detects few interactions but performs slightly better than the other metrics at high interaction strengths. LIMITS has low specificity, or a high false positive rate for all parameters. The false positive rate is affected by the detection threshold selected by the user.

The seven interaction networks with four species can be classified into three different types of systems based on qualitatively similar detection patterns using correlation based metrics (table 1). Logistic regression best predicts the probability of true network detection when using a model that includes interaction strength d, the network type, and reconstruction method as parameters. The model has a Nagelkerke Pseudo R^2 of 0.62. The magnitude of stochasticity γ is not a significant predictor of the probability of detection of the true network. These classifications by network type also align with the undirected network structure, which is either linear or star shaped. The linear networks (networks a,c,d) are best detected by correlation, the star networks with more top predators (networks b,g) are best detected by inverse correlation, and the star networks with more basal species (networks e,f) are best detected by covariance. In all cases, the networks are best detected when interspecific interactions are relatively strong ($d = 0.25$). Detection power is very low for the linear networks using any association metric. The linear-type network d is nearly undetectable with a highest predicted probability less than random.

There is a systematic relationship between the most likely network to be inferred and the true network, however the most likely network to be inferred is rarely the true network. The results for three case studies are synthesized in figure 3 the proportion of the 10 000 simulations in which each undirected network was detected with each reconstruction metric, with three different interaction strengths (0.001, 0.01 and 0.25 biomass^{-1}d^{-1}) and

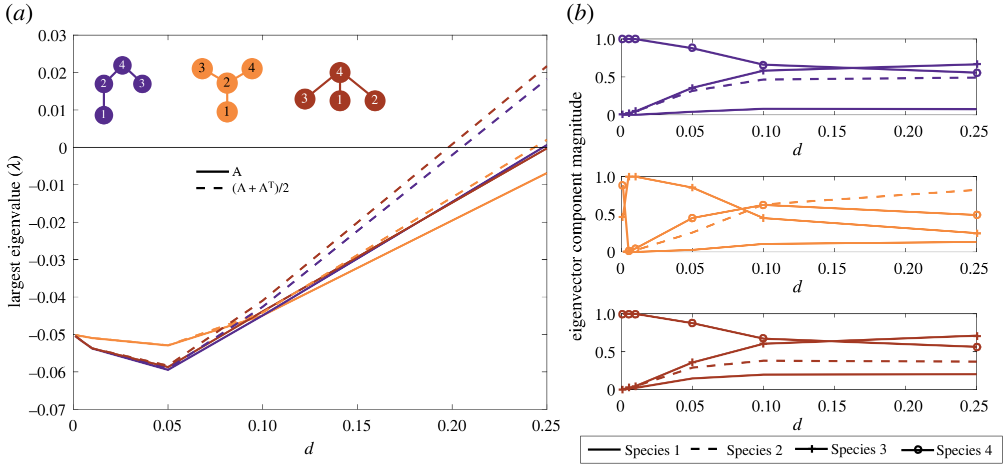

Figure 4. (*a*) Largest eigenvalues of the species interaction matrix (solid) and the symmetric part of the species interaction matrix (dashed) as a function of species interaction strength *d*. A positive maximum eigenvalue of the symmetric part of the species interaction matrix indicates non-normal growth. Growth occurs in the direction of the eigenvector that corresponds to the largest eigenvalue. (*b*) The direction of this eigenvector in species coordinates. (Online version in colour.)

two different stochasticity magnitudes (0.005 and 0.05 d^{-1}). For almost all experiments and parameter combinations, all 20 possible undirected networks were inferred with some non-zero probability (figure 3).

When the true network is a star-shaped network, the true network is detected across a wider parameter range when there are more basal species than top predators while a ring with three species connected and one species unconnected is inferred for motifs with more top predators (figure 3 top row and electronic supplementary material). For the star-shaped networks, the complement to the true network is a ring with one species unconnected. The network complement is purely indirect interactions. Covariance at times has a high probability of detecting the network complement. For example, with high stochasticity and large interaction coefficients, the complement of the network with two top predators (network b) is detected in 63% of the simulations and the complement of the network with three top predators (network g) is detected in over 99% of the simulations.

