Reconstruction of plant–pollinator networks from observational data

Jean-Gabriel Young,[1,*] Fernanda S. Valdovinos,[1,2,†] and M. E. J. Newman[1,3,‡]

[1] *Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Michigan, USA*
[2] *Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, USA*
[3] *Department of Physics, University of Michigan, Ann Arbor, Michigan, USA*

Empirical measurements of ecological networks such as food webs and mutualistic networks are often rich in structure but also noisy and error-prone, particularly for rare species for which observations are sparse. Focusing on the case of plant–pollinator networks, we here describe a Bayesian statistical technique that allows us to make accurate estimates of network structure and ecological metrics from such noisy observational data. Our method yields not only estimates of these quantities, but also estimates of their statistical errors, paving the way for principled statistical analyses of ecological variables and outcomes. We demonstrate the use of the method with an application to previously published data on plant–pollinator networks in the Seychelles archipelago, calculating estimates of network structure, network nestedness, and other characteristics.

I. INTRODUCTION

Network-based methods of analysis have contributed substantially to our understanding of ecological systems by helping us identify structure in the patterns of interaction between species [1–4]. Such patterns have been shown to be strongly linked to the dynamics and stability of ecosystems [5–7] and this is particularly the case for mutualistic networks such as plant–pollinator interactions—our focus in this paper—whose functions are crucial to terrestrial biodiversity [6, 8] and human food security [9, 10].

A central prerequisite for quantitative analysis of network structure and function is accurate network data, and significant effort has been invested in recent years in data gathering for ecological networks of many kinds, including mutualistic networks. There is, however, some debate over whether the observed structure of mutualistic networks represents the true interaction patterns produced by evolutionary and ecological mechanisms, at least to a good approximation [4, 6, 11], or whether, conversely, it is biased by incomplete sampling, for instance failing to detect the interactions of rare species [12, 13]. In this paper we describe a new technique that aims to give quantitative answers to these questions by applying methods of Bayesian inference to ecological network data. Treating the case of plant–pollinator networks, we show that it is possible to accurately infer interaction network structure from observational data while taking into account confounding variables such as varying species abundances. The output of our calculations includes not only an estimate of the true structure of the network but also an estimate of the certainty of each interaction, which allows us in turn to make precise statements about the accuracy of any further conclusions we draw from the network structure. Estimates of interaction certainty can also help us identify interactions that would benefit from greater sampling effort.

The structure of mutualistic networks is typified by several characteristic features [14]: *moderate connectance*, meaning that a modest fraction of all potential interactions are realized; *long-tailed degree distributions*, meaning that there are many specialist species with a small number of interactions and a few generalist species with many interactions; and *nestedness*, meaning that the interactions of the least-connected species are often subsets of the interactions of better-connected species. A significant volume of research has been devoted to explaining these features in terms of ecological and evolutionary mechanisms (see Vázquez *et al.* [11] and Bascompte and Jordano [6] for reviews). Other work, however, has suggested that they can also be generated merely as artifacts of skewed abundance distributions and incomplete sampling, both very common in ecological systems [12, 13]. In particular, Blüthgen *et al.* [12] have shown that nestedness and broad degree distributions can be a result of failure to observe interactions between rare species because of low sampling effort and/or the infrequency of the interactions in question. Findings like this have stimulated further investigations of the effects of sampling bias on network structure [4], both empirically by varying sampling effort in the field [15–19] and theoretically using models of network structure [12, 20–22]. These studies suggest that incomplete sampling strongly underestimates the number of interactions in networks and overestimates the degree of specialization. The approach described in this paper offers one way to address these shortcomings and obtain reliable estimates of the structure of mutualistic networks, free of measurement bias.

The paper is organized as follows. In Section II, we outline a first-principles statistical model of plant–pollinator interactions and show how it can be used to estimate network structure from error-prone observational data. Then in Section III we demonstrate these methods with an application to a typical plant–pollinator data set, showing how they give us not only the network structure itself but also statistically principled estimates of quantities such as nestedness. In Section IV we give some conclusions and directions for future work.

* jgyou@umich.edu
† fsvaldov@umich.edu
‡ mejn@umich.edu

II. NETWORK RECONSTRUCTION FROM OBSERVATIONAL DATA

The typical field study of plant–pollinator interactions involves observing instances of pollinators (such as insects) visiting plants within a prescribed observation area and over a prescribed period of time. The resulting data are reflective of the structure of the plant–pollinator network for the species involved but, for many reasons, they are not a perfect record of that network. First, there maybe observational errors. While the observers performing the work are usually highly trained individuals, they may nonetheless make mistakes. They may confuse one species for another, which is particularly easy to do for small-bodied insects, or smaller species may be overlooked altogether. Observers may make correct observations but record them wrongly. And there will be statistical fluctuations in the number of visits of an insect species to a plant species over any finite time. For rare interactions there may even be no visits at all if we are unlucky. The insects themselves may also make "mistakes" by visiting and appearing to pollinate a plant that in fact they typically do not pollinate. These and other factors mean that the record of observed visits is an inherently untrustworthy guide to the true structure of the plant–pollinator network.

Moreover, our data do not in any case tell us directly about network structure. The number of visits of a pollinator species to a plant species can vary widely, depending particularly on the abundance of the two species. How many visits do we take as evidence of a true plant–pollinator interaction? A single visit is probably not enough—it might well be an error or misobservation. Is two enough, or ten, or a hundred? And is this even the right question to ask? If we set the threshold at 100 visits, it seems absurd to then claim that 99 visits implies no interaction at all.

All of these objections can be overcome if we use a more realistic model of what our data mean.

A. Model of plant–pollinator data

Consider a typical plant–pollinator study, as described above, in which some number n_p of plant species, labeled by $i = 1 \ldots n_p$, and some number n_a of animal pollinator species, labeled by $j = 1 \ldots n_a$, are under observation for a set amount of time, producing a record of observed visits such that M_{ij} is the number of times plant species i is visited by pollinator species j. Collectively the M_{ij} can be regarded as a data matrix $\boldsymbol{M}$ with n_p rows and n_a columns. This is the input to our calculation.

The unknown quantity, the thing we would like to understand, is the network of plant–pollinator interactions. We can think of this network as composed of two sets of nodes, one representing plants and the other pollinators, with connections or edges joining each pollinator to the plants that it pollinates. In the language of network science this is a bipartite network, meaning that edges run only between nodes of unlike kinds—plants and pollinators—and never between two plants or two pollinators. Such a network can be represented by a second matrix $\boldsymbol{B}$, called the *incidence matrix*, with the same size as the data matrix, but with elements $B_{ij} = 1$ if plant i is pollinated by pollinator j and 0 otherwise.

