A quantitative framework for investigating the reliability of empirical network construction

Alyssa R. Cirtwill[1,2†], & Anna Eklöf[1], Tomas Roslin[3], Kate Wootton[3], Dominique Gravel[4]

[1] Department of Physics,
Chemistry and Biology (IFM)
Linköping University
Linköping, Sweden

[2] Current address: Department of Ecology,
Environment, and Plant Sciences (DEEP)
Stockholm University
Stockholm, Sweden

[3] Department of Ecology
P.O. Box 7044
Swedish University of Agricultural Sciences
SE-750 07 Uppsala, Sweden

[4] Département de biologie
Université de Sherbrooke
Sherbrooke, Canada [†] Corresponding author:

alyssa.cirtwill@gmail.com

Running head: Constructing quantitative interaction networks

Abstract

1. Descriptions of ecological networks typically assume that the same interspecific interactions occur each time a community is observed. This contrasts with the known stochasticity of ecological communities: community composition, species abundances, and link structure all vary in space and time. Moreover, finite sampling generates variation in the set of interactions actually observed. For interactions that have not been observed, most data sets will not contain enough information for the ecologist to be confident that unobserved interactions truly did not occur.

2. Here we develop the conceptual and analytical tools needed to capture uncertainty in the estimation of pairwise interactions. To define the problem, we identify the different contributions to the uncertainty of an interaction. We then outline a framework to quantify the uncertainty around each interaction by combining data on observed co-occurrences with prior knowledge. We illustrate this framework using perhaps the most extensively sampled network to date.

3. We found significant uncertainty in estimates for the probability of most pairwise interactions. This uncertainty can, however, be constrained with informative priors. This uncertainty scaled up to summary measures of network structure such as connectance and nestedness. Even with informative priors, we are likely to miss many interactions that may occur rarely or under different local conditions.

4. Overall, we demonstrate the importance of acknowledging the uncertainty inherent in network studies, and the utility of treating interactions as probabilities in pinpointing

areas where more study is needed. Most importantly, we stress that networks are best thought of as systems constructed from random variables, the stochastic nature of which must be acknowledged for an accurate representation. Doing so will fundamentally change networks analyses and yield greater realism.

Keywords

ecological networks; probabilistic interactions; Bayesian networks; sampling error; spatial variability; temporal variability; uncertainty

1 Introduction

2 Representing an ecological community as a network addresses both community composition

3 and interspecific interactions (Roslin & Majaneva, 2016). Identifying these interactions is

4 critical to understanding species' effects on each other and predicting community responses

5 to perturbations (Bartomeus *et al.*, 2016; Giron *et al.*, 2018). Importantly, information is

6 conveyed both by the presence and *absence* of interactions. Presences and absences are

7 not, however, equally certain. An observed link definitely occurred, but there are multiple

8 reasons why a given link may not be observed (Jordano, 2016).

9   First, ecological communities are stochastic (Gotelli, 2000). Community composition

10 and abundances vary over space (Baiser *et al.*, 2012) and time (Olesen *et al.*, 2011), leading

11 to variation in encounter probabilities and hence interactions (Vázquez *et al.*, 2005;

12 Poisot *et al.*, 2015). Species must co-occur in order to be observed interacting, so any

13 variation in co-occurrence will create variation in the set of observed

14 interactions (Gravel *et al.*, 2018; Graham & Weinstein, 2018). Even assuming constant

15 co-occurrence within a system, changing abundances can affect interactions if species which

16 would otherwise interact become too rare to detect each other (Tylianakis *et al.*, 2010;

17 Jordano, 2016). Moreover, interactions vary over space (Kitching, 1987; Baiser *et al.*,

18 2012), time (Olesen *et al.*, 2011; Lopez *et al.*, 2017), between individuals (Pires *et al.*, 2011;

19 Wells *et al.*, 2013; Fodrie *et al.*, 2015; Novak & Tinker, 2015), throughout life

20 cycles (Clegg *et al.*, 2018), and with environmental conditions (Poisot *et al.*, 2015).

21   Beyond "true" variation in network structure, several researchers note effects of

22 sampling intensity (e.g., Martinez *et al.*, 1999; Blüthgen *et al.*, 2006; Jordano, 2016).

Assessments of the accumulation of interactions with increasing sampling effort suggest that it is more challenging to correctly document interactions than species (Guimerà & Sales-Pardo, 2009; Poisot *et al.*, 2012; Bartomeus, 2013; Jordano, 2016; Giron *et al.*, 2018). As a result, it has been proposed that interactions should be described probabilistically and network metrics computed accordingly (Bartomeus, 2013; Poisot *et al.*, 2016). Such efforts usually take the form of creating model networks that approximate observed empirical networks (Allesina *et al.*, 2008; Guimerà & Sales-Pardo, 2009; Williams *et al.*, 2010; Rohr *et al.*, 2016). These models may be based on species traits and/or abundances (Rohr *et al.*, 2016; Weinstein & Graham, 2017b; Graham & Weinstein, 2018) or on simple abstract rules (Allesina *et al.*, 2008; Guimerà & Sales-Pardo, 2009; Williams *et al.*, 2010). Model networks provide an important "reality check" by allowing hypothesis testing about factors structuring ecological networks (Bartomeus, 2013; Graham & Weinstein, 2018). They can also allow estimation of the numbers unobserved of interactions (Jordano, 2016; Weinstein & Graham, 2017b) and predict which unobserved interactions are most likely (Guimerà & Sales-Pardo, 2009; Bartomeus, 2013).

To date, models incorporating uncertainty have either been very general (e.g., Guimerà & Sales-Pardo (2009); Gravel *et al.* (2018)) or very system-specific (e.g., Bartomeus (2013); Graham & Weinstein (2018)); in either case such a model may be a poor fit for a particular study system. Moreover, some of the models described above cannot accommodate species observed only once (Bartomeus, 2013; Weinstein & Graham, 2017a). As rare species are likely to be observed less frequently (Blüthgen *et al.*, 2006) and may be of particular interest (e.g., when modelling species loss), there is a need for a more widely-applicable framework that includes rarely-observed species.

