Finding missing links in interaction networks

J. Christopher D. Terry (1) james.terry@zoo.ox.ac.uk Corresponding author

Owen T. Lewis (1) owen.lewis@zoo.ox.ac.uk

1: Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK.

Running title: Finding missing links in interaction networks

Abstract

Documenting which species interact within ecological communities is challenging and labour-intensive. As a result, many interactions remain unrecorded, potentially distorting our understanding of network structure and dynamics. We test the utility of four structural models and a new coverage-deficit model for predicting missing links in both simulated and empirical bipartite networks. We find they can perform well, but that the predictive power of structural models varies with the underlying network structure. Predictions can be improved by ensembling multiple models. Sample-coverage estimators of the number of missed interactions are highly correlated with the number of missed interactions, but strongly biased towards underestimating the true number of missing links. Augmenting observed networks with most-likely missing links improves estimates of qualitative network metrics. Tools to identify likely missing links can be simple to implement, allowing the prioritisation of research effort and more robust assessment of network properties.

Keywords: ecological network, bipartite network, missing links, sampling completeness, network metrics

Introduction

Networks documenting trophic, competitive or mutualistic interactions among species are an essential tool for understanding the structure and dynamics of ecological communities (Memmott 2009, Poisot et al. 2016b, Dormann et al. 2017, Delmas et al. 2019). Interaction networks can be a better indicator of the resilience of communities than species counts (Valiente-Banuet et al. 2015) and are increasingly applied in conservation (Tylianakis et al. 2010) and biomonitoring (Gray et al. 2014).

Documenting the species present in a system can be difficult enough (Gotelli and Colwell 2010); accurately enumerating their interactions is even more challenging. As a result, ecological interaction networks are routinely under-sampled, with consequences for our understanding of communities (Goldwasser and Roughgarden 1997, Blüthgen 2010, Chacoff et al. 2012, Jordano 2016). Ecological networks are sparse and hence any unobserved interactions could be either 'missing links' (interactions that would be observed with complete sampling), or true negatives. Differences or deficiencies in sampling can lead to divergent conclusions about underlying ecological mechanisms (Lee and Guénard 2019). Furthermore, under-sampling will skew the apparent prevalence of interactions towards frequent interactions and may miss infrequent interactions entirely (Dormann et al. 2017).

Common solutions to this problem have included rarefying networks to a consistent level of sampling (Morris et al. 2014), sampling until metric values stabilise (Rivera-Hutinel et al. 2012, Costa et al. 2016), explicit observation models (Weinstein and Graham 2017), and the use of network metrics that are comparatively robust to under sampling (Vizentin-Bugoni et al. 2016). There has been an extensive literature testing these different approaches (Nielsen and Bascompte 2007, Blüthgen et al. 2008, Poisot et al. 2012b, Fründ et al. 2016, Henriksen et al. 2019, de

Aguiar et al. 2019), often arguing for an emphasis on quantitative rather than qualitative metrics, which reduce the influence of missed infrequent interactions.

Nonetheless, infrequent interactions may play an important role in ecological communities. Since networks are not static, there is value in distinguishing infrequent interactions from those that are genuinely absent. Rarely-observed links have the potential to grow in importance with changes in biotic (Terry et al. 2017) or abiotic conditions (Staniczenko et al. 2017). At equilibrium, interactions may occur at low frequency, yet be dynamically important. For instance, while interactions involving a species suppressed to a low level by a consumer may be observed infrequently, removal of the consumer can allow prey abundance to increase rapidly (Paine 1980). At a network level, the prevalence and distribution of apparently weak or infrequent interactions can have marked consequences for network stability and dynamics (McCann et al. 1998, Berlow 1999, Gellner and McCann 2016, Jacquet et al. 2016, Kadoya et al. 2018).

Ecological networks are manifestly non-random (Dunne and Pascual 2006, Montoya et al. 2006, Vázquez et al. 2009, Bascompte and Jordano 2013). In principle, it should be possible to use this underlying structure to help infer where links are most likely to have been missed (Bartomeus 2013, Morales-Castilla et al. 2015, Valdovinos 2019). To this end, a number of missing link inference approaches have been tested on ecological networks. These have included hierarchical structuring models (Clauset et al. 2008), linear filtering (Stock et al. 2017), matching-centrality models (Rohr et al. 2016), stochastic block models (Guimera and Sales-Pardo 2009) and *k*-nearest neighbour recommenders (Desjardins-Proulx et al. 2017). These models have shown that missing links can be inferred from partially-observed structured networks. However, although ecological networks are used as part of benchmarking tests, often the inference scenarios are not

orientated towards ecological needs and to date there has been limited uptake of these methods from the ecological community.

Furthermore, it is common in empirical ecological networks for species to have been sampled to markedly different extents (although see Novotny *et al.* 2010). This has the potential to give further information about the potential location of missing links, since missing links are more likely for poorly-sampled species (Blüthgen et al. 2008).

Here, we explore the potential to identify missing links in a wide range of ecological bipartite networks with a range of models, including a new species-level coverage-deficit approach which incorporates information on the completeness of sampling for individual species. We show that inferential models can perform well, especially when used in ensemble, and demonstrate how they could inform empirical work to document ecological networks and help evaluate hypotheses.

Methods

First we present a set of predictive models to detect missing links and introduce a suite of simulated and empirical networks. These are used to explore each model's capacity to identify missing links and improve the estimation of network-level metrics.

Predictive Models

Our predictive models seek to identify the true missing links within an observed bipartite adjacency matrix **O** composed of observation counts o_{ij}. This has rows representing each focal layer species i and columns detailing each potential interaction partner j. We assume that this set of observations is derived from a set of true interaction frequencies, where the relative rate of interaction between two species is given by a_{ij}. We use our statistical models to assign to each

unobserved interaction in a network a relative probability of being a 'missing link' (false negative), rather than a true negative: $P\left(a_{ij} > 0 \mid \mathbf{O}\right)$ for the cases where $\mathbf{O}_{ij} = 0$. We assume throughout that the species observed (the dimensions of **O**) are a complete list – we do not attempt to predict the identity of interactions with unobserved species. Here we introduce and detail a new 'coverage deficit' model and four structural models described in the literature. All models were fit using the *cassandRa* R package (Terry 2019). Example predictions on a simulated network of each of the core models are shown in Figure 1.

Sample size and Coverage-Deficit Models

Interactions are more likely to have been missed between species that are relatively poorly sampled. When ecological bipartite networks are constructed, it is common that one level (the focal level) of a network is sampled directly and the interaction partners in the other level counted. Examples include gathering insect hosts and rearing parasitoids from them, or observing pollinators that visit a set of plant species. This sampling may be opportunistic, matched to the local abundance, or standardised in some other manner. The result is a set of observed interaction counts with each focal layer species. Although interaction counts may be later rescaled to better represent local abundances, here we assume the raw observation data are available of discretisable relationships where individual interaction events can be observed. This includes many ecological interactions such as parasitism, pollination visits and feeding observations, but excludes diffuse facilitatory or competitive interactions. A simple approach assigns a relative probability of missed interactions based solely on the number of observations of each focal layer-species:

$$P\left(a_{ij} > 0 \mid \mathbf{O}\right) \sim \frac{1}{\sum_j \mathbf{o}_{ij}}.$$

We refer to this as our 'sample-size' model.

This approach can be refined using techniques from the extensive literature developed to assess the sample completeness of species inventories using information about the most infrequent observations to estimate the number of missing observations (Chao and Jost 2012). Rooted in information theory, these approaches can be applied to any case of estimating sample coverage, including interaction sample coverage (Chacoff et al. 2012, Jordano 2016). The core of our coverage-deficit model estimates the interaction partner sample completeness for each focal layer species, based on the observed occurrence frequencies and the 'Chao1' estimator (Chao 1984, Chao & Jost, 2012). Where data is sufficient, this coverage deficit estimator ($\hat{C}_{def}$) is an estimate of the proportion of interactions that occur between as-yet-unobserved partners, or alternatively the probability that the next interaction partner observed would not yet have been observed interacting with that species. It is defined as:

$$\hat{C}_{def} = \frac{f_1}{n} \frac{f_1\,(n-1)}{f_1(\mathrm{n}-1) + \,2f_2}$$

where f_1 is the number of interactions observed only once, f_2 is the number of interactions observed exactly twice, and n is the total observed interactions involving that focal species. For each focal layer species i, we use a $\hat{C}_{def_i}$ estimate to assign a relative probability to each potential missing interaction as:

$$P\left(a_{ij} > 0 \mid \mathbf{O}\right) \sim \frac{\hat{C}_{def_i}}{\sum_j (\boldsymbol{O}_{ij}=0)}\,.$$

Unfortunately, ecological interaction networks typically contain a number of focal layer species that are extremely poorly observed, resulting in situations where the Chao1 estimator for coverage deficit is undefined or inappropriate. For very poorly sampled species, the distribution

of interactions cannot itself provide much information (Colwell and Coddington 1994). Where observations are exclusively singletons ($f_1 = n$) the index will estimate zero coverage. At the other end of the scale, if a focal species is observed with no singleton interactions the index will estimate that there is zero sample coverage deficit.