The question we would like to answer is this: What is the best guess at the structure of the network, represented by $\boldsymbol{B}$, given the data $\boldsymbol{M}$? It is not straightforward to answer this question directly, but it is relatively easy to answer the reverse question. If we imagine that we know $\boldsymbol{B}$, then we can say what the probability is that we make a specific set of observations $\boldsymbol{M}$. And if we can do this then the methods of Bayesian inference allow us to invert the calculation and compute $\boldsymbol{B}$ from a knowledge of $\boldsymbol{M}$ and hence achieve our goal. The procedure is as follows.

Consider a specific plant–pollinator species pair i, j. How many times do we expect to see j visit i if j does, or does not, normally pollinate i? The answer will depend on several factors. First, and most obviously, we expect the number of visits to be higher if i is in fact a pollinator of j. That is, we expect M_{ij} to be larger if $B_{ij} = 1$ than if $B_{ij} = 0$. Second, we expect there to be more visits if the period of observation is longer or if the land area over which observations take place is larger, all other factors being equal. Third, we expect to see more visits for more abundant plant species than for less abundant ones, and similarly for more abundant pollinators. And fourth, as discussed at the start of Section II, we expect there to be some random variation in the number of visits, driven by fluctuations in individual behavior and the environment.

We can translate these factors into a mathematical model of plant–pollinator interaction as follows. Assuming that pollinator visits are independent—that the occurrence of one visit does not affect the timing or likelihood of another—the random variation in the number of visits will follow a Poisson distribution for each plant–pollinator pair i, j, parameterized by a single number, the distribution mean μ_{ij}. This mean value we expect to depend on the other three factors discussed above and we introduce additional parameters to represent this dependence.

First we introduce a parameter r to represent the larger number of visits when $B_{ij} = 1$, versus when it is 0. We write the factor by which the number is increased as $1 + r$ with $r \geq 0$, so that $r = 0$ implies no increase and successively larger values of r give us larger increases. Second, we represent the effect of the overall time or land area of observation by an overall constant C that multiplies the mean μ_{ij}. The same constant is used for all i and j, since all plant–pollinator pairs experience the same period and area of observation. Third, we assume that the number of visits is directly proportional to the abundance of the relevant plant and pollinator species: twice as many pollinators of species j, for instance, will mean twice as many visits by that species, and similarly for the abundance of the plant species. Thus the number of visits will be proportional to $\sigma_i \tau_j$, for some parameters σ_i and τ_j representing the abundances of plant i and pollinator j, respectively, in suitable units (which we will determine shortly).

Putting everything together, the mean number of observed visits to plant i by pollinator j is

$$\mu_{ij} = C\sigma_i \tau_j (1 + r B_{ij}), \tag{1}$$

and the probability of observing exactly M_{ij} visits is drawn from a Poisson distribution with this mean:

$$P(M_{ij}|\mu_{ij}) = \frac{\mu_{ij}^{M_{ij}}}{M_{ij}!}\, e^{-\mu_{ij}}. \tag{2}$$

This equation gives us the probability distribution of a single element M_{ij} of the data matrix. Assuming once again that visits are independent of one another, we can combine Eqs. (1) and (2) for all plant–pollinator pairs to get the likelihood of the complete data matrix $\boldsymbol{M}$ thus:

$$P(\boldsymbol{M}|\boldsymbol{B},\theta) = \prod_{i,j} \frac{\bigl[C\sigma_i\tau_j(1+rB_{ij})\bigr]^{M_{ij}}}{M_{ij}!}\, e^{-C\sigma_i\tau_j(1+rB_{ij})}, \tag{3}$$

where θ is a shorthand collectively denoting all the parameters of the model.

There are two important details that should be noted about this model. First, the definition in Eq. (1) does not completely determine C, σ, and τ because we can increase (or decrease) any of these parameters by a constant factor without changing the resulting value of μ_{ij} if we simultaneously decrease (or increase) one or both of the others. In the language of statistics we say that the parameters are not "identifiable." We can rectify this problem by fixing the normalization of the parameters in any convenient fashion. Here we do this by stipulating that σ_i and τ_j sum to one, thus:

$$\sum_{i=1}^{n_p} \sigma_i = \sum_{j=1}^{n_a} \tau_j = 1. \tag{4}$$

In effect, this makes σ_i and τ_j measures of relative abundance, quantifying the fraction of individual organisms that belong to each species, rather than the total number.

Second, there are other species-level effects on the observed number of visits in addition to abundance, such as the propensity for observers to overlook small-bodied pollinators. There is, at least within the data that we will be working with, no way to tell these effects from true variation in abundance—no way to tell for example if there are truly fewer individuals of a species or if they are just hard to see and hence less often observed. As a result, the abundance parameters in our model actually capture a combination of effects on observation frequency. This does not affect the accuracy of the model, which works just as well either way, but it does mean that we have to be cautious about interpreting the values of the parameters in terms of actual abundance. This point is discussed further in Section III.

B. Bayesian reconstruction

The likelihood of Eq. (3) tells us the probability of the data $\boldsymbol{M}$ given the network $\boldsymbol{B}$ and parameters θ. What we actually want to know is the probability of the network and parameters given the data, which we can calculate by applying Bayes' rule in the form

$$P(\boldsymbol{B},\theta|\boldsymbol{M}) = \frac{P(\boldsymbol{M}|\boldsymbol{B},\theta)P(\boldsymbol{B}|\theta)P(\theta)}{P(\boldsymbol{M})}. \tag{5}$$

This is the *posterior probability* that the network has structure $\boldsymbol{B}$ and parameter values θ given the observations that were made. There are three important parts to the expression: the likelihood $P(\boldsymbol{M}|\boldsymbol{B},\theta)$, the prior probability of the network $P(\boldsymbol{B}|\theta)$, and the prior probability of the parameters $P(\theta)$. The denominator $P(\boldsymbol{M})$ we can ignore because it depends on the data alone and will be constant (and hence irrelevant for our calculations) once $\boldsymbol{M}$ is determined by the observations.

Of the three non-constant parts, the first, the likelihood, we have already discussed—it is given by Eq. (3). For the prior $P(\theta)$ on the parameters we assume a simple uniform distribution, equivalent to saying that we know nothing about the parameter values in advance. This makes the prior a constant, which means we can ignore it as we did $P(\boldsymbol{M})$. For the prior on the network $P(\boldsymbol{B}|\theta)$ we make the conservative assumption—in the absence of any knowledge to the contrary—that all edges in the network are *a priori* equally likely. If we denote the probability of an edge by ρ, then the prior probability on the entire network is

$$P(\boldsymbol{B}|\theta) = \prod_{i,j} (1-\rho)^{1-B_{ij}} \rho^{B_{ij}}. \tag{6}$$

We consider ρ an additional parameter which is to be inferred from the data and which we will henceforth include, along with our other parameters, in the set θ. We also need to assume a prior probability on ρ and again we assume a uniform distribution, meaning the prior is constant and we can ignore it. With these choices, we now have everything we need to compute the posterior probability, Eq. (5).