Although stochasticity does not significantly influence the probability of detecting the true network, stochasticity may affect the detection of the true networks through interaction with the drift term. One way, this may happen is through noise-induced large transients away from equilibrium. Networks with the potential for large transients away from equilibrium can be identified by calculating the eigenvalues of the symmetric part of the community matrix *D*. If the largest eigenvalue of the symmetric part of the community matrix is positive (figure 4), there is the possibility for large transient growth [27,28]. The direction of this transient growth is given by the eigenvector that corresponds to the largest eigenvalues of the symmetric part of the community matrix. For network b, this vector points most in the direction of species 2, for networks c and f, this eigenvalue points equally in the direction of species 3 and 4. In all cases with long transients, one species becomes disconnected and the ring geometry is preferentially constructed. This is one example of a quantitative evaluation of the way that network structure interactions with the stochastic drift term may lead to an inferred network that is statistically significant but distinct from the true network.

The abundance distribution for each species in time is non-normal, with most distributions left skewed, especially for the top consumer (species 4), and increasingly so as the rate of interspecific interactions and stochasticity increases. An example abundance distribution for two interaction strengths is given for three of the four-species interaction networks geometries (figure 5). As this

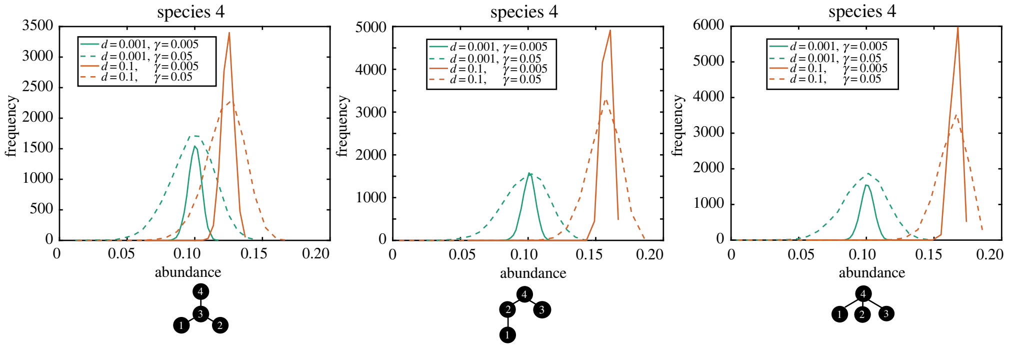

Figure 5. Distribution of each species abundances at the end of 10 000 simulations for different parameter combinations. (Online version in colour.)

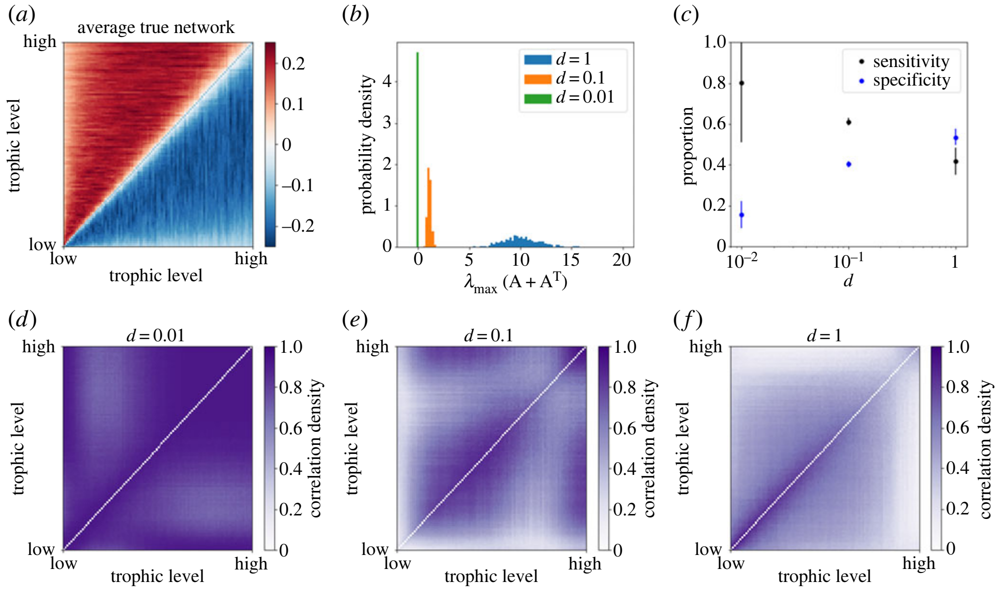

Figure 6. (*a*) Average interaction strength in the trophic networks used to simulate the many species communities. Darker shades indicate that the interaction is more frequent, lighter shades indicate less frequent interactions. All 100 species are shown on the axis arranged from low to high trophic levels. (*b*) Distribution of the maximum eigenvalue of the symmetric part of the species interaction matrix. Positive values indicate that the interaction matrix is susceptible to non-normal growth of perturbations. (*c*) Sensitivity and specificity for different values of interaction strength. The mean value for all 1000 simulations is shown with the confidence interval indicating the variance. (*d*–*f*). Constructed species interaction matrices using correlation. The darker shades indicate that the particular interaction is detected more frequently by correlation. (Online version in colour.)

figure shows, and all else being equal, network geometry affect the abundance distribution of species.