Here we introduce a simple, adaptable Bayesian framework. First, we briefly describe

the nested levels of uncertainty affecting ecological network construction. Understanding

this nesting is an important step towards formalising our interpretation of the different

drivers of uncertainty about interactions. Importantly, we distinguish between the

possibility that an interaction is unfeasible (cannot occur) and the possibility that, given

an interaction is feasible, it may not occur during a particular sampling event.

In the context of this uncertainty, which can be reduced but not eliminated by

high-quality sampling (Bartomeus, 2013)), we present a Bayesian framework which

combines observed data with a prior expectation of interaction probabilities. We conclude

with a worked example applying this framework to an intensively-sampled host-parasitoid

network. We find that even the highest-quality empirical network data we could find has

substantial uncertainty about whether many pairs of species interact, demonstrating the

need to include information about this uncertainty in published networks.

Why are some interactions not observed?

We start from the perspective of a community ecologist describing an interaction network.

This ecologist will want to describe the species present and the links between

them (Roslin & Majaneva, 2016). While an observed link definitely occurred (assuming

species are correctly identified), a link may not be observed for many reasons, *whether or*

not it truly occurred (Fig. 1). Moreover, given an unobserved interaction in an empirical

dataset, we often cannot determine why the interaction was not observed *post hoc*. The

detection of any interaction is a stochastic process subject to many levels of uncertainty.

As a conceptual guide, we describe three nested levels of uncertainty that roughly address the questions: "Could species i and j interact?", "Did they interact during sampling?", and "Did we observe the interaction?"

Could the species interact?

Fundamentally, some interactions are feasible while others are not (Poisot *et al.*, 2015). Assuming that the feasibility of interactions depends on the traits of species involved, we can define the probability of an interaction L_{ij} between species i and j given some function describing trait matching $\mathbf{T_{ij}}$ as $P(L_{ij}|\mathbf{T_{ij}}) = \lambda_{ij}$ (Bartomeus, 2013; Gravel *et al.*, 2013; Weinstein & Graham, 2017a). In some cases the traits prohibiting an interaction are known and there will be little uncertainty about λ_{ij}. Generally, however, the traits that might prohibit an interaction are unknown (Dormann *et al.*, 2017) and uncertainty about λ_{ij} will be high. An unobserved interaction may have $\lambda_{ij} > 0$, and the associated zero in an interaction matrix might be false. Trait-matching models can reduce uncertainty about λ_{ij} (Jordano, 2016). As these models improve, λ_{ij} should either tend to 0 or to 1. Nevertheless, every model is imperfect and lacks information that could be used to define constraints on at least some interactions (Dormann *et al.*, 2017). This leaves some uncertainty about whether some species pairs could interact.

Did the species interact during sampling?

A feasible interaction (i.e., $L_{ij} = 1$;), still may not occur at a particular place and time (Poisot *et al.*, 2015). This could be because constraints such as inclement weather or low abundance prevented the interaction (Jordano, 2016), a preferred interaction partner

7

was available (Weinstein & Graham, 2017b), or individual variation in traits meant that local individuals could not interact even if the species in general can (Gravel *et al.*, 2013; Wells *et al.*, 2013; Poisot *et al.*, 2015). For example, a rare galler may not parasitize a rare plant because they do not encounter each other. In metawebs, an interaction may not occur locally because two species do not co-occur at a particular site even if they do co-occur and interact elsewhere (Graham & Weinstein, 2018). We can define the local realisation of an interaction, X_{ij}, given that the interaction is feasible, as a stochastic process with associated probability $P(X_{ij}|L_{ij} = 1) = \chi_{ij}$. It is unlikely that $\chi_{ij} = 1$ except in the case of obligate specialists who always interact whenever they co-occur and always co-occur. When $\chi_{ij} < 1$, we cannot be sure whether an unobserved interaction cannot occur or whether it simply did not occur at a particular site.

Some of the uncertainty about χ_{ij} could be addressed by drawing on the rich literature about interaction contingencies. Phenological matching (Miller-Rushing *et al.*, 2010), species preferences (Pires *et al.*, 2011; Novak & Tinker, 2015), and fear effects (Luttbeg & Kerby, 2005) all affect the probability of an interaction and could be included in models similar to trait-matching models. Some studies attempt to reduce uncertainty by sampling intensively over a short period of time and restricting the scale of interest to interactions occurring at the sampling site and time (conflating χ_{ij} and λ_{ij}) (Bartomeus, 2013; Weinstein & Graham, 2017b). This approach does reduce local-scale uncertainty but is not appropriate if we wish to assemble the full set of interactions which occur in a community (beyond the specific site and time of sampling). In contrast, expanding sampling to cover a broader spatial or temporal range (e.g., sampling in a variety of micro-habitats or during a variety of weather conditions) will help

8

111 to reduce uncertainty about χ_{ij} by including a broader range of parameters affecting

112 interaction probability. It is unlikely, however, that resources will permit researchers to

113 sample extensively in the full range of sites and conditions covered by a community, and

114 some uncertainty about χ_{ij} will remain.

115 **Forbidden links**

116 Both whether species can ever interact and whether they interact at a particular place

117 and time have been discussed in the context of "forbidden links" (Jordano, 2016). In this

118 framework, all links which are prevented by spatio-temporal uncoupling, physiological

119 constraints, etc. are considered "structural zeros" that cannot be observed (Jordano, 1987,

120 2016). We conceptually distinguish between physiological constraints and spatio-temporal

121 uncoupling to allow for cases in which a species is introduced to a new habitat, expands its

122 range, or shifts phenology (Gravel *et al.*, 2013). In such cases, links which are impossible

123 due to physiological constraints remain forbidden but links previously "forbidden" by

124 spatio-temporal mismatch could potentially occur. When considering local realisations of a

125 metaweb, one can distinguish between interactions forbidden by traits and those forbidden

126 by lack of co-occurrence (Gravel *et al.*, 2018). This is not possible, however, when trying to

127 identify the metaweb.