Once we have the posterior probability there are a number of things we can do with it. The simplest is just to maximize it with respect to the unknown quantities $\boldsymbol{B}$ and θ to find the most likely structure for the network, and parameters, given the data. This, however, misses an opportunity for more detailed inference and can moreover give misleading results. In most cases there will be more than one value of $\boldsymbol{B}$ and θ with high probability under Eq. (5): there may be a unique maximum of the probability, a most likely value, but there are often many other values that have nearly as high probability and offer plausible network structures competitive with the most likely one. To get the most complete picture of the structure of the network we should consider all these plausible structures.

For example, if all plausible structures are similar to one another in their overall shape then we can be quite confident that that shape is reflective of the true plant–pollinator network. If plausible structures are widely varying, however, then we have many different candidates for the true structure and our certainty about that structure is correspondingly lower. In other words, by considering the complete set of plausible structures we can not only make an estimate of the network structure but also say how confident we are in that estimate, in effect putting "error bars" on the network.

How do we specify these errors bars in practice? One way is to place posterior probabilities on individual edges in the network. For example, when considering the edge connecting plant i and pollinator j, we would not ask "Is there an edge?" but rather "What is the probability that there is an edge?" Within the formulation outlined above, this probability is given

by the average

$$P(B_{ij} = 1|\boldsymbol{M}) = \sum_{\boldsymbol{B}} \int B_{ij} P(\boldsymbol{B},\theta|\boldsymbol{M})\, d\theta\,, \tag{7}$$

where the sum runs over all possible incidence matrices and the integral over all parameter values. More generally we can compute the average of any function $f(\boldsymbol{B},\theta)$ of the matrix $\boldsymbol{B}$ and/or the parameters θ thus:

$$\big\langle f(\boldsymbol{B},\theta)\big\rangle = \sum_{\boldsymbol{B}} \int f(\boldsymbol{B},\theta)\, P(\boldsymbol{B},\theta|\boldsymbol{M})\, d\theta. \tag{8}$$

Functions of both the matrix and the parameters can be interesting—the matrix tells us about the structure of the network, but the parameters, as we will see, can also reveal important information.

Computing averages of the form (8) is unfortunately not an easy task. A closed-form expression appears out of reach and the brute-force approach of performing the sums and integrals numerically over all possible networks and parameters is computationally intractable. Instead therefore we use an efficient Monte Carlo sampling technique to approximate the answers. We generate a sample of network/parameter pairs $(\boldsymbol{B}_1,\theta_1),\ldots,(\boldsymbol{B}_n,\theta_n)$, where each pair appears with probability proportional to the posterior distribution of Eq. (5). Then we approximate the average of $f(\boldsymbol{B},\theta)$ as

$$\big\langle f(\boldsymbol{B},\theta)\big\rangle \simeq \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{B}_i,\theta_i)\,. \tag{9}$$

Under very general conditions, this average will converge to its true value asymptotically as the number of Monte Carlo samples n becomes large. Full details of the computations are given in Materials and Methods.

C. Checking the model

Inherent in the discussion so far is the assumption that the data can be well represented by our model. In other words, we are assuming there is at least one choice of the network $\boldsymbol{B}$ and parameters θ such that the model will generate data similar to what we see in the field. This assumption could be violated if our model is a poor one, but there is nothing in the method of Section II B that would tell us so. To be fully confident in our results we need to be able not only to infer the network structure, but also to check whether that structure is a good match to the data. Luckily the Bayesian toolbox comes with a natural procedure for doing this. Given a set of high-probability values of $\boldsymbol{B}$ and θ, as described in Section II B, we can use them in Eq. (3) to compute the likelihood $P(\boldsymbol{M}|\boldsymbol{B},\theta)$ of a data set $\boldsymbol{M}$ and then sample possible data sets from this probability distribution, in effect recreating data as they would appear if the model were in fact correct. We can then compare these data to the original field data to see if they are similar: if they are then our model has done a good job of capturing the structure in the data.

In the parlance of Bayesian statistics this approach is known as a *posterior–predictive check*. It amounts to calculating the probability

$$P(\widetilde{M}_{ij}|\boldsymbol{M}) = \sum_{\boldsymbol{B}} \int P(\widetilde{M}_{ij}|\boldsymbol{B},\theta) P(\boldsymbol{B},\theta|\boldsymbol{M})\, d\theta \tag{10}$$

that there are $\widetilde{M}_{ij}$ observed visits of pollinator j to plant i in artificial data sets generated by the model, averaged over many sets of values of $\boldsymbol{B}$ and θ. We can then use this probability to calculate the average value of $\widetilde{M}_{ij}$ thus:

$$\langle \widetilde{M}_{ij}\rangle = \sum_{\widetilde{M}_{ij}} \widetilde{M}_{ij}\, P(\widetilde{M}_{ij}|\boldsymbol{M})\,. \tag{11}$$

The averages for all plant–pollinator pairs can be thought of as the elements of a matrix $\langle\widetilde{\boldsymbol{M}}\rangle$, which we can then compare to the real data matrix $\boldsymbol{M}$, or alternatively we can calculate a residue $\boldsymbol{M} - \langle\widetilde{\boldsymbol{M}}\rangle$. If $\langle\widetilde{\boldsymbol{M}}\rangle$ and $\boldsymbol{M}$ are approximately equal, or equivalently if the residue is small, then we consider the model a good one. Note that the calculation in Eq. (10) is of the same form as the one in Eq. (8), with $f(\boldsymbol{B},\theta) = P(\widetilde{M}_{ij}|\boldsymbol{B},\theta)$, which means we can calculate $P(\widetilde{M}_{ij}|\boldsymbol{M})$ in the same way we calculate other average quantities, using Monte Carlo sampling and Eq. (9).

III. RESULTS

To demonstrate how the method works in practice, we consider a large data set of plant–pollinator interactions gathered by Kaiser-Bunbury and collaborators [23] at a set of study sites on the island of Mahé in the Seychelles. The data describe the interactions of plant and pollinators species as observed over a period of eight months across eight different sites on the island. It also includes measurements of plant abundances for all observation periods and all sites. Our method for inferring network structure does not make use of the abundance measurements, but we discuss them briefly at the end of this section.

The study by Kaiser-Bunbury *et al.* focused particularly on the role of exotic plant species in the ecosystem and on whether restoring a site by removing exotic species would significantly impact the resilience and function of the plant–pollinator network. To help address these questions, half of the sites in the study were restored in this way while the rest were left unrestored as a control group.