(b) Niche model trophic network

With the larger, 100 species networks, we find that the relationship between the correlation network and the interaction networks depends on the interaction strength. With weak interactions, there is high sensitivity but low specificity due to overly dense correlation networks. With stronger interactions, by contrast, both the sensitivity and specificity are low, near 0.5

(figure 6*c*). In addition, we find that there is a non-random structure to the correlation networks. While the interaction networks have the highest interaction density away from the diagonals, we find that the correlation density is highest on the diagonal of the correlation matrix and near the edges of the matrix (figure 6*d–f*). The networks with large interaction strengths are prone to transients with positive maximum eigenvalues of the symmetric part of the interaction matrix.

5. Discussion

There is a low probability of detecting the true interaction network, but we find that it is possible to detect four-species interaction modules for a wide range of predation rates and especially for star-shaped interaction geometries. Detection of the true network is robust to stochasticity for all metrics and, counter intuitively, aided by stochasticity in some cases. We can explain some of the covariance behaviour through understanding the transient dynamics of the model system.

(a) Ability to detect interaction modules

For most parameter combinations where the true interaction matrix can be detected, it is detected in at most only 10–11% of the trials. This level of detection of whole modules yields specificity and sensitivity of about 0.5, which is consistent with the levels of sensitivity and specificity obtained by other simulation studies [8–10]. These values of sensitivity and specificity would seem to suggest that edges are selected randomly, however certain motifs are more likely to be detected than others, namely star-shaped geometries, and certain motifs are consistently constructed, namely motifs with three species completely connected. For all metrics, there is a bias against detecting linear networks, which are almost never detected. It is important that we found that there is some ability to detect the whole network structure with the inverse correlation technique because the maximum entropy method relies on a reconstruction of the most probable configuration for the whole network, whereas the correlation and covariance metrics are pairwise metrics.

(b) Network geometry

The motifs used in this study are prevalent in true ecological and genetic networks. A chain of length three and a motif with four nodes all connected in a loop was found to occur more often than expected by chance in ecological networks while an undirected chain of length four, which is nearly undetectable in our analysis, occurs less often and the star-shaped motif, which is more easily detectable, occurs more often than expected by chance in a protein interaction network [29]. The networks used in this study differ from true networks in that true networks are thought to be more sparse, meaning they have a much lower connectance [30]. This difference might be expected to affect the total number of edges and hence the specificity and sensitivity.

The network geometry affects which networks are inferred, even if the inferred networks are not the true network. Researchers may conclude that species networks inferred statistically (‘association networks’ [8]) are interesting and useful objects to study that have some relationship to the underlying interaction network structure even if there is not direct correspondence between the observed associations and true interactions. This result of non-random generation of a network holds for large interaction networks as well.

(c) Transience and other effects of stochasticity

There is an implicit assumption in the literature that increased noise decreases the ability to detect interactions within networks [11,31]. Noise or stochasticity in either abundance or occurrence, however, is an essential assumption behind the use of correlation-based analysis, since in deterministic environments there is no correlation. Therefore, noise can be exploited to make inferences about network structure [32]. We find that the magnitude of the stochasticity is not a predictor of the success in inferring network structure.

Propagation of stochasticity and of biomass through the web affects the inferred structures. A few of the networks used (networks b, c and f) have the potential for non-normal growth of perturbations (as indicted by positive eigenvalues of the symmetric part of the community matrix), especially with strong interactions between species. Although all species have identical and independent stochastic diffusion terms, the interaction network directs perturbations through the web such that certain species may be perturbed away from equilibrium more than others. This is particularly true of networks that have the potential for non-normal growth. Stochasticity can have an apparent organizing effect because perturbations are amplified in a specific direction and recovery to equilibrium is constrained to occur along certain pathways. In biological terms, we find that the effects of top predators seem to propagate through food webs, affecting all species present in the system. For example, the network with species 1, 2 and 4 connected in a ring is commonly constructed for a motif with two intermediate consumers (networks c and d). In these networks, species 1, 2 and 4 are connected in a chain in the true network. Connection of species 1 and species 4 is meaningful in that it represents a trophic cascade. Even more strikingly, the network in which all species are connected to species 4 (the top predator) is constructed by covariance across a wide range of parameters for network d. Species 4 is only connected to species 2 in the true food web. This strong influence of top predators could be one reason that linear chains are almost never detected—the top predator is likely to covary with more species than just its immediate prey.