128 **Did we observe the interaction?**

129 Lastly, measurement errors are a pervasive source of uncertainty about ecological processes.

130 Even a feasible interaction which occurs during sampling ($L_{ij} = 1$, $X_{ij} = 1$) may still go

131 undetected (Jordano, 2016; Weinstein & Graham, 2017a). We can define the probability of

132 detecting of an interaction, D_{ij}, given the interaction is feasible and occurs, as a stochastic

process with probability $P(D_{ij}|X_{ij} = 1, L_{ij} = 1) = \delta_{ij}$. Note that D_{ij} is a stochastic variable where $D_{ij} = 1$ if the interaction observed and $D_{ij} = 0$ otherwise, while $P(D_{ij})$ is a continuous probability. Detection failure could happen for many reasons (Wirta *et al.*, 2014). Some sources of detection failure can be minimised (e.g., by combining multiple sampling methods (Wirta *et al.*, 2014; Jordano, 2016)), but other sources are difficult to reduce. For example, it may not be possible to sample some species directly due to conservation concerns (Lagrue & Poulin, 2015). The probability of detecting an interaction also depends on the duration of the interaction and its evidence, the species involved, and the individuals sampled (Wells *et al.*, 2013; Lagrue & Poulin, 2015; Cirtwill *et al.*, 2016; Weinstein & Graham, 2017a). As sampling improves D_{ij} should tend towards 1 but, importantly, D_{ij} will vary between interactions and it is very likely that some $D_{ij} < 1$. For example, when analysing gut contents interactions involving prey with hard parts will be more detectable than interactions with quickly-digested, soft prey (Alberdi *et al.*, 2018).

In summary, an unobserved interaction could be unfeasible $[P(L_{ij}) = 0]$, be feasible but not occur during sampling $(P(X_{ij}|L_{ij} = 1) = 0]$, or be feasible and occur during sampling but not be observed $[P(D_{ij}|X_{ij} = 1, L_{ij} = 1) = 0]$. Researchers might be interested in identifying feasible interactions (i.e., reducing uncertainty about λ_{ij}) or identifying feasible interactions that occurred locally (i.e., reducing uncertainty about χ_{ij}). Here we consider the former problem while noting that uncertainty about both χ_{ij} and δ_{ij} contribute to uncertainty about λ_{ij}. Future sampling should aim to characterise all three probabilities, as is possible in some studies focused on local realisations of a metaweb (e.g., Graham & Weinstein (2018); Gravel *et al.* (2018)), but for now it is often difficult to distinguish between them.

10

156   When considering a metaweb, we wish to separate unobserved interactions where $L_{ij} = 0$

157 (unfeasible interactions) from interactions which were feasible but did not occur during

158 sampling or were not detected (Fig. 1). An empirical ecologist will measure the marginal

159 probability $P(L_{ij}) = k_{ij}/n_{ij}$, where k_{ij} is the number of observed interactions between

160 species i and j and n_{ij} the number of observed co-occurrences. Given this information, how

161 can we reduce the uncertainty around our estimated interaction probability?

162   One obvious way to reduce uncertainty is to increase sampling. High sample sizes will

163 reduce uncertainty about the upper bound of λ_{ij} and increase the probability of detecting

164 unlikely or cryptic interactions (where $L_{ij} = 1$ but χ_{ij} or δ_{ij} is low). In some systems,

165 targeted sampling could also reduce especially high uncertainties about λ_{ij}, but this will

166 not be possible everywhere (e.g., in host-parasitoid systems where parasitoid identities are

167 unknown during sampling). High-quality sampling is crucial in studies of empirical

168 networks, but there are limits to the ability of sampling to reduce uncertainty (*Appendix*

169 *S2, Fig. S2*). As a complement to improved sampling, we can also reduce uncertainty by

170 leveraging prior knowledge about the system using a Bayesian approach, as detailed below.

Approach

A Bayesian framework for interaction probabilities

173 The basis of a Bayesian approach to modelling the probability that an interaction between

174 species i and j occurs (λ_{ij}) is combining the maximum likelihood estimate (MLE) of λ_{ij}

175 with a prior distribution (described in the next section) and a normalising function. The

MLE of λ_{ij} can be computed from the number of observed interactions k_{ij} and observed co-occurrences n_{ij}: $\lambda_{ij} = \frac{k_{ij}}{n_{ij}}$. The most appropriate prior distribution for λ_{ij} is the beta distribution:

$$\lambda_{ij} \sim Beta(\alpha_{ij}, \beta_{ij}), \tag{1}$$

which has two shape parameters, α_{ij} and β_{ij}.

The shape parameters, or hyperparameters, may be set to particular values or derived from data (see below). They may be the same for all interactions, allowed to vary independently across interactions, or incorporate non-independence between interactions (e.g., when the shape of the prior distribution depends upon species' abundances or traits). For example, a prior could account for the tendency for abundant species to interact more frequently with all possible partners than rare species.

Combining the prior distribution with the MLE of λ_{ij} derived from observed data, we obtain a posterior distribution of λ_{ij} that also follows a beta distribution with new hyperparameters $\alpha'_{ij} = \alpha_{ij} + k$ and $\beta'_{ij} = \beta_{ij} + n - k$ (*Appendix S3*). These definitions illustrate the influence of the prior on the posterior distribution: the difference between the posterior and the prior will increase with k_{ij} and $n_{ij} - k_{ij}$. The distribution of λ_{ij} for better-sampled pairs of species relies less on the information used to build the prior distribution and more on the observed data. As n_{ij} increases, the total uncertainty about λ_{ij} decreases (Fig. 3). Note that these models of λ_{ij} are designed to i) quantify and ii) reduce the total uncertainty about λ_{ij}. They do not distinguish between sources of uncertainty except through the choice of prior.