As an illustration of our method we apply it to data from one of the restored sites in the Mahé study, as observed over the course of a single month in December 2012 (the smallest time interval for which data were available). We pick the site named "Trois-Frères" because it is relatively small but also well sampled. Our calculation then proceeds as shown in Fig. 1. There were 8 plant and 21 pollinator species observed at the site during the month, giving us an 8×21 data matrix $\boldsymbol{M}$ as shown in Fig. 1a. (Following common convention, the plots of matrices in this paper are drawn with rows and columns ordered by decreasing numbers of observed interactions, so that the

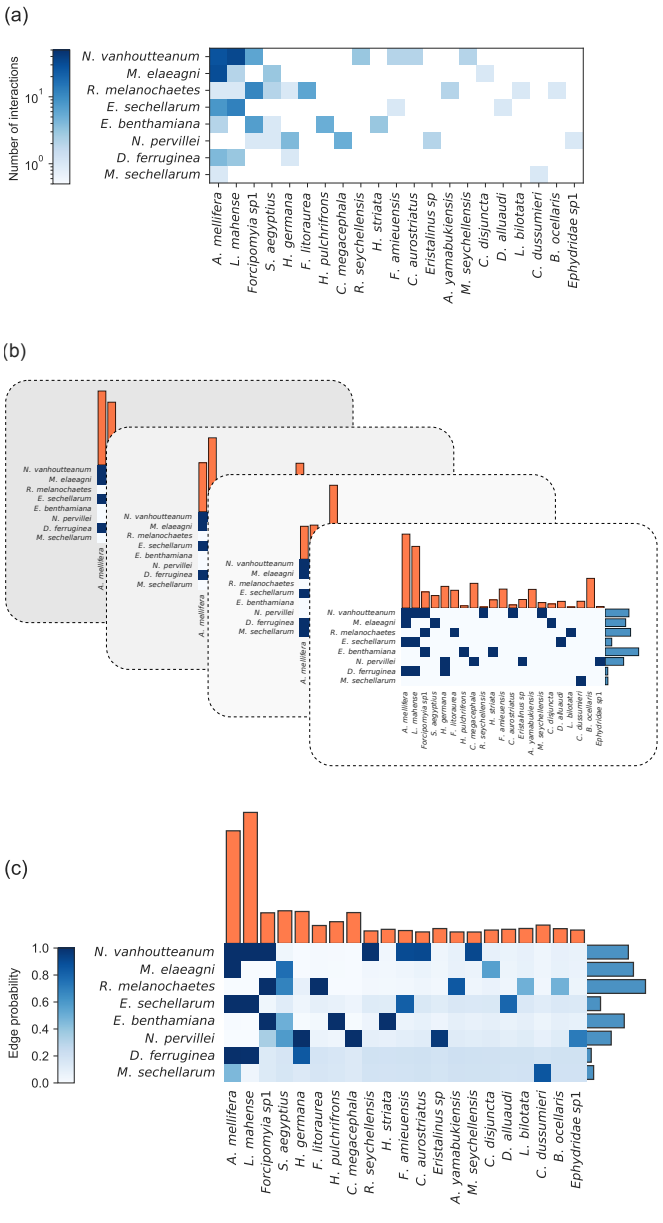

FIG. 1. Illustration of our inference method applied to an example data set from Kaiser-Bunbury *et al.* [23]. (*a*) We start with a data matrix $\boldsymbol{M}$ that records the number of interactions between each plant species and pollinator species. Missing interactions $M_{ij} = 0$ are shown in white. (*b*) We then draw 2 000 samples from the distribution appearing in Eq. (5), four of which are shown in the figure. Each sample consists of a binary incidence matrix $\boldsymbol{B}$, values for the relative abundances σ and τ (shown as the orange and blue bar plots, respectively), and values for the parameters C, r, and ρ (not shown). (*c*) We combine the samples using Eqs. (7)–(9) to give an estimate of the probability of each edge in the network and the complete parameter set θ. For the data set studied here our estimates of the expected values of the parameters C, r, and ρ are $\langle C\rangle = 52.2$, $\langle r\rangle = 58.5$, and $\langle\rho\rangle = 0.208$.

largest elements of the data matrix—the darkest squares—are in the top and left of the plot.)

Now we use our Monte Carlo procedure to draw 2000 sets of incidence matrices $\boldsymbol{B}$ and parameters θ from the posterior distribution of Eq. (5) (Fig. 1b). These samples vary in their structure: some edges, like the one connecting the plant *N. vanhoutteanum* and the pollinator *A. mellifera*, are present in nearly all samples, while others, like the one between *M. sechellarum* and *A. mellifera*, appear only a small fraction of the time. Some others never occur at all. Averaging over these sampled networks we can estimate the probability, Eq. (7), that each connection exists in the true plant–pollinator network—see Fig. 1c. Some connections have high probability, close to 1, meaning that we have a high degree of confidence that they exist. Others have probability close to 0, meaning we have a high degree of confidence that they do not exist. And some have intermediate probabilities, meaning we are uncertain about them (such as the *M. sechellarum*–*A. mellifera* connection, which has probability around 0.45). This is not a failing of our methodology, but rather one of its strengths. It is telling us that the data are not sufficient to reach a firm conclusion about these connections. Indeed, if we compare with the original data matrix $\boldsymbol{M}$ in Fig. 1a, we find that most of the uncertain connections are ones for which we have very few observations, say $M_{ij} = 1$ or 2.

As discussed in Section II C, we also need to check whether the model is a good fit to the data by performing a posterior-predictive test. Figure 2 shows the results of this test. The main plot in the figure compares the values of the 40 largest elements of the original data matrix $\boldsymbol{M}$ with the corresponding elements of the generated matrix $\widetilde{\boldsymbol{M}}$. In each case, the original value is well within one standard deviation of the average value generated by the test, confirming the accuracy of the model. The inset of the figure shows the residue matrix $\boldsymbol{M} - \widetilde{\boldsymbol{M}}$, which reveals no systematic bias unaccounted for by the model.

In addition to inferring the structure of the network itself, our method allows us to estimate many other quantities from the data. There are two primary methods by which we can do this. One is to look at the values of the fitted model parameters, which represent quantities such as the preference r and species abundances σ, τ. The other is to compute averages of quantities that depend on the network structure or the parameters (or both) from Eq. (9).

As an example of the former approach, consider the parameter ρ, which represents the average probability of an edge, also known as the connectance of the network. Figure 3a shows the distribution of values of this quantity over our set of Monte Carlo samples, and neatly summarizes our overall certainty about the presence or absence of edges. If we were certain about all edges in the network, then ρ would take only a single value and the distribution would be narrowly peaked. The distribution we observe, however, is somewhat broadened, indicating significant uncertainty. The most likely value of ρ, the peak of the distribution, turns out to be quite close to the value one would arrive at if one were simply to assume that every pair of species that interacts even once is connected in the network. This does not mean, however, that one could make this assumption and get good results. As we show below, the

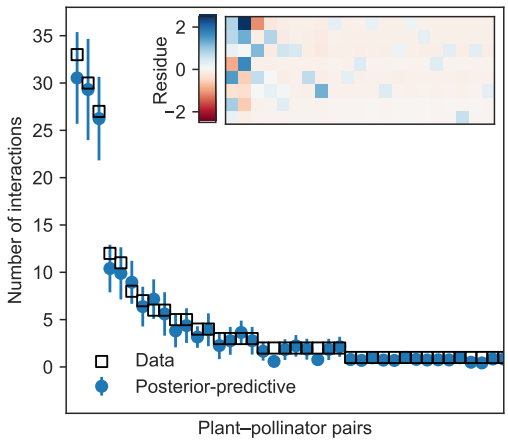

FIG. 2. Results of a posterior–predictive test on the data matrix $\boldsymbol{M}$ for the example data set analyzed in Fig. 1. The main plot shows the error on the 40 largest entries of $\boldsymbol{M}$, while the inset shows the residue matrix $\boldsymbol{M} - \langle \tilde{\boldsymbol{M}} \rangle$. Because the actual data $\boldsymbol{M}$ are well within one standard deviation of the posterior–predictive mean, the test confirms that the model is a good fit in this case.