In these cases, we use a simple binomial model to estimate $\hat{C}_{def}$, the probability that the next interactor drawn will not yet have been observed interacting with the focal species. The likelihood of not having yet observed any of the unobserved missing links is: $L = \left(1 - \hat{C}_{def}\right)^n$. While the maximum likelihood estimate of $\hat{C}_{def}$ will be 0, with a Bayesian approach and assuming a flat prior the posterior mean can be found directly as: $\mathbb{E}(\hat{C}_{def}|n) = 1 - \frac{1}{n+2}$. We use this alternative estimate for $\hat{C}_{def}$ when $n \leq 5$ (cut-off selected via accuracy tests on simulated samples), all the observations are singletons ($f_1 = n$), or if there are no singletons or doubletons ($f_1 + f_2 = 0$).

Latent-Trait Model

Species do not interact at random with each other – similar species tend to interact with a similar set of interaction partners. It should be possible to use this information to guide our expectations of missing links through a trait-based framework (Morales-Castilla et al. 2015, Bartomeus et al. 2016, Laigle et al. 2018). These similarities can be in directly-measured traits, such as corolla and tongue length (Vizentin-Bugoni et al. 2014, Klumpers et al. 2019) or through inferred 'latent' traits that are not directly measurable. Latent traits can be identified whether the community is structured by phylogeny, microhabitat or physiological differences. A survey of published networks found that the number of latent traits required may be quite low (Eklöf et al. 2013).

Intraspecific trait variation caused by spatial, temporal or ontogenetic differences can pose considerable challenges (González-Varo and Traveset 2016). Nonetheless, such approaches have had successes and a suite of approaches has been developed including matching models (Rohr et al. 2016), Dirichlet-multinomial regression (Crea et al. 2016), matrix factorisation (Seo and Hutchinson 2018) and fourth-corner analyses (Spitz et al. 2014). While it is possible to fit any number of latent traits, here we explore use of a simple single latent trait model to reduce the potential of overfitting (Rohr et al. 2016).

In this model, each focal layer species i is assigned a trait value m_i and each partner-layer species j a value m_j. The probability of the interaction existing between each i and j is determined from the difference in trait values through a logistic model:

$$\mathrm{logit}\left(P\left(a_{ij} > 0 \mid \mathbf{O}\right)\right) = -\lambda\left(m_i - m_j\right)^2 + k_M \,.$$

where λ, k_M and the vector of trait parameters $\mathbf{m}$ are found through maximum likelihood optimisation (details given in SI 3). We constrain λ to be positive, such that greater trait differences between potential interactors reduce the probability of an interaction. k_M is an intercept parameter to capture the average probability of an interaction between two species with perfectly matching traits. We penalise strongly divergent traits by introducing a Cauchy-distributed prior with mean 0 on $\mathbf{m}$.

Multiple trait distributions may have very similar likelihoods, especially with several specialised interactions or where species bridge network components. To account for this, we optimise 10 differently initialised models and average the predictions of the five models with the greatest likelihood. This model averaging approach is conceptually similar to, but computationally far cheaper than, sampling from posterior distributions of trait vectors.

Degree Model

It is readily observable that some species have many interaction partners while other species interact with very few. This manifests itself in skewed degree distributions within ecological networks (Jordano et al. 2003). All else being equal, it would be less surprising to discover a missed interaction involving a known generalist species than one with few observed interactions. Many bipartite interaction networks, especially mutualistic networks, are observed to be 'nested' - the range of specialists is a subset of the generalists (Bascompte et al. 2003, Ulrich et al. 2009). Simple degree-models can fit binary nestedness and preferential patterns well. Although nestedness can be generated by a variety of processes (Song et al. 2017) and patterns observable in binary networks may not be reflected in the quantitative networks (Staniczenko et al. 2013), our principle objective here is to identify missing links.

To fit a degree model, we assign each species a connectance term, c, and determine the probability of an interaction between focal species i and interactor j as:

$$\text{logit}\left(P\left(a_{ij} > 0 \mid \mathbf{O}\right)\right) = c_i + c_j + k_C$$

where k_C is a constant intercept term and all parameters are found by maximum likelihood. When both species have high degrees, their connectance parameters will be high, and the probability of those species interacting will be high.

Matching – Centrality Model

The matching – centrality approach (Rohr et al. 2016, 2018) combines a latent-trait and a degree model and is fit in a similar manner. This dual approach has been shown to have a high capacity

to fit tightly to diverse network structures (Rohr et al. 2016). The probability of each interaction existing is found by:

$$\text{logit}\Big(P\big(a_{ij} > 0 | \mathbf{O}\big)\Big) = c_i + c_j - \lambda\big(m_i - m_j\big)^2 + k_B.$$

As in the latent-trait model, λ is defined to be positive, k_B is a constant intercept term and all parameters are found by maximum likelihood. We again apply a weak Cauchy prior centred at 0 onto the latent trait terms **m** and fit 10 models, averaging predictions of the best five.

Block Model

Ecological networks, especially antagonistic networks, frequently show comparatively discrete compartments or modularisation, where subsets of species interact more strongly within their group than with outsiders (Olesen et al. 2007, 2008, Allesina and Pascual 2009, Schleuning et al. 2014). This may reflect fundamental incompatibilities in physiological traits, temporal mismatch (such as nocturnal/diurnal partitioning) or spatial segregation.

This grouping can be represented by stochastic block models (SBMs), which have been shown to perform well on ecological data in comparison to other clustering algorithms (Leger et al. 2015) and are used increasingly in ecology (Sander et al. 2015, Kéfi et al. 2016). In SBMs, each species is assigned to a group, G_x, in a defined set: $x \in \{1, \ldots, g\}$. The probability of interaction between two species is determined based on their group membership:

$$P(a_{ij} > 0 | \mathbf{O}) = \omega_{G_i G_j}.$$

The elements of ω are the between-group interaction probabilities and are directly specified as the fraction of observed interactions between each of the groups. We find optimal group assignations and fit the model using a degree-corrected bipartite-SBM specific algorithm

(Larremore *et al.* 2014, detailed in SI 3). To capture inherent indeterminacy in underlying group structure, we fit 10 differently-initialised models, and use the average prediction of the five models with the highest likelihood.

Combinations of Models

Individually the above models each capture discrete pieces of information about the identity of missing links. However, the structure of ecological networks is the product of many separate drivers. To capture this diversity, we combine the predictions of multiple models into ensembles. We test combining the matching-centrality model with the block model, and each of the 'structural' models with the coverage-deficit model.

We test two ensembling approaches, multiplication and averaging. Multiplying the relative probabilities assigned to each putative missing link, $P_{ens}(a_{ij} > 0) \sim P_1(a_{ij} > 0) \times P_2(a_{ij} > 0)$, emphasises the extreme probabilities of the constituent models. Averaging the relative probabilities, $P_{ens}\left(a_{ij} > 0\right) \sim P_1\left(a_{ij} > 0\right) + P_2(a_{ij} > 0)$, highlights possible interactions that are consistently identified by multiple models. Before combining models we standardise each set of probabilities assigned to unobserved interactions to sum to 1.

<u>*Datasets*</u>

Testing the efficacy of extrapolations requires knowledge of the 'true' network. We take two complimentary approaches to defining our 'true' networks. First, we generate a set of simulated networks that vary over a wide range of network properties. Second, we use a large and diverse set of empirical networks from literature sources. In both cases 'observed' networks were generated by taking draws from a multinomial distribution parameterised by the 'true' interaction

frequency. This replicates ecological sampling processes where the most frequent interactions are highly likely to have been observed and the rarest missed.

We initially generated 2000 simulated networks from a probabilistic two-trait niche model (modified from Fründ *et al.* 2016) described in full in SI 1. Key network properties were calculated using the *bipartite* R package (Dormann et al. 2008): connectance, weighted nestedness (Galeano et al. 2009), cluster coefficient (Watts and Strogatz 1998), distribution of interaction frequency (Shannon diversity, Bersier *et al.* 2002) and network level specificity H2' (Blüthgen et al. 2006). Distributions of these metrics and the (low) correlations between them are depicted in SI 1.

Our objective was to generate interaction matrices that represent a wide range of ecologically plausible bipartite networks. However, models used to generate plausible interaction networks have many similarities to the predictive models described above. This is for a good reason – models are chosen because they are thought to represent ecological networks parsimoniously. To reduce circularity, our generative model is considerably more complex than our predictive models and we examine how predictive model performance changes in response to network properties.

Using the 'true' networks we took 300-2000 samples per network to generate our 'observed' networks. We discarded occasional cases where all true interactions were observed. This resulted in a low mean true sample coverage deficit of 2.4%, corresponding to an average network completeness of 64% due to skewed interaction strength distributions. Our 'observed' networks therefore represent reasonably 'well-sampled' systems that are nonetheless missing a large proportion of the interactions present.