12

Choosing a prior distribution

The prior used reflects a study's goal, the amount of prior knowledge available, and the strength of researchers' belief in that knowledge. Note that there is no 'correct' prior as long as the true distribution of interaction probabilities is unknown (Spiegelhalter *et al.*, 2000). The goal is to select a prior that reflects reasonable assumptions about the true distribution and then adjust it with observed data as it becomes available. When in doubt, it is also possible to model interaction probabilities using a variety of priors and compare results (Spiegelhalter *et al.*, 2000).

Some researchers are reluctant to adopt Bayesian approaches because of their perceived subjectivity (i.e., a metaweb simulated using a Bayesian distribution will depend upon the assumptions inherent in the prior). We argue that one of the advantages of the Bayesian approach is that it makes this subjectivity explicit (Spiegelhalter *et al.*, 2000). Researchers compiling empirical networks face many decisions about defining community boundaries, the set of species to include in the focal community (Jordano, 2016), the most appropriate type of sampling (Wirta *et al.*, 2014), etc. These choices all affect the network obtained and most do not have clear answers, making all empirical networks (and all theoretical networks, which likewise depend upon the assumptions used to build the network) somewhat subjective. In addition, researchers identifying "forbidden links" inherently rely on prior knowledge to identify these links (Jordano, 2016). Stating that an interaction is "forbidden" is, in effect, applying a very strong prior to reduce/eliminate uncertainty about the non-observation of that interaction. This is especially true when researchers assume that any species not observed co-occurring could not ever interact. Given the reality of

subjective network construction, we feel it is better to make the assumptions used to build a network as explicit as possible so that these assumptions can be questioned, compared between studies, tested, and perhaps validated.

Uninformative priors

If researchers wish to capture uncertainty in their data without applying any assumptions or prior knowledge, they can select α_{ij} and β_{ij} to produce an intentionally uninformative prior (Leyland & Davies, 2005; Berger, 2006). This prior can be given a low weight, such that any observed data will outweigh the prior distribution. For pairs of species with little or no observed data, an uninformative prior will leave large variance about λ_{ij} and reflect our lack of knowledge about the pair. Since we model λ_{ij} as a Bernoulli trial, the appropriate uninformative prior is the Jeffreys prior: a beta distribution with $\alpha_{ij}=\beta_{ij}=0.5$. Note that this prior assumes that interaction probabilities are uniformly distributed between 0 and 1, giving an estimated connectance of 0.5; this is likely unrealistic for most networks. The sum of α_{ij} and β_{ij} is roughly equivalent to the weight of the prior, so the Jeffreys prior has equal weight to a single observation.

Informative priors

We often have some prior knowledge which can inform uncertainty about interactions. One common example is the well-justified assumption that plants never consume animals in a food web; excluding carnivorous plants it is not contentious to say that we are certain that unobserved links indicating plants feeding on animals truly do not occur. Similarly, if we observed two species which are normally obligate mutualists co-occurring but not interacting, we would be very doubtful that a zero reflecting this interaction was true (i.e., high uncertainty about this zero). It is straightforward to extend the intuition behind these

14

special cases to the full set of species in a network using trait-based models if the traits likely to affect interaction probabilities are known (e.g., Gravel *et al.* (2013); Bartomeus *et al.* (2016); Weinstein & Graham (2017a)). Informative models could also include information on abundances, phenological matching, co-occurrence, etc. as available (Jordano, 2016; Weinstein & Graham, 2017b; Gravel *et al.*, 2018). Multiple sources of information can be combined into a single model (e.g., the the K nearest neighbour (KNN) algorithm, which identifies likely interaction partners based on similarity to known partners (Desjardins-Proulx *et al.*, 2017)). This would result in a highly informative prior that would, moreover, account for some of the non-independence in interaction probabilities.

Using an informative prior tailors the level of uncertainty surrounding each interaction and may account for non-independence of interactions (as in abundance-based or KNN models). Despite these benefits, informative priors often involve many assumptions and carry a risk of over-fitting. Also note that researchers wishing to test whether a particular trait influences interaction probabilities should either exclude that trait from their prior or model the network using priors which include and exclude the trait of interest (as in Weinstein & Graham (2017b)).

Alternatively, if information on the traits affecting interaction probabilities is lacking, one can develop a prior based on the properties of a set of published networks ("empirical" or "reference Bayes" Spiegelhalter *et al.* (2000)). This type of prior does not vary between species pairs but entails only one assumption: that the focal network has similar structural properties to those of some published networks. For example, connectance ($C = N/S^2$, where N is the number of interactions and S is the number of species) is available for many

networks. The distribution of published connectances can be used to predict the distribution of interaction probabilities in a new network that is similar to those used to construct the prior (same type of interaction, similar taxonomic resolution, habitat, etc.). If we know the mean $\overline{C}$ and variance σ_C^2 of a distribution of connectances, then the hyperparameters are:

$$\alpha_{ij} = \overline{C}(\frac{\overline{C}(1 - \overline{C})}{\sigma_C^2} - 1),\tag{2}$$

$$\beta_{ij} = (1 - \overline{C})(\frac{\overline{C}(1 - \overline{C})}{\sigma_C^2} - 1).\tag{3}$$

Note that here the prior distribution will be the same for all interaction probabilities. It is also possible to interpret other network properties (e.g., species degrees) in a similar network(s) as interaction probabilities and use these to create a prior (worked example in *Appendix S4*). When using network summary statistics or degree distributions, it is important to consider whether the available information reflects what we believe about the network. Given the likelihood that most published networks are undersampled (Jordano, 2016), it may be wise to include only the highest-quality networks available in a prior. Likewise, only the degree distribution for a similar system is likely to be a reasonable prior for a focal system. For example, if the focal system is large and contains a diverse array of species, it would be inappropriate to use a prior distribution drawn from a small network describing closely-related taxa (demonstrated in *Appendix S7*). Similarly, data from a network where nodes are resolved to different taxonomic levels than the focal network are likely to have quite different properties and would not be an appropriate prior.