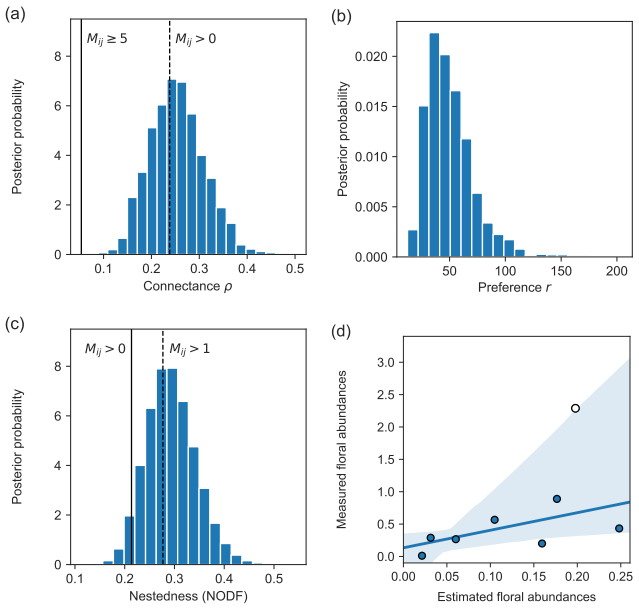

FIG. 3. Examples of the types of analysis that can be performed using samples from the posterior distribution of Eq. (5). (*a*) Distribution of the connectance ρ. Connectance values for binary networks obtained by thresholding the data matrix at $M_{ij} > 0$ and $M_{ij} \geq 5$ are shown as vertical lines for reference. (*b*) Distribution of the preference parameter r. The mean of this distribution is $\langle r \rangle = 58.5$ and its mode close to 50, but individual values as high as 100 are possible. (*c*) Distribution of the nestedness measure NODF. Values obtained by thresholding the data matrix at $M_{ij} > 0$ and $M_{ij} > 1$ are shown for reference. (*d*) Estimated and measured abundances for each of the plant species ($R^2 = 0.29$). The white dot is treated as an outlier.

network one would derive by doing so would be badly in error in other ways.

Figure 3b shows the distribution of another of the model parameters, the parameter r, which measures the extent to which pollinators prefer the plants they normally pollinate over the ones they do not. For this particular data set the most likely value of r is around 60, meaning that pollinators visit their preferred plant species about 60 times more often than they do the average species, an impressive level of selectivity on the part of the pollinators.

For the calculation of more complicated network properties we can perform an average over the value of any function $f(\boldsymbol{B}, \theta)$ as described in Section II B. As an example, Fig. 3c shows a calculation of the quantity known as "Nestedness based on Overlap and Decreasing Fill" (NODF), a measure of the nestedness property discussed in the introduction. This quantity measures the extent to which the specialist species—those with relatively few interactions—tend to interact with a subset of the partners of the generalist species [24]. While NODF is hard to compute in closed form [25], it is straightforward to calculate within our framework: we simply calculate the value for each sampled network $\boldsymbol{B}$ and plot the resulting distribution. Interestingly, the most likely value of NODF is significantly different from the one we would calculate had we assumed, as discussed above, that a single interaction is sufficient to consider two species connected. We instead find that the system is almost certainly more nested than this simpler analysis would conclude.

In Fig. 3d, we compare the values of our estimated plant abundance parameters σ to the measured abundances reported by Kaiser-Bunbury *et al.* As discussed in Section II A, our parameters are not true measures of abundance because they combine actual abundance with other characteristics such as ease of observation. We do find some correlation between the estimated and observed abundances, but it is relatively weak ($R^2 = 0.29$), signaling significant disagreement. This indicates that the frequency of observed interaction between plants and pollinators is not in fact proportional to their plain abundances, but to a combination of abundance, ease of observation, and potentially other factors as well. One candidate for a possible additional factor that could play a role is adaptive foraging by pollinators, which has been shown to influence the structure of ecological networks [4, 26]. Adaptive foraging occurs, for example, when pollinators deliberately visit less abundant plants more often if those plants contain more food (such as nectar or pollen) relative to more abundant plants with less food [7]. Our estimated "abundance" parameters automatically include such factors where traditional field measurements of abundance do not, and analyses that use such traditional measurements, as in [12, 13], may as a result fail to control for significant species-level effects on observed visitation rates. We would therefore argue that best practice calls for the use of estimated rather than measured abundances, as in the methodology proposed here.

Finally, as we have mentioned, the connections in the network about which we are most uncertain tend naturally to be those for which we have the smallest amount of data. In an ideal world we could address this problem by taking more data,

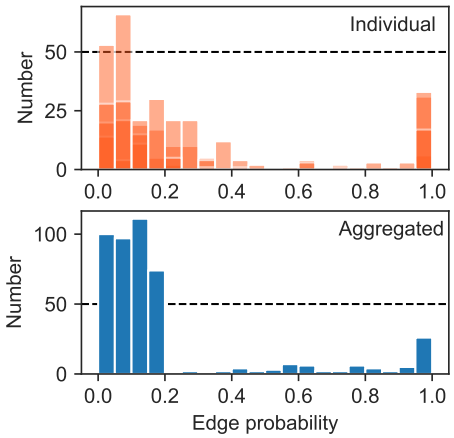

FIG. 4. Illustration of the effect of data aggregation on edge uncertainty. In the top panel, we show a histogram of the edge probabilities $P(B_{ij} = 1|\boldsymbol{M})$ for the four restored sites as observed in October 2012, analyzed individually. In the bottom panel, we show the equivalent histogram obtained by aggregating the data over the sites and then estimating a single network from the resulting data matrix. The horizontal lines, both drawn at fifty observations—are added merely as a guide to the eye. Note how the upper histogram has more mass near the middle of the plot, while the lower one has most of its mass close to probability zero or one, indicating greater certainty in the positions of the edges in the aggregated data.

but the more common situation is the one we face here where the data have already been gathered and we are tasked with performing the best analysis we can. There are nonetheless some remedies open to us, such as aggregating data over different geographical areas or time windows. In Fig. 4 we compare the edge probabilities estimated from data recorded individually at the four "restored" sites in October 2012 to the edge probabilities we obtain when we aggregate these observations into a single data matrix and only then estimate the network. (We use restored sites observed during the same month because they are likely to be ecologically similar, meaning the data are measuring approximately the same system.) Comparison of the two distributions shows—as we would hope—that there are fewer uncertain edges in the the aggregated network than in its disaggregated parts, i.e., there are fewer edges with probabilities in the middle of the distribution and more with probabilities close to zero or one.