We collated a diverse set of 113 empirical networks representing antagonistic, mutualistic and commensalistic interaction types (SI 2). We collated quantitative single-class bipartite networks from the *Web of Life* (www.web-of-life.es) repository with 30-200 species and 40-431 observed interactions. This included 48 plant-pollinator (19 sources), 23 mammal-flea (one source, but across a wide geographic range), 2 plant-ant mutualism (2 sources) and 10 seed disperser (from 7 sources) networks. We supplemented this with 25 mammal-dung beetle interaction networks (Frank *et al*. 2018, selecting those with ≥30 total species, ≥40 observed interactions, and ≥5 mammal species) and 5 host-parasitoid networks (Tylianakis et al. 2007).

These empirical networks are not exhaustively sampled - the Chao1 estimator indicates that the observed interactions represent on average only 76% of the underlying network (SI 2). Nonetheless, we took the empirical interaction frequency to be the 'true' network to define a multinomial distribution. We subsampled each empirical network 20 times, each time taking a different fraction between 0.2 and 1 of the original sample size, taken to be the sum of the observations. We excluded cases where the observed web included more than 95% of the 'true' interactions.

Model Testing

Using each model or combination of models we assign each unobserved interaction a relative probability of being a missing link. We assume that all observations are true positives (i.e. there have been no misidentifications) although in principle similar methods could be used to identify candidates for re-evaluation. Where a species in the 'true' network was not included in the 'observed' network, we did not attempt to predict its interactions and reduced the size of the 'true' network correspondingly for all purposes.

Identifying missing links

We assess performance at identifying missing links with the area under receiver operating characteristic curve (AUC) metric to assess the information content of a signal, using the pROC R package (Robin et al. 2011). AUC can be considered the chance that a 'false negative' (missing link) was assigned a higher relative probability than a 'true negative'. A value of 0.5 indicates no useful signal while a value of 1 indicates perfect discrimination. To determine if $P(a_{ij} > 0)$ is reflective of true relative interaction frequency, we calculated the mean Spearman's rank-correlation with the 'true' relative interaction strengths of the unobserved interactions for each model.

Improving estimates of network metrics

We used the *estimateR* function in the *vegan* R package (Oksanen et al. 2018) to estimate the number of missing links based on the total distribution of observed interaction counts using the bias-corrected Chao1 index. We then augmented the observed interaction networks by adding the estimated number of the most probable missing links to the network based on the average of the block, matching-centrality and coverage-deficit models. Each selected missing link was taken to be 'observed' once, and network-level metrics calculated for the observed, augmented and true networks. We tested three qualitative network metrics (connectance, nestedness and niche overlap) identified as being susceptible to misestimation in under-sampled networks (Fründ et al. 2016), alongside two metrics considered to be comparatively resistant to under-sampling (H2' and weighted NODF nestedness).

Results

Identifying missing links - Simulated networks

In general, the network models showed impressive performance at identifying likely missing links. The best-performing model overall was the block model, closely followed by the other three network models (Fig. 2a). The coverage deficit model lagged behind the structural models, but was more informative than simply ranking focal species by sample size. Combining the predictions of multiple models tended to improve performance relative to individual models (Table 1). The overall best-performing model averaged the predictions of the block model, the matching-centrality model and the coverage-deficit model. However, the marginal improvements from combining models were often small.

Model performance and the relative hierarchy of models was affected by the underlying true network structure (Fig. 3). On average, missing links were identified best in nested, unclustered networks where interaction frequencies were more evenly distributed. The coverage-deficit and sample size models were the least responsive to network structure, since our simulated network generator did not result in a pronounced abundance-generality correlation. Average model performance did not change appreciably with increased sample size, overall interaction coverage deficit or samples per host (Fig. S4). However, when a greater fraction of interactions had been observed the structural models increased their performance, while performance of the coverage deficit model remained unchanged (Fig. S4).

Identifying missing links - Empirical networks

The performance of the different predictive models varied considerably between the different categories of empirical network (Fig. 2b). As expected, the degree model performed best for

mutualistic networks, and the latent-trait model for antagonistic networks. In all six network categories the matching-centrality model had the best overall performance. The 'true network' properties (Fig. S3) or the degree of sampling rarefaction (Fig. S4) had relatively little effect on predictive performance, except for networks that show 'anti-nestedness'. In these cases, the predictive capacity of the degree model was low, although the matching-centrality model maintained its advantage. The coverage-deficit model was very poor at predicting missing links in the empirical datasets and combining it with the network-structure models led to a *reduction* in overall predictive power (Table 1).

Identifying interaction frequency of missing links

Correlations between probabilities assigned to missing links $P\left(a_{ij} > 0 \middle| \mathbf{O}\right)$ were related to their 'true' interaction frequency. The relationship was stronger in the simulated dataset than the empirical dataset (Table S2). Of the structural models, the degree model showed the strongest positive correlation (ρ_{Sim} = 0.51, ρ_{Emp} = 0.15), while the block model showed the weakest (ρ_{Sim} = 0.28, ρ_{Emp} = 0.06). The coverage-deficit model was negatively correlated with true frequency (ρ_{Sim} = -0.35, ρ_{Emp} = -0.05) i.e. missing links successfully identified via this model tended to be infrequent.

Improving estimates of network metrics

Integrating inferred interactions into observed networks could improve estimated of key metrics. The 'augmented' networks showed increased correlation, reduced bias and reduced error in qualitative metrics, compared to the original observed networks (Fig. 4a). The quantitative network metrics (H2'and weighted NODF, Fig. 4b) were already well-estimated by the under-sampled network, and the inclusion of additional interactions made little difference. The Chao1

estimator for the number of missing interactions appeared to perform well for the empirical networks but consistently underestimated the true interaction count and coverage deficit in the simulated data (Fig. S5).

Discussion

We have shown that is possible to infer missing links in range of ecological networks with a reasonable degree of accuracy and this approach can improve estimates of network metrics. The challenge now is to put these tools into practice to leverage additional ecological insight.

Choosing between predictive models

Ideally, a model used to infer missed interactions would capture the true ecological drivers, but this is not essential for all purposes. Given the diversity of ecological networks, there will never be a single 'best' model and victors in comparisons will depend on the data set. In our simulated data our block model appears to perform best. In our empirical datasets, the matching-centrality model performed consistently well (as found by Rohr *et al.* 2016). However, we have reason to distrust the empirical datasets (discussed below). The pronounced, likely artefactual, abundance-generality relationship will favour degree-models.

Nonetheless, splitting hairs over the best-performing model is not necessarily a productive route. Networks are structured by multiple processes and in our simulated data sets the very best performance comes from combining different models, which identify different missing interactions. Structure-based models pick out the more frequent interactions while coverage deficit models highlight comparatively infrequent interactions that would be the hardest to determine through further undirected sampling. Future progress will come from operationalising estimated missing links, rather than from further incremental model refinements.

Distinctions between empirical and simulated networks

The empirical and simulated datasets overlapped substantially in key network metrics. The main relevant difference between these datasets is the stronger correlation in empirical networks between species' marginal totals and generality (mean Spearman's ρ_{Sim} = 0.43, ρ_{Emp} = 0.76). This can account for the poor performance of the coverage-deficit model in these cases. Disentangling the extent to which apparent specialism of rarely observed species is a sampling artefact is challenging (Blüthgen et al. 2006), since there are ecological explanations for such a relationship (Fort et al. 2016, Dormann et al. 2017, Simmons et al. 2019). For our purposes, this likely bias in the structure of empirical ecological networks has the consequence that predictive models may identify missing links introduced both by our subsampling procedure, and due to gaps in the original dataset. We therefore place more weight on the results from the simulated data, while noting that, despite the obstacles, the structural models are still able to perform reasonably well on the sparse empirical data.

Estimating the quality of sampling and the number of missing interactions

Sample coverage estimators offer promise for estimating the number of missed interactions (Chacoff et al. 2012, Traveset et al. 2015, Jordano 2016, MacGregor et al. 2017). The distinction between sample-coverage and network completeness is worth emphasising. Apparently high values of sample coverage alone should not be used as a simple proxy for 'network completeness', although they are related to the reliability of certain quantitative network metrics (Henriksen et al. 2019). We found that the whole-network-level Chao1 estimate of missed interactions was strongly correlated with the 'true' number of missing links in our simulated networks, but notably underestimated this number, a pattern also observed by Fründ *et al.* (2016). Although there appeared to be no systematic bias in estimated network completeness in

the empirical datasets, these 'true' networks are themselves highly incomplete, suggesting an overall underestimation also in these cases. Our focus in this paper is on the location rather than the number of missing links. Future work should investigate how long tails of infrequent interactions relate to this underestimation, and consider whether species rank abundance distribution estimators (Chao et al. 2014) can be usefully generalised to assign a frequency distribution to estimated missing links.