16

If a similar dataset is not available, it is also possible to use data from the study at hand to develop a prior (i.e., "approximate Bayes" Gelman (2016)) (Spiegelhalter *et al.*, 2000). This approach is used in medical testing (Spiegelhalter *et al.*, 2000) and disease modelling (Leyland & Davies, 2005) using information from pilot studies and can easily be applied to ecological systems. Returning to the example of degree distributions, approximate Bayes allows us to use information from abundant and easily-sampled species to predict interaction probabilities for species which were observed only rarely. Despite its practicality, approximate Bayes has been criticised for being overly subjective (Berger, 2006) and because estimating hyperparameters from data serves as an approximation of some true hyperparameters which remain unknown (Gelman, 2016). We therefore encourage researchers to carefully consider whether external data or an appropriate trait model is available before opting for a prior based on some of their own data. Nevertheless, there are likely to be many situations where approximate Bayes provides the best way to describe the uncertainty in the dataset. We therefore illustrate this approach in the following empirical example.

Empirical example

To illustrate the process of constructing a Bayesian network to quantify uncertainty about interactions, we use the comprehensively-sampled system of willows (*Salix*), herbivorous gallers, and their natural enemies described by Kopelke *et al.* (2017). This dataset consists of a single community type sampled across Europe over 29 years and at 374 unique locations. The meta-network consists of 1,173 different interactions between 52 *Salix*

17

nodes, 92 herbivore nodes, and 126 natural enemy nodes (*Appendix S5*). The high spatiotemporal resolution of this dataset makes it ideal for illustrating the difficulties in completely sampling a network; even with such an unusually high sampling effort, there were many pairs of species which were rarely or never observed together and about which we therefore have high uncertainty about interaction probabilities (Fig. 2). Using the Bayesian framework above, we can identify which potential interactions are more and less uncertain, allowing us to better predict the true structure of the metaweb. We calculated an empirical prior and computed the posterior distribution of the probability of an as-yet-unobserved interaction being feasible (λ_{ij}). We analysed both the *Salix*-galler and galler-natural enemy components of the network but, for brevity, present only the latter here (see *Appendix S6* for *Salix*-galler results).

Computing the prior and posterior distributions

To treat each interaction as a Bayesian probability, we combine observed data with a distribution based on prior information. As we lack a trait-based model or similar published network, we use data from a single sub-network from the middle of the geographical distribution of the Kopelke *et al.* (2017) dataset to inform our prior distribution. This simulates an empirical Bayes model using a pilot site to inform the prior distribution. A site in the middle of the geographic distribution was selected in order to minimise bias towards species at any geographic extreme. To demonstrate how a poor choice of prior can give unreasonable posterior distributions, we repeated our analyses using priors derived from a much smaller system (*Appendix S7*).

To obtain priors, we estimated interaction frequencies based on the normalised degree of

each species (details and code in *Appendix S8*). Using these prior parameters, we then estimated the posterior distributions of interaction probabilities λ_{ij}. For species without observed interactions ($n = 0$), the posterior distribution is identical to the prior distribution. For species where $n > 0$, we can update the prior distribution with data. If we consider only pairs of species which were observed co-occurring but not interacting, k_{ij} is always 0 and only n_{ij} will vary between species pairs. This gives $\alpha'_{ij} = \alpha_{ij}$ and $\beta'_{ij} = \beta_{ij} + n_{ij}$. We calculated posterior distributions and 95% credible intervals (function "credible_interval"; *Appendix S9*) for species with n ranging between 0 and 374, the total number of sites in our dataset.

Rather than credible intervals for a posterior distribution after collecting data, we be interested in the number of observations necessary to be confident that two co-occurring species do not interact (note that the beta distribution can never equal 0 or 1). The number of samples needed will depend on our desired level of confidence and the threshold below which we assume that two species are unlikely to ever interact. We calculated the number of samples required to reach 95% confidence that λ_{ij} was below thresholds of 0.1, 0.05, and 0.01 as examples (function "samples_for_threshold"; *Appendix S9*). Finally, we may be interested in measures of network structure. We can use posterior distributions of interactions to generate plausible "true" networks that would be obtained if sampling uncertainty could be eliminated and use these networks to estimate many network properties (worked example in *Appendix S10*).

19

Empirical results

In the *Salix*-based food webs sampled by Kopelke *et al.* (2017), most pairs of gallers and natural enemies (9,794/12,096) never co-occurred and, for species that did occur together, the total number of co-occurrences was generally low (mean=3.87, median=2; Fig. 2A). The bulk (92.24%) of these co-occurring species pairs were never observed interacting. Of those pairs that did interact, the incidence of interaction was also low (mean=4.04, median=2; Fig. 2B) and was lower than the number of observed co-occurrences (Fig. 2C).

We obtained prior parameters of α_{ij}=0.700, β_{ij}=8.49, giving a beta distribution strongly skewed towards 0. Where $n = 0$, these parameters gave a posterior distribution with $\bar{\lambda}_{ij}$=0.076, var(λ_{ij})=0.008. When $n = 374$, we obtained a posterior distribution with $\bar{\lambda}_{ij}$=1.83 $\times$ 10^{-3}, var(λ_{ij})=4.76$\times10^6$. This distribution is very close to 0 with small variance about λ_{ij}; if species i and j co-occurred 374 times without interacting, they are extremely unlikely to do so at other sites or times.

For most pairs of species i and j, n_{ij} was much less than 374 and our posterior mean and variance therefore retain more of the influence of the prior. We can see this in the increasing means and variances as we decrease n_{ij} (Fig. 3; *Appendix S11*). To be 95% confident that the probability of interaction is below 0.1, 0.05, or 0.01 would require 15, 39, and 229 observed co-occurrences, respectively. Note that these are relatively large sample sizes compared to currently-available empirical networks (e.g., Morris *et al.*, 2014).