IV. DISCUSSION

In this paper, we have proposed a statistical model of plant–pollinator interactions and shown how it can be used to infer the structure and properties of real plant–pollinator networks from noisy, error-prone measurements. The model employs elementary ecological insights to create an expressive and versatile structure that can capture the pattern of interactions in a wide range of ecosystems. We have used the toolbox of Bayesian statistics to develop both an inference algorithm and a model checking procedure for the model. Our methods explicitly allow for the possibility that there are multiple plausible networks that could fit a given set of observations, a hallmark of Bayesian analysis. Doing this allows us to make accurate deductions even in cases where data sets are small and the number of model parameters large.

To show how the method works in practice, we have presented a case study of a plant–pollinator data set gathered on the island of Mahé in the Seychelles. We have shown that our method is able to reconstruct the plant–pollinator network itself and we have validated the reconstruction using a posterior-predictive check that compares the observed data against those we would expect to see if the reconstruction were correct. We have further shown that the method can be used to calculate estimates of arbitrary network properties without significant additional work. As an example we have calculated the nestedness measure known as NODF, though other network properties can also be easily estimated.

As a byproduct of the fitting procedure we have also shown that we can estimate a range of model parameters that are of potential ecological interest, including parameters that correspond to network connectance and species abundance, and a parameter that measures the interaction specificity of pollinator species, the amount by which they favor their preferred species of plants over others.

There could be a number of possible extensions of the work presented here. The method as described assumes an ecosystem that is more or less static, but real ecosystems can change rapidly with the seasons. One could imagine a dynamic variant of the model that allows parameters to evolve over time. On the applications side, it would be interesting to investigate previously studied plant–pollinator data sets anew, and verify whether careful statistical treatment and a better delineation between networks and measurements leads to new conclusions. These developments, however, we leave for future work.

ACKNOWLEDGMENTS

We thank Alec Kirkley, George Cantwell, and Maria Riolo for helpful discussions. This work was funded in part by the James S. McDonnell Foundation (JGY) and the US National Science Foundation under grants DEB–1834497 (FSV) and DMS–1710848 (MEJN), as well as University of Michigan MICDE grant U061182 (FSV).

Appendix A: Materials and Methods

As outlined in the main text, our method relies on a generative network model in which observed visits to plants by pollinators are considered noisy measurements of an unobserved underlying plant–pollinator network. This formulation allows us to frame the task of determining the network structure as a Bayesian inference problem [27–30] in which the probability of the network having incidence matrix $\boldsymbol{B}$ given a

data matrix $\boldsymbol{M}$ is

$$P(\boldsymbol{B},\theta|\boldsymbol{M}) = \frac{P(\boldsymbol{M}|\boldsymbol{B},\theta)P(\boldsymbol{B}|\theta)P(\theta)}{P(\boldsymbol{M})}, \tag{A1}$$

where θ are model parameters and $P(\boldsymbol{M})$ is an unimportant normalizing constant. The element M_{ij} of matrix $\boldsymbol{M}$ records the number of times insects of species j are seen to pollinate plant species i, while $B_{ij} = 0, 1$ encodes the presence or absence of an edge between the two species in the plant–pollinator network. Both matrices are of dimension $n_p \times n_a$ where n_p is the number of plants and n_a is the number of pollinators.

We model the number of visits M_{ij} as a Poisson random variable with mean

$$\mu_{ij} = C\sigma_i\tau_j(1 + rB_{ij}), \tag{A2}$$

and, assuming uniform priors on all the parameters and edges that are *a priori* equally likely with probability ρ, we find that

$$P(\boldsymbol{B},\theta|\boldsymbol{M}) \propto \prod_{ij}(1-\rho)^{1-B_{ij}}\rho^{B_{ij}}\frac{\mu_{ij}^{M_{ij}}}{M_{ij}!}e^{-\mu_{ij}}. \tag{A3}$$

1. Bayesian network reconstruction

Given the probability distribution in Eq. (A3) there are a number of approaches we could take. Following [28, 29] we could employ an expectation–maximization (EM) algorithm to calculate the distribution over potential network structures and a point-estimate of θ or, following [30], we could integrate out the parameters θ and then sample from the resulting marginal distribution on $\boldsymbol{B}$. Neither of these approaches is completely satisfactory here however, the first because point estimates of the parameters can be unreliable for large models such as ours, and the second because the values of the model parameters are actually of interest to us, so we would prefer not to eliminate them.

Instead therefore we make use of a technique from the literature on finite mixture models [31] to sample efficiently from the joint distribution of both $\boldsymbol{B}$ and θ and hence estimate both. The main idea is to first sample values of the parameters θ from their marginal distribution

$$P(\theta|\boldsymbol{M}) = \sum_{\boldsymbol{B}} P(\boldsymbol{B},\theta|\boldsymbol{M}). \tag{A4}$$

The sum over $\boldsymbol{B}$ can, it turns out, be carried out in closed form for the particular $P(\boldsymbol{B},\theta|\boldsymbol{M})$ defined in Eq. (A4) and gives

$$P(\theta|\boldsymbol{M}) \propto e^{-C}\prod_{ij}(C\sigma_i\tau_j)^{M_{ij}}\left[1-\rho+\rho(1+r)^{M_{ij}}e^{-C\sigma_i\tau_j r}\right]. \tag{A5}$$

We can then sample from this distribution using standard methods such as Hamiltonian Monte Carlo—see below. This gives us our estimates of the parameter values themselves.

For given values of the parameters we can then estimate the network $\boldsymbol{B}$ by sampling from the distribution

$$P(\boldsymbol{B}|\boldsymbol{M},\theta) = \frac{P(\boldsymbol{M}|\boldsymbol{B},\theta)P(\boldsymbol{B}|\theta)}{P(\boldsymbol{M}|\theta)}. \tag{A6}$$

Using our previous values for the likelihood $P(\boldsymbol{M}|\boldsymbol{B},\theta)$ and $P(\boldsymbol{B}|\theta)$, and noting that the denominator $P(\boldsymbol{M}|\theta)$ is proportional to Eq. (A5), we find

$$\begin{aligned} P(\boldsymbol{B}|\boldsymbol{M},\theta) &= \frac{\prod_{ij}(1-\rho)^{1-B_{ij}}\left[\rho(1+r)^{M_{ij}}e^{-C\sigma_i\tau_j r}\right]^{B_{ij}}}{\prod_{ij}\left[1-\rho+\rho(1+r)^{M_{ij}}e^{-C\sigma_i\tau_j r}\right]} \\ &= \prod_{ij} Q_{ij}^{B_{ij}}(1-Q_{ij})^{1-B_{ij}}, \end{aligned} \tag{A7}$$

where

$$Q_{ij} = P(B_{ij}=1|\boldsymbol{M},\theta) = \frac{\rho(1+r)^{M_{ij}}e^{-C\sigma_i\tau_j r}}{1-\rho+\rho(1+r)^{M_{ij}}e^{-C\sigma_i\tau_j r}} \tag{A8}$$

is the posterior probability of an edge between species i and j, given the parameters θ.