The way forward

Most of the models used here, and indeed our fitting approaches, likely have further room for refinement, and there are many further classes of predictive models that we did not test, such as nearest neighbour algorithms (Desjardins-Proulx et al. 2017). Furthermore, we based our models entirely on the information contained in the network itself. In most cases, ecologists constructing a network have access to (sometimes considerable) additional information and expertise on the study system. Introducing pre-existing knowledge about likely or forbidden interactions, or pooled knowledge from multiple sample sites offers considerable opportunities to maximise the information in a network (Gray et al. 2015, Pomeranz et al. 2018). External information would be particularly useful in separating 'forbidden links' (Olesen et al. 2011) that never occur due to fundamental incompatibilities from those that occur elsewhere but not within those spatio-temporal boundaries. For instance, species may not be observed to interact because of a fundamental mismatch in body-sizes (a forbidden link) or because they interact at night but data were gathered only during the day.

Because the naïve models used here do not use external information, or the judgment of the ecologists constructing the network, the quality of the estimation very much represents a lower bound of what is possible with the approach in practice. As a result of the underlying

idiosyncrasies of ecological systems, often euphemistically labelled ‘species identity effects’, there will always be an upper limit to the predictive capacity of any statistical model. We suggest that, rather than developing new models to gain marginal improvements in predictive capacity, a more productive focus would be to develop frameworks exploiting this information to test the robustness of conclusions derived from ecological networks.

Complete network inventories are not a realistic or necessarily useful goal (Jordano 2016). Enormous sample sizes and strictly deliminated boundaries are required to exhaustively sample all interactions (Martinez et al. 1999). There is undoubtedly a long tail of extremely infrequent interactions that will never be observed in a realistic sampling regime. Furthermore, there is rarely a clear line defining network boundaries (Poisot et al. 2015). Ecological networks are a constantly moving target (Rasmussen et al. 2013, CaraDonna et al. 2017) and there may be little value in including certain rare, occasional visitors and unusual interactions. It can be challenging to determine if infrequent interactions are simply a curiosity (e.g. Mercer 1966) or have significant consequences (e.g. Dudley et al. 2016). The utility of increasing the accuracy with which one type of interaction is understood will plateau as the impact of other interaction types and ecological guilds predominate (Lafferty et al. 2008). Despite the challenges, we see two principle uses for inferred missing links.

First, inferring missing links will direct further sampling where the goal is a descriptive network. In many cases the topography of the network is of principle interest, given the potential for interaction strengths to vary through time and in response to perturbation. For example, when tracing the relationships of species that may be the target for eradication measures, such as disease-vector mosquitoes (Collins et al. 2019), a critical first step is an inventory of interactions in the community. Since in most cases investment in sampling is a finite resource, directing

sampling towards high-likelihood interactions could more efficiently increase the overall network completeness. However, we caution that directed sampling will introduce biases in the generated network (Gibson et al. 2011, Falcão et al. 2016) and may hinder cross-community comparability (Poisot et al. 2012a). In certain cases, independently-sampled networks (even if less complete) may be more useful than comparisons between 'augmented' networks. The use of information from other networks may lead them to appear more homogenous than they really are and lead to underestimates of interaction β-diversity through time and space (Poisot et al. 2015, Graham and Weinstein 2018).

Second, the identification of probable missing links can improve tests of the reliability of conclusions drawn from networks, potentially through explicitly probabilistic networks (Poisot et al. 2016a). Statistical resampling and rarefaction procedures that can only work backwards or by adding random interactions may both introduce biases. While particular quantitative metrics are relatively robust to under-sampling (Blüthgen et al. 2008, Fründ et al. 2016, Henriksen et al. 2019), these metrics may not align with the goals of the study. If the conclusions are robust to the inclusion of probable missing links, then confidence in the results would be increased. Conversely, if including most-likely missing links alters conclusions, this would signal that these conclusions are unlikely to be robust. In these cases, care should be taken when selecting a predictive model to avoid a clash with the properties under investigation.

Conclusion

The identification of likely missing links is a powerful tool to include in the arsenal of network ecologists. We have shown that models to identify missing links can perform well across a wide range of network types. Importantly, inferential models need not be technically challenging to implement – for example the *cassandRa* package allows all the models discussed here to be fit

with a single function call. We hope that such approaches become more widely used in guiding ecological sampling and testing the robustness of conclusions drawn from bipartite networks.

Acknowledgments

Thanks to the CERO group for comments on this project, and two anonymous reviewers for their constructive comments. Funding was provided by NERC Standard Grant NE/N010221/1.

Data accessibility

R functions to run the predictive models are publicly available in the R package *cassandRa* on CRAN. All code and necessary data to generate simulated datasets and carry out analyses are available on GitHub (github.com/jcdterry/PredictingMissingLinks). Should the manuscript be accepted, this will also be archived in a public repository and the data doi included.

Statement of Authorship

JCDT initiated the research, in discussion with OTL. JCDT conducted all analyses and wrote the first draft of the manuscript. Both authors contributed to revisions.

References

de Aguiar, M. A. M., E. A. Newman, M. M. Pires, J. D. Yeakel, C. Boettiger, L. A. Burkle, D. Gravel, P. R. Guimarães, J. L. O'Donnell, T. Poisot, M.-J. Fortin, and D. H. Hembry. 2019. Revealing biases in the sampling of ecological interaction networks. PeerJ 7:e7566.

Allesina, S., and M. Pascual. 2009. Food web models: a plea for groups. Ecology Letters 12:652–662.

Bartomeus, I. 2013. Understanding Linkage Rules in Plant-Pollinator Networks by Using Hierarchical Models That Incorporate Pollinator Detectability and Plant Traits. PLoS ONE

8:1–8.

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

Bascompte, J., and P. Jordano. 2013. Mutualistic Networks. Princeton University Press, Princeton.

Bascompte, J., P. Jordano, C. J. Melian, and J. M. Olesen. 2003. The nested assembly of plant-animal mutualistic networks. Proceedings of the National Academy of Sciences 100:9383–9387.

Berlow, E. L. 1999. Strong effects of weak interactions in ecological communities. Nature 398:330–334.

Bersier, L.-F., C. Banašek-Richter, and M.-F. Cattin. 2002. Quantitative descriptors of food web matrices. Ecology 83:2917–2926.

Blüthgen, N. 2010. Why network analysis is often disconnected from community ecology: A critique and an ecologist's guide. Basic and Applied Ecology 11:185–195.

Blüthgen, N., J. Fründ, D. P. Vázquez, and F. Menzel. 2008. What Do Interaction Network Metrics Tell Us about Specialization and Biological Traits. Ecology 89:3387–3399.

Blüthgen, N., F. Menzel, and N. Blüthgen. 2006. Measuring specialization in species interaction networks. BMC Ecology 6:9.

CaraDonna, P. J., W. K. Petry, R. M. Brennan, J. L. Cunningham, J. L. Bronstein, N. M. Waser, and N. J. Sanders. 2017. Interaction rewiring and the rapid turnover of plant–pollinator

networks. Ecology Letters 20:385–394.

Chacoff, N. P., D. P. Vázquez, S. B. Lomáscolo, E. L. Stevani, J. Dorado, and B. Padrón. 2012. Evaluating sampling completeness in a desert plant-pollinator network. Journal of Animal Ecology 81:190–200.

Chao, A. 1984. Nonparametric estimation of the number of classes in a population. Scandinavian Journal of Statistics 11:265–270.

Chao, A., T. C. Hsieh, R. L. Chazdon, and R. K. Colwell. 2014. Unveiling the Species-Rank Abundance Distribution by Generalizing Good-Turing Sample Coverage Theory. Ecology 96:1189–1201.

Chao, A., and L. Jost. 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. Ecology 93:2533–2547.

Clauset, A., C. Moore, and M. E. J. Newman. 2008. Hierarchical structure and the prediction of missing links in networks. Nature 453:98–101.

Collins, C. M., J. A. S. Bonds, M. M. Quinlan, and J. D. Mumford. 2019. Effects of the removal or reduction in density of the malaria mosquito, Anopheles gambiae s.l ., on interacting predators and competitors in local ecosystems. Medical and Veterinary Entomology 33:1–15.

Colwell, R. K., and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society B-Biological Sciences 345:101–118.

Costa, J. M., L. P. da Silva, J. A. Ramos, and R. H. Heleno. 2016. Sampling completeness in

seed dispersal networks: When enough is enough. Basic and Applied Ecology 17:155–164.

Crea, C., R. A. Ali, and R. Rader. 2016. A new model for ecological networks using species-level traits. Methods in Ecology and Evolution 7:232–241.

Delmas, E., M. Besson, M.-H. Brice, L. A. Burkle, G. V. Dalla Riva, M.-J. Fortin, D. Gravel, P. R. Guimarães, D. H. Hembry, E. A. Newman, J. M. Olesen, M. M. Pires, J. D. Yeakel, and T. Poisot. 2019. Analysing ecological networks of species interactions. Biological Reviews 94:16–36.

Desjardins-Proulx, P., I. Laigle, T. Poisot, and D. Gravel. 2017. Ecological interactions and the Netflix problem. PeerJ 5:e3644.