Discussion

Real interaction networks vary over several dimensions (Kitching, 1987; Olesen *et al.*, 2011; Pires *et al.*, 2011; Baiser *et al.*, 2012; Fodrie *et al.*, 2015; Novak & Tinker, 2015), leading to pervasive under-estimation of interactions in published networks (Jordano, 2016). Even in the most extensive data set that we could find, there was very little empirical data for each species pair. Most pairs of species were not observed co-occurring even once, less than 10% of species pairs were observed interacting, and no pairs were observed to co-occur frequently enough to conclude that their probability of interacting was below 0.01. This suggests that limited sampling is a major source of uncertainty in empirical networks, in agreement with Jordano (2016); Weinstein & Graham (2017b).

Although 374 sampling sites were not enough to obtain sufficient observations of all species pairs, we were able to reduce uncertainty about many potential interactions using a simple Bayesian framework. This framework allowed us to set confidence intervals about the probability of an unobserved interaction occurring based on the distribution of interactions in one "pilot" site. This gives us more reasonable estimates of interaction probabilities than assuming that no unobserved interactions actually occur. Such approaches are particularly useful when considering interactions involving species entering new ranges due to climate change or introductions. Our framework could, for example, be used to predict the probability of interaction between a galler and parasitoid at a new site on the frontier of their ranges. If the species have not been observed co-occurring at other sites, we would expect them to interact with a probability of approximately 0.1 rather than assuming that they will not interact because they have not been observed interacting

21

elsewhere (Fig. 3). Understanding species' interactions in novel or changing communities is important for a variety of conservation questions (Bartomeus, 2013; Gravel *et al.*, 2013), and a Bayesian approach using a trait-matching model or data from species' current ranges could help us to anticipate how species will integrate into new communities.

Note that, as in all Bayesian analyses, our results do depend on the prior chosen. To demonstrate this, we repeated our analyses using a prior derived from a study of gallers found on several genotypes of *Salix hookeriana* and the parasites which emerged from them (Barbour *et al.*, 2016; Barbour, 2016; *Appendix S7*). Although the study system is similar to that in Kopelke *et al.* (2017), the network is quite different due to using different genotypes of a single *Salix* species rather than several *Salix* species as the basis for sampling. While this had a relatively small effect on our expectations for the galler-parasitoid community (*Appendix S7*), the prior based on (Barbour *et al.*, 2016) resulted in very high probabilities of interaction between *Salix* and galler pairs that were not observed interacting (Fig. 4). No amount of additional sampling would allow us to conclude that a given *Salix*-galler pair did not interact with a threshold interaction probability of 0.01. This is reasonable for the situation described in Barbour *et al.* (2016), as it is very likely that gallers which can interact with one *S. hookeriana* genotype can interact with most others, but is not reasonable for the more diverse community in Kopelke *et al.* (2017). This demonstrates the importance of conducting a "sanity check" on the posterior distribution obtained from any given prior.

We conclude with some recommendations for improving descriptions of interactions. While researchers are not likely to observe all interactions in a network (Jordano, 2016), they can take steps to address the uncertainty this implies. First, sample sizes should be as

large as possible. Second, researchers should acknowledge the varying levels of confidence surrounding interspecific interactions. Including the n and k values for each interaction will clearly indicate which unobserved interactions are most likely to be observed with further sampling and which estimates are more reliable, as well as indicating potential sampling biases. Where there are strong prior expectations about pairs of species that will not interact, these should be explicitly stated so that readers know which zeros in an interaction matrix are based on observed data and which are based primarily upon expert knowledge. Third, the uncertainty around interactions should be incorporated in calculations of network properties. Re-sampling networks based on a probabilistic understanding of networks is straightforward and gives distributions for network properties rather than point estimates. This not only acknowledges the fact that interactions vary over time and space but will also facilitate comparisons between networks by adding confidence intervals to estimates of network properties (worked example in *Appendix S9*). This will allow us to say whether networks have different structures *and* whether those differences are greater than we would expect given the inherent variability of interactions. To facilitate the practical application of these recommendations, we provide all code used in this paper in the supplementary material.

Acknowledgements

Authors' contributions

DG designed the analytical approach. AC and DG wrote the code. AC performed

statistical analyses. All authors contributed to writing and revising the manuscript.

Data accessibility

Data from Kopelke *et al.* (2017) used in this manuscript is available from the Data Paper

https://doi.org/10.1002/ecy.1832. Data from Barbour *et al.* (2016) used in this manuscript

is available from *https://doi.org/10.5061/dryad.g7805*. All code used in this manuscript is

available in the online supporting information or can be accessed via

github.com/cirtwill/Constructing-quantitative-interaction-networks or

https://doi.org/10.5281/zenodo.2593695.

Figures

Figure 1: Nested levels of uncertainty mean that an observed interaction matrix is unlikely to capture all of the interactions that truly occur. Some feasible interactions (black squares in True matrix) are less likely than others (grey squares; lighter grey is less likely) based on species traits. For example, lions are less likely to predate upon elephants than zebras. If our trait model is incomplete (e.g., if we neglect group hunting in lions), we might assume that elephants are too large to be lion prey and assign this interaction a probability of 0. Interactions occur with different probabilities in any given sample. We show six example networks representing interactions observed over two days at three sites. Interactions which occur during fewer sampling days are less likely to be observed.

Not all interactions are equally detectable. For example, interactions involving cryptic species are less likely to be observed.

After combining these layers of uncertainty, some interactions which truly occur are very unlikely to be observed. The observed matrix (bottom) is a subsample of the true matrix where the probability of observing each interaction depends upon multiple layers of uncertainty. Note that some low-probability interactions (e.g., lion predation on zebra) are included while others (e.g., lion predation on elephants) are not. For an illustration of the way in which these levels of uncertainty combine in simulated data, and the effects of uncertainty on network structure, see *Appendix S1*. Attributions for images are given in *Appendix S1-A*.