Stochasticity only affects detection if there is interaction between the stochasticity and the drift terms. In the numerical experiments, we find that γ, the strength of the stochastic noise, is not a significant predictor of the ability to detect the true network (electronic supplementary material, table S1).

The Brownian motion used in this study is on a particular (fast) time scale. The functional form used results in larger magnitude stochastic jumps at intermediate abundances, which could be another organizing force on the communities because movement towards equilibrium can generate correlation. Species should on average be at equilibrium and move about it randomly, but they are knocked farther away due to stochasticity when they are near equilibrium. While we use uncorrelated Brownian motion in these models, it is possible that the stochasticity alone could generate covariance if the diffusion terms are not orthogonal.

(d) Comparison of metrics

Covariance is in many cases the most successful metric at detecting species interactions. However, we find that it is in general the metric that is most likely to detect non-random structure, including false networks. Consequently, correlation and inverse correlation, which are normalized by the single species variance at times outperform covariance in detecting the true network, with the caveat that these metrics infer each of the possible networks with more uniform probability. The LIMITS algorithm had lower sensitivity than the other methods when interactions were weak, but increased as interaction strength increases, but within similar ranges as for the other methods. Specificity, on the other hand remained low, due to the low threshold necessary for detecting the interspecific interactions when they are weaker than or the same magnitude as intraspecific interactions. It is important to note that covariance and correlation only detect symmetric matrices while LIMITS can detect asymmetric matrices. This may affect the performance of the metrics and affects interpretation of the results. Evaluating a large network simulated using the niche model [22], which generates networks that are similar in their properties to observed food webs, and the correlation method, shows that sensitivity decreases and specificity increases as interactions strength increases. The large values of associated eigenvalues implies a larger role for transients in affecting the correlation structure.

Inverse correlation removes indirect interactions by minimizing the large-scale trends. In the simulations presented here, in which there are no deterministic external forcing or ‘hubs’ controlling species interactions, correlation and inverse correlation perform similarly. The inverse correlation method might perform better with a more complex network of species interactions. The covariance metric is able to make use of species relative abundances while the normalization used by correlation removes this information. As we show in figure 5, species relative abundances do provide meaningful information on network geometry to the extent that the species abundance distribution is affected by the network structure. As shown in this figure, the same species under the same parameters changes in abundance distribution due to network geometry. It is important to note that the interaction coefficients are symmetric and that adding additional complexity to this model by working with asymmetric community matrices may alter the reported simulation results. However, the mathematical exposition is agnostic to the structure of the interaction matrix.

This study suggests a few paths forward for the development of improved metrics of species interactions. The metrics currently in use perform best when variables are normally distributed. While use of correlation does not require that sample values are normally distributed, it is only an exhaustive measure of association if the joint distribution of the samples is a multivariate normal. The maximum entropy method used here similarly makes the approximation that the samples values are drawn from a normal distribution, however the maximum entropy method could be generalized to be appropriate for the observed abundance distributions. In our case, the samples (and hence the joint distribution) are not normally distributed, which is true for most communities [18,33]. Metrics based on the abundance distribution of biological species across different trophic levels might be better suited for network detection. For example, a linkage disequilibrium metric in genetics is specialized for beta distributed observations [34]. In addition, this system is in a non-equilibrium steady state and the absence of detailed balance means that in addition to maximizing the entropy, understanding entropy production could help to detect the true network. A distinct approach would be to use parameter estimation or constrained optimization to fit a model to observed time series.

6. Conclusion

Whole static networks of species interactions are detectable using existing methods for network inference including a maximum entropy method, but with low probability. Our finding that specificity and sensitivity do not differ significantly from random while there is non-random selection of network motifs demonstrates that it is important to consider not only the success in detecting pairwise interactions but also the way in which correlation metrics may systematically select certain sets of edges. Counterintuitively, increased stochasticity does not necessarily make detection of interactions between species less likely. Instead, the path that the system takes to equilibrium once perturbed is determined by the links between species, however this path does not necessarily facilitate detection by the existing metrics. Existing metrics have systematic biases. Indirect interactions may be more likely to be detected than direct interactions, even using inverse correlation. This is particularly true for systems prone to long transients.