We can now simply average Q_{ij} over our set of sample values of the parameters θ to get the expected probability of an edge between any pair of nodes. More generally, we can calculate an estimate of any function $f(\boldsymbol{B},\theta)$ by drawing m samples θ_k of the parameter set and n random incidence matrices $\boldsymbol{B}_l(\theta_k)$ for each set with edges appearing independently with probabilities Q_{ij} given by (A8), then averaging:

$$\big\langle f(\boldsymbol{B},\theta)\big\rangle \simeq \frac{1}{mn}\sum_{k=1}^{m}\sum_{l=1}^{n} f\big(\boldsymbol{B}_l(\theta_k),\theta_k\big). \tag{A9}$$

2. Implementation

In our implementation of this approach we sample parameters θ from the distribution of Eq. (A5) using the technique known as Hamiltonian Monte Carlo, in which one defines an inertial mechanics with a position space equivalent to the space of allowed values of the parameters and auxiliary momenta chosen so that the dynamics under the corresponding Hamilton's equations samples from the desired distribution [32]. We implement the calculation in Stan,[1] a probabilistic programming language that automatically performs Hamiltonian Monte Carlo sampling for arbitrary target distributions [33]. In practice, the program operates on the log of the posterior probability, which for our distribution (A5) has the form $\log P(\theta|\boldsymbol{M}) = -C + \sum_{ij}(X_{ij} + Y_{ij})$ where

$$X_{ij} = M_{ij}\log C\sigma_i\tau_j\,, \tag{A10}$$

$$Y_{ij} = \log\big(1-\rho+\rho(1+r)^{M_{ij}}e^{-C\sigma_i\tau_j r}\big). \tag{A11}$$

To avoid potential over- or underflow and ensure numerical stability we rewrite the latter expression slightly by defining

$$\mu_{ij} = \log(1-\rho), \quad \nu_{ij} = \log\rho + M_{ij}\log(1+r) - Cr\sigma_i\tau_j, \tag{A12}$$

[1] Code implementing our model is available online at https://github.com/jg-you/plant-pollinator-inference.

and then writing

$$Y_{ij} = \begin{cases} \mu_{ij} + \log\bigl(1 + e^{\nu_{ij}-\mu_{ij}}\bigr) & \text{if } \mu_{ij} > \nu_{ij}, \\ \nu_{ij} + \log\bigl(1 + e^{\mu_{ij}-\nu_{ij}}\bigr) & \text{otherwise,} \end{cases} \tag{A13}$$

which ensures that Y_{ij} is always a manageable number.

An important practical consideration is verifying the convergence of the Monte Carlo algorithm. Hamiltonian Monte Carlo mixes rapidly, but, like all Monte Carlo methods, it can sometimes become trapped at local optima. To ensure representative sampling of the posterior distribution, we therefore perform multiple Monte Carlo runs from random initial states and if any of the runs converges to a region of significantly smaller probability than the others then we repeat the entire calculation. In the example calculations given in the paper we perform four runs, with an equilibration period of 5000 Monte Carlo steps each, followed by taking 500 samples.

3. Quantifying error using posterior predictive checks

A crucial part of the model fitting process is assessing whether the model is a good fit to the data. In the main text we argue that a so-called posterior predictive test is a good way of making this assessment. The idea is to generate a new artificial data set $\widetilde{\boldsymbol{M}}$ from the model using the values of the model parameters derived from the fit to the input data $\boldsymbol{M}$. If we find that $\widetilde{\boldsymbol{M}}$ looks similar to the input then our model has done a good job of capturing the structure of the data.

To carry out this procedure we need to calculate the posterior predictive distribution for species pair i, j given by

$$P(\widetilde{M}_{ij}|\boldsymbol{M}) = \sum_{\boldsymbol{B}} \int P(\widetilde{M}_{ij}|\boldsymbol{B},\theta) P(\boldsymbol{B},\theta|\boldsymbol{M})\, d\theta. \tag{A14}$$

Since the likelihood $P(\widetilde{M}_{ij}|\boldsymbol{B},\theta)$, Eq. (3), factors into separate terms for each plant–pollinator pair i, j, this expression can with only a little work be simplified to

$$P(\widetilde{M}_{ij}|\boldsymbol{M}) = \int P(\theta|\boldsymbol{M}) \bigl[Q_{ij} P(\widetilde{M}_{ij}|B_{ij}=1,\theta) + (1-Q_{ij}) P(\widetilde{M}_{ij}|B_{ij}=0,\theta) \bigr]\, d\theta, \tag{A15}$$

and the integral can then be approximated by simply averaging over the set of sampled values of θ.

Two particularly useful statistics for the posterior predictive test are the mean and the variance of $\widetilde{M}_{ij}$, which in this case are equal since $\widetilde{M}_{ij}$ by definition has a Poisson distribution for given $\boldsymbol{B}$ and θ. Both are, to a good approximation, given by

$$\lambda_{ij} \approx \frac{1}{n} \sum_{k=1}^{n} \bigl[Q_{ij}(\theta_k) \mu_{ij}(B_{ij}=1) + \bigl(1 - Q_{ij}(\theta_k)\bigr) \mu_{ij}(B_{ij}=0) \bigr], \tag{A16}$$

where μ_{ij} is the mean defined in Eq. (A2).

4. Description of the data set

The data analyzed in Section III were gathered by Kaiser-Bunbury *et al.* [23] in inselbergs (steep-sided monolithic rocky outcroppings) on the tropical granitic island of Mahé, located in the Indian Ocean. Mahé is the largest of the Seychelles islands, the oldest extant island group in the world, which originated after splitting from Gondwanaland about 70 million years ago. Because of the islands' age and geographical isolation, the native species in the Seychelles are mostly endemic, arising more by evolution than immigration. An influx of exotic species, coupled with a habitat loss, has nonetheless had lasting effects on the islands' biological communities, but the vegetation on the inselbergs of Mahé has been less impacted than the rest of the island. The vegetation on the inselbergs is characterized by short trees, shrubs, and an absence of flowering herbs.

The data we analyze includes records of the visits of pollinator species to all plant species found in each of the eight inselbergs, observed between September 2012 and April 2013 during the island's eight-month-long tropical flowering season. Species visiting flowers were recorded as pollinators if they touched the sexual parts of the flowers within a standard observation window of 30 minutes [34]. Floral abundances were obtained by counting flowers in 1-meter cubes randomly located along transects spanning the inselbergs. The visit data were used to generate 64 data matrices of plant–pollinator interactions, one for each period and location. Our primary analysis focuses on the matrix for the site known as Trois-Frères as observed during the month of December 2012. We choose this data set primarily because it is relatively small and hence easy to visualize.