Figure 2: Despite the high replication in our empirical dataset, most galler-natural enemy pairs were never observed co-occurring and those that co-occurred rarely interacted. **A)** Histogram of the number of pairs of species observed co-occurring at least once. 9794 galler-enemy pairs were never observed co-occurring (not shown). **B)** Histogram of the number of observed interactions between co-occurring species. Species which co-occurred but never interacted are included. **C)** For each species pair, the number of observed interactions plotted against the number of observed co-occurrences. The red, dashed line indicates a 1:1 relationship.

Figure 3: Posterior distributions for the interaction probability (λ_{ij}) for two species that have not been observed interacting ($k_{ij} = 0$), based on a single site (Zillis in Graubünden, Switzerland) from Kopelke *et al.* (2017). Posterior distributions (curves) and 95% credible intervals (lines at top of panel) for λ_{ij} narrow and approach zero as the number of observed co-occurrences (n_{ij}) increases. Diamonds indicate the maximum likelihood estimator for the mean probability of interaction. Dashed lines indicate threshold probabilities of 0.01, 0.05, and 0.1. The number of samples required to obtain a 95% credible interval below each threshold increases rapidly. It takes just over 20 observed co-occurrences to be 95% confident that $\lambda_{ij} < 0.10$ but over 100 co-occurrences to be 95% confident that $\lambda_{ij} < 0.01$.

Figure 4: Posterior distributions for the interaction probability (λ_{ij}) for two species that have not been observed interacting ($k_{ij} = 0$), A) based on a single site (Zillis in Graubünden, Switzerland) from Kopelke *et al.* (2017) and B) based on a network describing interactions between gallers and several genotypes of *Salix hookeriana* from Barbour *et al.* (2016). Posterior distributions (curves) and 95% credible intervals (lines at top of panel) for λ_{ij} narrow and approach zero as the number of observed co-occurrences (n_{ij}) increases. Diamonds indicate the maximum likelihood estimator for the mean probability of interaction. Dashed lines indicate threshold probabilities of 0.01, 0.05, and 0.1. Note that both priors constrain the MLE of λ_{ij} above 0.1 and that the posterior probability distributions become similar for very high n.

References

Alberdi, A., Aizpurua, O., Bohmann, K., Gopalakrishnan, S., Lynggaard, C., Nielsen, M. & Gilber, M.T.P. (2018). Promises and pitfalls of using high-throughput sequencing for diet analysis. *Molecular Ecology*, Online Early View.

Allesina, S., Alonso, D. & Pascual, M. (2008). A general model for food web structure. *Science*, 320, 658–661.

Baiser, B., Gotelli, N.J., Buckley, H.L., Miller, T.E. & Ellison, A.M. (2012). Geographic variation in network structure of a nearctic aquatic food web. *Global Ecology and Biogeography*, 21, 579–591.

Barbour, M.A. (2016). Data from: *Genetic specificity of a plant-insect food web: implications for linking genetic variation to network complexity.* Dryad Digital Repository. Available at: https://doi.org/10.5061/dryad.g7805.

Barbour, M.A., Fortuna, M.A., Bascompte, J., Nicholson, J.R., Julkunen-Tiitto, R., Jules, E.S. & Crutsinger, G.M. (2016). Genetic specificity of a plant-insect food web: implications for linking genetic variation to network complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 2128–2133.

Bartomeus, I. (2013). Understanding linkage rules in plant-pollinator networks by using hierarchical models that incorporate pollinator detectability and plant traits. *PLoS ONE*, 8, e69200.

Bartomeus, I., Gravel, D., Tylianakis, J.M., Aizen, M.A., Dickie, I.A. & Bernard-Verdier, M. (2016). A common framework for identifying linkage rules across different types of interactions. *Functional Ecology*, 30, 1894–1903.

Berger, J. (2006). The case for objective bayesian analysis. *Bayesian Analysis*, 1, 385–402.

Blüthgen, N., Menzel, F. & Blüthgen, N. (2006). Measuring specialization in species interaction networks. *BMC Ecology*, 6, 9.

Cirtwill, A.R., Stouffer, D.B., Poulin, R. & Lagrue, C. (2016). Are parasite richness and abundance linked to prey species richness and individual feeding preferences in fish hosts? *Parasitology*, 143, 75–86.

Clegg, T., Ali, M. & Beckerman, A.P. (2018). The impact of intraspecific variation on food web structure. *Ecology*, 99, 2712–2720.

Desjardins-Proulx, P., Laigle, I., Poisot, T. & Gravel, D. (2017). Ecological interactions and the netflix problem. *PeerJ*, 5, e3644.

Dormann, C.F., Fründ, J. & Schaefer, H.M. (2017). Identifying causes of patterns in ecological networks: opportunities and limitations. *Annual Review of Ecology, Evolution, and Systematics*, 48, 559–584.

Fodrie, F.J., Yeager, L.A., Grabowski, J.H., Layman, C.A., Sherwood, G.D. & Kenworthy, M.D. (2015). Measuring individuality in habitat use across complex landscapes: approaches, constraints, and implications for assessing resource specialization. *Oecologia*, 178, 75–87.

Gelman, A. (2016). Data-dependent prior as an approximation to hierarchical model. `andrewgelman.com/2016/03/25/28321/`. Accessed: 2018-11-29.

Giron, D., Dubreuil, G., Bennett, A., Dedeine, F., Dicke, M., Dyer, L.A., Erb, M., Harris, M.O., Huguet, E., Kaloshian, I., Kawakita, A., Lopez-Vaamonde, C., Palmer, T.M., Petanidou, T., Poulsen, M., Sallé, A., Simon, J., Terblanche, J.S., Thiéry, D., Whiteman, N.K., Woods, H.A. & Pincebourde, S. (2018). Promises and challenges in insect-plant interactions. *Entomologia Experimentalis et Applicata*, 166, 319–343.

Gotelli, N.J. (2000). Null model analysis of species co-occurrence patterns. *Ecology*, 81, 2606–2621.

Graham, C.H. & Weinstein, B.G. (2018). Towards a predictive model of species interaction beta diversity. *Ecology Letters*, 21, 1299–1310.