Dormann, C. F., J. Fründ, and H. M. Schaefer. 2017. Identifying Causes of Patterns in Ecological Networks: Opportunities and Limitations. Annual Review of Ecology, Evolution, and Systematics 48:559–584.

Dormann, C. F., B. Gruber, and J. Frund. 2008. Introducing the bipartite Package: Analysing Ecological Networks. R News 8:8–11.

Dudley, J. P., B. M. Hang'Ombe, F. H. Leendertz, L. J. Dorward, J. de Castro, A. L. Subalusky, and M. Clauss. 2016. Carnivory in the common hippopotamus Hippopotamus amphibius□: implications for the ecology and epidemiology of anthrax in African landscapes. Mammal Review 46:191–203.

Dunne, J. A., and M. Pascual. 2006. Ecological Networks: Linking Structure to Dynamics in Food Webs. Oxford University Press, Oxford.

Eklöf, A., U. Jacob, J. Kopp, J. Bosch, R. Castro-Urgal, N. P. Chacoff, B. Dalsgaard, C. de Sassi,

M. Galetti, P. R. Guimarães, S. B. Lomáscolo, A. M. Martín González, M. A. Pizo, R. Rader, A. Rodrigo, J. M. Tylianakis, D. P. Vázquez, and S. Allesina. 2013. The dimensionality of ecological networks. Ecology Letters 16:577–583.

Falcão, J. C. F., W. Dáttilo, and V. Rico-Gray. 2016. Sampling effort differences can lead to biased conclusions on the architecture of ant-plant interaction networks. Ecological Complexity 25:44–52.

Fort, H., D. P. Vázquez, and B. L. Lan. 2016. Abundance and generalisation in mutualistic networks: Solving the chicken-and-egg dilemma. Ecology Letters 19:4–11.

Frank, K., F.-T. Krell, E. M. Slade, E. H. Raine, L. Y. Chiew, T. Schmitt, C. S. Vairappan, P. Walter, and N. Blüthgen. 2018. Global dung webs: high trophic generalism of dung beetles along the latitudinal diversity gradient. Ecology Letters 21:1229–1236.

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

Galeano, J., J. M. Pastor, and J. M. Iriondo. 2009. Weighted-Interaction Nestedness Estimator (WINE): A new estimator to calculate over frequency matrices. Environmental Modelling & Software 24:1342–1346.

Gellner, G., and K. S. McCann. 2016. Consistent role of weak and strong interactions in high- and low-diversity trophic food webs. Nature Communications 7:11180.

Gibson, R. H., B. Knott, T. Eberlein, and J. Memmott. 2011. Sampling method influences the structure of plant-pollinator networks. Oikos 120:822–831.

Goldwasser, L., and J. Roughgarden. 1997. Sampling Effects and the Estimation of Food-Web Properties. Ecology 78:41–54.

González-Varo, J. P., and A. Traveset. 2016. The Labile Limits of Forbidden Interactions. Trends in Ecology and Evolution 31:700–710.

Gotelli, N. J., and R. K. Colwell. 2010. Estimating species richness. Pages 416–422 *in* A. Magurran and B. J. McGill, editors. Biological Diversity: Frontiers in Measurement and Assessment. Open University Press.

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

Gray, C., D. J. Baird, S. Baumgartner, U. Jacob, G. B. Jenkins, E. J. O'Gorman, X. Lu, A. Ma, M. J. O. Pocock, N. Schuwirth, M. Thompson, and G. Woodward. 2014. Ecological networks: the missing links in biomonitoring science. Journal of Applied Ecology 51:1444–1449.

Gray, C., D. H. Figueroa, L. N. Hudson, A. Ma, D. Perkins, and G. Woodward. 2015. Joining the dots: An automated method for constructing food webs from compendia of published interactions. Food Webs 5:11–20.

Guimera, R., and M. Sales-Pardo. 2009. Missing and spurious interactions and the reconstruction of complex networks. Proceedings of the National Academy of Sciences 106:22073–22078.

Henriksen, M. V., D. G. Chapple, S. L. Chown, and M. A. McGeoch. 2019. The effect of network size and sampling completeness in depauperate networks. Journal of Animal Ecology 88:211–222.

Jacquet, C., C. Moritz, L. Morissette, P. Legagneux, F. Massol, P. Archambault, and D. Gravel. 2016. No complexity–stability relationship in empirical ecosystems. Nature Communications 7:12573.

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

Jordano, P., J. Bascompte, and J. M. Olesen. 2003. Invariant properties in coevolutionary networks of plant-animal interactions. Ecology Letters 6:69–81.

Kadoya, T., G. Gellner, and K. S. McCann. 2018. Potential oscillators and keystone modules in food webs. Ecology Letters 21:1330–1340.

Kéfi, S., V. Miele, E. A. Wieters, S. A. Navarrete, and E. L. Berlow. 2016. How Structured Is the Entangled Bank? The Surprisingly Simple Organization of Multiplex Ecological Networks Leads to Increased Persistence and Resilience. PLOS Biology 14:e1002527.

Klumpers, S. G. T., M. Stang, and P. G. L. Klinkhamer. 2019. Foraging efficiency and size matching in a plant-pollinator community: the importance of sugar content and tongue length. Ecology Letters 22:469–479.

Lafferty, K. D., S. Allesina, M. Arim, C. J. Briggs, G. De Leo, A. P. Dobson, J. A. Dunne, P. T. J. Johnson, A. M. Kuris, D. J. Marcogliese, N. D. Martinez, J. Memmott, P. A. Marquet, J. P. McLaughlin, E. A. Mordecai, M. Pascual, R. Poulin, and D. W. Thieltges. 2008. Parasites in food webs: the ultimate missing links. Ecology Letters 11:533–546.

Laigle, I., I. Aubin, C. Digel, U. Brose, I. Boulangeat, and D. Gravel. 2018. Species traits as drivers of food web structure. Oikos 127:316–326.

Larremore, D. B., A. Clauset, and A. Z. Jacobs. 2014. Efficiently inferring community structure in bipartite networks. Physical Review E 90:012805.

Lee, R. H., and B. Guénard. 2019. Choices of sampling method bias functional components estimation and ability to discriminate assembly mechanisms. Methods in Ecology and Evolution 10:867–878.

Leger, J.-B., J.-J. Daudin, and C. Vacher. 2015. Clustering methods differ in their ability to detect patterns in ecological networks. Methods in Ecology and Evolution 6:474–481.

MacGregor, C. J., D. M. Evans, and M. J. O. Pocock. 2017. Estimating sampling completeness of interactions in quantitative bipartite ecological networks: incorporating variation in species specialisation. bioRxiv:195917.

Martinez, N. D., B. A. Hawkins, H. A. Dawah, and B. P. Feifarek. 1999. Effects of Sampling Effort on Characterization of Food-Web Structure. Ecology 80:1044–1055.

McCann, K., A. Hastings, and G. R. Huxel. 1998. Weak trophic interactions and the balance of nature. Nature 395:794–798.

Memmott, J. 2009. Food webs: a ladder for picking strawberries or a practical tool for practical problems? Philosophical Transactions of the Royal Society B: Biological Sciences 364:1693–1699.

Mercer, A. J. 1966. Turnstones feeding on human corpse. British Birds 59:307.

Montoya, J. M., S. L. Pimm, and R. V Solé. 2006. Ecological networks and their fragility. Nature 442:259–64.

Morales-Castilla, I., M. G. Matias, D. Gravel, and M. B. Araújo. 2015. Inferring biotic

interactions from proxies. Trends in Ecology & Evolution 30:347–356.

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

Nielsen, A., and J. Bascompte. 2007. Ecological networks, nestedness and sampling effort. Journal of Ecology 95:1134–1141.

Novotny, V., S. E. Miller, L. Baje, S. Balagawi, Y. Basset, L. Cizek, K. J. Craft, F. Dem, R. a I. Drew, J. Hulcr, J. Leps, O. T. Lewis, R. Pokon, A. J. a Stewart, G. A. Samuelson, and G. D. Weiblen. 2010. Guild-specific patterns of species richness and host specialization in plant-herbivore food webs from a tropical forest. The Journal of animal ecology 79:1193–203.

Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. Henry, H. Stevens, E. Szoecs, and H. Wagner. 2018. vegan: Community Ecology Package.

Olesen, J. M., J. Bascompte, Y. L. Dupont, H. Elberling, C. Rasmussen, and P. Jordano. 2011. Missing and forbidden links in mutualistic networks. Proceedings of the Royal Society B: Biological Sciences 278:725–732.

Olesen, J. M., J. Bascompte, Y. L. Dupont, and P. Jordano. 2007. The modularity of pollination networks. Proceedings of the National Academy of Sciences 104:19891–19896.

Olesen, J. M., J. Bascompte, H. Elberling, and P. Jordano. 2008. Temporal dynamics in a pollination network. Ecology 89:1573–1582.

Paine, R. T. 1980. Food Webs: Linkage, Interaction Strength and Comunity Infrastructure. The Journal of Animal Ecology 49:666–685.