[1] N. D. Martinez, Artifacts or attributes? Effects of resolution on the Little Rock Lake food web. *Ecological Monographs* **61**, 367–392 (1991).

[2] J. Bascompte, P. Jordano, C. J. Melian, and J. M. Olesen, The nested assembly of plant-animal mutualistic networks. *Proc. Natl. Acad. Sci. USA* **100**, 9383–9387 (2003).

[3] E. Thèbault and C. Fontaine, Stability of ecological communities and the architecture of mutualistic and trophic networks. *Science* **329**, 853–856 (2010).

[4] F. S. Valdovinos, Mutualistic networks: Moving closer to a predictive theory. *Ecol. Lett.* (2019).

[5] U. Brose, R. J. Williams, and N. D. Martinez, Allometric scaling enhances stability in complex food webs. *Ecol. Lett.* **9**, 1228–1236 (2006).

[6] J. Bascompte and P. Jordano, *Mutualistic Networks*. Princeton University Press, Princeton, NJ (2014).

[7] F. S. Valdovinos, B. J. Brosi, H. M. Briggs, P. Moisset de Espanés, R. Ramos-Jiliberto, and N. D. Martinez, Niche partitioning due to adaptive foraging reverses effects of nestedness and connectance on pollination network stability. *Ecol. Lett.* **19**,

1277–1286 (2016).

[8] J. N. Thompson, *The Coevolutionary Process*. University of Chicago Press, Chicago, IL (1994).

[9] S. G. Potts, V. Imperatriz-Fonseca, H. T. Ngo, M. A. Aizen, J. C. Biesmeijer, T. D. Breeze, L. V. Dicks, L. A. Garibaldi, R. Hill, J. Settele, and A. J. Vanbergen, Safeguarding pollinators and their values to human well-being. *Nature* **540**, 220–229 (2016).

[10] J. Ollerton, Pollinator diversity: Distribution, ecological function, and conservation. *Annu. Rev. Ecol. Evol. Syst.* **48**, 353–376 (2017).

[11] D. P. Vázquez, N. Blüthgen, L. Cagnolo, and N. P. Chacoff, Uniting pattern and process in plant-animal mutualistic networks: A review. *Ann. Bot.* **103**, 1445–1457 (2009).

[12] N. Blüthgen, J. Fründ, D. P. Vázquez, and F. Menzel, What do interaction network metrics tell us about specialization and biological traits? *Ecology* **89**, 3387–3399 (2008).

[13] N. Blüthgen, Why network analysis is often disconnected from community ecology: A critique and an ecologist's guide. *Basic Appl. Ecol.* **11**, 185–195 (2010).

[14] J. Bascompte and P. Jordano, Plant-animal mutualistic networks: The architecture of biodiversity. *Annu. Rev. Ecol. Evol. Syst.* **38**, 567–593 (2007).

[15] A. Nielsen and J. Bascompte, Ecological networks, nestedness and sampling effort. *J. Ecol.* **95**, 1134–1141 (2007).

[16] T. Petanidou, A. S. Kallimanis, J. Tzanopoulos, S. P. Sgardelis, and J. D. Pantis, Long-term observation of a pollination network: Fluctuation in species and interactions, relative invariance of network structure and implications for estimates of specialization. *Ecol. Lett.* **11**, 564–575 (2008).

[17] S. J. Hegland, J. Dunne, A. Nielsen, and J. Memmott, How to monitor ecological communities cost-efficiently: The example of plant-pollinator networks. *Biol. Cons.* **143**, 2092–2101 (2010).

[18] N. P. Chacoff, D. P. Vázquez, S. B. Lomáscolo, E. L. Stevani, J. Dorado, and B. Padrón, Evaluating sampling completeness in a desert plant-pollinator network. *J. Anim. Ecol.* **81**, 190–200 (2012).

[19] A. Rivera-Hutinel, R. O. Bustamante, V. H. Marín, and R. Medel, Effects of sampling completeness on the structure of plant-pollinator networks. *Ecology* **93**, 1593–1603 (2012).

[20] D. P. Vázquez, C. J. Melián, N. M. Williams, N. Blüthgen, B. R. Krasnov, and R. Poulin, Species abundance and asymmetric interaction strength in ecological networks. *Oikos* **116**, 1120–1127 (2007).

[21] I. Bartomeus, Understanding linkage rules in plant-pollinator networks by using hierarchical models that incorporate pollinator detectability and plant traits. *PLOS One* **8**, e69200 (2013).

[22] J. Fründ, K. S. McCann, and N. M. Williams, Sampling bias is a challenge for quantifying specialization and network structure: Lessons from a quantitative niche model. *Oikos* **125**, 502–513 (2016).

[23] C. N. Kaiser-Bunbury, J. Mougal, A. E. Whittington, T. Valentin, R. Gabriel, J. M. Olesen, and N. Blüthgen, Ecosystem restoration strengthens pollination network resilience and function. *Nature* **542**, 223–227 (2017).

[24] M. Almeida-Neto, P. Guimaraes, P. R. Guimaraes Jr, R. D. Loyola, and W. Ulrich, A consistent metric for nestedness analysis in ecological systems: Reconciling concept and measurement. *Oikos* **117**, 1227–1239 (2008).

[25] C. Payrató-Borràs, L. Hernández, and Y. Moreno, Breaking the spell of nestedness: The entropic origin of nestedness in mutualistic systems. *Physical Review X* **9**(3), 031024 (2019).

[26] F. S. Valdovinos, R. Ramos-Jiliberto, L. Garay-Narváez, P. Urbani, and J. A. Dunne, Consequences of adaptive behaviour for the structure and dynamics of food webs. *Ecology Letters* **13**, 1546–1559 (2010).

[27] I. Brugere, B. Gallagher, and T. Y. Berger-Wolf, Network structure inference, a survey: Motivations, methods, and applications. *ACM Computing Surveys* **1**, 1 (2016).

[28] M. E. J. Newman, Network structure from rich but noisy data. *Nature Physics* **14**, 542–545 (2018).

[29] M. E. J. Newman, Estimating network structure from unreliable measurements. *Phys. Rev. E* **98**, 062321 (2018).

[30] T. P. Peixoto, Reconstructing networks with unknown and heterogeneous errors. *Phys. Rev. X* **8**, 041011 (2018).

[31] A. Gelman, H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC, New York, 3rd edition (2013).

[32] M. Betancourt, A conceptual introduction to Hamiltonian Monte Carlo. Preprint arxiv:1701.02434 (2017).

[33] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, Stan: A probabilistic programming language. *J. Stat. Softw.* **76**(1) (2017).

[34] C. N. Kaiser-Bunbury, T. Valentin, J. Mougal, D. Matatiken, and J. Ghazoul, The tolerance of island plant–pollinator networks to alien plants. *J. Ecol.* **99**, 202–213 (2011).