Gravel, D., Baiser, B., Dunne, J.A., Kopelke, J.P., Martinez, N.D., Nyman, T., Poisot, T., Stouffer, D.B., Tylianakis, J.M., Wood, S.A. & Roslin, T. (2018). Bringing Elton and Grinnell together: a quantitative framework to represent the biogeography of ecological interactions. *Ecography*, Online Early View.

Gravel, D., Poisot, T., Albouy, C., Velez, L. & Mouillot, D. (2013). Inferring food web structure from predator-prey body size relationships. *Methods in Ecology and Evolution*, 4, 1083–1090.

Guimerà, R. & Sales-Pardo, M. (2009). Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 22073–22078.

Jordano, P. (1987). Patterns of mutualistic interactions in pollination and seed dispersal: connectance, dependence asymmetries, and coevolution. *The American Naturalist*, 129, 657–677.

Jordano, P. (2016). Sampling networks of ecological interactions. *Functional Ecology*, 30, 1883–1893.

Kitching, R.L. (1987). Spatial and temporal variation in food webs in water-filled treeholes. *Oikos*, 48, 280–288.

Kopelke, J.P., Nyman, T., Cazelles, K., Gravel, D., Vissault, S. & Roslin, T. (2017). Food-web structure of willow-galling sawflies and their natural enemies across Europe. *Ecology*, 98, 1730.

Lagrue, C. & Poulin, R. (2015). Bottom-up regulation of parasite population densities in freshwater ecosystems. *Oikos*, 124, 1639–1647.

Leyland, A.H. & Davies, C.A. (2005). Empirical Bayes methods for disease mapping. *Statistical Methods in Medical Research*, 14, 17–34.

Lopez, D.N., Camus, P.A., Valdivia, N. & Estay, S.A. (2017). High temporal variability in the occurrence of consumer–resource interactions in ecological networks. *Oikos*, 126, 1699–1707.

Luttbeg, B. & Kerby, J.L. (2005). Are scared prey as good as dead? *Trends in Ecology and Evolution*, 20, 416–418.

Martinez, N.D., Hawkins, B.A., Dawah, H.A. & Feifarek, B.P. (1999). Effects of sampling effort on characterization of food-web structure. *Ecology*, 80, 1044–1055.

Miller-Rushing, A.J., Høye, T.T., Inouye, D.W. & Post, E. (2010). The effects of phenological mismatches on demography. *Philosophical Transactions of the Royal Society B*, 365, 3177–3186.

Morris, R.J., Gripenberg, S., Lewis, O.T. & Roslin, T. (2014). Antagonistic interaction networks are structured independently of latitude and host guild. *Ecology Letters*, 17, 340–349.

Novak, M. & Tinker, M.T. (2015). Timescales alter the inferred strength and temporal consistency of intraspecific diet specialization. *Oecologia*, 178, 61–74.

Olesen, J.M., Stefanescu, C. & Traveset, A. (2011). Strong, long-term temporal dynamics of an ecological network. *PLoS ONE*, 6, e26455.

Pires, M.M., Guimarães Jr, P.R., Araújo, M.S., Giaretta, A.A., Costa, J.C.L. & dos Reis, S.F. (2011). The nested assembly of individual-resource networks. *Journal of Animal Ecology*, 80, 896–903.

Poisot, T., Canard, E., Mouillot, D., Mouquet, N. & Gravel, D. (2012). The dissimilarity of species interaction networks. *Ecology Letters*, 15, 1353–1361.

Poisot, T., Cirtwill, A.R., Cazelles, K., Gravel, D., Fortin, M.J. & Stouffer, D.B. (2016). The structure of probabilistic networks. *Methods in Ecology and Evolution*, 7, 303–312.

Poisot, T., Stouffer, D.B. & Gravel, D. (2015). Beyond species: why ecological interaction networks vary through space and time. *Oikos*, 124, 243–251.

Rohr, R.P., Naisbit, R.E., Mazza, C. & Bersier, L.F. (2016). Matching–centrality decomposition and the forecasting of new links in networks. *Proceedings of the Royal Society B - Biological Sciences*, 283, 20152702.

Roslin, T. & Majaneva, S. (2016). The use of DNA barcodes in food web construction–terrestrial and aquatic ecologists unite! *Genome*, 59, 603–628.

Spiegelhalter, D.J., Myles, J.P., Jones, D.R. & Abrams, K.R. (2000). Bayesian methods in health technology assessment: a review. *Health Technology Assessments*, 4.

Tylianakis, J.M., Laliberté, E., Nielsen, A. & Bascompte, J. (2010). Conservation of species interaction networks. *Biological Conservation*, 143, 2270–2279.

Vázquez, D.P., Morris, W.F. & Jordano, P. (2005). Interaction frequency as a surrogate for the total effect of animal mutualists on plants. *Ecology Letters*, 8, 1088–1094.

Weinstein, B.G. & Graham, C.H. (2017a). On comparing traits and abundance for predicting species interactions with imperfect detection. *Food Webs*, 11, 17–25.

Weinstein, B.G. & Graham, C.H. (2017b). Persistent bill and corolla matching despite shifting temporal resources in tropical hummingbird-plant interactions. *Ecology Letters*, 20, 326–335.

Wells, K., O'Hara, R.B., Pfeiffer, M., Lakim, M.B., Petney, T.N. & Durden, L.A. (2013). Inferring host specificity and network formation through agent-based models: tick-mammal interactions in Borneo. *Oecologia*, 172, 307–316.

Williams, R.J., Anandanadesan, A. & Purves, D. (2010). The probabilistic niche model reveals the niche structure and role of body size in a complex food web. *PLoS ONE*, 5, e12092.

Wirta, H.K., Hebert, P.D.N., Kaartinen, R., Prosser, S.W., Várkonyi, G. & Roslin, T. (2014). Complementary molecular information changes our perception of food web structure. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 1885–1890.