Poisot, T., E. Canard, D. Mouillot, N. Mouquet, and D. Gravel. 2012a. The dissimilarity of species interaction networks. Ecology Letters 15:1353–1361.

Poisot, T., E. Canard, N. Mouquet, and M. E. Hochberg. 2012b. A comparative study of ecological specialization estimators. Methods in Ecology and Evolution 3:537–544.

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

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

Poisot, T., D. B. Stouffer, and S. Kéfi. 2016b. Describe, understand and predict: why do we need networks in ecology? Functional Ecology 30:1878–1882.

Pomeranz, J. P. F., R. M. Thompson, T. Poisot, and J. S. Harding. 2018. Inferring predator–prey interactions in food webs. Methods in Ecology and Evolution 2018:1–12.

Rasmussen, C., Y. L. Dupont, J. B. Mosbacher, K. Trøjelsgaard, and J. M. Olesen. 2013. Strong Impact of Temporal Resolution on the Structure of an Ecological Network. PLoS ONE 8:e81694.

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

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12:77.

Rohr, R. P., R. E. Naisbit, C. Mazza, and L.-F. Bersier. 2016. Matching–centrality

decomposition and the forecasting of new links in networks. Proceedings of the Royal Society B: Biological Sciences 283:20152702.

Rohr, R. P., R. E. Naisbit, C. Mazza, and L.-F. Bersier. 2018. Statistical Approaches for Inferring and Predicting Food-Web Architecture. Pages 178–192 *in* J. C. Moore, P. C. de Ruiter, K. S. McCann, and V. Wolters, editors. Adaptive Food Webs, Stability and Transistions of Real and Model Ecosystems. Cambridge University Press.

Sander, E. L., J. T. Wootton, and S. Allesina. 2015. What Can Interaction Webs Tell Us About Species Roles? PLoS Computational Biology 11:e1004330.

Schleuning, M., L. Ingmann, R. Strauß, S. A. Fritz, B. Dalsgaard, D. Matthias Dehling, M. Plein, F. Saavedra, B. Sandel, J.-C. Svenning, K. Böhning-Gaese, and C. F. Dormann. 2014. Ecological, historical and evolutionary determinants of modularity in weighted seed-dispersal networks. Ecology Letters 17:454–463.

Seo, E., and R. A. Hutchinson. 2018. Predicting Links in Plant-Pollinator Interaction Networks Using Latent Factor Models With Implicit Feedback. Pages 808–815 Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18).

Simmons, B. I., J. Vizentin-Bugoni, P. K. Maruyama, P. A. Cotton, O. H. Marín-Gómez, C. Lara, L. Rosero-Lasprilla, M. A. Maglianesi, R. Ortiz-Pulido, M. A. Rocca, L. C. Rodrigues, B. Tinoco, M. F. Vasconcelos, M. Sazima, A. M. M. González, J. Sonne, C. Rahbek, L. V. Dicks, B. Dalsgaard, and W. J. Sutherland. 2019. Abundance drives broad patterns of generalisation in plant–hummingbird pollination networks. Oikos:oik.06104.

Song, C., R. P. Rohr, and S. Saavedra. 2017. Why are some plant–pollinator networks more nested than others? Journal of Animal Ecology 86:1417–1424.

Spitz, J., V. Ridoux, and A. Brind'Amour. 2014. Let's go beyond taxonomy in diet description: Testing a trait-based approach to prey-predator relationships. Journal of Animal Ecology 83:1137–1148.

Staniczenko, P. P. A., J. C. Kopp, and S. Allesina. 2013. The ghost of nestedness in ecological networks. Nature Communications 4:1391.

Staniczenko, P. P. A., O. T. Lewis, J. M. Tylianakis, M. Albrecht, V. Coudrain, A.-M. Klein, and F. Reed-Tsochas. 2017. Predicting the effect of habitat modification on networks of interacting species. Nature Communications 8:792.

Stock, M., T. Poisot, W. Waegeman, and B. De Baets. 2017. Linear filtering reveals false negatives in species interaction data. Scientific Reports 7:45908.

Terry, J. C. D. 2019. cassandRa R package.

Terry, J. C. D., R. J. Morris, and M. B. Bonsall. 2017. Trophic interaction modifications: an empirical and theoretical framework. Ecology Letters 20:1219–1230.

Traveset, A., J. M. Olesen, M. Nogales, P. Vargas, P. Jaramillo, E. Antolín, M. M. Trigo, and R. Heleno. 2015. Bird–flower visitation networks in the Galápagos unveil a widespread interaction release. Nature Communications 6:6376.

Tylianakis, J. M., E. Laliberté, A. Nielsen, and J. Bascompte. 2010. Conservation of species interaction networks. Biological Conservation 143:2270–2279.

Tylianakis, J. M., T. Tscharntke, and O. T. Lewis. 2007. Habitat modification alters the structure of tropical host-parasitoid food webs. Nature 445:202–5.

Ulrich, W., M. Almeida-Neto, and N. J. Gotelli. 2009. A consumer's guide to nestedness

analysis. Oikos 118:3–17.

Valdovinos, F. S. 2019. Mutualistic networks: moving closer to a predictive theory. Ecology Letters 22:1517–1534.

Valiente-Banuet, A., M. A. Aizen, J. M. Alcántara, J. Arroyo, A. Cocucci, M. Galetti, M. B. García, D. García, J. M. Gómez, P. Jordano, R. Medel, L. Navarro, J. R. Obeso, R. Oviedo, N. Ramírez, P. J. Rey, A. Traveset, M. Verdú, and R. Zamora. 2015. Beyond species loss: The extinction of ecological interactions in a changing world. Functional Ecology 29:299–307.

Vázquez, D. P., N. Bluthgen, L. Cagnolo, and N. P. Chacoff. 2009. Uniting pattern and process in plant-animal mutualistic networks: A review. Annals of Botany 103:1445–1457.

Vizentin-Bugoni, J., P. K. Maruyama, V. J. Debastiani, L. da S. Duarte, B. Dalsgaard, and M. Sazima. 2016. Influences of sampling effort on detected patterns and structuring processes of a Neotropical plant-hummingbird network. Journal of Animal Ecology 85:262–272.

Vizentin-Bugoni, J., P. K. Maruyama, and M. Sazima. 2014. Processes entangling interactions in communities: forbidden links are more important than abundance in a hummingbird-plant network. Proceedings of the Royal Society B: Biological Sciences 281:20132397.

Watts, D. J., and S. H. Strogatz. 1998. Collective dynamics of “small-world” networks. Nature 393:440–442.

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

	Simulated Dataset Accuracy Score (AUC)			**Empirical Dataset** Accuracy Score (AUC)		
Model	Structural Model-Only	Ensembled with Coverage-Deficit model		Structural Model-Only	Ensembled with Coverage-Deficit model	
		(×)	(+)		(×)	(+)
Focal Layer Sampling						
Coverage-Deficit	0.600			0.589[abc]		
Observation Count	0.580			0.603[de]		
Network Models						
Latent-Trait	0.660 [a]	0.649	0.654 [a]	0.631[f]	0.588[abc]	0.590[abc]
Degree	0.676	0.708[efg]	0.683 [b]	0.721 [i]	0.651[h]	0.663
Matching-Centrality	0.671	0.702[de]	0.684 [b]	0.718 [i]	0.642[g]	0.658
Block	0.691 [bcd]	0.689 [bc]	0.695 [cd]	0.630[f]	0.584[a]	0.590[c]
Combined Models						
Block and Matching-Centrality (Multiplied)	0.702 [de]	0.705[ef]	0.711[f]	0.647 [gh]	0.594[c]	0.604[de]
Block and Matching-Centrality (Averaged)	0.705[ef]	0.709[f]	0.713 [g]	0.648 [gh]	0.600[d]	0.603[d]