While we focus on the problem of ecological network inference in this paper, network inference is an important tool in other domains including genetic networks [35,36]. There are many network inference techniques for researchers to choose including those based on machine learning [37]. We recommend that the mathematical and scientific basis of each of these techniques be evaluated carefully before their application in new domains.

Data accessibility. The simulation codes are available as electronic supplementary material.

Authors' contributions. M.A.F., R.R. and P.A.M. designed the study, analysed the results and wrote the manuscript; M.A.F. and R.R. designed the stochastic model and performed the theoretical derivations; M.A.F. ran the numerical experiments; M.A.F. and D.C. performed the data analysis.

Funding. This work was supported by a Fulbright Student Grant awarded to M.A.F. P.A.M. acknowledges support from project FONDECYT 1161023, FONDECYT 1200925, and AFB 170008 from CONICYT, R.R. acknowledges partial support of CONICYT-Proyecto Redes 180018 and CIMFAV-CIDI.

Acknowledgements. We thank the Fulbright Chile program for research support and the Santa Fe Institute for hosting a research visit. The authors also appreciate insightful conversations with Sergio Rojas. We acknowledge constructive comments from two anonymous reviewers.

References

1. Dunne JA. 2006 The network structure of food webs. In *Ecological networks: linking structure to dynamics in food webs* (eds M Pascual, JA Dunne), pp. 27–86. New York, NY: Oxford University Press.
2. Bascompte J, Jordano P. 2013 *Mutualistic networks*. Princeton, NJ: Princeton University Press.
3. Stephens CR, Heau JG, González C, Ibarra-Cerdeña CN, Sánchez-Cordero V, González-Salazar C. 2009 Using biotic interaction networks for prediction in biodiversity and emerging diseases. *PLoS ONE* **4**, e5725. (doi:10.1371/journal.pone.0005725)
4. Steele JA *et al.* 2011 Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J.* **5**, 1414–1425. (doi:10.1038/ismej.2011.24)
5. Faust K, Raes J. 2012 Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* **10**, 538–550. (doi:10.1038/nrmicro2832)
6. Faust K, Lahti L, Gonze D, de Vos WM, Raes J. 2015 Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr. Opin Microbiol.* **25**, 56–66. (doi:10.1016/j.mib.2015.04.004)
7. Barner AK, Coblentz KE, Hacker SD, Menge BA. 2018 Fundamental contradictions among observational and experimental estimates of non-trophic species interactions. *Ecology* **99**, 557–566. (doi:10.1002/ecy.2133)
8. Freilich MA, Wieters E, Broitman BR, Marquet PA, Navarrete SA. 2018 Species co-occurrence networks: can they reveal trophic and non-trophic interactions in ecological communities? *Ecology* **99**, 690–699. (doi:10.1002/ecy.2142)
9. Weiss S *et al.* 2016 Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* **10**, 1669–1681. (doi:10.1038/ismej.2015.235)
10. Berry D, Widder S. 2014 Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front. Microbiol.* **5**, 219. (doi:10.3389/fmicb.2014.00219)
11. Lezon TR, Banavar JR, Cieplak M, Maritan A, Fedoroff NV. 2006 Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc. Natl Acad. Sci. USA* **103**, 19 033–19 038. (doi:10.1073/pnas.0609152103)
12. Azaele S, Muneepeerakul R, Rinaldo A, Rodriguez-Iturbe I. 2010 Inferring plant ecosystem organization from species occurrences. *J. Theor. Biol.* **262**, 323–329. (doi:10.1016/j.jtbi.2009.09.026)
13. Yodzis P. 1988 The indeterminacy of ecological interactions as perceived through perturbation experiments. *Ecology* **69**, 508–515. (doi:10.2307/1940449)
14. Gotelli NJ, Graves GR. 1996 *Null models in ecology*. Washington, DC: Smithsonian Institution Press.
15. Freilich MA, Connolly SR. 2015 Phylogenetic community structure when competition and environmental filtering determine abundances. *Global Ecol. Biogeogr.* **24**, 1390–1400. (doi:10.1111/geb.12367)
16. Gilarranz LJ, Hastings A, Bascompte J. 2015 Inferring topology from dynamics in spatial networks. *Theor. Ecol.* **8**, 15–21. (doi:10.1007/s12080-014-0231-y)
17. Coenen AR, Weitz JS. 2018 Limitations of correlation-based inference in complex virus-microbe communities. *mSystems* **3**, e00084–18. (doi:10.1128/mSystems.00084-18)
18. Marquet PA, Espinoza G, Abades SR, Ganz A, Rebolledo R. 2017 On the proportional abundance of species: integrating population genetics and community ecology. *Sci. Rep. (Nature Publisher Group)* **7**, 1–10. (doi:10.1101/223529)
19. Fisher CK, Mehta P. 2014 The transition between the niche and neutral regimes in ecology. *Proc. Natl Acad. Sci. USA* **111**, 13 111–13 116. (doi:10.1073/pnas.1405637111)
20. Haegeman B, Loreau M. 2011 A mathematical synthesis of niche and neutral theories in community ecology. *J. Theor. Biol.* **269**, 150–165. (doi:10.1016/j.jtbi.2010.10.006)
21. Rebolledo R, Navarrete SA, Kéfi S, Rojas S, Marquet PA. 2019 An open-system approach to complex biological networks. *SIAM J. Appl. Math.* **79**, 619–640. (doi:10.1137/17M1153431)
22. Williams RJ, Martinez ND. 2000 Simple rules yield complex food webs. *Nature* **404**, 180–183. (doi:10.1038/35004572)