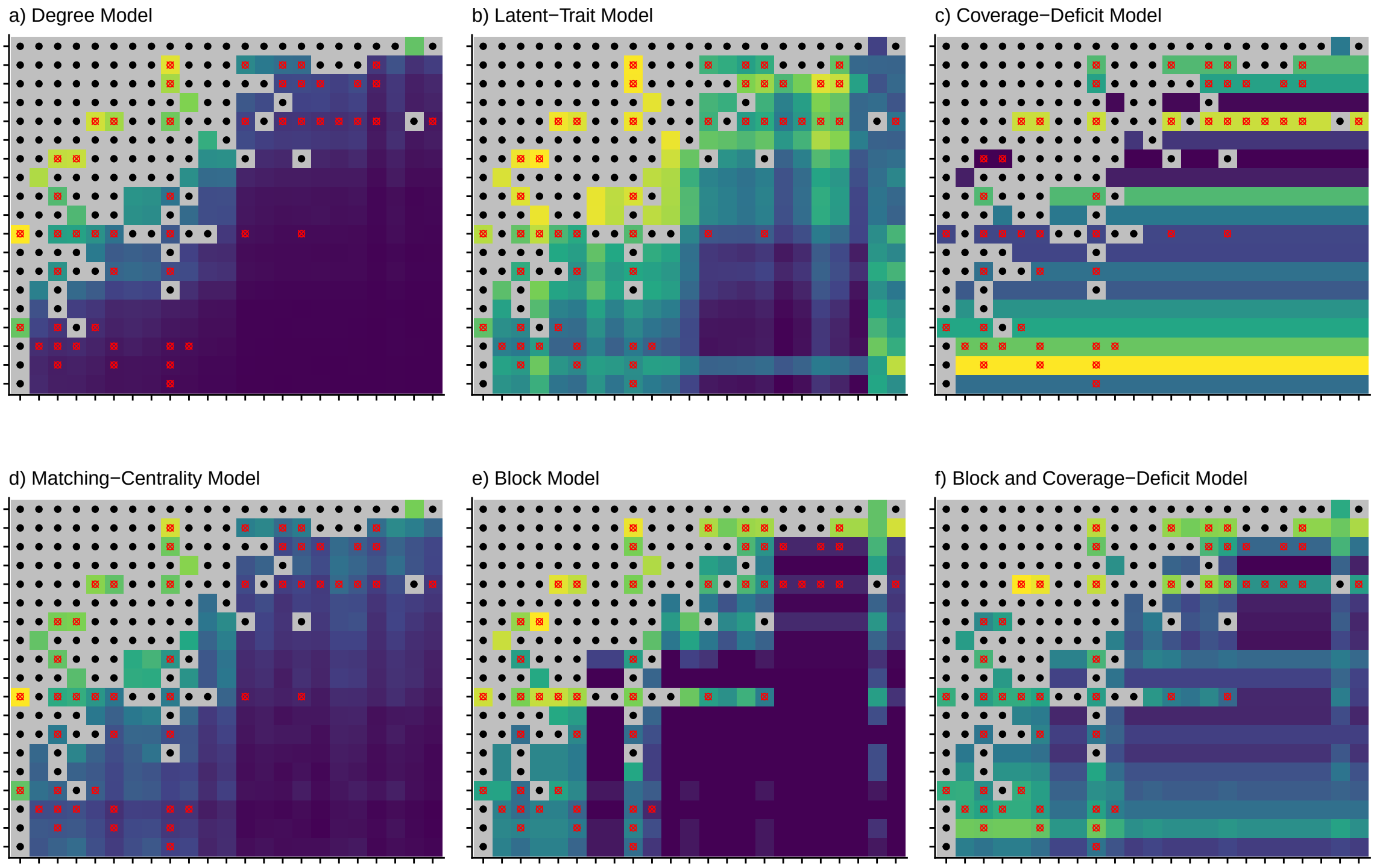
Table 1 Mean AUC values using different predictive models or combinations of models across the whole simulated (n=1985) and empirical (n=113, 2112 subsamples) network sets. Within each dataset, results are shown with just the network-structure based model (left column), ensembled with the coverage-deficit model, by multiplying the probabilities (centre column) or by averaging the assigned probabilities (right column). Incorporating coverage deficit tends to lead to modest improvements in the simulated networks but not in the empirical networks. Sets of results found not to differ significantly (Holm-correct t-tests, paired by network ID) are indicated with grouping letters, applied separately for simulated and empirical results.

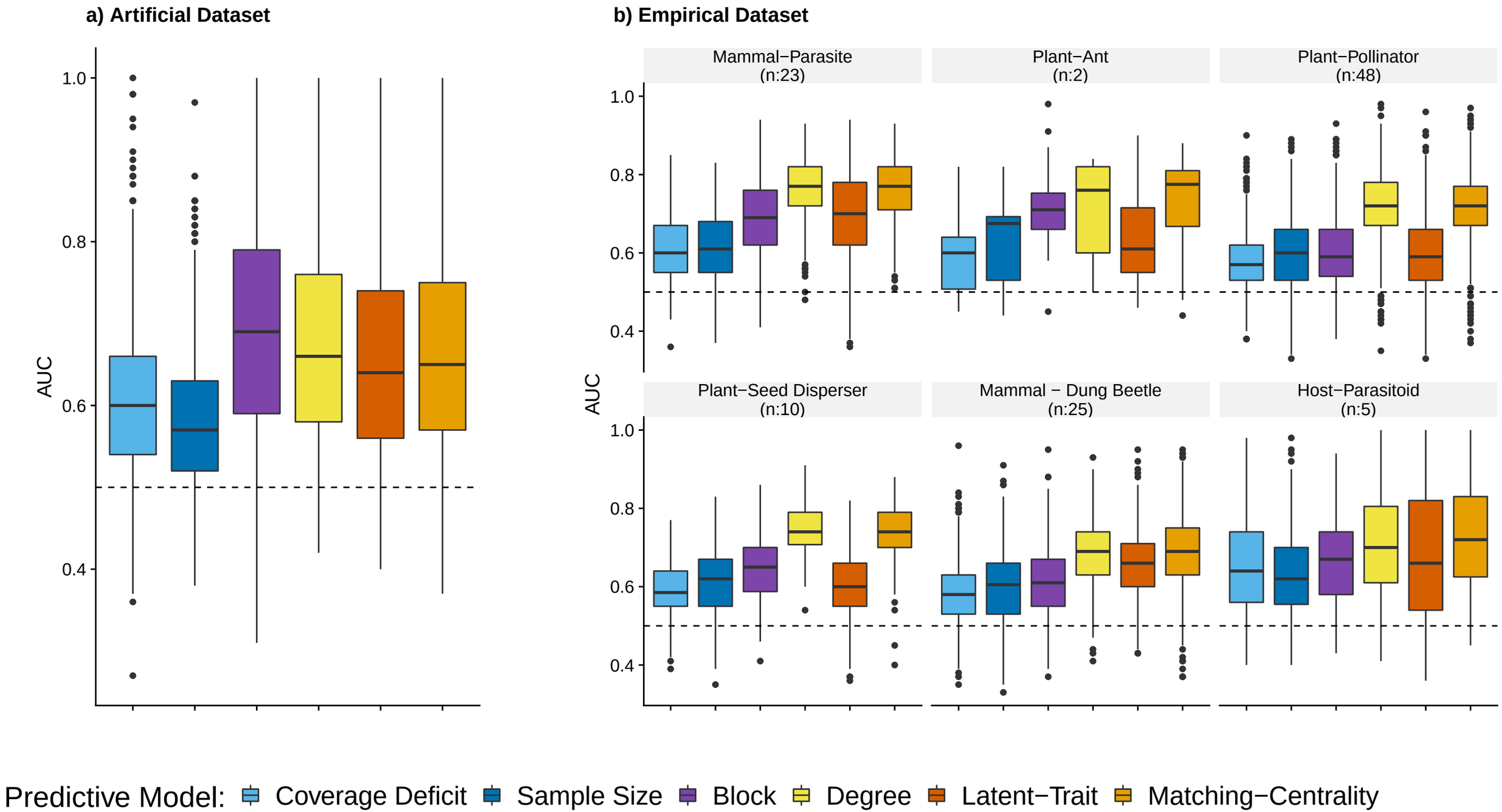
Figure 1. Demonstration of the fits of predictive models to a subsampled simulated bipartite network. Brighter colours indicate a higher probability of an unobserved interaction being present according to that model. Black dots (●) indicate observed interactions and red targets (⍰) indicate unobserved missing links that we wish to identify. AUC values for this example were as follows: Degree: 0.80, Latent-Trait: 0.76, Matching-Centrality: 0.81, Block: 0.88, Coverage-Deficit: 0.7, Averaging the Block and the Coverage-Deficit model 0.92.

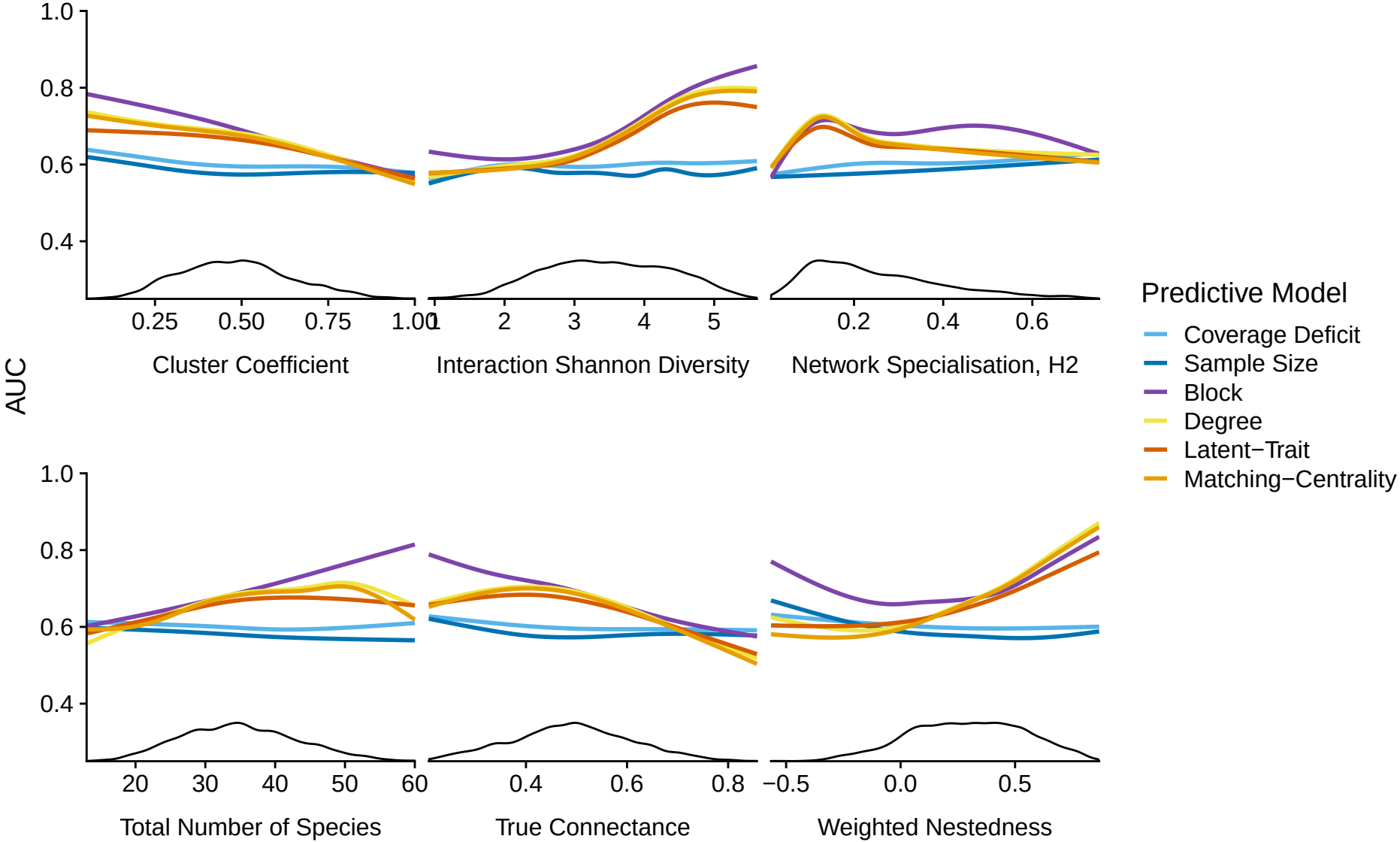
Figure 2. a) Model performance across 1985 simulated networks. Mean differences between the statistical models were small, although given the large sample size all pairwise differences were statistically significant (Holm-corrected t-tests, paired by network ID, $p<0.0001$). **b)** Model performance on the empirical data sets, subdivided by network class. Each empirical network was subsampled 20 times to different fractions of the original sample size. Models perform differently on different classes of data (Model comparison, accounting for dataset ID as a random effect, shows model:class interaction term is significant: $p< 0.0001$, n_{obs} = 12672). In all cases boxes show first and third quartiles.

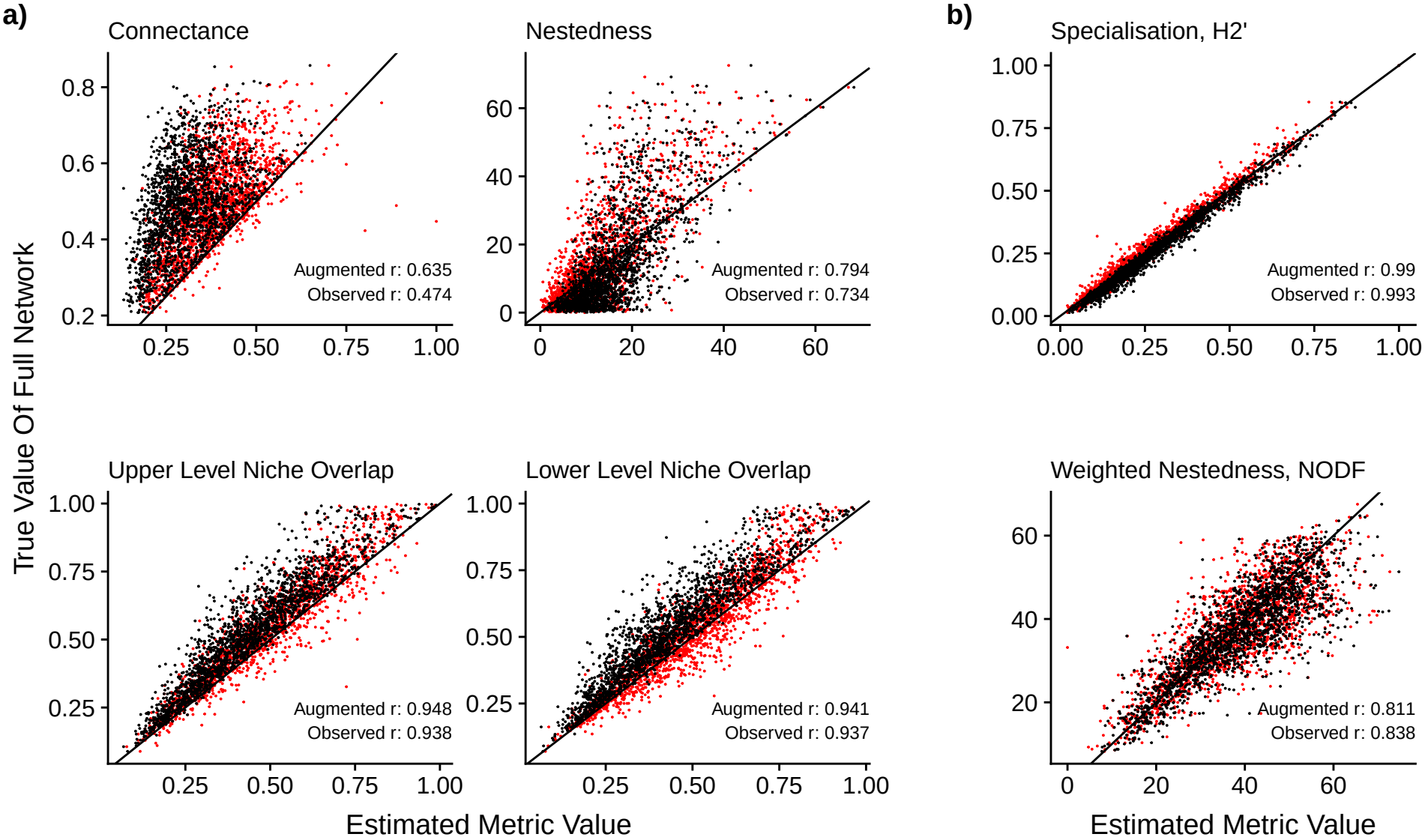
Figure 3. Response of model performance to 'true' network structure. Smoothed line shows a GAM fit for each predictive model. Lower black line shows the local density distribution of that property across the simulated test dataset.

Figure 4. Relationship between network metrics calculated on the observed (black) and 'augmented' (red) interaction networks to the true network value in the simulated networks. The line shows the 1:1 relationship. Pearson's correlation coefficient (r) is shown for each metric. Qualitative metrics (a) show notable improvements in correlation with the true values when augmented, while quantitative metrics (b) are largely unchanged.


a) Degree Model
b) Latent−Trait Model
c) Coverage−Deficit Model
d) Matching−Centrality Model
e) Block Model
f) Block and Coverage−Deficit Model


a) Artificial Dataset
b) Empirical Dataset
AUC
1.0
0.8
0.6
0.4
Mammal−Parasite (n:23)
Plant−Ant (n:2)
Plant−Pollinator (n:48)
Plant−Seed Disperser (n:10)
Mammal − Dung Beetle (n:25)
Host−Parasitoid (n:5)
Predictive Model:
Coverage Deficit
Sample Size
Block
Degree
Latent−Trait
Matching−Centrality


AUC
1.0
0.8
0.6
0.4
0.25
0.50
0.75
1.00
Cluster Coefficient
1
2
3
4
5
Interaction Shannon Diversity
0.2
0.4
0.6
Network Specialisation, H2
20
30
40
50
60
Total Number of Species
0.4
0.6
0.8
True Connectance
−0.5
0.0
0.5
Weighted Nestedness
Predictive Model
Coverage Deficit
Sample Size
Block
Degree
Latent−Trait
Matching−Centrality


a)
Connectance
Augmented r: 0.635
Observed r: 0.474
Nestedness
Augmented r: 0.794
Observed r: 0.734
Upper Level Niche Overlap
Augmented r: 0.948
Observed r: 0.938
Lower Level Niche Overlap
Augmented r: 0.941
Observed r: 0.937
b)
Specialisation, H2'
Augmented r: 0.99
Observed r: 0.993
Weighted Nestedness, NODF
Augmented r: 0.811
Observed r: 0.838
True Value Of Full Network
Estimated Metric Value
Estimated Metric Value