23. Fisher CK, Mehta P. 2014 Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS ONE* **9**, e102451. (doi:10.1371/journal.pone.0102451)
24. Röttjers L, Faust K. 2018 From hairballs to hypotheses–biological insights from microbial networks. *FEMS Microbiol. Rev.* **42**, 761–780. (doi:10.1093/femsre/fuy030)
25. McCullagh P. 1984 Generalized linear models. *Eur. J. Oper. Res.* **16**, 285–292. (doi:10.1016/0377-2217(84)90282-0)
26. Angulo MT, Moreno JA, Lippner G, Barabási AL, Liu YY. 2017 Fundamental limitations of network reconstruction from temporal data. *J. R. Soc. Interface* **14**, 20160966. (doi:10.1098/rsif.2016.0966)
27. Farrell BF, Ioannou PJ. 1996 Generalized stability theory. Part I: autonomous operators. *J. Atmos. Sci.* **53**, 2025–2040. (doi:10.1175/1520-0469(1996)053<2025:GSTPIA>2.0.CO;2)
28. Neubert MG, Caswell H. 1997 Alternatives to resilience for measuring the responses of ecological systems to perturbations. *Ecology* **78**, 653–665. (doi:10.1890/0012-9658(1997)078[0653:ATRFMT]2.0.CO;2)
29. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. 2002 Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827. (doi:10.1126/science.298.5594.824)
30. Proulx SR, Promislow DE, Phillips PC. 2005 Network thinking in ecology and evolution. *Trends Ecol. Evol.* **20**, 345–353. (doi:10.1016/j.tree.2005.04.004)
31. Deng Y, Jiang YH, Yang Y, He Z, Luo F, Zhou J. 2012 Molecular ecological network analyses. *BMC Bioinf.* **13**, 113. (doi:10.1186/1471-2105-13-113)
32. Lipinski-Kruszka J, Stewart-Ornstein J, Chevalier MW, El-Samad H. 2015 Using dynamic noise propagation to infer causal regulatory relationships in biochemical networks. *ACS Synth. Biol.* **4**, 258–264. (doi:10.1021/sb5000059)
33. Harte J, Kinzig A, Green J. 1999 Self-similarity in the distribution and abundance of species. *Science* **284**, 334–336. (doi:10.1126/science.284.5412.334)
34. Gianola D, Manfredi E, Simianer H. 2012 On measures of association among genetic variables. *Anim. Genet.* **43**, 19–35. (doi:10.1111/j.1365-2052.2012.02326.x)
35. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ. 2009 Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* **7**, 129. (doi:10.1038/nrmicro1949)
36. Dondelinger F, Mukherjee S. 2019 Statistical network inference for time-varying molecular data with dynamic Bayesian networks. In *Gene regulatory networks* (eds G Sanguinetti, VA Huynh-Thu), pp. 25–48. Berlin, Germany: Springer.
37. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. 2018 Next-generation machine learning for biological networks. *Cell* **173**, 1581–1592. (doi:10.1016/j.cell.2018.05.015)